

New statistical approaches for the assessment of metabolomics data

M van Reenen



orcid.org 0000-0002-5856-3258

Thesis submitted in fulfilment of the requirements for the degree
Doctor of Philosophy in Statistics at the North West University

Promoter: Prof JA Westerhuis

Co-Promoter: Prof CJ Reinecke

Graduation October 2018

12791733

DEDICATION

To my girls, Emma & Stephanie

*"Great minds discuss ideas, average minds discuss events,
small minds discuss people" (Eleanor Roosevelt)*

ACKNOWLEDGEMENTS

To me there is no doubt that all gratitude is due to God. The many tumultuous life events that ran parallel to the writing of this thesis, if listed here, would convince any soul that grace was abounding. But it is also my firm belief that God's hand moves through people and so I would like to thank every person I encountered on this journey who offered a shoulder or a word of encouragement. Yet 'thank you' becomes an insufficient phrase when I have to single out some of the great minds and hearts that shaped this thesis and my character:

Prof Johan, thank you for your incredible patience. I have often wondered if you are as surprised as I am that we've accomplished this task. Thank you for your willingness to let me explore and make this thesis my own. Thank you for opening the doors of your esteemed academia and your home to me. I know there is still so much more I can learn from you and so I hope to continue on with you as a mentor and friend.

Prof Carools, you have been my co-supervisor, manager, mentor and even at times, my counsellor. You have allowed me to challenge you and calmly corrected - bending without breaking. I will always admire how you remain in awe of the human body, your enthusiasm towards life and science, and your ability to reinvent yourself. Indeed, I will strive to copy you in these respects. Thank you for building my career, but above all, my character.

Prof Hennie, thank you for sharing your ideas and for your continued guidance to turn those ideas into novel methods. Thank you for being a fantastic teacher, for not becoming frustrated by my limited time, and for never letting me leave your office without expanding my knowledge. I stand in admiration of you: such a great mind, enthralled by countless great ideas, and yet such a humble human being. What a privilege it's been, I know I am the envy of many.

To my parents-in-law for making the personal cost of this endeavour so much less. Ma Rina, I do not deserve a mother-in-law as loving and self-sacrificing as you. Thank you for being the mother to my girls I so often could not be. Pa Fanie, thank you for your open arms, words of wisdom and for supporting my husband when I could not.

To my mother Irene and sister Anna-Marie, thank you for all the proofreading and babysitting at the bitter end to make this submission possible. Thank you for your unwavering belief in me and for carrying me in prayer, always.

Last, but not least, my husband, Maarten. You have had to bear the brunt of my ups and downs, you have been put last and made least on numerous occasions. Thank you. I know the sacrifices made for the dreams we have will result in the beautiful and impactful lives we strive for.

ABSTRACT

The aim of this PhD study was primarily to develop statistical methods that can accommodate the characteristics of metabolomics data, as well as assist in answering the underlying biological questions. However, to identify where a contribution could be made required an understanding of metabolomics data and the statistical methods applied in practice. This, in turn, required interaction with a metabolomics investigation and so the novel application and/or combination of existing statistical methods became a secondary aim. A longitudinal, intervention-based metabolomics study, with a crossover design, was selected for this purpose.

To make the primary aim of this thesis achievable, it was necessary to understand the different approaches to research in statistics. New statistical theory can be developed without reflexion on the application of such developments, that is, for whom a new or expanded method may be of use. However, if research in statistics commences separately from an area of application, developments may not cater for the specific requirements of that area. New statistical theory can also be developed to accommodate specific characteristics of data or to answer questions specific to a given area of application or discipline, that is, context centred statistical research. This thesis then firstly, explores the implications of these two diverse approaches from a theoretical perspective. Context centred statistical research is explored in greater depth as a transdisciplinary approach in the context of metabolomics. Metabolomics is the study of the interactions between endo- and exogenous stimuli (such as lifestyle or disease) and metabolic pathways of a living organism through the metabolites formed. The interactions between statistics and metabolomics are explored next, for the various steps in the knowledge production process, to understand how such transdisciplinary endeavours may be executed.

Metabolomics data are known to: (i) have many times more variables than cases; (ii) exhibit severe multicollinearity; (iii) have unequal sample sizes for experimental groups; (iv) have large proportions of missing values; (v) present with skewed distributions; and (vi) exhibit high levels of natural variation. These characteristics make the statistical analysis of metabolomics data challenging. To illustrate this and to achieve the secondary aim of this thesis, three publications, describing the design and analysis of data sets relating to a longitudinal crossover alcohol intervention study, are included.

The challenging nature of metabolomics data and the limited number of statistical methods to accommodate such data presents many opportunities to develop or expand upon statistical methods. To achieve the primary aim of this thesis, two publications are included to demonstrate how interaction with contextualized data can generate new ideas, culminating in new methods.

The thesis culminates in a reflection on the dynamics of transdisciplinary research to conclude that a context centred approach to research in statistics does not only favour the context or the end-users of statistical methods, but can also act as a muse, inspiring new innovations in statistics. Finally, the thesis concludes with an outlook on future avenues that may be explored given the work presented here.

Keywords

Metabolomics: Crossover longitudinal intervention; Acute alcohol consumption; Nuclear magnetic resonance (NMR) spectroscopy; gas chromatography–mass spectrometry (GC–MS)

Statistics: Multivariate analysis; Binary classification; Variable selection; Non-parametric statistics; Significance testing.

Publications

This thesis is presented in article format and includes the following accepted publications:

van Reenen, M., Reinecke, C.J., Westerhuis, J.A. & Venter, J.H. 2016. Variable selection for binary classification using error rate p-values applied to metabolomics data. *BMC bioinformatics*, 17(1):33.

Irwin, C., van Reenen, M., Mason, S., Mienie, L.J., Westerhuis, J.A. & Reinecke, C.J. 2016. Contribution towards a metabolite profile of the detoxification of benzoic acid through glycine conjugation: an intervention study. *PloS one*, 11(12):e0167309.

van Reenen, M., Westerhuis, J.A., Reinecke, C.J. & Venter, J.H. 2017. Metabolomics variable selection and classification in the presence of observations below the detection limit using an extension of ERp. *BMC bioinformatics*, 18(1):83.

Irwin, C., Mienie, L.J., Wevers, R.A., Mason, S., Westerhuis, J.A., van Reenen, M. & Reinecke, C.J. 2018. GC–MS metabolomics reveals multiple dysregulated metabolic pathways following experimental acute alcohol consumption. *Scientific reports*, 8:5775

Irwin, C., van Reenen, M., Mason, S., Mienie, L.J., Wevers, R.A., Westerhuis, J.A. & Reinecke, C.J. 2018. The ¹H-NMR-based metabolite profile of acute alcohol consumption: a metabolomics intervention study. *PloS one*. (Still in press)

TABLE OF CONTENTS

DEDICATION	I
ACKNOWLEDGEMENTS	II
ABSTRACT.....	III
CHAPTER 1 – BACKGROUND	1
1.1 Towards Metabolomics	1
1.2 Aim	4
1.3 Objectives	4
1.4 References.....	5
CHAPTER 2 – THE BIGGER PICTURE: A THESIS IN CONTEXT	6
2.1 The Art of Statistics – An Opinion.....	6
2.1.1 The Advent of Statistics	6
2.1.2 Metaphor 1: Spectrum Statistics	7
2.1.3 Metaphor 2: Silo Statistics.....	8
2.1.4 Metaphor 3: Circular Statistics	10
2.1.5 Metaphor 4: Hyphenated Statistics	13
2.2 Statistics at ‘The Gap’ – an overview of current literature	15
2.2.1 An Interpretation of Knowledge Production in Metabolomics.....	15
2.2.2 ‘Gaps’ in the Research Cycle – Some Examples.....	18
2.3 Roadmaps.....	33
2.4 References.....	35
CHAPTER 3 – EXAMPLES OF THE ADVANCE APPLICATION OF STATISTICS IN HYPOTHESIS GENERATING RESEARCH	41

3.1	Background	41
3.2	Accepted Publication: Contribution towards a Metabolite Profile of the Detoxification of Benzoic Acid through Glycine Conjugation: An Intervention Study.....	45
3.2.1	Title Page	45
3.2.2	Abstract	46
3.2.3	Introduction	47
3.2.4	Materials and methods.....	50
3.2.5	Results	57
3.2.6	Discussion.....	66
3.2.7	Additional Information	70
3.3	Supplementary Information: Contribution towards a Metabolite Profile of the Detoxification of Benzoic Acid through Glycine Conjugation: An Intervention Study.....	72
3.3.1	Original ¹ H-NMR spectral data for intervention 1	72
3.3.2	Threshold value and normalization	73
3.3.3	Data Pre-processing	74
3.3.4	Statistical Analysis	77
3.3.5	Graphs on Excretion Kinetics.....	89
3.3.6	NMR spectra on the excretion of six substances	91
3.3.7	NMR analysis on guanidinoacetic acid	93
3.4	Accepted Publication: The ¹H-NMR-based metabolite profile of acute alcohol consumption: a metabolomics intervention study	95
3.4.1	Title Page	95
3.4.2	Abstract.....	96

3.4.3	Introduction	97
3.4.4	Materials and methods.....	101
3.4.5	Results	104
3.4.6	Discussion.....	118
3.4.7	Additional Information	123
3.5	Supplementary Information: The ¹H-NMR-based metabolite profile of acute alcohol consumption: a metabolomics intervention study	124
3.5.1	Methods for sample treatment, storage, preparation and ¹ H-NMR analysis	124
3.5.2	Data Processing.....	125
3.5.3	Uric acid analysis	128
3.5.4	Original ¹ H-NMR spectral data.....	130
3.6	Accepted Publication: GC–MS-based urinary organic acid profiling reveals multiple dysregulated metabolic pathways following experimental acute alcohol consumption	131
3.6.1	Title Page	131
3.6.2	Abstract	132
3.6.3	Introduction	132
3.6.4	Results	135
3.6.5	Discussion.....	148
3.6.6	Conclusions.....	153
3.6.7	Materials & Methods	153
3.6.8	Statistical Analysis	158
3.6.9	Additional Information	159

3.7	Supplementary Information: GC–MS-based urinary organic acid profiling reveals multiple dysregulated metabolic pathways following experimental acute alcohol consumption	161
3.7.1	Sample Preparation, Organic Acid Extraction, and GC–MS Analysis.....	161
3.7.2	Alcohol Intervention Study Data.....	164
3.7.3	Identified Features	164
3.7.4	Statistical Power.....	170
3.7.5	Data Pre-processing	171
3.7.6	Statistical Methods & Additional Results.....	173
3.7.7	The Hippuric Acid Effect.....	175
3.7.8	Example of Informed Consent Form	177
3.8	Statistical Software	181
3.9	Comments on: GC–MS-based urinary organic acid profiling reveals multiple dysregulated metabolic pathways following experimental acute alcohol consumption.....	182
3.10	References	186
 CHAPTER 4 – NOVEL STATISTICAL APPROACHES FOR THE ANALYSIS OF METABOLOMICS DATA.....		
4.1	Background	201
4.2	Accepted Publication: Variable selection for binary classification using error rate <i>p</i>-values applied to metabolomics data	203
4.2.1	Title Page	203
4.2.2	Abstract	204
4.2.3	Background	205
4.2.4	Methods	207

4.2.5	Results & Discussion	212
4.2.6	Conclusion	226
4.2.7	Additional Information	226
4.3	Supplementary Information: Variable selection for binary classification using error rate p-values applied to metabolomics data.....	228
4.3.1	Estimating error rates from data.....	228
4.3.2	Using classification error rates as test statistics.....	229
4.3.3	Calculation of the null distributions by simulation.....	230
4.3.4	Asymptotic estimation of the null distribution of the directional full sample error rate test statistic.....	231
4.3.5	Tests based on leave-one-out error rates	233
4.3.6	Summary of the ERp Approach	235
4.3.7	Comparing null distributions.....	238
4.3.8	Power comparisons of test statistics	240
4.3.9	Comparing threshold estimators	254
4.3.10	Comparisons of LOO and FS error rate estimates per se.....	267
4.3.11	Complete results for metabolomics study	279
4.4	Accepted Publication: Metabolomics variable selection and classification in the presence of observations below the detection limit using an extension of ERp.....	286
4.4.1	Title Page	286
4.4.2	Abstract	287
4.4.3	Background	288
4.4.4	Methods	291
4.4.5	Results & Discussion	298

4.4.6	Conclusion	309
4.4.7	Additional Information	310
4.5	Supplementary Information: Metabolomics variable selection and classification in the presence of observations below the detection limit using an extension of ERp.....	312
4.5.1	The null distribution of <i>erdown</i> *and <i>ermin</i> *	312
4.5.2	Simulating the null distribution	313
4.5.3	The null hypothesis probability of getting a zero error rate	314
4.5.4	XERp Software.....	317
4.5.5	Comparison of the p-values under the null hypothesis	320
4.5.6	Comparison of the p-values under the alternative hypothesis	322
4.5.7	Comparison to random imputation	326
4.6	References.....	328
CHAPTER 5 – DISCUSSION & FUTURE PROSPECTS.....		333
5.1	Concluding Remarks	333
5.1.1	Valid, valuable and generalizable results – a view on evaluating Chapter 3 and similar contributions	333
5.1.2	Statistics for purpose or prestige - a view on evaluating Chapter 4 and similar contributions	339
5.2	Future Prospects.....	340
5.2.1	Continuing the ALC2013 investigation	340
5.2.2	Experimental design at ‘The Gap’	341
5.2.3	Extending and expanding ERp beyond XERp	342
5.3	In conclusion	342

5.4	References.....	344
	ADDENDUMS	347
6.1	Permission from Co-authors.....	347
6.2	Copyright Licences of Journals	347
6.2.1	PloS one.....	347
6.2.2	Scientific Reports	353
6.2.3	BMC Bioinformatics.....	355
6.3	Journals' Instructions to Authors	357
6.3.1	PloS one.....	357
6.3.2	Scientific Reports	357
6.3.3	BMC Bioinformatics.....	357

LIST OF TABLES

Table 2-1:	Popular methods used to compare two groups based on metabolomics data.....	32
Table 3-1:	Description of the Time factor in the ALC2013 intervention study.....	41
Table 3-2:	Quantified data on metabolically important metabolites.	65
Table 3-3:	Excerpt of raw ¹ H-NMR spectral data.....	73
Table 3-4:	PLS-DA Fit Statistics Pairwise Time Point Comparisons with time 0.	85
Table 3-5:	A summary of the experimental designs, data analysis approaches and main metabolic conclusions of some NMR-based ethanol administration studies.	100
Table 3-6:	Quantified data of important metabolites following alcohol consumption.	111
Table 3-7:	A small extract from the file containing the raw ¹ H-NMR spectral data.	130
Table 3-8:	Univariate, multivariate and descriptive statistics for the most perturbed metabolites following alcohol consumption.	141
Table 3-9:	An excerpt of the data used for the alcohol intervention study.....	164
Table 3-10:	Identification, classification and reference ranges of 172 identified variables.	169
Table 3-11:	List of statistical software used throuout this chapter.	181
Table 4-1:	Significant variables based on weight set 1 and 2.....	219
Table 4-2:	Group classification and outlier detection using significant variables based on weight set 1 and 2.....	225
Table 4-3:	List of Scenarios.	238
Table 4-4:	Results using <i>elmin</i> with w0 = w1 = 12	280
Table 4-5:	Results using <i>elmin</i> with w0 = 13 and w1 = 23	281
Table 4-6:	Results using the Mann-Whitney Test.....	282

Table 4-7:	Classification results using the median LOO threshold with $w_0 = w_1 = 12$	284
Table 4-8:	Classification results using the median LOO threshold with $w_0 = 13$ and $w_1 = 23$	285
Table 4-9:	XERp Results for TBM vs Healthy Controls.	305
Table 4-10:	LOO XERp Results for TBM vs Healthy Controls.	307
Table 4-11:	Classification Results for Sick Controls.	309
Table 4-12:	Null hypothesis probabilities of the event that <i>erup</i> $\neq 0$	317
Table 4-13:	XERp significant variables, performance & proportion of zeroes.	326

LIST OF FIGURES

Figure 1-1:	An overview of the genesis of this PhD thesis.....	1
Figure 1-2:	The funding value chain from basic research to industry and TIA's role in it (adapted from an internal TIA brochure circulated in 2008).	3
Figure 2-1:	Metabolomics as a transdisciplinary research field.	14
Figure 2-2:	An interpretation of the metabolomics knowledge production process.	16
Figure 3-1:	Representation of all elements of the experimental design.	51
Figure 3-2:	Flow diagram indicating the main lines of activity following data generation, identification and quantification of important metabolites on the intervention towards the proposed biological interpretation.	55
Figure 3-3:	500-MHz ¹ H-NMR spectra of urine.	56
Figure 3-4:	Group separation among experimental groups through dendrograms and Volcano plots based on equidistant binning spectral data.....	58
Figure 3-5:	Unfolded PCA Scores Plots.....	60
Figure 3-6:	Application of ASCA to the 344 NMR spectral bins for 21 individuals over the period of the intervention.	62
Figure 3-7:	Venn diagram displaying counts of variables selected by various techniques.	63
Figure 3-8:	Metabolite profile of benzoic acid biotransformation with hippuric acid as outcome.....	67
Figure 3-9:	QC Outlier Detection.	76
Figure 3-10:	Volcano Plots of Pairwise Time Point Comparisons with time 0.	79
Figure 3-11:	Dendrograms of Pairwise Time Point Comparisons with time 0.....	80
Figure 3-12:	PCA Score Plots of Pairwise Time Point Comparisons with time 0.....	82
Figure 3-13:	PLS-DA Score Plots of Pairwise Time Point Comparisons with time 0.	84
Figure 3-14:	Unfolding Data.....	87

Figure 3-15:	Sum of Squared Loadings of ASCA model.	88
Figure 3-16:	Urinary Excretion kinetics of important metabolites.....	90
Figure 3-17:	500 MHz ¹ H-NMR spectra of minor components from urine.	92
Figure 3-18:	1D and 2D JRES NMR spectra.	94
Figure 3-19:	Representative ¹ H-NMR spectrum of urine collected one hour following the 'alcohol plus vehicle' intervention.	105
Figure 3-20:	Confirmation of sorbitol annotation.....	107
Figure 3-21:	Group separation between participants, based on bins data from the 'vehicle only' and the 'alcohol plus vehicle' interventions.	109
Figure 3-22:	Changes in the concentrations of the six up-regulated metabolites from time 0 to time 4 following alcohol consumption.	114
Figure 3-23:	Indications of differences in the average levels of hypoxanthine and sorbitol across the four interventions and six time points.	117
Figure 3-24:	Model of the metabolite profile based on the important metabolites up-regulated following alcohol consumption.....	119
Figure 3-25:	PCA scores plot illustrating the variation and correlation within and between bins for QC and experimental samples.....	127
Figure 3-26:	PCA and PLS–DA scores plots following alcohol consumption.	137
Figure 3-27:	Multivariate approaches to indicate the time effect.	139
Figure 3-28:	Correlation matrix over the full time period for the 13 metabolites listed in Table 3-8.	144
Figure 3-29:	Unfolded PCA scores and selected bi-plots.	147
Figure 3-30:	Proposed model indicating some important metabolic pathways affected by acute alcohol consumption.	149
Figure 3-31:	Representation of the metabolomics workflow to investigate the effect of acute alcohol consumption.	156
Figure 3-32:	Power curves.....	172

Figure 3-33:	PCA and PLS–DA plots following alcohol consumption.	176
Figure 4-1:	Algorithm to simulate the null cumulative distribution functions.	211
Figure 4-2:	The null cumulative distribution functions.	213
Figure 4-3:	Simulation comparison of the different error rate test statistics.	215
Figure 4-4:	CART variable importance.	222
Figure 4-5:	The ERp method for variable selection.	236
Figure 4-6:	A graphically comparison of approaches to calculating the null CDF.	239
Figure 4-7:	Power comparison between test statistics for scenario 1a: $N0 = 20$ and $N1 = 20$ with $w0 = 12$ and $w1 = 12$	242
Figure 4-8:	Power comparison between test statistics for scenario 1b: $N0 = 10$ and $N1 = 20$ with $w0 = 12$ and $w1 = 12$	244
Figure 4-9:	Power comparison between test statistics for scenario 1c: $N0 = 20$ and $N1 = 10$ with $w0 = 12$ and $w1 = 12$	245
Figure 4-10:	Power comparison between test statistics for scenario 2a: $N0 = 20$ and $N1 = 20$ with $w0 = 34$ and $w1 = 14$	246
Figure 4-11:	Power comparison between test statistics for scenario 2b: $N0 = 10$ and $N1 = 20$ with $w0 = 34$ and $w1 = 14$	247
Figure 4-12:	Power comparison between test statistics for scenario 2c: $N0 = 20$ and $N1 = 10$ with $w0 = 34$ and $w1 = 14$	249
Figure 4-13:	Power comparison between test statistics for scenario 3a: $N0 = 20$ and $N1 = 20$ with $w0 = 14$ and $w1 = 34$	250
Figure 4-14:	Power comparison between test statistics for scenario 3b: $N0 = 10$ and $N1 = 20$ with $w0 = 14$ and $w1 = 34$	252
Figure 4-15:	Power comparison between test statistics for scenario 3c: $N0 = 20$ and $N1 = 10$ with $w0 = 14$ and $w1 = 34$	253
Figure 4-16:	Comparing threshold estimates for scenario 1a: $N0 = 20$ and $N1 = 20$ with $w0 = 12$ and $w1 = 12$	257

Figure 4-17:	Comparing threshold estimates for scenario 1b: $N0 = 10$ and $N1 = 20$ with $w0 = 12$ and $w1 = 12$	258
Figure 4-18:	Comparing threshold estimates for scenario 1c: $N0 = 20$ and $N1 = 10$ with $w0 = 12$ and $w1 = 12$	259
Figure 4-19:	Comparing threshold estimates for scenario 2a: $N0 = 20$ and $N1 = 20$ with $w0 = 34$ and $w1 = 14$	260
Figure 4-20:	Comparing threshold estimates for scenario 2b: $N0 = 10$ and $N1 = 20$ with $w0 = 34$ and $w1 = 14$	261
Figure 4-21:	Comparing threshold estimates for scenario 2c: $N0 = 20$ and $N1 = 10$ with $w0 = 34$ and $w1 = 14$	263
Figure 4-22:	Comparing threshold estimates for scenario 3a: $N0 = 20$ and $N1 = 20$ with $w0 = 14$ and $w1 = 34$	264
Figure 4-23:	Comparing threshold estimates for scenario 3b: $N0 = 10$ and $N1 = 20$ with $w0 = 14$ and $w1 = 34$	265
Figure 4-24:	Comparing threshold estimates for scenario 3c: $N0 = 20$ and $N1 = 10$ with $w0 = 14$ and $w1 = 34$	266
Figure 4-25:	Simulation comparison of error rate estimates with the population error rate.	268
Figure 4-26:	Comparing error rate estimates for $N0 = 20$ and $N1 = 20$ with $w0 =$ 12 and $w1 = 12$	270
Figure 4-27:	Comparing error rate estimates for $N0 = 10$ and $N1 = 20$ with $w0 =$ 12 and $w1 = 12$	271
Figure 4-28:	Comparing error rate estimates for $N0 = 20$ and $N1 = 10$ with $w0 =$ 12 and $w1 = 12$	272
Figure 4-29:	Comparing error rate estimates for $N0 = 20$ and $N1 = 20$ with $w0 =$ 34 and $w1 = 14$	273
Figure 4-30:	Comparing error rate for $N0 = 10$ and $N1 = 20$ with $w0 = 34$ and $w1 =$ 14	275

Figure 4-31:	Comparing error rate estimates for $N0 = 20$ and $N1 = 10$ with $w0 = 34$ and $w1 = 14$	276
Figure 4-32:	Comparing error rate estimates for $N0 = 20$ and $N1 = 20$ with $w0 = 14$ and $w1 = 34$	277
Figure 4-33:	Comparing error rate estimates for $N0 = 20$ and $N1 = 10$ with $w0 = 14$ and $w1 = 34$	278
Figure 4-34:	An illustration of the CDFs discussed.....	292
Figure 4-35:	Null CDF for <i>erup</i> * with π taking on different values.	296
Figure 4-36:	Bias, MSE and size of the three p-value alternatives for <i>erup</i> *.....	300
Figure 4-37:	Measures of power of the three p-value alternatives for <i>erup</i> *.....	302
Figure 4-38:	Overview of the XERp software.....	319
Figure 4-39:	Size results of the three p-value alternatives.....	321
Figure 4-40:	Average p-value ($p \leq 0.1$) for the dissonant case.	323
Figure 4-41:	The proportion of null hypotheses rejected given $\alpha = 5\%$	324
Figure 4-42:	The proportion of null hypotheses rejected given $\alpha = 1\%$	325

CHAPTER 1 – BACKGROUND

1.1 Towards Metabolomics

On 11 June 2001, the South African Minister of Arts, Culture, Science and Technology, Dr. Ben Ngubane, published a policy document entitled *A National Biotechnology Strategy for South Africa* (Department of Arts, Culture, Science and Technology, 2001). The policy was a comprehensive declaration outlining many details on the implementation of the biotechnology strategy and laying out a course of development that included capacity building through advanced education (PhDs), of which, amongst others, the present PhD thesis is an example (Figure 1-1).

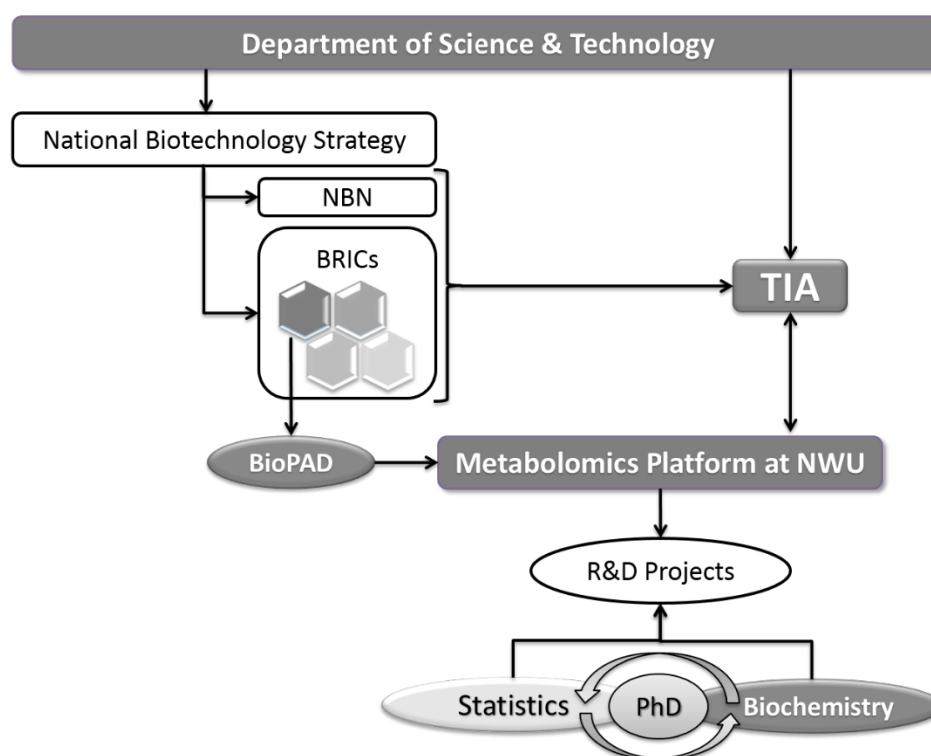


Figure 1-1: An overview of the genesis of this PhD thesis.

Two key components of the biotechnology strategy were: (i) the establishment of the National Bioinformatics Network (NBN) at the University of the Western Cape, to meet the needs of South African science and biotechnology in the field of Bioinformatics; and (ii) the creation of four Biotechnological Innovation Centres (BRICs) responsible for the regional development of biotechnology in South Africa, including the establishment of BioPADs (Biotechnology

Partnership and Development units) that should focus on the development of metabolomics technology. Dr Butana Mboniswa, Chief Executive Officer of BioPAD, took the initiative to approach North-West University (NWU) in 2006 to establish a Metabolomics Platform at its Potchefstroom Campus. BioPAD required an associated three-year Project Plan (BioPAD Project BPP007) which would be funded by BioPAD. The University agreed to the conditions and an agreement of cooperation was signed, formalizing Phase 1 of the BioPAD initiative. In 2008 the South African Department of Science and Technology took a further initiative (Phase 2) by the formation of a public entity, called the Technological Innovation Agency (TIA), established in terms of the TIA Act (26 of 2008). TIA was formed by a merger of the previous four BRICs, NBN, and other related biotechnology structures. The ultimate aim of TIA was and remains to support and intensify technology innovation in order to stimulate economic growth and improve the quality of life for all South Africans by exploiting technological innovations. TIA's core business objective is to assist the development and commercialization of competitive technology-based services and products. For this it primarily uses the South African science and technology base to develop industries, create sustainable jobs, and help diversify the economy. TIA marketed its new vision and approach through a flow-diagram (Figure 1-2), indicating the continuity between basic research at universities (funded by the Department of Higher Education and Training, the National Research Foundation (NRF) and others such as the Medical Research Council (MRC)), through the innovation phase (funded by TIA) to the final phase of commercialization (funded through the Industrial Development Cooperation (IDC) and others in industry)

Amongst other activities, TIA invests in industrial biotechnology and initiatives towards health improvement, including drug development. TIA again approached the management of the NWU to support the further development of the Metabolomics Platform to become a national facility, hosted at its Potchefstroom Campus, to which the University agreed.

Expertise in statistics, geared towards the analysis of data generated by metabolomics technologies, is an integral part of metabolomics, necessitating the appointment of a full-time statistician. Dr Gerhard Koekemoer fulfilled this role, supported by the Statistical Consultation Services of the NWU as co-researcher to the metabolomics platform. The University advanced the metabolomics endeavour by establishing a Centre for Human Metabolomics of the NWU (CHM) as an associated TIA Metabolomics Platform and, following the appointment of Dr Koekemoer as Statistician at Sastech (Sasol), created a full-time position for a statistician at CHM, in which I was appointed. This opportunity opened the possibility for me to become involved in

the development of bioinformatics expertise at CHM and the Metabolomics Platform, including research for a PhD degree, which culminated in the present thesis

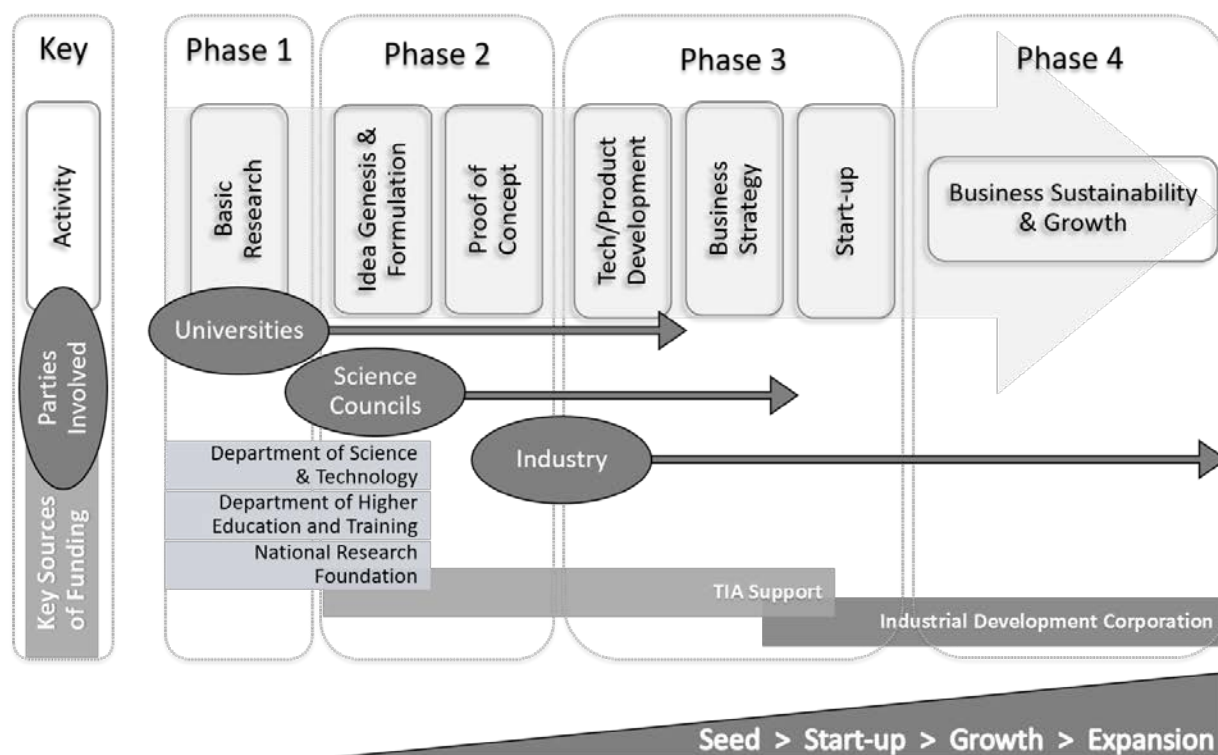


Figure 1-2: The funding value chain from basic research to industry and TIA's role in it (adapted from an internal TIA brochure circulated in 2008).

The day-to-day responsibilities of the statistician appointed at CHM and the Metabolomics Platform revolve around specific projects. These projects include many different single-stimuli approaches, for example, the effects of diseases such as pulmonary tuberculosis, tuberculosis meningitis, and fibromyalgia. One project at the time was of particular interest - a longitudinal multiple intervention study. This design was relatively new to metabolomics studies during that period and the first of its kind for CHM or the Metabolomics Platform. These projects, specifically the last, presented opportunities to define an aim and derived objectives for this PhD study.

1.2 Aim

The advent of metabolomics as a field of research made contributions from bioinformatics in conjunction with metabolomics a rather open-ended process. This PhD therefore had a dual aim: (i) to combine existing statistical approaches in novel ways to enable more comprehensive biological interpretation of results; and (ii) to develop new statistical approaches directed to challenging real metabolomics data characteristics.

Specifically, this thesis focussed on the binary discrimination problem, that is, to determine whether two groups differ significantly, which variables drive these differences and how can models be constructed to predict group membership. This problem is of great importance to metabolomics investigations not just because of its prevalence, but because it forms the cornerstone of biological interpretation even in studies with more elaborate designs.

Regardless of the project, metabolomics data all exhibit characteristics which are difficult for most traditional statistical methods to accommodate, making metabolomics very attractive for innovative research in statistics. Characteristics inherent to the metabolome (the total set of metabolites measured) include: (i) multicollinearity; (ii) large natural variation between and within cases; (iii) highly skewed distributions; (iv) large proportions of missing values, which are not missing at random; (v) many times more variables than observations; (vi) analytical limitations leading to complex correlation structures between measurements; and (vii) small and unequal group sizes.

1.3 Objectives

To achieve the aim discussed above, various objectives, indicated below, were formulated and are presented as chapters in this thesis:

- (i) To gain insight into the drivers of development in the field of statistics and the role of statistics as an intrinsic pillar of metabolomics, in order to determine a knowledge production process to explore further through case studies in the remainder of my thesis (Chapter 2).
- (ii) To apply advanced statistical methods to an intervention-based, longitudinal, crossover human metabolomics case study, to illustrate and emphasize how essential a

transdisciplinary approach is, as well as to demonstrate the use of and need for multiple statistical methods to maximize the discovery and interpretation of findings (Chapter 3).

- (iii) To develop a new statistical approach which can take some of, and be developed to take more of, the characteristics of metabolomics data into account. Simultaneously, this approach should be applicable when the aim of the investigation is biomarker discovery (i.e. to identify metabolites that can discriminate between two groups) or biological interpretation (Chapter 4).
- (iv) To produce a view on the way forward, not only with respect to the further enhancement of the method developed in Objective (iii), but also with respect to statistics as a core discipline in metabolomics (Chapter 5).

1.4 References

Department of Arts, Culture, Science and Technology **see** South Africa.

South Africa. Department of Arts, Culture, Science and Technology. 2001. A National Biotechnology Strategy for South Africa. Pretoria.

South Africa. 2008. Technology Innovation Agency Act 2008 (Act 26 of 2008).

CHAPTER 2 – THE BIGGER PICTURE: A THESIS IN CONTEXT

In this chapter, by way of orientation, I would like to convey some of my views and experiences, which I will support with literature, on the role of statistics in metabolomics. They are, to some extent, my personal opinions on the state of the art that is statistics, but limited to the context of metabolomics. I start my discourse by touching on what I view as an identity crisis faced by statisticians, shared by some statisticians and missed by others. I do so by using four metaphors on statistics that best express my day-to-day experience. By unpacking the definition, it becomes apparent that statistics, and indeed most disciplines, are practised along a spectrum with advanced theory at the one extreme and advanced application at the other. Both these two aspects are illustrated by the original research presented in this thesis – Chapter 3 illustrates the advanced application of statistics, while Chapter 4 illustrates the advancement of the theory of statistics, motivated by a distinct unexplored approach for analysing metabolomics data. This chapter, therefore, binds this thesis into a whole as it places into context the parts to follow, hence the title – providing context to consolidate the contributions presented in this thesis.

2.1 The Art of Statistics – An Opinion

“Too often people do not treat statistics as an art, but as a recipe that can passively be followed... Turning data into clear and compelling information for an audience requires many technical skills and much training, but at its soul, it is an art” (Wright, 2006).

2.1.1 The Advent of Statistics

Statistics originated from the mathematical modelling of games of chance. Today mathematics, as well as statistics, are recognized as disciplines in their own right. In my opinion, this resulted due to a core philosophical difference: statistics builds models from context, mathematics builds models regardless of context. Poincaré said that *“mathematicians do not study objects, but the relationship between objects”* (QuoteHD, 2017) - this implies that objects are interchangeable as long as the relationship remains unchanged. In contrast, statistics requires *“...some maturity in understanding of the world”* (Hand, 2009). The objects themselves are important, phrased

differently, context matters. In statistics data, being a representation of objects and the relationships among them, should not and cannot be separated from their origin; doing so would make data lose their meaning and power to change the way we see the world. Policy changes, business strategies, industrial innovation, budgetary spending, the significance of research findings, to name but a few, are all examples of the power of statistics to drive decision-making when applied to data in context.

I align my definition with that of Chambers, who defined statistics, in the broadest sense of the word, as: “...*everything related to learning from data, from the first planning or collection to the last presentation or report*” (Chambers, 1993). Statisticians are essentially data transformers, using data and context (i.e. the design through which the data were generated; who or what the data represent; and with which instrument they were measured) to disclose interpretable information - summarizing without compromising. Importantly, Chambers’ definition extends more traditional views on statisticians to include practitioners of computer learning, data mining, data analytics, business analytics etc. as they all have a common aim of gaining insight into data that inform on the context from which the data emerged. Yet all have a different area of expertise and should provide complementary insights on this common aim. Unfortunately, it appears that these different labels have resulted in a dilution of skills.

2.1.2 Metaphor 1: Spectrum Statistics

*“Applied disciplines operate in a contested space between the quest for ‘pure’ knowledge in basic research and the quest for tangible results in practice”
(Korte, 2016).*

The search for a relatable description of a statistician prompted me to view statistics more holistically, as a discipline practised along a broad spectrum from highly theoretical advancements, focused on developing new theory by almost returning to mathematics at times, moving steadily along to the application of any data analysis tool with foundations in fields as seemingly far removed as psychology and engineering. This led me to conclude likewise - it is necessary to make known the unifying definition of statistics. There are many reasons for this, but two are primary. First, advanced application will become superficial if practitioners are not rooted in the fundamentals of statistics. Second, if practitioners at the theoretical end are not aware of the data issues and complexities faced by practitioners at the opposite end of the

spectrum, advancement in the 'needed' directions will be slow or duplicity will occur. Though the context driving the theoretical development will dictate its functionality, it is reasonable to assume that some core theoretical elements will be replicated, resulting in the advancement of theory in parallel - in width rather than in depth (Hand, 2009).

2.1.3 Metaphor 2: Silo Statistics

"Profounder men than I have failed to diagnose, let alone cure, the disease that has infected us all. If I may be orphic for a moment, one should say that the ostensible goals have obliterated the real origins of our search"
(Chargaff, 1975).

To unite statistics it is necessary to understand how the family of data transformers became estranged from data reductionists in the first place. The literature and personal experience prompts me to conclude that at least one reason for this is the creation of disciplines. To support this view we need to take a step back. If research is a process of investigation or experimentation aimed at discovery, insight, clarity, and application to address a social need, in its widest sense, then research culminated in a body of knowledge. Only in more recent times was this body cast into silos representing various disciplines: *"The scientific discipline... is an invention of nineteenth century society"* (Stichweh, 2001).

The formation of disciplines through an unintentional ring-fencing of knowledge redirected research and questions started to sprout from the disciplines themselves rather than societal needs. Placing the discipline centrally evolved into the modernist approach to research, which aims to solve problems from the discipline, for the discipline, to be applied by the discipline, and culminated in extreme mono-disciplinarity. Other disciplines are rarely consulted or even considered. It is then no wonder that disciplines can exist separate from each other and yet strive towards the same goals. So why then were these silos created if they are so isolating? Silos are not all bad – a focused approach to research is responsible for the immense depth of understanding of the basic building blocks that underpin a discipline. Unfortunately, this benefit has become progressively overemphasized. Hand (Hand, 2009) lists two motives for this: (i) how research is funded and rewarded; and (ii) basic human nature (my terminology, not Hand's). Research in disciplinary silos is easier to evaluate and to budget for, while humans like to form communities with comforting ways of life. The rigid structure of the silo resulted in what Cohen

and Lloyd liken to disciplinary inbreeding effectively 'weakening the species' as a whole and the reductionist way of pursuing science within a silo, viewed by Chargaff (Chargaff, 1975) as the disease that infected us all.

Though these may be rather generalizing statements, I believe there is truth in them. Statistics grew from the great minds of Fisher, Laplace, Poisson, Youden, Box, Tukey and others (Stigler, 1973; MacGregor, 1997), who were primarily scientists, geneticists, chemists, and engineers: *"They understood the way scientists and engineers thought, and they understood their problems. As a result, they developed statistical methods to treat these real world problems, not by starting with the statistical theory and trying to find a problem it could treat, but by starting from the very real problems and developing methods to treat them"* (MacGregor, 1997). In these early years statistics was a verb, a practical application, described by Fisher as *"fruitful labor"* (Fisher, 1922). However, Fisher did recognize that fundamental theoretical problems also required resolution. But is there value in the precise definition of concepts in a field revolving around uncertainty and error? Here is my summary of Fisher's concern in layman's terms - if the aim of statistics is data reduction to chewable chunks of summarizing information, then at least some proof of adequacy is required - *"...the function of Theoretical Statistics is to show how such adequate statistics may be calculated"* (Fisher, 1922). At the very least we need to prove that summarizing does not equate to compromising.

The pendulum seems to have swung towards the other extreme in more recent years – *"The last two decades have seen statistics grow as a mathematical discipline. However, this period has seen much less interesting growth in applied statistics, not because there were no new problems, but because the leadership in the statistical disciplines passed on to a new generation of mathematical statisticians"* (MacGregor, 1997). It is my view that many statistical theorists have become too interested in the *"fundamental paradoxes"* (Fisher, 1922) and too comfortable within their silo of peers that they have forgotten how mere flirtations between data reductionists and data generators produce only superficiality. The ironic result is that statistics, a context-driven discipline, has started losing context. Statistics fell into the trap of severe mono-disciplinarity as it started to solve questions from statistics for statistics, with no real intent to explore whether the solution related to a real-world problem. The spectrum of statistics became unbalanced with too much weight being placed on the theoretical end. I dare to say that this one-sidedness led to statistics boarding the 'big data' train a little late: *"The leadership in developing statistical methods for these data rich problems appears to have again returned to the owners of the problems..."* (MacGregor, 1997). While statistics was fine-tuning its arsenal, the 'data-bang' happened and

data evolved from limited, manually captured observations to enormous amounts of streaming data. This left a gap which the proverbial market was eager to fill. Computer scientists and engineers jumped at the opportunities statisticians were more hesitant to explore (Breiman, 2002). The unstructured heaps of messy data made many theoretically focused statisticians weary of making any assumptions in the fear of drawing an incorrect conclusion. More recent additions to the data-transformer family were less risk averse, they jumped in and started swimming – drifting or leaning towards scholarship as they go along. This then poses a philosophical question – how does one balance the discipline's spectrum, meaning how does one give equal weight to both ends of the spectrum? Does one split the discipline in two to ensure sufficient practitioners as well as theorists? This seems counter intuitive given the preceding discussion. I propose that we move the ends towards one another, we don't try to balance the spectrum, but rather, to bend it.

2.1.4 Metaphor 3: Circular Statistics

"The real problem might not be an obsession with methodology as much as it is the neglected state of discovery and imagination in our work to build theories that actually help people do something to improve the world"
(Korte, 2016).

It is my belief that the spectrum along which statistics should be practiced, must be circular. Let me explain, understanding the context or source of the data is just as much a prerequisite as knowledge of the inner workings and, more importantly, limitations of techniques. Therefore, where advanced application ends, theoretical development begins again. Malley and Moore stated that *"...novel and big problems should compel novel solutions and not persistence of historical artifact"* (Malley & Moore, 2013). Finding a balance is then no longer the trick of the trade, it has become a matter of finding depth. Statisticians are no longer just required to become specialists in the fundamental theories, but also generalists. Statisticians must be sufficiently familiar with the pressing questions of a context to make a fresh onslaught - mechanizing, modernizing, or creating methods that can provide satisfactory answers. Coming to this conclusion, I have embraced the idea of one specialization and one generalization (i.e. fluency in a data generating context). I believe statisticians should stay firmly rooted in the core discipline, but also gain sufficient insight into a chosen field of application. I will explain why I believe the specific application of statistics is better than the general application, in relation to three

observations or rather experiences: (i) application breeds appreciation; (ii) context gives meaning; and (iii) investments incentivize.

My arguments require some personal details for which I ask some leeway. I would not have known that excitement can accompany statistics if I had not moved into an applied field after completing my undergraduate studies. The purely theoretical exposure to statistics during my early student years left me wanting for purpose. I now risk a question: Has statistical education not 'regressed' and become too mathematical, avoiding the messiness of real-world data, avoiding the context? I find that many agree. According to George Box: *"...statistical core research and graduate education went wrong by ignoring the history of the influence of important practical problems in the development of general statistical methods"*, as referenced by Parzen (Parzen, 1998). Rao reiterates this even more directly, *"...lack of contact with live problems has impeded the expansion of statistics in desired directions or sharpening of existing tools"* (Rao, 2001).

My first argument is this: it is precisely through the application of statistical theory that its power and ability to change how people view the world and themselves unfolds and can be appreciated. Once this realization sets in, the theory behind the application will become of even greater rather than lesser importance, as many may fear. Practitioners who, by linking to a specific context will experience first-hand the impact of results, are bound to realize anew how essential the accurate selection and application of statistical methods are. Also, practitioners in this setting are best positioned to recognise which methods should be further developed to ensure correct inference.

This brings me to my second argument - context gives meaning. Parzen phrased this need for meaning so well when he said that: *"...the purpose of statistical computing is insight, not numbers"* (Parzen, 1998). During my tenure as a statistical consultant, I came to the same conclusion as Kimball did many years earlier - answers without context lead to *"errors of the third kind"*, i.e. *"...the error committed by giving the right answer to the wrong problem"* (Kimball, 1957). If statisticians are not collaborators their role is diminished to that of a service provider or consultant, providing answers rather than solutions due to a lack of understanding of the problem behind the questions. Breiman concluded that before you can successfully model data you have to live with them: *"...the emphasis needs to be on the problem and on the data"* (Breiman, 2001).

This led me to my third argument - investment incentivizes. Once I selected a field of application, I experienced 'some skin in the game', so moving towards hyphenated statistics. Investopedia attributes the phrase *"skin in the game"* to the celebrated investor Warren Buffett (Investopedia, 1999). The phrase describes a scenario where you buy stock in your own company. Hyphenated

statistics is a scenario where the statistician moves beyond being a service provider to a true collaborator, invested in the quality of the research and not just the accuracy of the statistics. This requires investment, a belief in the value of the research being conducted, and ownership of the management of factors that can diminish the truth or impact of subsequent findings.

There is also some monetary motivation here, that is, you have to become invested to get investors. I dislike the funding conversation as much as the next scientist, but as societal pressures diversify and increase, the funding of research from the co-funding model between industry and government will also increase, expressed in the term translational research. I suspect that a larger and larger proportion of research funding will come from the private sector. Private funding inadvertently comes with a different set of demands - demands linked to the commercial application of research to serve communities of paying customers. Collaborative efforts are key to bring about new perspectives required to address society's needs. The aim shifts dramatically, from expanding individual disciplines to solving societal problems. These current trends in research should perhaps be more appropriately labelled as *"the democratization of science"* (Silka, 2013). The aim is for closer interaction between science and society (in both directions), where *"...communities become the architects of rather than merely the objects of study"* (Silka, 2013). This signals the emergence of a new kind of science: contextualized or context-sensitive science (Nowotny *et al.*, 2003). In the present South Africa, at least five main traits are discernible:

- Priority-driven research to help grow the country's economy.
- Research priorities closely linked to social needs, for example, addressing the severe HIV/TB epidemic in South Africa.
- Commercialization to help ensure adequate funding for educational and research activities.
- Accountability with respect to applied resources and subsequent innovations.
- Globalization of research activities.

Assuming my arguments were convincing, I would like next to introduce my hyphen more formally.

2.1.5 Metaphor 4: Hyphenated Statistics

"Ignore your mother. Play in the intersection" (Waters, 2012).

I was and still am convinced that the choice to work where the spectrum's ends meet, where advanced application creates new theory, will be rewarding. Here research questions are loaded with context and the resulting data are complex, requiring innovative data analysis strategies. So only one question remained – what would my hyphen be? What discipline could I collaborate with to bring statistics full-circle? I chose metabolomics.

Metabolomics can be described as the resulting field of research when the traditional disciplines of biology, statistics, and chemistry are its foundation and collaborate to explore metabolic processes in biological systems. Metabolomics aims to understand a biological system by investigating it at a rather basic level through exploring the products of its metabolic pathways - its metabolites. More formally, metabolomics investigates metabolic pathways through the metabolites formed, to understand differences between phenotypes or interactions of the biological system with stimuli, such as from exogenous chemicals or disease.

Metabolomics has evolved into *"...a distinct and very active, multi-disciplinary research field"* (Madsen *et al.*, 2010), because to understand the implication of the presence of a given metabolite at a given concentration requires expertise from various fields - such as clinical science, analytical chemistry, and data analysis (Figure 2-1).

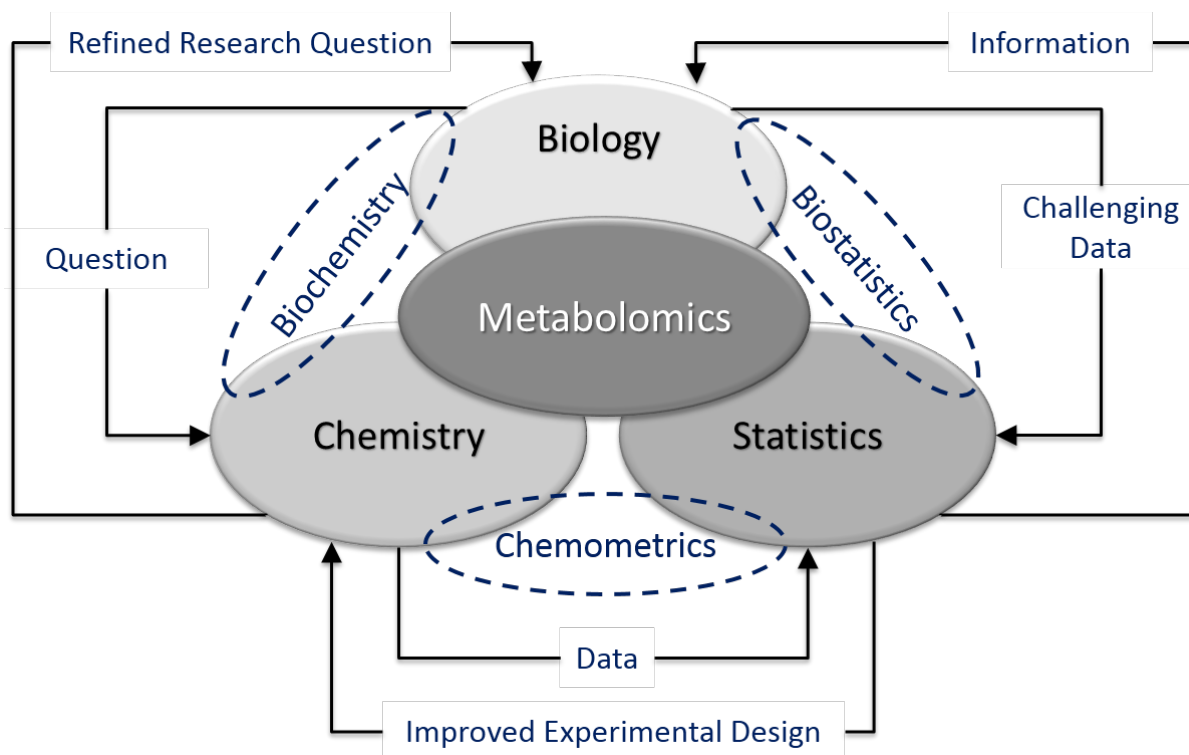


Figure 2-1: Metabolomics as a transdisciplinary research field.

I would venture beyond a multi-disciplinary description towards a trans-disciplinary one (Ciesielski *et al.*, 2016). To move from a question to new knowledge in metabolomics implies that the underpinning disciplines cannot simply provide input at some stage during the research process. These disciplines must work as a team to allow researchers question, plan, evaluate, and interpret together. As Goodacre stressed: *“Raw data is nothing but a poor relative of information and information is itself a giant leap away from knowledge”* (Goodacre, 2005). Such leaps and bounds have never been made in isolation, teams of people bring them about. To study the metabolome these teams must be as diverse as the technologies employed and so cohesive that new disciplines are born - for example metabolomics - whose gestation came about through trans-disciplinary ventures (Krastanov, 2014).

2.2 Statistics at ‘The Gap’ – an overview of current literature

“...data- and technology-driven programmes are not alternatives to hypothesis-led studies in scientific knowledge discovery but are complementary and iterative partners with them” (Kell & Oliver, 2004).

I have now established my opinion that statistics is practised along a spectrum with advancements in theory at the one extreme and advanced application, within a given context, at the other. I have also discussed that the ends of the spectrum should not be extremes of each other, but rather neighbours with the one lending to the other. This ‘circular spectrum’ is not a closed loop, as this would imply the tools required to extract all the information from a set of context-specific data exist and are sufficient. ‘The Gap’ is then the challenge to complete the circle by developing new or improving existing techniques appreciative of the application context and of transdisciplinarity. This section explores ‘The Gap’ when working in metabolomics, specifically human metabolomics, by looking at some key challenges encountered throughout the metabolomics knowledge production process and justified from the metabolomics literature as required for the context of this thesis.

2.2.1 An Interpretation of Knowledge Production in Metabolomics

Figure 2-2 depicts my interpretation of the metabolomics knowledge production process as I have observed it. In my experience, though limited, I have been exposed to many different research projects and all these projects can be classified into one of three ‘stages’ of research, namely, stage 0, the hypothesis generating stage, and the hypothesis testing stage. This classification is my own, though I draw from other authors, and is based on: (i) the depth of prior knowledge of the research aim; (ii) how specific research objectives are; and (iii) the sample size or expected power of the study. The different research stages make use of the same research cycle to gain more insight into the research question. The insight gained through each stage allows for a more refined study in the next, until research findings can become new knowledge.

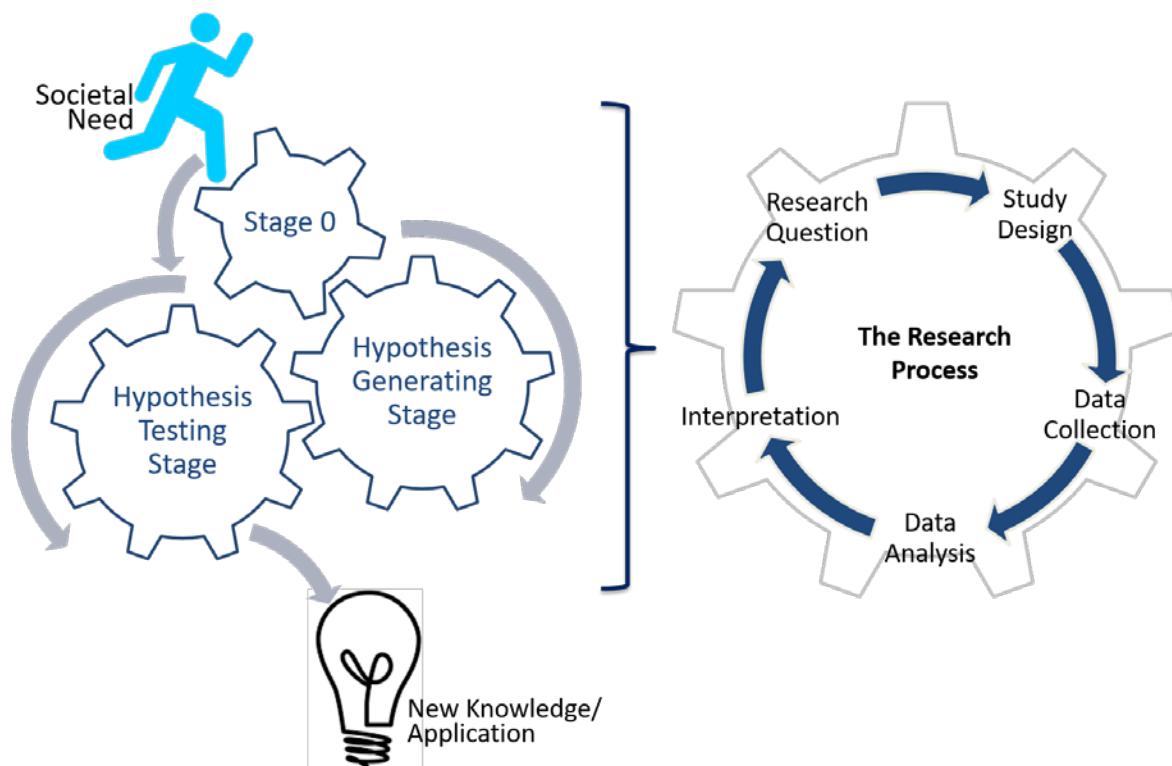


Figure 2-2: An interpretation of the metabolomics knowledge production process.

I include this idea of knowledge production in my thesis to show that the stage within which data are generated must guide the choice of statistical technique and most importantly the interpretation of results. I include the research cycle to contextualize some examples of 'The Gap', i.e. opportunities for statistics to contribute to metabolomics. The elements of Figure 2-2 will now be discussed in more detail:

- The source of the research question:

Metabolomics research is context driven, that is, research questions are formed from the real-world needs of a biological system, in this instance human beings or society. These questions, needs or problems trigger the knowledge production process in search of not only answers but solutions.

- The first stage of research:

Stage 0 represents research projects that are purely explorative in nature as little is known about any aspect of a societal need or problem. *"Exploratief onderzoek is nodig als theorie*

niet voor handen is, en wordt gebruikt om een bepaald fenomeen beter te leren begrijpen, en juist de theorie te ontwikkelen” (Timmerman, 2016): roughly translated, exploratory research is needed when theory is not yet established to better understand a phenomenon and develop sound theory. Stage 0 projects are often only interested in some quantification of variation, either naturally occurring, or variation in measurement, or from one condition to the next. The aim is purely to understand the phenomena better and so these studies are often based on small sample sizes. Once sufficient stage 0 research cycles have been completed to give some context to the problem, a more structured research question can be formulated as we progress to the next stage.

- The second stage of research:

The hypothesis generating stage involves additional and more controlled investigations as we start to progress from exploratory to confirmatory research. The hypothesis generating stage generates new questions, should previous stage 0 research findings be proven questionable. Alternatively or in addition, the hypothesis generating stage moves stage 0 findings, which are in essence only ‘hunches’, to ‘expert opinions’ and potentially even to hypotheses. Such hypotheses may then require further refinement or can be tested more rigorously, so moving to hypothesis testing. The hypothesis generating stage is required to reduce bias. More controlled experiments will reveal the presence of confounding factors, discussed later in the chapter.

- The third stage of research:

The hypothesis testing stage occurs only after the validation of findings from the preceding stages have led to a formal hypothesis which can now be tested. The aim of hypothesis testing research is then to solidify findings so they can become accepted new knowledge and, one hopes, produce new applications. Defining a population may not be essential in the preceding stages, given that a lack of generality does not disqualify the acceptance of results for the group studied, but it is crucial in hypothesis testing to clarify for whom the findings are valid. That is, the hypothesis testing stage proves or disproves the idea that findings can be generalized to other groups or even populations.

- A recursive process:

I purposefully distinguish between three stages because I believe that research cannot be limited to clinical trials (i.e. hypothesis testing). I disagree with the idea that underpowered yet well-designed studies serve no purpose. That said, I strongly assert that findings from stage 0 or even the hypothesis generating stage should not flow back to society without explicitly stating the need for further validation. These beliefs may seem contradictory but the view they support is singular: neither stage 0 nor the hypothesis generating stage can be avoided as the information they provide is key to accurately accept or reject hypotheses posed in more formal and larger trials. The knowledge production process I describe then assumes validation of findings through more than one independent investigation or research stage. In so doing I build on the view of Broadhurst and Kell (Broadhurst & Kell, 2006) and assume that to truly assess a model's validity requires independent data and to truly assess its generalizability requires independent study.

- The gear and its teeth:

Figure 2-2 also shows an abbreviated form of a more generic research cycle, repeating itself within each of the three stages of knowledge production. The remainder of this chapter will provide some insight into the research cycle by discussing steps in the cycle, or the teeth of the gears, linking to this and other chapters of my thesis – that is, I give an overview of study design to be elaborated on next, the complex characteristics of metabolomics data requiring new statistical approaches, as developed in Chapter 4, and popular data analysis methods, demonstrated in Chapter 3.

2.2.2 'Gaps' in the Research Cycle – Some Examples

Study Design

The design of research studies comprises three aspects: the research aim, the experimental design, and the measurement design. The literature on study design addresses these aspects but often by assuming: (i) some hypothesis is being tested; (ii) extensive knowledge of the expected size of the effect being measured; (iii) confounding factors are known; and (iv) measurement error can be randomized away or perfectly corrected for. Unfortunately, these assumptions do not hold water in most human metabolomics research studies. Research into the human metabolome is still highly exploratory and demands a more strategic approach to design,

one where compromise is central. This demands a fourth and formidable aspect: the design team. Each of the four design aspects are discussed in the subsections to follow with specific reference to human metabolomics research.

The Research Aim

Experimental and measurement design aspects can be viewed as a set of questions requiring answers. The context of and therefore answers to these questions are principally defined by the overall research aim and how absolute it is in nature. The research aim can be as explicit as a formal hypothesis statement or a more exploratory set of objectives. For example, one could set out to prove that a specific metabolite differs between two groups or one could simply be interested in understanding how the metabolism responds to a change in conditions. The stage of knowledge production will as such affect the design of a study. The same aspects may be considered whether the research is purely exploratory or confirmatory, but the attention and weight they receive will differ as the interpretation and implication of results will also differ.

The metabolome is a complex component of human life, expressing both genetic and environmental factors that produce metabolic profiles/fingerprints, at least to some extent, unique to each individual. The ambition of metabolomics is *“to understand how the overall metabolism of an organism has been changed under different conditions”* (Ren *et al.*, 2015). Human metabolomics research is then primarily hypothesis generating as we have only limited insight into how the human metabolome responds to different conditions. In addition, confounding factors are sundry and their bearing can be difficult to gauge, making it risky and even incorrect to set absolute research aims.

The design of research studies has received much attention in the hypothesis testing literature - focussing mostly on the estimation of a sufficient sample size to ensure the proposed statistical model is not under-powered; see for example Guo *et al.* (2010), Nyamundanda *et al.* (2013) and Saccenti & Timmerman (2016). Designing a study to understand the human metabolome is complex. It is not only a question of adequate sampling, but strategic sampling to uphold ethics and reason. Little may be known about the expected size of fluctuations in metabolic levels and even more so how the size of the effect compares with individual variation and measurement error. Designing a study to detect changes in the metabolome given a change in condition is not only a question of how many individuals to measure, but which individuals. The sections that follow on experimental and measurement design explore some of the more malleable aspects of

design for generating hypotheses. First, however, it is necessary to understand how a design can be malleable and who must mould it.

The Design Team

This aspect may not seem as unique to metabolomics as proposed earlier, but is in my opinion much more challenging in transdisciplinary settings such as metabolomics and therefore worth emphasising. Two challenges are primary. First, questions regarding design are posed to the design team, the members of which should include representatives from all disciplines required to collect samples, generate data, analyse data, and interpret results. In the case of metabolomics this can become a rather large group of people with often conflicting views and priorities. Second, though there exists a 'right answer' for most design questions, especially given the vast literature on study design, the right answer may not be realistically achievable. These challenges make designing a metabolomics study an exercise in compromise: how to find a design, perhaps less than optimal but achievable, without compromising the validity of the findings.

Such a negotiation is best understood through a case study, illustrated by the study on acute alcohol consumption (ALC2013). The ALC2013 study was an investigation into the effect of supplemented NAD (Nicotinamide adenine dinucleotide) in the presence of alcohol on the human metabolic system and generated the data analysed in Chapter 3, therefore making it appropriate for discussion here. The research aim was to determine how the human metabolome responds to exposure to NAD and alcohol in combination. The ALC2013 study illustrates the real-life dilemmas faced by a design team. Its members consisted of a PhD student, responsible for the collection and analysis of biological samples, an analytical chemist (advising and supervising the student on the chemical analysis of sample material), a statistician (responsible for data analysis), a biochemist (responsible for the interpretation of results), and a medical doctor (responsible for the safe collection of biological samples). In addition, experts from all fields represented in the design team were consulted to weigh in on design decisions. Examples of the difficult decisions faced by the team will be presented in the section to follow, while the consequence of the decisions made will become evident in Chapter 3 and will be discussed in Chapter 5.

Experimental Design

Experimental design is concerned with accurately capturing factors of interest while obtaining sufficient information on confounding factors to control their effect through the design. The idea of experimental design is to place as much emphasis as possible on the variation of interest by

optimally grouping observations or exposing research subjects to the optimal set of interventions in the optimal order and timing. This can only be achieved if noise and confounding factors are differentiated from one another and appropriately controlled or corrected for. Sources of noise should be identified and their impact minimized by randomizing across these sources as far as possible, through the design of measurements. Some confounding factors are, however, true confounders and play some role in the interpretation of the data and should be more carefully considered and, if possible, accurately captured to allow for bias corrections during statistical modelling.

Potentially the most influential source of noise and confounding in human metabolomics is the 'individual effect'. The term individual effect, as used here, represents two sources of variation: (i) the variation between individuals under similar conditions (e.g. variation at steady state or differences in response to treatment); and (ii) the variation between measurements on the same individual under similar conditions. These sources of variation can be classed as noise or truly confounding and the distinction is both important and tricky.

Individuals may differ from one another in negligible ways, for example, by exhibiting different metabolic responses due to differences in general lifestyle. But they may differ also in non-negligible ways, for example, exhibiting different metabolic responses due to underlying genetic variation. The severity of and consequently classification of these sources of variation as noise or true confounders depend very much on the specific individuals selected, making it difficult to design successful studies as the classification of confounders impacts on the number of measurements and participants required. The most challenging part of study design for hypothesis generating research in human metabolomics is the limited number of cases that can reasonably be obtained (i.e. the sample size).

Here we focus on two mainstream alternative approaches. One approach is to assume that differences between individuals under the same conditions are simply noise. Researchers make assumptions like diseased individuals all have 'the same disease' and controls are all in comparable health. However, it is well established that doing so may result in the loss of valuable information, large biases, and potentially wrong conclusions. This makes the alternative presented preferable, albeit more cumbersome.

The alternative approach I recommend considers the classification of variation a design challenge. For example, given the comments above, using one group of individuals as a control for another becomes dubious. So, wherever possible, individuals should be used as their own

controls. The design should then be further refined by collecting control and test samples from the same individual at each intervention because metabolomics is “...*the integrative readout of both genetic and environmental impacts...*” (Li *et al.*, 2017). In other words, the difference between measurements on the same individual under similar conditions is usually classed as noise by default and assigned to metabolic regulation, for example, yet environmental influences may be the drivers of the change and so add valuable information, thus qualifying for further investigation.

Statistics plays a vital role when making this decision but cannot override ethical consideration. In the ALC2013 study, for example, individuals were exposed to high levels of alcohol, sufficient to cause intoxication, with well-established disadvantages to health and well-being. Sample size therefore would inevitably be lower than required for multivariate statistical analysis. The design team decided on an intervention-based crossover design to increase power in another way, by controlling as many sources of unwanted variation as possible. A crossover design implied each participant could be his or her own control for each intervention, reducing the impact of intra- and inter-individual variation and the need for a larger sample size. Though this was a theoretically obvious decision, it did have cost implications. Participants would probably need some incentive to partake in all interventions. A total of four interventions measured over six hours each with a sufficient wash-out period implied a commitment from participants spanning several weeks. Ethical approval could be obtained for only 24 individuals so inter-individual variation had to be further reduced. This meant limiting the generalizability of results by setting strict selection criteria: (i) males only; (ii) 20 to 25 years of age; (iii) students living at the same university hostel and assumed to eat primarily at the hostel's cafeteria, to control diet and increase traceability of participants; and (iv) subjects in good general health while being habitual yet casual consumers of alcohol, to limit the possibility of any adverse effects.

Replication could also have been used to increase the statistical power without actually increasing the number of participating individuals. Replication can remove some of the unwanted variation by presenting the average of replicates for data analysis and in so doing emphasize the effect and indirectly increase the power. There are various levels of replication: at the lowest level the same sample is analysed multiple times (i.e. platform repeatability), that is, a single prepared sample or aliquot is analysed more than once by the same platform to reduce the impact of measurement error, as discussed in the next section. At the highest level, the same experiment is repeated on the same individuals. A higher level of replication is much more useful as more

sources of unwanted variation can be controlled for, but is not always possible due to ethical (as was the case here), time, cost and other constraints.

Individual variation is one aspect of study design for human metabolomics, but it should suffice to convey the complexity of experimental design, the difficulty faced by design teams, and the need for well-founded yet realistic statistical guidance. Finally, we consider measurement design.

Measurement Design

The nature and magnitude of variation in the data generation or measurement process can place metabolomics investigations back in the basic research stage - struggling to reproduce results and questioning whether any distinction between noise and effect is possible. Here only a specific selection of aspects is discussed to provide insight into the complexities of designing metabolomics studies to control measurement variation.

The measurement design is the set of steps planned to obtain biological samples and extract data from them. Measurement design is particularly concerned with managing unwanted variation - how to measure an experimental design without introducing noise that masks the variation of interest. The design should encompass, among other things: the choice of biological matrix; sample collection; sample transportation; sample storage; sample handling; sample analysis order; the number and size of analytical batches; the use and placement of quality control, replicate and blank samples; and the inclusion and choice of internal standard(s).

The collection, transportation, and storage of samples are easily taken for granted, even though these are essential aspects requiring careful planning, and are therefore used as a first example. These aspects can have a detrimental effect if even the smallest detail goes unplanned, especially given experimental designs which incorporate multiple collection sites and times. Sample handling is complementary to transportation and storage, but can also form part of sample preparation. For example, some understanding of the allowable storage temperature range and number of freeze/thaw cycles is definitely required as these factors have been shown to affect different biological matrixes and measurement platforms to a large extent. Selecting an appropriate biological matrix should be carefully considered as well, not only in terms of the classes of metabolites that can be measured, but also in terms of sample stability and the invasive nature of sample collection. A different matrix may be more informative, but if it cannot be collected, transported and stored in a standardized way, this benefit, as well as the variation of interest, may be lost.

The same can be said for the choice of analytical platform. Larger studies may be better served by less sensitive yet highly repeatable platforms. The benefit of measuring more of the metabolome should be carefully weighed against the risk of having no informative data due to noise. Van Batenburg *et al.* (2011) discuss the intricate measurement error structures that can occur for the most popular analytical platforms used in metabolomics research; the reader is referred to their publication for more details.

One approach to obtaining data that are statistically more powerful is through the use of replicates, as noted earlier. Using the average over multiple replicates, at whatever level, rather than a single value, reduces analytical variation. It is unfortunately not effective in all circumstances. If, for example, there is a platform drift present (i.e. variation increases for later analyses or with time), averaging may increase rather than decrease the inaccuracy of measurement. Drift can be corrected for, however, if the drift is not in and of itself accurately measured; such a correction will also contribute to the inaccuracy of measurement.

In the ALC2013 study the design team faced the harsh reality of analysing 576 samples to be collected over several weeks. The study was associated with a PhD student and was therefore subject to a limited time frame, as most projects are. The harsh reality was that analysing 576 on the available gas chromatography–mass spectrometry (GC–MS) platform would take roughly 480 hours. Samples would have to be split into batches as the student was not the only user of the platform. Working with such a large number of samples made sample mix-up a concern, so to assist the student all samples of an individual were analysed as a batch. This was justifiable since the intra-individual variation was the primary interest, but necessitated the inclusion of quality control (QC) samples to correct for any drifts within batches and shifts across batches. The benefits of replications could not be reaped given the sheer size of the analysis that would result. Ultimately, the analysis spanned 1176 samples and was performed over several months.

A collective term for the (im)perfection of measurement is ‘measurability’. The theory of measurability aims to reconstruct a signal as “...*there are no perfect instruments*” (Smilde *et al.*, 2002), but any reconstruction model of a signal is only as good as the data provided are representative of the true signal. In other words, though we may be able to correct for these unwanted sources of variation, these corrections are only as accurate as the measurements they are based on.

This conundrum was also encountered in the ALC2013 study. A pooled QC was considered the best option, but implied that some metabolites would become highly diluted. To overcome this,

randomly selected samples were analysed in duplicate in different batches to estimate correction factors better for low abundance peaks and assess the impact of QC corrections. After sample analysis it became apparent that the pooling effect of so many samples and the size of batches, requiring almost 48 hours per batch to analyse, resulted in large variations across QC samples, raising serious doubts about their ability to remove shifts and drifts from the data. Unfortunately, the selective use of replicates did not make these samples a viable alternative in isolation. Of the 24 GC–MS batches, 12 complied with the selection criteria based on the available QC and replicate data and were used in further analysis. It became clear that in order to reliably measure such a large number of samples (1176) over such a long period of time (10 months) would require a more controlled (i.e. minimal sample preparation) and repeatable (i.e. smaller shifts and drifts across and within batches) method of measurement, albeit less sensitive. These aspects, GC–MS batch selection and platform comparison, are fully described in the PhD thesis of Irwin (2018). Accordingly, nuclear magnetic resonance (NMR) was employed as a second platform as it not only meets the above criteria, but is also complementary to GC–MS with respect to metabolome coverage. The NMR data were indeed much more comparable across batches and allowed for the statistical analysis of the entire experimental design. A selection of publications resulting from both the NMR and GC–MS data sets is presented in Chapter 3.

Data Characteristics

The ability to measure the metabolome comprehensively is a great advantage. Unfortunately, the characteristics of the data are quite challenging to model statistically. These characteristics include:

- Multicollinearity:

Trainor *et al.* (2017) explain the origins of relationships among metabolites as follows: *“Metabolites in the same reaction or linked reactions function as substrates, intermediates, and products thus generating complex correlation structures”* (see also Camacho *et al.*, 2005). Metabolites can also regulate other reactions, for example, the rate of the reaction that produces one metabolite for another is influenced by the availability of a third. Steuer explains that these correlations or association are complex as they can be intrinsic (meaning no perturbation is required in their formation), non-linear, and change from one physiological state to the next (Steuer, 2006). Multicollinearity is also not restricted to the signal, but is also found in the noise, i.e. errors are correlated. Some reasons for this include metabolites that

co-elude, as well as deterioration in the performance of the analytical platform (van Batenburg *et al.*, 2011).

- High-dimensionality:

Metabolomics technologies are now able to measure with facility many hundreds, even thousands, of metabolites (Xia *et al.*, 2012). These numbers are also continually growing as analytical platforms become more sensitive. Though this is an essential advantage to truly capture the whole metabolome it does introduce more challenges. Most data sets consist of many times more variables (p) than observations (n), i.e. $n \ll p$ (Shulaev, 2006). This is problematic because: (i) many traditional statistical methods require at least as many observations as variables, i.e. $n \geq p$; (ii) multivariate models that can accommodate the $n < p$ scenario, are still prone to overfit (Hendriks *et al.*, 2011), especially when using supervised methods (Xi *et al.*, 2014); (iii) multivariate models are difficult to interpret, different models do not always agree (Kjeldahl & Bro, 2010); (iv) studies may be under-powered, i.e. unable to detect effects, or become unrealistic, making false discoveries; (v) even robust machine learning techniques may not be able to find a good model fit if observations are few and far between (Broadhurst & Kell, 2006); and (vi) if sample sizes are small due to disease prevalence, for example, models that are not correctly weighted become biased.

Combining data sets would bring n closer to p , although, such mergers are not trivial. For example, assigning metabolite names to observed spectral peaks is not a seamless nor an infallible step in the sample processing pipeline: “... *high confidence analyte assignments/identifications may not be made owing to the fundamental challenges of the metabolomic identification process*” (Schrimpe-Rutledge *et al.*, 2016). This means that even though different data sets contain the same names, they could still represent different metabolites. Re-aligning the raw spectra is a feasible practice in that peaks would be consistently identified across sets, but this is very time consuming and requires sharing much more information than most researchers are willing to do.

- Data types unusual to most data analysis methods:

Some analytical platforms, NMR technology for example, do not measure metabolites or complete variables, but rather fingerprints of variables. For example, one approach to align NMR spectra is to chop them into equal parts called bins, rather than selecting sections that represent metabolites. These bins are then aligned, quantified, and presented for data

analysis. Only after the statistically significant bins have been identified, are they converted to metabolite concentrations. The problem, in terms of data analytics, is that this is not a one-to-one conversion, meaning that once converted the resulting metabolites may no longer be statistically significant or vice versa. Nevertheless, NMR is one of the favoured approaches as measurement error is less, partly due to the limited sample preparation required, and samples can be re-analysed on other platforms as they are not destroyed during NMR analysis (Ren *et al.*, 2015). Alternative alignment strategies are available but binning is often preferred as it reduces the complexity of the data (Euceda *et al.*, 2015). New approaches to the analysis of these data types are therefore needed.

- Heteroscedasticity:

The absence of homoscedasticity (i.e. uncorrelated and uniform modelling errors) will result in the incorrect calculation of standard errors for modelling parameters. Metabolomics data are complex as variation in the quantification of peak areas is often related to intensity, that is, measurement errors are “... *proportional to intensity level at large intensity levels but constant at low intensity levels*” (van Batenburg *et al.*, 2011). Rocke-Lorenzato error models assume a constant variance for low intensities and increasing variance as intensity increases. That said, peaks closer to the detection limit are also problematic. As peaks drop below the limit of detection, values are set to ‘missing’ or assumed zero, which increases variability and introduces other difficulties, as discussed next. Modelling covariance structures can be altered to account for different associations between error terms, but necessitate the estimation of many additional parameters and so call for larger sample sizes. More primitive strategies in turn cause other problems. For example, assuming less complex and less accurate covariance structures can lead to false discoveries (Scariano & Davenport, 1987), while estimating parameters using an insufficient number of observations implies that models will not generalize well due to wide confidence intervals or the size of effects may be overstated (Ioannidis, 2008).

- Large proportions of missing values:

Data sets, especially when generated by gas chromatography–mass spectrometry, can contain large proportions of missing values. Many metabolites may have more missing values than non-missing ones. The source of these missing values can be analytical, biological, or both. Values can be missing at random, with a specific structure, or both. This makes imputing missing values exceedingly difficult since neither the source nor the structure of a specific

missing value is known. Imputation can also have a substantial effect given the large proportions of observations requiring imputation, especially if the associations between metabolites cannot be accounted for. Missing values, their origins, and a new approach that can serve as a remedy in some instances, will be discussed further in Chapter 4; for now, it is important to note that the accuracy of estimates can be adversely affected by large proportions of missing values or inappropriate missing value imputation strategies.

- General uncertainty:

The data may contain other artefacts that can cause erroneous results. The quantified levels of a metabolite may not always be associated with its importance in explaining observed variations. Some metabolites are more prone to variation, be it natural variation - "...metabolites in the central metabolism are relatively constant, while... metabolites... of the secondary metabolism usually show much larger differences..." - or induced variation where metabolites fluctuate more dramatically than others under identical experimental conditions (van den Berg *et al.*, 2006). Analytical errors may also be correlated with time and concentration, meaning that the analytical error is confounded within the biological variation. Finally, the presence of a matrix effect, where the accuracy of measuring one metabolite is affected by the presence of another, is somewhat inevitable (Hendriks *et al.*, 2011).

What makes metabolomics data so challenging is that these characteristics are often all present in a single data set. These structures, hidden or known, will result in unreliable statistical conclusions if not appropriately pre-treated or modelled. Unfortunately, such inaccuracies may come to light only when attempting to validate findings.

Data Analysis

There is a wide range of statistical methods that can accommodate different data characteristics, but there is no single method that can accommodate all the characteristics of metabolomics data. Table 2-1 provides an overview of statistical methods popular in metabolomics data analysis, specifically when considering the binary classification problem noted in Section 1.2. For a less specific overview of methods, the reader is referred to Worley & Powers, 2013.

Methods were included if they were noted in the literature overviewing data analysis techniques for metabolomics or in a software package developed specifically for the analysis of metabolomics data. For a more in-depth description and discussion of methods, the reader is referred to the

publications of Barnes *et al.* (2016), Ren *et al.* (2015), Sugimoto *et al.* (2012), Hendriks *et al.* (2011), and Madsen *et al.* (2010). For an overview of available software packages, the reader is referred to the publications of Costa *et al.* (2016) and Xia *et al.* (2015).

Table 2-1 lists, groups, and briefly discusses some of the aspects of data analysis methods popular for the analysis of metabolomics data. The first column (Approach) groups the methods listed in the second column (Method) by the number of variables considered per analysis, as described in the short definition provided. The third column (Comments) notes some of the assumptions and limitations of each method.

From Table 2-1 it is evident that the data analysis methods used range from basic univariate statistics to highly sophisticated machine learning techniques. It is common practice to apply a wide range of methods to uncover and understand different structures in the data, as well as to allow for biological interpretation (Saccenti *et al.*, 2013): *"...based on different goals and designs of studies, it may be necessary to use a variety of data analysis methods or a combination of them in order to obtain an accurate and comprehensive result"* (Ren *et al.*, 2015). Each method has its own assumptions, however few or easy to overcome, and limitations (as touched on in the third column of Table 2-1), which one cannot lose sight of. These assumptions and limitations, along with the characteristics of metabolomics data, emphasize the existence of 'The Gap' in metabolomics research, i.e. the need for modified or new approaches to data analysis.

Each method also places emphasis on a set of variables with particular characteristics so that the order in which these techniques are applied cannot be standardized. Statistical data processing and analyses are then *"...both very powerful and very dangerous, and it is all too easy to produce clusters or models that are simply statistical artefacts"* (Kell, 2007). It is then essential to consider both the biological and analytical contexts to ensure the right decisions are made: *"A good metabolomics study requires knowledge in many areas, such as clinical science, analytical chemistry, data analysis and pathology, thus close collaboration between several researchers is therefore critical"* (Madsen *et al.*, 2010).

Approach	Methods	Comments
Univariate (Considers one variable at a time)	Fold Change	Assumes the mean is a good measure of centrality within each group (i.e. assumes distributional symmetry) and does not take variation around the mean into account. The size of the sample is also not taken into consideration.
	Paired or Independent t-Test	Variable transformation required or the use of non-parametric alternatives (e.g. Mann-Whitney or Wilcoxon signed rank tests). It is essential to correct for multiple testing by either controlling the family-wise error rate or false discovery rate.
	Effect Size	Effect sizes give an indication of practical significance as opposed to statistical significance, making them comparable across studies, but to do so the size of the sample is not considered. Importantly, effect sizes are associated with a significance test (parametric or non-parametric) and as such must be used on the same data (albeit transformed data).
	Volcano Plot	Combines t-test (or Mann-Whitney) p-values with fold change values as variable selection criteria. Importantly, p-values must be corrected for multiple testing and users should consider substituting fold change with effect size based on the limitations noted above.
Multivariate (Considers two or more variables at a time but p cannot exceed n)	Pearson's Correlation Coefficient	Only detects linear relationships between two variables at a time. Transformation of variables is required as normality is assumed and non-parametric alternatives are usually based purely on rank, ignoring the size of differences between observations.
	(RM) MANOVA	(RM) MANOVA, i.e. (Repeated Measures) Multivariate Analysis of Variance: Assumes a linear relationship between variables. Assumes predictor variables are not highly correlated. Prior supervised variable selection may be required. Normality is assumed so variable transformation or the use of a non-parametric alternative may be required. Applying ANOVA models to the binary classification problem may provide a different view on the data as it is not based on a ratio of means to standard deviations, like a t-test, but on the ratio of between sample variance to within sample variance
	Logistic Regression	Assumes independent variables are not highly correlated. Prior supervised variable selection may be required.
	Mixed Effects	Error structures should be carefully selected and not over-simplified.
	Linear Regression	Assumes a linear relationship between variables. Assumes predictor variables are not highly correlated and normality. Applying linear regression a binary classification problem is possible, but logistic regression or other classification methods are preferred.

Approach	Methods	Comments
Multivariate (Considers two or more variables at a time but p cannot exceed n)	L(Q)DA	L(Q)DA, i.e. Linear (or Quadratic) Discriminant Analysis Assumes linearity, multivariate normality, and that group covariance matrices are equal and of full rank (i.e. no missing values). Supervised method and so validation is needed as models are prone to overfit. QDA is similar to LDA but more flexible as it does not assume equal covariance matrices and thus can learn linear, as well as quadratic boundaries.
	CCA	CCA (Canonical Correlation Analysis) Primary aim: to integrate two sets of variables (measured for the same set of cases). For binary classification – discriminant analysis is a special case of CCA (continuous independent variable set and a binary dependent set). Can be sensitive to collinearity in the independent variable set (Gromski <i>et al.</i>, 2015).
	Heatmaps	Variable scaling required since larger magnitudes will dominate. Top 3 guidelines for using heatmaps by (Bojko, 2009): (i) "Generate heatmaps only if they add value to the research"; (ii) "Use heatmaps for data visualization instead of data analysis"; and (iii) "Use heatmaps to support quantitative analysis rather than on their own". Though p can exceed n, it becomes almost impossible to interpret a mega-variate heatmap and so it is viewed as a multivariate method
Mega-variate (Considers more than one variable at a time and p can exceed n)	Cluster Analysis	Variable scaling is required since larger magnitudes will dominate. Take care to use distance measures that compensate for the correlations between variables. All influential variables must be included to avoid misleading solutions.
	PCA	Principal Component Analysis (PCA): Will not be sufficient to explore variation when correlations are not linear. Also, given metabolomics data, latent variable selection should explore lower ranking components since they may not necessarily be projections of noise, as expected in PCA. That is, lower ranking components may not explain the largest amount of variation, but may well explain the variation of interest as this method is unsupervised. Variable transformation required since larger magnitudes will dominate and orthogonal components implicitly assume multivariate normality.
	(O)PLS(-DA)	(O)PLS(-DA), i.e. (Orthogonal) Partial Least Squares (-Discriminant Analysis): Variable scaling is required since larger magnitudes will dominate. This is a supervised method and so is prone to overfit, making model validation for prediction essential.

Approach	Methods	Comments
Mega-variate (Considers more than one variable at a time and p can exceed n)	ASCA	ANOVA Simultaneous Component Analysis (ASCA): Variable scaling required as larger magnitudes will dominate.
	Neural Networks	Less restrictive in terms of assumptions such as linearity, normality, and independence of predictor variables. Black box with limited information on the model. It is difficult to determine what a sufficient sample size would be as it depends on the complexity of a model not yet constructed. These characteristics imply that it is difficult then to decide if the model is bad or just requires more training data.
	Random Forest	Used to build predictive rather than descriptive models. Supervised method and so prone to overfit. Associations between variables are still predominantly modelled as linear. It is difficult to determine what a sufficient sample size would be as it depends on the complexity of a model not yet constructed.
	SOM	Self-Organizing Maps (SOM): Time consuming in high dimensional settings. Prone to converge on local minima. Predominantly an exploratory method.
	SVM	Support Vector Machines (SVM): Mahadevan <i>et al.</i> (2008) highlight three strengths: (i) models generalize well; (ii) models can be extended to take nonlinear relationships between variables into account; (iii) models are robust to noise and outliers. Supervised method and so prone to overfit. It is difficult to determine what a sufficient sample size would be as it depends on the complexity of a model not yet constructed.

Table 2-1: Popular methods used to compare two groups based on metabolomics data.

2.3 Roadmaps

"The application of metabolomics is growing rapidly, but many challenges must first be addressed before it can reach its true potential"
(Tripathy K, 2011).

I conclude this chapter with a roadmap of the remainder of this thesis, as well as the route I envision to a brighter future in metabolomics and statistics research.

Most metabolomics studies are either more exploratory and *"aim to understand biological processes"* or more confirmatory and aim *"to develop biomarkers"* (Xia *et al.*, 2013). To achieve either demands a transdisciplinary approach, therefore the thesis continues by exploring the workings of a metabolomics team. Chapter 3 will show that the statistical analysis of data collected for exploratory research can be more flexible and combines multiple approaches to aid biological interpretation. Chapter 4 describes new statistical methods which can be applied at any stage, but are more powerful for confirmatory research as model outputs can be directly implemented in a clinical setting as biomarkers.

A condemning criticism of metabolomics is that findings cannot be validated or do not validate in subsequent investigations. The generalizability of findings can be negatively affected by well-known factors such as small sample sizes and large biological variation between groups. What must not be underestimated are the effects of more intricate details such as compromises in the study design and the complexity of a disease which does not always comply with a one-rule-fits-all approach. The consequence has been that few results from metabolomics investigations have persevered beyond proof of concept and even fewer are actually being used routinely. Unfortunately, most of the discoveries made through metabolomics research, though promising, have remained exactly that, promises. This is most surely metabolomics' greatest challenge, but is in truth an amalgamation of the many smaller hurdles just discussed.

In my view the primary reason underpinning all others is this - the accurate interpretation of results requires the active participation and expertise of all the disciplines involved in its production. Our approach to generating and understanding data is not yet truly transdisciplinary. Yener's comment on reductionist science in a holistic science like metabolomics highlights this for me: *"...we've taken our biological machines apart but can't put them back together - our ability*

to accumulate reductionist data has outstripped our ability to understand it" (Yener *et al.*, 2008). The reliance on a biological understanding of results, in my opinion, hinders the clinical application of it. For example, univariate and multivariate methods are usually applied in combination and this is a good approach as the one complements the other (Saccenti *et al.*, 2014). However, the emphasis is often placed on univariate results simply because they are easier to interpret and communicate. Yet, by its own definition, metabolomics *"...recognizes that these biological phenomena can only be characterized by the interrelation of metabolites"* (Saccenti *et al.*, 2014). We need to trust the analytical chemists and statisticians just as much as we do the biologists. All three disciplines should play equal parts in the metabolomics team, allowing the use of complex models with high predictive ability even though we might still not understand fully the underlying biology.

The aim of this thesis is to show how statistics (assuming the more inclusive definition advocated here) can be applied in novel ways to play its part in the metabolomics team to overcome some of the challenges faced, thereby emphasizing that metabolomics is a discipline in its own right, formed through the collaborative efforts of a combination of disciplines. It is not merely an application rooted in one discipline, simply making reference to other disciplines, but truly a combination of biology, analytical chemistry, and statistics. In so doing I hope also to convincingly substantiate my view on statistics, that it is truly beneficial to bend the spectrum along which statistics is practised so that advanced application (within a given context) can result in the advancement of theory.

2.4 References

- Barnes, S., Benton, H.P., Casazza, K., Cooper, S.J., Cui, X., Du, X., Engler, J., Kabarowski, J.H., Li, S., Pathmasiri, W., Prasain, J.K., Renfrow, M.B. & Tiwari, H.K. 2016. Training in metabolomics research. II. Processing and statistical analysis of metabolomics data, metabolite identification, pathway analysis, applications of metabolomics and its future. *Journal of mass spectrometry*, 51(8):535-548.
- Bojko, A. 2009. Informative or misleading? Heatmaps deconstructed. (*In* International Conference on Human-Computer Interaction. Berlin: Springer. p. 30-39.)
- Breiman, L. 2001. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199-231.
- Breiman, L. 2002. Statistical fraud detection: a review - comments and rejoinders. *Statistical science*, 17(3):249-255.
- Broadhurst, D.I. & Kell, D.B. 2006. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics*, 2(4):171-196.
- Camacho, D., De La Fuente, A. and Mendes, P. 2005. The origin of correlations in metabolomics data. *Metabolomics*, 1(1):53-63.
- Chambers, J.M. 1993. Greater or lesser statistics: a choice for future research. *Statistics and computing*, 3:182-184.
- Chargaff, E. 1975. A fever of reason: the early way. *Annual review of biochemistry*, 44(1):1-20.
- Ciesielski, T.H., Aldrich, M.C., Marsit, C.J., Hiatt, R.A. & Williams, S.M. 2016. Transdisciplinary approaches enhance the production of translational knowledge. *Translational research*, 182:123-134.
- Coetzer, R.L.J. & de Jongh, P.J. 2016. Discussion of "Industrial statistics: The challenges and the research". *Quality Engineering*, 28(1):63-68.
- Costa, C., Maraschin, M. & Rocha, M. 2016. An R package for the integrated analysis of metabolomics and spectral data. *Computer methods and programs in biomedicine*, 129:117-124.

Euceda, L.R., Giskeødegård, G.F. & Bathen, T.F. 2015. Preprocessing of NMR metabolomics data. *Scandinavian journal of clinical and laboratory investigation*, 75(3):193-203.

Fisher, R.A. 1922. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society London A: Mathematical, physical and engineering sciences*, 222(594-604):309-368.

Goodacre, R. 2005. Making sense of the metabolome using evolutionary computation: seeing the wood with the trees. *Journal of experimental biology*, 56(410):245-254.

Gromski, P.S., Muhamadali, H., Ellis, D.I., Xu, Y., Correa, E., Turner, M.L. & Goodacre, R., 2015. A tutorial review: Metabolomics and partial least squares-discriminant analysis – a marriage of convenience or a shotgun wedding. *Analytica chimica acta*, 879:10-23.

Guo, Y., Graber, A., McBurney, R.N. & Balasubramanian, R. 2010. Sample size and statistical power considerations in high-dimensionality data settings: a comparative study of classification algorithms. *BMC bioinformatics*, 11:447-466.

Hand, D.J. 2009. Modern statistics: the myth and the magic. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(2):287-306.

Hendriks, M.M.W.B., Van Eeuwijk, F.A., Jellema, R.H., Westerhuis, J.A., Reijmers, T.H., Hoefsloot, H.C.J. & Smilde, A.K. 2011. Data-processing strategies for metabolomics studies. *TrAC trends in analytical chemistry*, 30(10):1685-1698.

Investopedia. 1999. Skin in the game. <https://www.investopedia.com/terms/g/gaming-industry-etf.asp> Date of access: 20 March 2018.

Ioannidis, J.P. 2008. Why most discovered true associations are inflated. *Epidemiology*, 19(5):640-648.

Irwin, C. 2018. A metabolomics investigation on experimental interventions of acute alcohol consumption. Potchefstroom: NWU. (Dissertation –Ph.D.).

Kell, D.B. 2007. Metabolomic biomarkers: search, discovery and validation. *Expert review of molecular diagnostics*, 7(4):329-333.

- Kell, D.B. & Oliver, S.G. 2004. Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *Bioessays*, 26(1):99-105.
- Kimball, A. 1957. Errors of the third kind in statistical consulting. *Journal of the American Statistical Association*, 52(278):133-142.
- Kjeldahl, K. & Bro, R. 2010. Some common misunderstandings in chemometrics. *Journal of chemometrics*, 24(7-8):558-564.
- Korte, R. 2016. From methodology to imagination. *Human resource development review*, 15(1):123-126.
- Krastanov, A. 2010. Metabolomics - the state of art. *Biotechnology & biotechnological equipment*, 24(1):1537-1543.
- Li, B., He, X., Jia, W. & Li, H. 2017. Novel applications of metabolomics in personalized medicine: a mini-review. *Molecules*, 22(7):1173.
- MacGregor, J.F. 1997. Using on-line process data to improve quality: challenges for statisticians. *International statistical review*, 65(3):309-323.
- Madsen, R., Lundstedt, T. & Trygg, J. 2010. Chemometrics in metabolomics - a review in human disease diagnosis. *Analytica chimica acta*, 659(1-2):23-33.
- Mahadevan, S., Shah, S.L., Marrie, T.J. & Slupsky, C.M. 2008. Analysis of metabolomic data using support vector machines. *Analytical chemistry*, 80(19):7562-7570.
- Malley, J.D. & Moore, J.H. 2013. The disconnect between classical biostatistics and the biological data mining community. *BioData mining*, 6:12.
- Nowotny, H., Scott, P. & Gibbons, M. 2003. Introduction: Mode 2' Revisited: the new production of knowledge. *Minerva*, 41(3):179-194.
- Nyamundanda, G., Gormley, I.C., Fan, Y., Gallagher, W.M. & Brennan, L. 2013. MetSizeR: selecting the optimal sample size for metabolomic studies using an analysis based approach. *BMC bioinformatics*, 14:338-346.

Parzen, E. 1998. Data mining, statistical methods mining, and history of statistics. (*In* Scott, D. Computing science and statistics. Fairfax Station, Va.: Interface Foundation of North America. p. 365-374.)

QuoteHD. 2017. Henri Poincare quotes. <http://www.quotehd.com/quotes/henri-poincare-mathematician-mathematicians-do-not-study-objects-but-relations>. Date of access: 20 March 2018.

Rao, C.R. 2001. Statistics: reflections on the past and visions for the future. *Communications in statistics - theory and methods*, 30(11):2235-2257.

Ren, S., Hinzman, A.A., Kang, E.L., Szczesniak, R.D. & Lu, L.J. 2015. Computational and statistical analysis of metabolomics data. *Metabolomics*, 11(6):1492-1513.

Saccenti, E., Hoefsloot, H.C.J., Smilde, A.K., Westerhuis, J.A. & Hendriks, M.M.W.B. 2014. Reflections on univariate and multivariate analysis of metabolomics data. *Metabolomics*, 10(3):361-374.

Saccenti, E. & Timmerman, M.E. 2016. Approaches to sample size determination for multivariate data: applications to PCA and PLS-DA of omics data. *Journal of proteome research*, 15(8):2379-2393.

Scariano, S.M. & Davenport, J.M. 1987. The effects of violations of independence assumptions in the one-way ANOVA. *American statistician*, 41(2):123-129.

Schrimpe-Rutledge, A.C., Codreanu, S.G., Sherrod, S.D. & McLean, J.A. 2016. Untargeted metabolomics strategies - challenges and emerging directions. *Journal of the American Society for Mass Spectrometry*, 27(12):1897-1905.

Shulaev, V. 2006. Metabolomics technology and bioinformatics. *Briefings in bioinformatics*, 7(2):128-139.

Silka, L. 2013. "Silos" in the democratization of science. *DEMESCI: International Journal of deliberative mechanisms in science*, 2(1):1-14.

Smilde, A.K., Van den Berg, F.W.J. & Hoefsloot, H.C.J. 2002. Peer reviewed: How to choose the right process analyzer. *Analytical chemistry*, 74(13):368A-373A.

- Steuer, R. 2006. Review: on the analysis and interpretation of correlations in metabolomic data. *Brief bioinformatics*, 7(2):151-158.
- Stichweh, R. 2001. History of scientific disciplines. (In Smelser, N.J. & Baltes, P.B., eds. International encyclopedia of the social and behavioral sciences. Amsterdam: Elsevier. p. 13727-13731.)
- Stigler, S.M. 1973. Studies in the history of probability and statistics. XXXII: Laplace, Fisher, and the discovery of the concept of sufficiency. *Biometrika*, 60(3):439-445.
- Sugimoto, M., Kawakami, M., Robert, M., Soga, T. & Tomita, M. 2012. Bioinformatics tools for mass spectroscopy-based metabolomic data processing and analysis. *Current bioinformatics*, 7(1):96-108.
- Timmerman, M.E. 2016. Van optimalisatie naar begrip en weer terug. Groningen: Rijksuniversiteit Groningen.
- Trainor, P.J., DeFilippis, A.P. & Rai, S.N. 2017. Evaluation of classifier performance for multiclass phenotype discrimination in untargeted metabolomics. *Metabolites*, 7(2):30.
- Tripathy, K.N.T., Chinnapatlola, S. & Sudharani, O.V. 2011. Metabolomics and its value in today's era. *Journal of computer science & systems biology*, S13(001).
- Van Batenburg, M.F., Coulier, L., Van Eeuwijk, F., Smilde, A.K. & Westerhuis, J.A. 2011. New figures of merit for comprehensive functional genomics data: the metabolomics case. *Analytical chemistry*, 83(9):3267-3274.
- Van den Berg, R.A., Hoefsloot, H.C., Westerhuis, J.A., Smilde, A.K. & Van der Werf, M.J. 2006. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC genomics*, 7(1):142.
- Waters, D.J. 2012. The paradox of tethering: key to unleashing creative excellence in the research-education space. *Informing science, the international journal of an emerging transdiscipline*, 15:229-245.
- Worley, B. & Powers, R., 2013. Multivariate analysis in metabolomics. *Current Metabolomics*, 1(1):92-107.

Wright, D. 2006. The art of statistics: a survey of modern statistics. 2nd ed. Mahwah, N.J.: Erlbaum.

Xi, B., Gu, H., Baniasadi, H. & Raftery, D. 2014. Statistical analysis and modeling of mass spectrometry-based metabolomics data. *Mass spectrometry in metabolomics*, 1198:333-353.

Xia, J., Broadhurst, D.I., Wilson, M. & Wishart, D.S. 2013. Translational biomarker discovery in clinical metabolomics: an introductory tutorial. *Metabolomics*, 9(2):280-299.

Xia, J., Mandal, R., Sinelnikov, I.V., Broadhurst, D. & Wishart, D.S. 2012. MetaboAnalyst 2.0 - a comprehensive server for metabolomic data analysis. *Nucleic acids research*, 40(W1):W127-W133.

Xia, J., Sinelnikov, I.V., Han, B. & Wishart, D.S. 2015. MetaboAnalyst 3.0 - making metabolomics more meaningful. *Nucleic acids research*, 43(W1):W251-W257.

Yener, B., Acar, E., Aguis, P., Bennett, K., Vandenberg, S.L. & Plopper, G.E. 2008. Multiway modeling and analysis in stem cell systems biology. *BMC systems biology*, 2:63-69.

CHAPTER 3 – EXAMPLES OF THE ADVANCE APPLICATION OF STATISTICS IN HYPOTHESIS GENERATING RESEARCH

3.1 Background

This chapter contains three publications resulting from the ALC2013 intervention study, rooted in different subsets of primarily NMR (Nuclear Magnetic Resonance), but also GC–MS (Gas chromatography–mass spectrometry) data. The ALC2013 study aimed to identify significant changes in the human metabolic system over a set period of time (4 hours) due to the intentional consumption of flavoured water with or without alcohol and/or NAD. Specifically, a longitudinal crossover design was employed to assess changes over time due to four interventions: (i) Vehicle (flavoured water); (ii) Vehicle & Alcohol; (iii) Vehicle & NAD; and (iv) Vehicle, NAD & Alcohol. Six urine samples were collected from each participant after each intervention, as described in Table 3-1.

Time (hours from consumption of vehicle and/or alcohol)	Description with respect to	
	Vehicle and Alcohol & Vehicle	Vehicle & NAD and Alcohol, Vehicle & NAD
Time -1	1 hour prior to intervention	Just prior to NAD consumption
Time 0	Just before consumption	1 hour after NAD consumption
Time 1	1 hour after consumption	2 hours after NAD consumption
Time 2	2 hours after consumption	3 hours after NAD consumption
Time 3	3 hours after consumption	4 hours after NAD consumption
Time 4	4 hours after consumption	5 hours after NAD consumption

Table 3-1: Description of the Time factor in the ALC2013 intervention study.

NAD was consumed one hour earlier, where relevant, so as to achieve an expected maximum effect at Time 1, on average, which was expected to coincide with the expected maximum effect of alcohol, on average, also at Time 1. Samples were collected over multiple weeks subject to participant availability and a required minimum washout period of seven days.

Given ethical, financial, and time constraints a total of 24 individuals ultimately participated in the study, each participating in four interventions and providing six samples per intervention, culminating in 24 samples in total for each participant. The power of the study, though limited,

was sufficient to detect many effects, as is evident from the publications to follow, while ensuring the necessary precautions were taken to avoid false discoveries.

Overview of publications

- The effect of flavoured water was not primarily of interest, however it was consumed throughout the study and was as such, a confounding factor, making an investigation into its effect obligatory. A vehicle was required for the following reasons: (i) participants would find it difficult to consume undiluted alcohol; (ii) the NAD tablet would be difficult to ingest unaccompanied by a liquid; and (iii) participants had to be allowed to consume liquids throughout the course of the experiment to avoid dehydration and ensure sufficient urine production. Flavoured water was preferred to water since the alcohol used, vodka, has a very strong taste making it difficult to consume in large quantities. Interestingly, though the investigation was intended more to identify and correct for the interfering effect of the vehicle, the results proved so informative that it justified a publication in its own right. From a statistical point of view, the most interesting aspect was the inter- and intra-individual variation, how delicately this should be investigated and how risky such investigative work becomes when ventured at by statisticians alone. This then is the first publication presented in this chapter (Section 3.2): Accepted Publication: Contribution towards a Metabolite Profile of the Detoxification of Benzoic Acid through Glycine Conjugation: An Intervention Study (Irwin *et al.*, 2016).
- The second publication started to investigate changes in the human metabolism due to the sudden consumption of a large volume of alcohol with or without NAD. The NMR data was explored again, but from a different perspective using a different subset of data. Time 0 was used as a point of reference to measure metabolic change at an individual level. This and the first publication noted above, resulted in part due to the application of a wide range of statistical techniques, some far more advanced than others, but all essential to deliver a full-colour picture of the data for biological interpretation. The effect of time and multiple interventions is especially intricate to analyse from all perspectives - chemical, statistical and biological - and only through the team effort presented here could the NMR publications have resulted. I purposefully use the word 'started' in the first sentence of this paragraph as we barely scratched the surface with these initial investigations, much more can and will be done (refer to Chapter 5). That said, so much was already revealed that it justified a second publication (Section 3.4): Accepted Publication: The ^1H -NMR-based metabolite profile of

acute alcohol consumption: a metabolomics intervention study (Irwin *et al.*, 2018 – still in press).

- Finally, for the students involved in this project, which include myself, it was important to also explore the GC–MS data even though the NMR data was not yet mined to its full potential. A GC–MS approach is so distinct in terms of sample and data analysis that exposure to this provided too valuable a learning opportunity to pass by. The most obvious challenge from a statistical point of view, was the large proportion of zero-valued observations in the data. In the resulting publication this was addressed through imputation. Specifically, the sources of the zero-valued observations were assumed to result from the limit of detection of the GC–MS platform, thus truncating the distribution of values of each variable at an unknown lower limit. A beta distribution was used to ‘restore’ the lower tail of the distribution of each variable based on the non-zero observations and their minimum. As some fit estimation was required, only variables with sufficient non-zero values were retained and imputed. The imputation was successful in the sense that biologically important information could be extracted from the data. Nonetheless, a different approach could just as well have revealed different, but equally important biological information. Working with GC–MS data was then the inspiration behind the development of the novel statistical approaches presented in Chapter 4. The small sample size of this data set should not lead to the underestimation of its value, especially for the planning of future studies (to be discussed further in Chapter 5). The statistical and biological analysis of the GC–MS data (resulting from careful chemical analysis) beautifully illustrates the benefit of ‘hyphenated’ work (as described in Chapter 2) as it lead to the following publication in a Nature affiliated journal (Section 3.6): Accepted Publication: GC–MS-based urinary organic acid profiling reveals multiple dysregulated metabolic pathways following experimental acute alcohol consumption (Irwin *et al.*, 2018).

This chapter then serves firstly as a continuation of the discussion on transdisciplinarity, initiated in Chapter 2. This chapter illustrates the fine line between mindful statistical analysis and data dredging, which can only be avoided through a multidisciplinary team. The transdisciplinary workings of a metabolomics team to design metabolomics studies, analyse the resulting data, and interpret important findings are presented here as a final proof-of-concept of the ideas advocated in Chapter 2.

Secondly, it illustrates the advanced, yes, but also the broad spectrum of statistical methods required in isolation, but more in combination, to aid interpretation. The ALC2013 data sets follow

a rather complex design especially when considering the vast number of variables measured by both NMR and GC–MS platforms. Various approaches to data analysis could have been and could still be followed, for example, the entire design could have been analysed as a whole. A very limited number of data analysis methods exist that can achieve this. In fact, perhaps only one method is capable of this at a mega-variate level, the ASCA method. That said, ASCA too has its limitations in that it is a linear model, which cannot as yet be used to perform mixed model analysis. Furthermore, biological interpretation based on this model alone would have been quite difficult, so prompting new and new combinations of approaches. This chapter is then an excellent example of statistics on the ‘advanced application’ side of ‘The Gap’, as defined in Chapter 2.

Note that the publications presented in this chapter were altered somewhat to conform to the template used for this thesis.

3.2 Accepted Publication: Contribution towards a Metabolite Profile of the Detoxification of Benzoic Acid through Glycine Conjugation: An Intervention Study

3.2.1 Title Page

Contribution towards a Metabolite Profile of the Detoxification of Benzoic Acid through Glycine Conjugation: An Intervention Study

Authors:

Cindy Irwin^{1#}, Mari van Reenen^{1,2#}, Shayne W. Mason¹, Lodewyk J. Mienie¹, Johan A. Westerhuis^{2,3} and Carolus J. Reinecke^{*1}

#Set of equal first authors:

These authors contributed equally to this work:

Cindy Irwin email: cindyirwin74@gmail.com

Mari van Reenen email: 12791733@nwu.ac.za

***Corresponding author:**

Carolus J. Reinecke email: carools.reinecke@nwu.ac.za

Addresses:

¹Centre for Human Metabolomics, Faculty of Natural Sciences, North-West University (Potchefstroom Campus), Potchefstroom, South Africa

²Department of Statistics, Faculty of Natural Sciences, North-West University (Potchefstroom Campus), Potchefstroom, South Africa

³Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam, The Netherlands

Other email addresses:

Shayne W. Mason: nmr.nwu@gmail.com

Lodewyk J. Mienie: japie.mienie@gmail.com

Johan A. Westerhuis: j.a.westerhuis@uva.nl

3.2.2 Abstract

Benzoic acid is widely used as a preservative in food products and is detoxified in humans through glycine conjugation. Different viewpoints prevail on the physiological significance of the glycine conjugation reaction and concerns have been raised on potential public health consequences following uncontrolled benzoic acid ingestion. We performed a metabolomics study which used commercial benzoic acid containing flavoured water as vehicle for designed interventions, and report here on the controlled consumption of the benzoic acid by 21 cases across 6 time points for a total of 126 time points. Metabolomics data from urinary samples analysed by nuclear magnetic resonance spectroscopy were generated in a time-dependent crossover study. We used ANOVA-simultaneous component analysis (ASCA), repeated measures analysis of variance (RM-ANOVA) and unfolded principal component analysis (unfolded PCA) to supplement conventional statistical methods to uncover fully the metabolic perturbations due to the xenobiotic intervention, encapsulated in the metabolomics tensor (three-dimensional matrices having cases, spectral areas and time as axes). Identification of the biologically important metabolites by the novel combination of statistical methods proved the power of this approach for metabolomics studies having complex data structures in general. The study disclosed a high degree of inter-individual variation in detoxification of the xenobiotic and revealed metabolic information, indicating that detoxification of benzoic acid through glycine conjugation to hippuric acid does not indicate glycine depletion, but is supplemented by ample glycine regeneration. The observations lend support to the view of maintenance of glycine homeostasis during detoxification. The study indicates also that time-dependent metabolomics investigations, using designed interventions, provide a way of interpreting the variation induced by the different factors of a designed experiment – an approach with potential to advance significantly our understanding of normal and pathophysiological perturbations of endogenous or exogenous origin.

Keywords

Crossover intervention study; Proton nuclear magnetic resonance (^1H -NMR) metabolomics; Metabolomics data tensor; ANOVA-simultaneous component analysis (ASCA); Xenobiotic detoxification; Glycine-N-acyltransferase (GLYAT).

3.2.3 Introduction

Applications of metabolomics to intervention or challenge studies greatly enhance the holistic understanding of the effects of consumed substances on metabolic pathways (Gibney *et al.*, 2005; Wishart, 2008). Data sets from intervention studies are, however, complex, as these investigations aspire to measure multiple metabolites in a biofluid, obtained from several experimental subjects, collected at different points in time and subjected to interventions from different consumed substances. In addition, these studies call for methods of data analysis specifically designed for longitudinal (time-dependent), multi-subject (data from several experimental participants), multi-group (intervention studies) and multivariate data (Antti *et al.*, 2002; Smilde *et al.*, 2005). In this paper, we present the experimental design for an intervention study which includes the complex aspects mentioned above. The interventions were consumption of alcohol in the presence or absence of NAD, using flavored water as a vehicle. We generated matched-sample series through a crossover study of participating subjects, collecting samples over a distinct time frame. The biochemical responses to the interventions were distinctly different: responses to alcohol and NAD intake resided in the intermediary metabolism, whereas those to exogenous substances in the vehicle involved detoxification through biotransformation mechanisms. Here, we present the full experimental design of the intervention study, but focus on the contribution from the biotransformation response by presenting the outcomes of vehicle consumption only. The results on the alcohol and NAD interventions will be published in a separate paper.

Benzoic acid was an important constituent in the vehicle used in the present study. Benzoic acid and its derivatives are routinely used as preservatives and flavoring agents in food products. Consequently, human exposure to them is quite common, and has raised concerns about potential public health consequences (Park *et al.*, 2011). Evidence that benzoic acid is excreted as hippuric acid after enzymatic conjugation to glycine dates back to the 1950s (Schachter & Taggart, 1953), but different viewpoints seem to prevail on the physiological significance of the glycine conjugation reaction. Traditionally, glycine conjugation became part of the paradigm of detoxification, with the critical role of glycine conjugation for aromatic acids (Conti & Bickel, 1977). More recently, new views were proposed, shifting the focus to glycine homeostasis to assist in the regulation of body stores of glycine and other amino acids which are key neurotransmitters in the central nervous system (CNS) (Beyoğlu *et al.*, 2011) or to serve as a molecular escort in the glycine deportation system to excrete excess glycine into urine as hippuric acid (Beyoğlu & Idle, 2012).

Detoxification pathways – the traditional viewpoint – can directly affect the integrity of multiple organs and hence can be widely involved in a variety of human conditions, such as health (Wen *et al.*, 2013), co-metabolism in humans with the gut microbiome (Nicholson *et al.*, 2005), disease therapy (Jana & Mandlekar, 2009) and aging (Zimniak, 2008). Lipophilic endogenous or exogenous xenobiotics are first metabolized by the phase 1 detoxification system, which converts the compounds into substances having a hydrophilic functional group for increased solubility. The phase 2 detoxification system involves conjugation reactions (Holtzclaw *et al.*, 2004). The phase 2 system comprises several enzymes, with glycine-N-acyltransferase (GLYAT) being the key enzyme in glycine conjugation.

Invoking the roles of aromatic acids and gut metabolites to regulate blood levels of glycine – the new viewpoint – relates to that of glycine as neurotransmitter in the CNS. A putative role for it as a neurotransmitter dates back to the observation that the concentration of glycine in spinal cord tissue is far higher than elsewhere in the brain (Aprison & Werman, 1965). The intracellular concentration of glycine is regulated mainly by a glycine transporter, acting in synergy with glutamic acid decarboxylase (Tanaka & Ezure, 2004). The proposed glycine deportation system, moreover, functions as a homeostatic regulator of the GLY central pool and, by definition, also the GLY tissue pool (Beyoğlu & Idle, 2012)

GLYAT (reviewed by Badenhorst *et al.*, 2013) is located in the mitochondria of mammalian liver and kidney and is a member of the superfamily of N-acyl-transferases. It has been demonstrated that several GLYAT species [glycine N-acyltransferase (EC 2.3.1.13), glutamine N-acyltransferase (EC 2.3.1.68) and glycine N-benzoyltransferase (EC 2.3.1.71)] are closely related but distinct mitochondrial enzymes (Nandi *et al.*, 1979), although originally regarded as being a single enzyme (Schachter & Taggart, 1954). Benzoyl-CoA, salicyl-CoA, and certain short-, straight- and branched-chain fatty acyl-CoA esters are substrates for the former enzyme, and glycine is the acyl acceptor for all these enzymes. As glycine and glutamic acid are known to cross the blood–brain barrier freely, it is assumed that removal of GLY from the GLY central pool will lower levels of GLY in the CNS (Beyoğlu & Idle, 2012).

Commercial flavored water, containing benzoic acid (the preservative and co-substrate in glycine conjugation), was used as vehicle for the present intervention study, which opened the opportunity for a metabolomics investigation on controlled benzoic acid consumption. The intervention study consisted of 24 experimental cases (healthy males between 20 and 24 years of age), and used a metabolomics approach for the experimental design and data analysis. The study included four

treatments and six urine samples per experimental case were collected over a period of five hours and used as biofluid for the metabolomics measurements. We selected proton nuclear magnetic resonance (^1H -NMR) spectroscopy as the technology for the generation of the metabolomics data as it is the recognized method of choice for an untargeted coverage of the metabolome. The outcomes offered several intriguing views on the metabolites formed in each of these interventions. The results using the vehicle as control provided novel insights into the metabolite profile following benzoic acid consumption.

Time-dependent intervention studies require more technical statistical methods for data analysis, forming another central aspect of this paper. Metabolomics data are mostly represented as a matrix of controls and cases (rows) measured over the metabolites that reflect the perturbation under investigation (columns). Such data are traditionally processed through multivariate statistical methods, using mostly unsupervised principal component analysis (PCA) and a supervised method such as partial least-squares discriminant analysis (PLS-DA), to identify metabolites that differ significantly between the test groups studied (Lindon *et al.*, 2001). In human intervention studies, individual variation tends to dominate the often less pronounced metabolic perturbations due to the intervention. A reasonably large number of cases is mostly required as well as a sufficient number of time points to cover the effect of the intervention. We used a crossover experimental design, which generated several hundred samples which requires a different approach from the traditional multivariate methods for processing the data and a more extensive validation of the results than for standard metabolomics studies.

For optimal insight into the intervention involving vehicle consumption, we followed a novel approach of complementary statistical methods: (1) the traditional multivariate and univariate methods of analysis were included as the first approximation of the biological profile following the intervention; (2) ASCA (ANOVA-simultaneous component analysis), developed for analyzing designed metabolomics data (Smilde *et al.*, 2005), was used next as a method that deals with multivariate data sets based on a defined experimental design, including time-dependent data; (3) RM ANOVA (repeated measures analysis of variance), which is similar to ASCA as it accounts for the experimental design of the data (although RM ANOVA was used to assess each metabolite individually to find those metabolites that changed significantly with time); finally, (4) unfolded PCA (Villegz *et al.*, 2009) was applied to gain insight into the global (i.e. multivariate) effect of the vehicle over time. The combination of these complementary statistical methods and transdisciplinary approach of metabolomics to the research lead to a more holistic view of the

data. This proved essential in elucidating the effect of the intervention and provided novel insights into benzoic acid biotransformation, not observed by more reductionist methods of analysis.

3.2.4 Materials and methods

Chemicals and reagents

The substances used for the intervention reported here were flavored water (aQuellé lemon-flavored sparkling water: fructose and citric acid flavoring; sodium benzoate preservative; sodium cyclamate, aspartame, acesulfame K sweeteners; vitamin C – www.aquelle.co.za, product of South Africa), commercial water (Valpré still spring water, inorganic contents specified – product of South Africa). The internal standard for the ^1H -NMR analysis was trimethyl-2,2,3,3-tetradeuteriopropionic acid (TSP, sodium salt; Sigma Aldrich).

Experimental design

The intervention study was based on an observation in a preliminary metabolomics investigation on the effect of acute alcohol consumption in healthy cases (Mason, 2010) that indicated some NAD depletion. The experimental group consisted of 24 medically confirmed healthy males, between 20 and 24 years of age, residing in the same student hostel of North-West University (South Africa). They were neither alcohol addicts nor total abstainers, but confirmed their use of alcohol on a moderate social level. No cases took any medication, all were asked to refrain from vitamins, minerals, and other supplementation, and were requested to follow a similar dietary and lifestyle pattern for the duration of the study. The protocol was approved by the Health Sciences Ethical Committee of North-West University (Ethical approval number: NWU-00045-12-S1), conducted in accordance with guidelines for good clinical practice and performed at the Health Clinic of the university. A medical doctor and nurse were present during the entire period of intervention and all cases could leave the premises only after approval by the doctor.

The experiments were conducted on Saturday mornings between 08:00 and 12:00. All subjects had to abstain from breakfast and had to provide an early morning urine sample, collected one hour before the start of the experiment (time –1). The treatments consisted of a measurement of the baseline effect of imbibing the vehicle (500 mL flavored water) that was also used in the remaining three interventions. All the cases were randomly assigned to a treatment/intervention group until all 24 participated in all four interventions. All cases were provided with 1.5 L pure spring water, which was the only substance that could be taken over the period of the experiment. High diurnal variation in urine is well established, and a validated dietary exposure biomarker

discovery protocol demonstrated that top-ranked signals discriminating between fasting and 2–4 hour postprandial urine samples could be linked to metabolites abundant between components of a standardized dietary intervention (Favé *et al.*, 2011). Thus, urine samples were collected at time zero, just prior to consumption, followed by four further samples at 1, 2, 3 and 4 hours after imbibing the vehicle, providing six samples in total for each case. Owing to commitments of some participants, the experiments were performed over a period of 7 consecutive Saturdays, and all samples were treated and stored as described below.

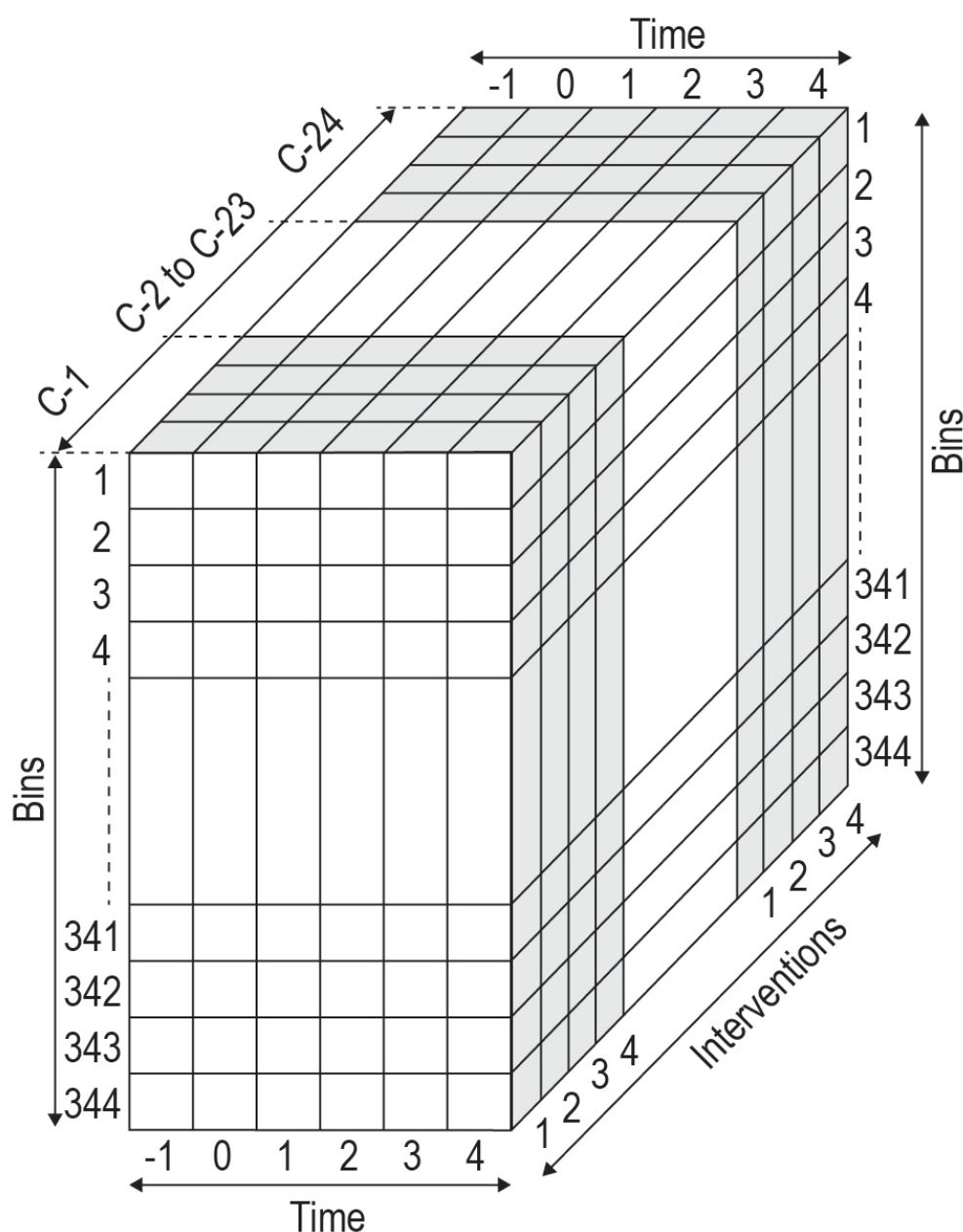


Figure 3-1: Representation of all elements of the experimental design.

The tensor for the complete design consisted of four dimensions: the experimental cases (24 in total, listed as C-1 to C-24 in Figure 3-1), time points (6, hourly), interventions (4, including the vehicle only) and spectral bins (344, following data pretreatment). The total number of 198 144 data points ($= 24 \times 6 \times 4 \times 344$) thus required bioinformatics analysis to uncover the information from the four interventions. All 24 experimental subjects agreed to participate in each of the four interventions (i.e. a crossover design), indicated by 1 (vehicle only), 2, 3 and 4. For each of the 24 cases, urine samples (from which the NMR spectral bins were generated) were collected at one hour prior to the intervention (time -1), just before the intervention (time 0) and then at hourly intervals for 4 hours (times 1–4). The results reported in this paper apply to only one of the treatments (consumption of flavored water), which yielded 49 536 data points.

Measurement design

Because of the large number of samples produced by this study, their analysis spanned a long time, which necessitated splitting samples across multiple analysis blocks or batches. Analytical assessment of metabolomics data generation proved the highly repeatable nature of NMR data measurements (Vian *et al.*, 2009). Although repeatability and reproducibility are not concerns in NMR analyses, the measurement design included the use of pooled quality control (QC) samples to estimate any batch effect or other interfering analytical aspect. Each batch contained the 24 samples (corresponding to 4 treatments across 6 time points) of a single case and three QC samples.

The 27 samples from each batch were analysed in the following order:

$QC_1 [S_{-1}S_0S_1S_2S_3S_4]_{Vehicle} [S_{-1}S_0S_1S_2S_3S_4]_{Alcohol} QC_2 [S_{-1}S_0S_1S_2S_3S_4]_{NAD} [S_{-1}S_0S_1S_2S_3S_4]_{NAD+Alcohol} QC_3$

where S_{-1} represents the sample collected at time -1 , S_0 represents the sample collected at time 0, and so on.

Sample collection, characterization and storage

Six samples per intervention were collected from each student. This gave a total of 24 urine samples from each case over the course of the study. One 5 mL and two 1 mL vials were used to provide aliquots of each of the urine samples; these aliquots, together with the remainder of the bulk urine samples, were stored at -80°C . Once all the urine samples were collected, one 1 mL aliquot of each was thawed and combined to prepare a QC sample for the experiment as a whole. This QC sample was then divided into 15 mL aliquots and once again stored at -80°C .

Sample preparation and ^1H -NMR analysis

Spectral analyses were conducted at the NMR facility of the Centre for Human Metabolomics at North-West University. Prior to analysis, an aqueous 1.5 M KH_2PO_4 deuterated buffer solution at pH 7.4 was prepared (Dona *et al.*, 2014). This solution served to lock the signal during analysis, maintained a stable pH in the sample and contained TSP as the internal standard to provide a chemical shift reference of $\delta = 0.00$. The urine samples, stored at -80°C , were thawed at room temperature for analysis. A 600 μL volume of each sample was centrifuged at $12\,000 \times g$ for 5 min to remove any sediments or debris. A 60 μL volume of buffer solution was added to 540 μL of the supernatant, vortexed and transferred to a 5-mm NMR tube.

Each sample so prepared was analyzed on a Bruker Avance III HD 500 MHz NMR spectrometer equipped with a triple-resonance inverse (TXI) $^1\text{H}\{^{15}\text{N}, ^{13}\text{C}\}$ probe head and x, y, z gradient coils. ^1H spectra were acquired as 128 transients in 32K data points with a spectral width of 6002 Hz. The sample temperature was maintained at 300 K and the H_2O resonance was pre-saturated by single-frequency irradiation during a relaxation delay of 4 s, with a 90° excitation pulse fixed at 8 μs . Shimming of the sample was performed automatically on the deuterium signal. The resonance line widths for TSP and metabolites were <1 Hz (measurements at half the height of the peak). Fourier transformation and phase and baseline correction were done automatically. The software used was Bruker Topspin (V3.2) and Bruker AMIX (V3.9.9) (Ellinger *et al.*, 2013).

All urine samples were normalized with reference to the creatinine CH_2 peak at 4.05 ppm. We employed two methods of spectral analysis: (1) the first method consisted of equidistant binning, using a bin width of 0.02 ppm applied to the spectral region of 0.5–10 ppm (344 bins). This gave a total of 467 integrated units per NMR spectrum, excluding the water region, for each sample for multivariate analysis. Based on previous empirical experience with NMR spectral analysis, we defined a threshold value of 2×10^6 , being approximately the limit of detection of metabolomic substances presumed to be present in a spectral bin. All values below this threshold were set to zero.

The methodology of processing NMR spectra is well known. Powers (Powers, 2009) describes two schools of thought in this method: peak alignment and binning, the latter of which we used in this study. The equal-binning procedure masks subtle chemical shift differences and hides potentially significant changes of low-intensity peaks, but incurs the risk of splitting peaks or spectral features between bins. The second method used quantified metabolites, derived from variably-sized binning of the original spectral peaks above the noise level. This alternative method

prevents peak division between multiple bins, avoiding the problems incurred in the first method. This second approach was specifically applied for the accurate identification and quantification ($\mu\text{mol}/\text{mmol}$ creatinine) of discernible metabolites, generating data for univariate analysis.

In addition, we used both the 1D and 2D J-resolved (JRES) approach to further characterize guanidinoacetic acid observed in some urine samples (3.3.6 of the SI and Figure 3-17).

Data analysis

The original spectral data, derived from the cases following the vehicle intervention, sampling at six time points and including the QCs (original data in Table 3-3), yielded a total of 144 samples, as shown in the flow diagram (Figure 3-2) and illustrated through a representative NMR spectrum for one QC sample (Figure 3-3). However, during the experiment one participant did not complete all four interventions, thus yielding 138 samples obtained from the remaining 23 cases, which were used in the data analysis. The original spectral data were pre-processed by normalization relative to the creatinine content (between 4.05 and 4.07 ppm); replacing very low values with zero and performing a 50% zero-filter. Case reduction was based on batch comparisons of the QC samples (detail presented in the SI) – two batches were found to be outliers and were removed. A tensor (Figure 3-1) was used for three modes of statistical analyses: (1) cross-sectional analysis of time points using traditional multivariate (PCA and PLS-DA) as well as univariate analyses – Wilcoxon signed rank test (WC p-value) and associated fold changes (FC), generated for each bin; (2) ASCA performed across all times; and (3) univariate RM ANOVA for each bin. Further detail on these methods is presented in the SI. VIP refers to variables important in projection, based on a PLS-DA; SSL refers to the sum of the squared loadings of the ASCA model; p refers to the RM-ANOVA p-values for the time effect.

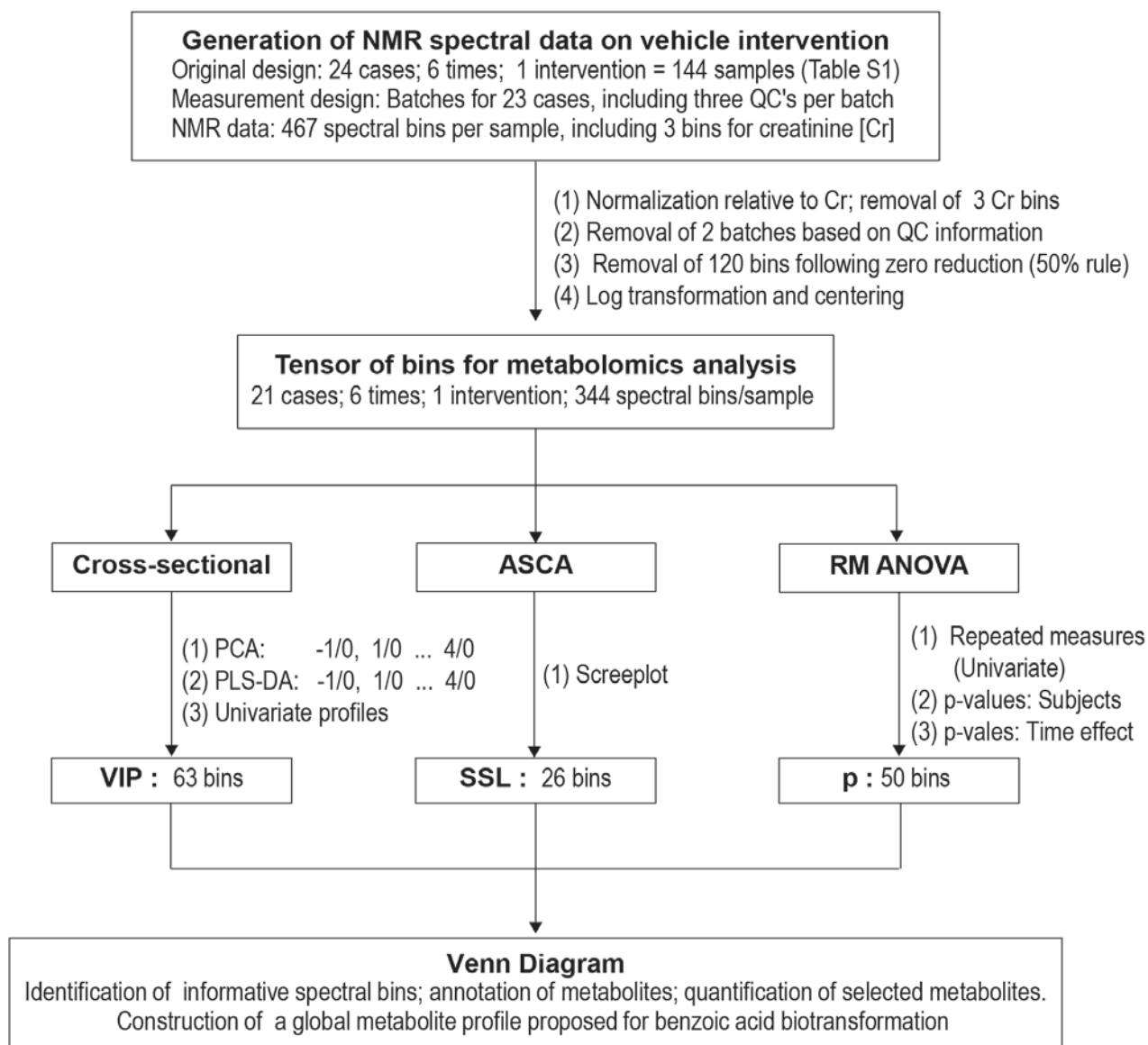


Figure 3-2: Flow diagram indicating the main lines of activity following data generation, identification and quantification of important metabolites on the intervention towards the proposed biological interpretation.

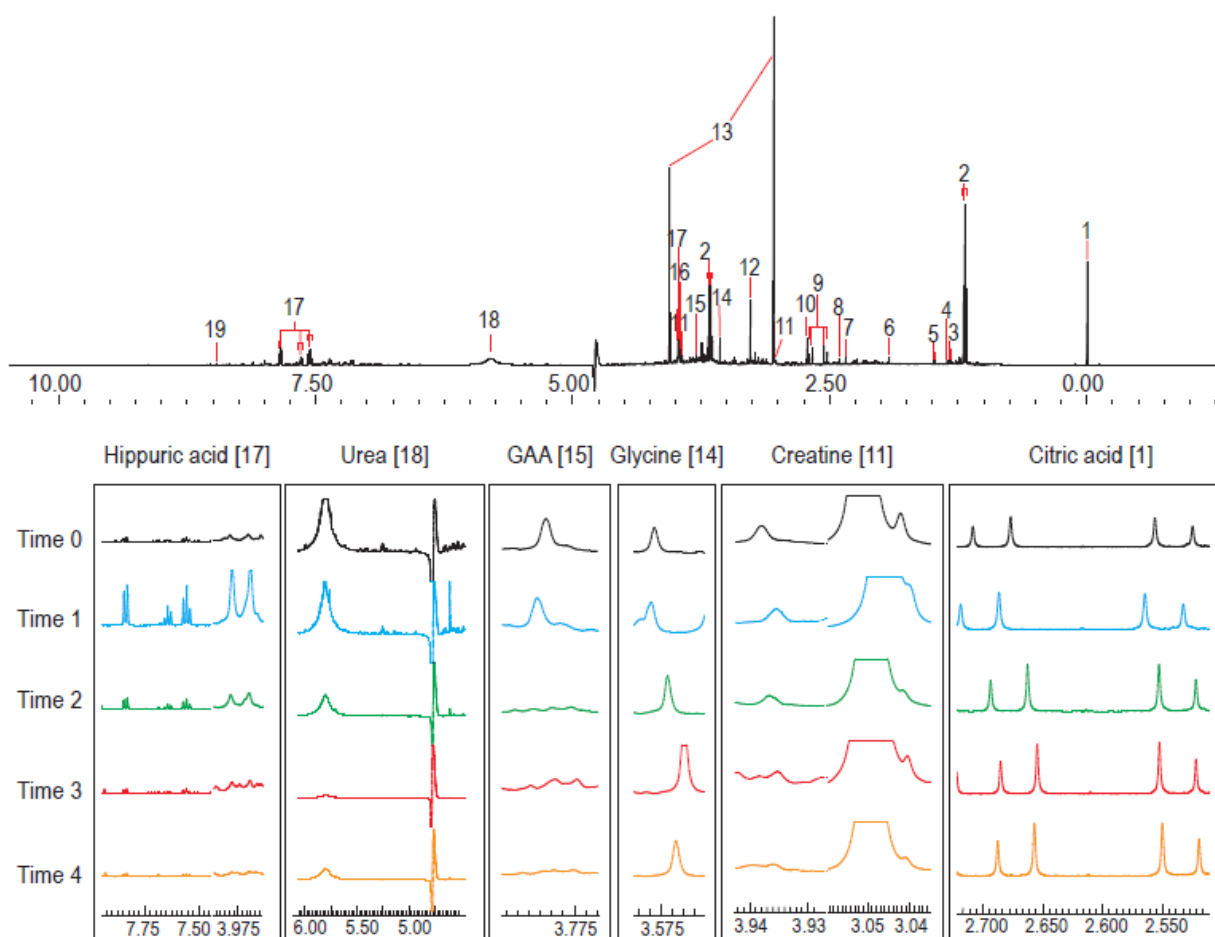


Figure 3-3: 500-MHz ^1H -NMR spectra of urine.

Top spectrum taken from QC sample; numbers indicate the following metabolites: 1: TSP; 2: ethanol; 3: lactic acid; 4: 2-hydroxyisobutyric acid; 5: alanine; 6: acetic acid; 7: pyruvic acid; 8: succinic acid; 9: citric acid; 10: dimethylamine; 11: creatine; 12: trimethylamine-N-oxide (TMAO); 13: creatinine; 14: glycine; 15: guanidinoacetic acid (GAA); 16: glycolic acid; 17: hippuric acid; 18: urea; 19: formic acid. The six important metabolites (high VIP values) are highlighted to qualitatively show differences over time 0 (black), 1 (blue), 2 (green), 3 (red) and 4 (orange) – scaled according to creatinine.

The spectrum, shown in Figure 3-3 was taken from the QC sample. Six metabolites with high VIP values, are highlighted to qualitatively show differences over time 0 (black), 1 (blue), 2 (green), 3 (red) and 4 (orange). Apart from citric acid and hippuric acid, the detoxification product of benzoic acid, no other constituents from the vehicle appeared to be detectable in the NMR analysis.

A list of important bins was generated by each of the three modes and combined using a Venn diagram approach to identify a final shortlist of important spectral bins. This shortlist contained information on the main metabolites that reflected the intervention.

3.2.5 Results

Cross-sectional analysis

One student did not complete the full intervention program, and was thus removed from the group. All results were finally based on the 344 ^1H -NMR profiled bins for 21 cases, following the elimination of two outlier batches (i.e. data from two cases), across the 6 time points. General time-dependent changes in the spectral data were first evaluated with regard to the baseline, i.e. time 0. Cross-sectional comparisons of time points –1, 1, 2, 3 and 4 hours vs. time 0 were performed using three multivariate approaches (unsupervised Euclidian and Ward hierarchical cluster analyses presented as dendrograms, unsupervised PCA and supervised PLS-DA models) and the combination of two univariate approaches (the p-values from Wilcoxon signed rank test and fold change (FC) values presented as Volcano plots). The data were log transformed (shifted natural logarithm with shift parameter set to 1) and centered prior to performing PCA and PLS-DA. The remaining methods were applied to untransformed data. The statistical techniques used and the results of all time points for the cluster analyses, Volcano plots as well as for the PCA and PLS-DA, are discussed in detail in Section 3.3.4.

Figure 3-4 indicates the results from time –1 hour, 1 hour and 4 hours relative to time 0 for the cluster analysis and Volcano plots. Figure 3-4(a–c) are dendrograms from cluster analysis of 21 cases are shown. The dendrograms were constructed based on subsets of the data representing time 0 along with the data obtained one hour before the intervention (time –1); one hour; and four hours following the intervention (times 1 and 4), respectively. Data for all cases at time 0 are shown as black dots and at the other time slots as red dots. Time points from the same individuals that clustered close together are indicated by squared brackets, linking them.

Figure 3-4(d–f) are volcano plots for the same time points as in a–c indicate the distribution of the individual bins, based on FC and WC p-values. Demarcation of important bins is shown by the horizontal and vertical dotted lines. The number of influential bins is shown as red dots in the rectangles of the upper left and right segments of each Volcano plot.

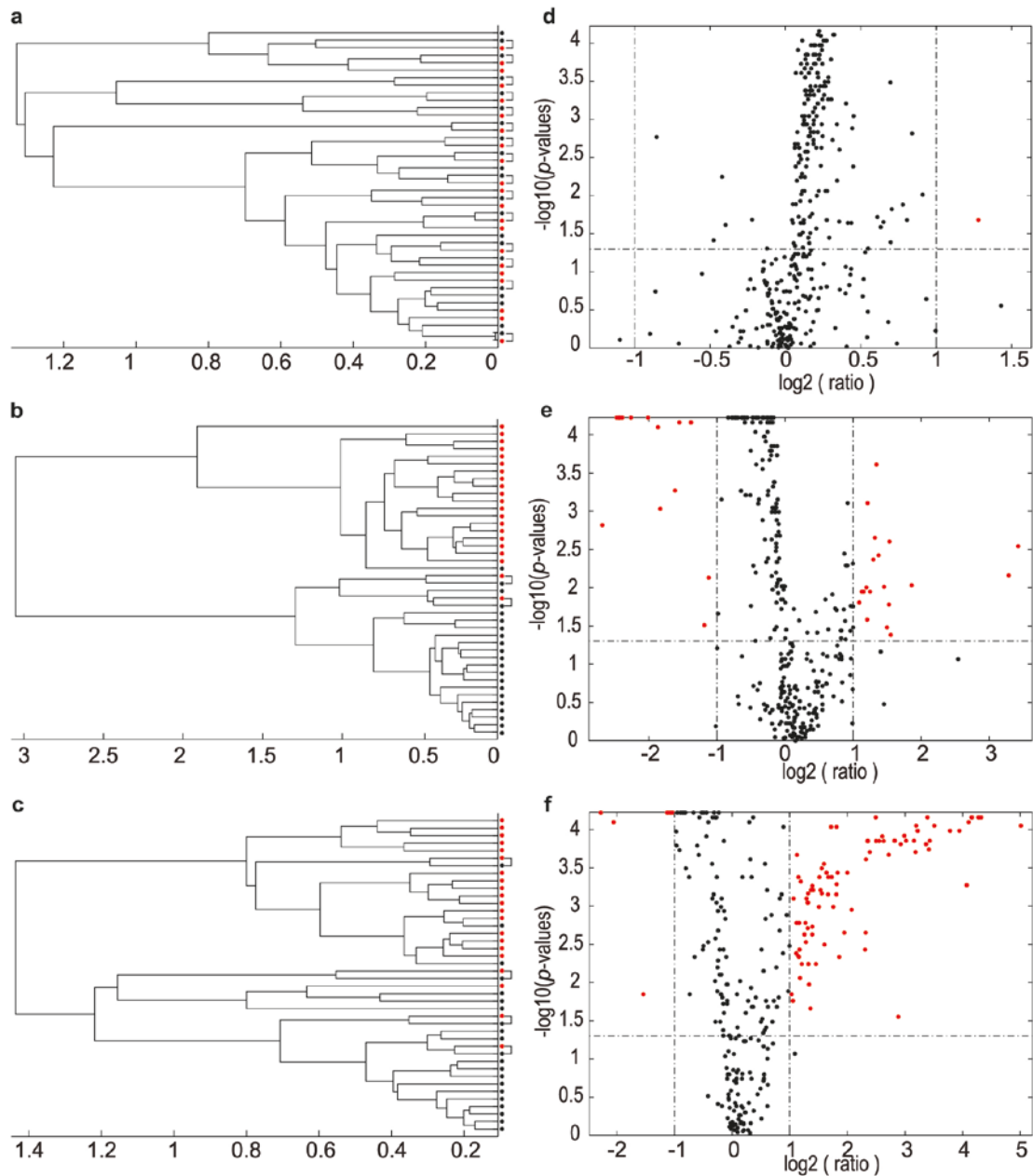


Figure 3-4: Group separation among experimental groups through dendrograms and Volcano plots based on equidistant binning spectral data.

Figures a to c represent dendrograms and d to f represent volcano plots resulting from cluster and univariate analysis of 21 cases, respectively. The figures were constructed based on subsets of the data representing time 0 and -1 (a and d); time 0 and 1 (b and e); and time 0 and 4 (c and f). In figures a to c observations at time 0 are shown as black dots and at the other time slots as red dots, with time points from the same individuals that clustered close together indicated by squared brackets. Figures d to f represent volcano plots based on FC and WC p-values. The horizontal and vertical dotted lines show the demarcation of important bins, with significant bins shown as red dots and non-significant bins shown as black dots.

Figure 3-4a shows little separation (time -1 relative to time 0) between the individual cases, indicating a perceived closeness of spectral data encapsulated in the NMR bins for the individual cases. The metabolite profiles detected in the urine of most samples collected immediately prior to consumption of the vehicle thus represent the profile following 12 hours fasting. This conclusion also holds for the case indicated as an outlier bin within the group as a whole (indicated by a red dot in Figure 3-4d). The distinct group separation after vehicle consumption (Figure 3-4 b and c) clearly indicates a metabolic perturbation. The results for the other two multivariate methods (PCA: Figure 3-12; PLS-DA: Figure 3-13) confirmed the main observations shown in the dendrograms: no group separation between the data obtained at time zero and one hour before the intervention (Figure 3-12A and Figure 3-13A), clear separation one hour after consumption, i.e. time 0 vs. time 1 (Figure 3-12B and Figure 3-13B). Characteristic of multivariate methods, the group separations were more evident when applying a supervised (PLS-DA) as opposed to an unsupervised method (PCA). The Volcano plots revealed that a single bin indicated a significant ($p \leq 0.05$) up-regulated value ($FC \geq 2.0$) at time 0 with respect to time -1 (Figure 3-4d), which increased to a total of 31 up- and down-regulated values (Figure 3-4e) one hour after the intervention ($p \leq 0.05$ and $|FC| \geq 2$) and becoming abundant after 4 hours (Figure 3-4f), as compared to the baseline measurement (time 0).

Inter-individual variation following vehicle consumption

Figure 3-5 represents the scores based on the first two principal components (PC1 and PC2) of the PCA model for the unfolded data tensor. The data tensor was unfolded in time as described and illustrated in the SI (Villez *et al.*, 2009). The unfolding transforms a three-dimensional tensor into a two-dimensional matrix and thus allows for PCA.

PCA of the unfolded tensor provides insight into the effect of the vehicle in time on the bins (indicated by the ellipses and centroids in Figure 3-5) as well as for individual cases, shown by the dots and trajectories in Figure 3-5. It is evident from Figure 3-5 that there is a change in the global profile (i.e. across all cases) over time. Times -1, 0, 2, 3 and 4 appear to be quite similar, with a dramatic change visible after 1 hour (time 1). It was evident from the prior analyses (Figure 3-4 and Figure 3-5) that the urinary spectral profiles following the consumption of flavored water suggested notable changes in the metabolite profiles of all individuals in response to the intervention. To illustrate the inter-individual variation, the centroids of the PCA for the group (all 21 cases) and individual cases were tracked and highlighted over the period from time 0 to time 4 hours (Figure 3-5).

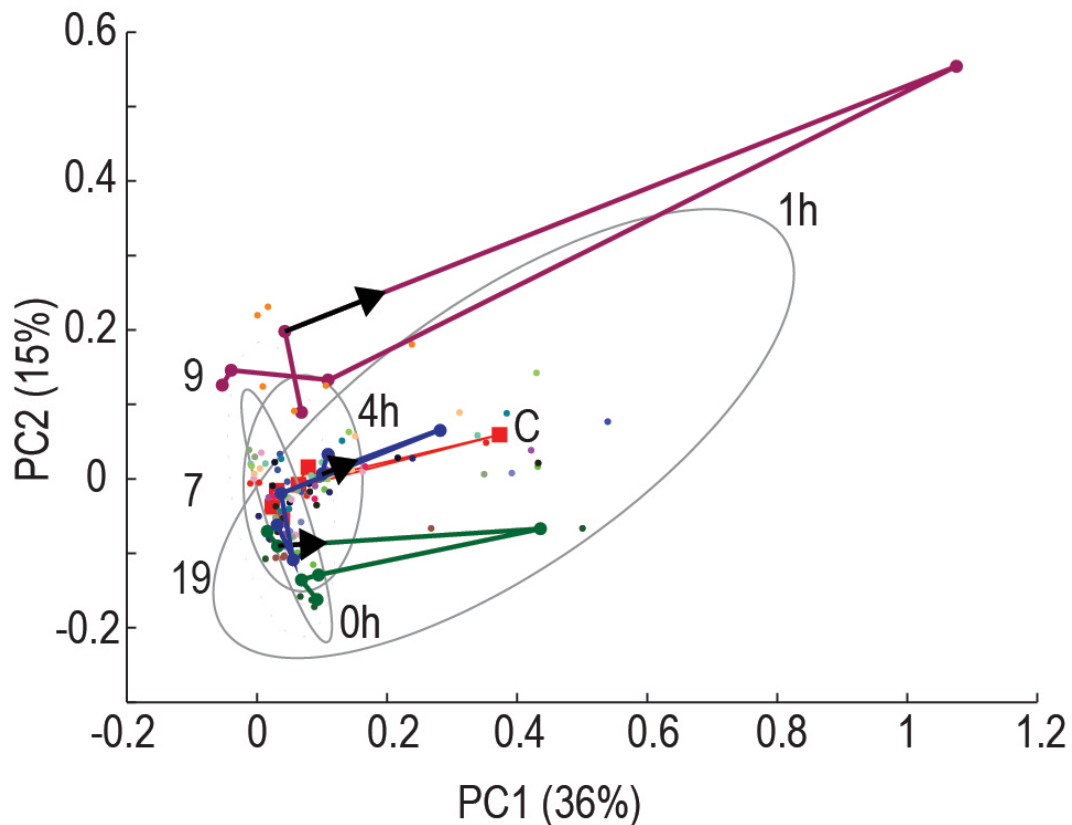


Figure 3-5: Unfolded PCA Scores Plots.

PC1 and PC2 for all individual cases (dots in a different color for each individual) and indication of the 95% confidence ellipsoids for scores of PC1 and PC2 for times -1 , 1 and 4 hours. The centroids for these three clusters are indicated as red squares. The trajectories of the spectral profiles of three individuals are also shown. The trajectories of three cases (numbers 9 , 7 and 19 are shown as purple, blue and green lines, respectively) illustrate the inter-individual responses to the intervention. The direction of the trajectories is indicated by the short black arrow, starting from time 0 , for clarity.

The trajectories drawn in Figure 3-5 represent only three selected individuals, for clarity. The selected trajectories (as well as for the group as a whole, Figure 3-5) indicate clear similarities in the individual responses to vehicle consumption in time: the profiles of the trajectory from time 0 to 4 hours were comparable in some respects. However, distinct differences were also noted: the response at time 1 hour following vehicle consumption in case 9 gives the impression of being an outlier. However, since the QC samples for this individual (linked to a given batch) were not outliers, we attribute this variation to the unique response of the case to the vehicle. Since we are interested in these kinds of responses, the observation was retained. Overall, the observations shown in Figure 3-5 indicate that the cases themselves are a noteworthy source of variation.

Longitudinal response to vehicle consumption

The results presented thus far suggest that the variation in the total data set was based on the intervention and superimposed by the normal inter-individual heterogeneity and time effects and by their interactions. For optimal insight into the effect of the intervention, and specifically to identify the most important metabolites reflecting the metabolic perturbation induced by the flavored water, we used three complementary approaches. The traditional PLS-DA (illustrated in Figure 3-13) was used to find an approximation of the biological profile following the intervention. We identified 63 bins as important based on a maximum VIP over the five comparisons greater than 2.

Second, ASCA was applied as it provided an approach which was developed to deal with multivariate data sets based on a defined experimental design, including time-dependent data (Smilde *et al.*, 2005). Figure 3-6 plots of the sum of the ASCA effects and projected residuals are colour coded according to time following the intervention (–1, 0, 1, 2, 3 and 4 hours shown in red, black, blue, light blue, green and pink, respectively), with the arrow showing the time-dependent trend, using the same discriminating colour sequence.

Again, the data were log transformed and centered. Similar to the unfolded PCA result, the plot of the sum of the ASCA effects and projected residuals onto the first two components of the effects subspace (Zwanenburg *et al.*, 2011), shown in Figure 3-6, indicate that observations at times –1, 0 are quite similar, indicating a state of homeostasis that existed after the fasting period. A marked change became visible one hour after the intervention. This was followed by comparable profiles, especially between times 3 and 4, indicating a return to a state of homeostasis, which was not identical to the state prior to the intervention, as will be discussed below. The sum of the squared loadings (SSL) of the ASCA model were ranked in decreasing order and used as a scree plot (Figure 3-15) to identify 26 bins with notably higher SSL values.

Third, RM ANOVA was used to assess each bin individually to identify bins that changed significantly in time, from which we selected the top 50 bins (p-values less than 0.00001) which appeared to be most informative. Data, generated from complex experimental designs such as the intervention study presented here, produce information that intersect in various ways.

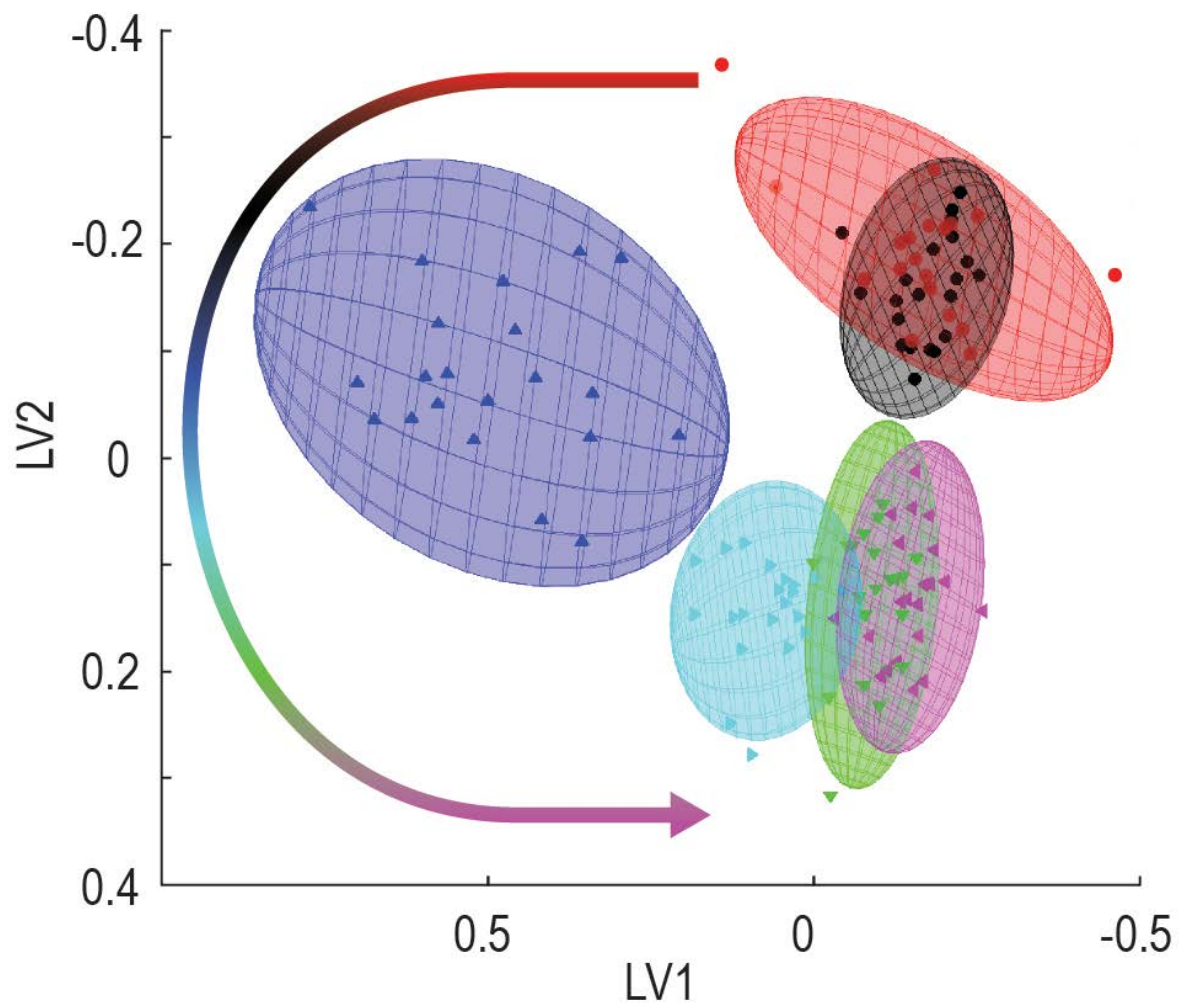


Figure 3-6: Application of ASCA to the 344 NMR spectral bins for 21 individuals over the period of the intervention.

Figure 3-6 plots of the sum of the ASCA effects and projected residuals are colour coded according to time following the intervention (–1, 0, 1, 2, 3 and 4 hours shown in red, black, blue, light blue, green and pink, respectively), with the arrow showing the time-dependent trend, using the same discriminating colour sequence.

Identification of important variables

A Venn diagram (Figure 3-7) was used to visualize the counts (number of bins) of the lists of important bins sharing some properties. Three properties were selected: maximum VIP ≥ 2 from the PLS-DA, SSL > 0.01 from the ASCA, and p-values from the RM ANOVA as defined above.

A total of 29 bins were shortlisted, derived from being present in at least two of the three lists and which were unique in the intersection reflected by the shaded areas in Figure 3-7. It is interesting to note the power of the ASCA approach as it was able to identify 26 of the 29 bins, but this does not imply that variables not selected (e.g. the 29 RM ANOVA metabolites) are all unimportant as each method selects variables in its own way.

For our purpose we concentrated, however, on the shared metabolites; how this was achieved is described in more detail in Section 3.3.4 of the SI. The list of 29 bins provided the final selection of bins to be identified as important metabolites and quantified to concentrations. Detailed information on the application of ASCA, RM ANOVA and unfolded PCA is presented in 3.3.4.

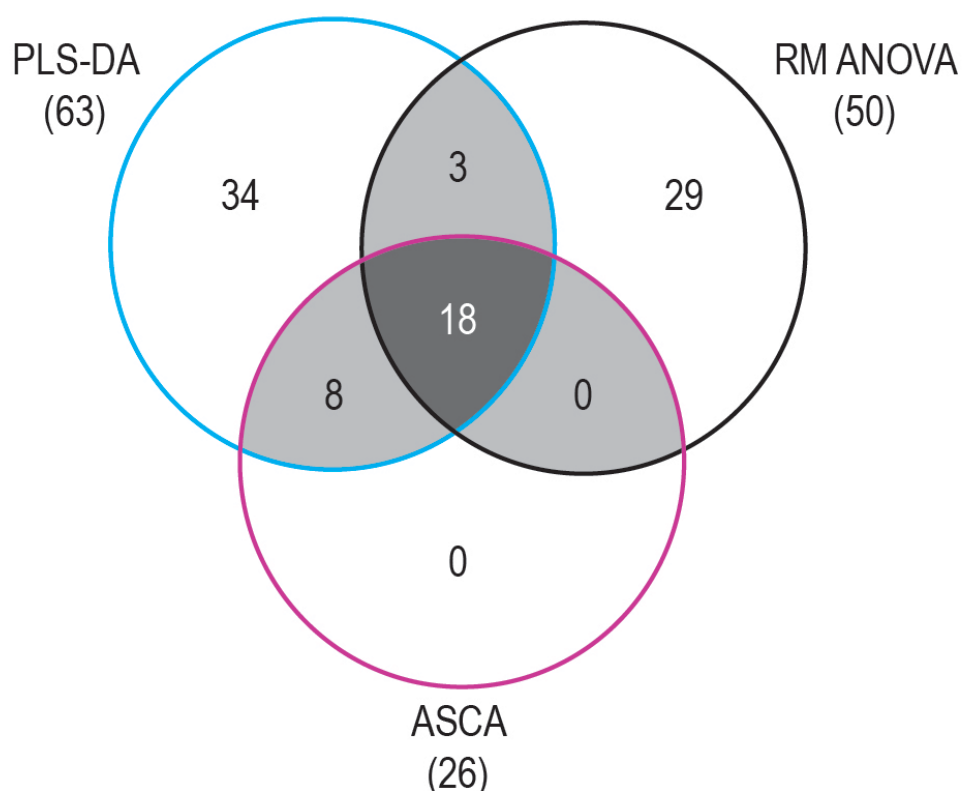


Figure 3-7: Venn diagram displaying counts of variables selected by various techniques.

Three properties were selected to create the Venn diagram: maximum VIP ≥ 2 from the PLS-DA, SSL > 0.01 from the ASCA, and p-values from the RM ANOVA as defined above. Bins which are common to any two statistical approaches are indicated by the lightly shaded areas, and the 18 bins that are common to all three are shown in the central, heavily shaded area.

Analysis of the spectral characteristics of the 29 bins indicated that 23 bins related to endogenous human metabolites. Spectral information from four bins could not be assigned to any known chemical substance and two bins indicated exogenous contaminants. Some bins (indicated in brackets) had structural information on the same metabolite, which resulted in a list of six key metabolites: urea (10 bins), hippuric acid (8 bins) and citric acid (2 bins) as well as creatine, 3-methylhistidine and guanidinoacetic acid each with one bin. One bin for each of the six metabolites, as well as for glycine (not identified as a highly perturbed metabolite), were quantified for all time points, as summarized in Table 3-2.

Excretion kinetics for all six these metabolites are shown in Figure 3-16 of the SI. The most conspicuous change – indicating the highly efficient detoxification of benzoic acid – occurred in the more than fourfold ($p < 0.0005$) increased urinary excretion of hippuric acid within one hour of consuming the flavored water, and its return to the same level as before the intervention (Table 3-2 and Figure 3-16A). As indicated by the unfolded PCA scores (Figure 3-5), the trajectories of individual cases (illustrated for cases 7, 9 and 19) tended to return to positions close, but not identical, to the pre-intervention positions. The excretion kinetics of the six metabolites (Figure 3-16) and the Wilcoxon signed rank p-values (Table 3-2) gave strong numerical indications for this observation: (1) Hippuric acid dominated the profiles, and returned to practically the same state as prior to the intervention ($p=0.689$, Time 0 vs Time 4). (2) The levels of guanidinoacetic acid ($p=0.028$) and creatine (0.003) were significantly lower than prior to the intervention (Time 0 vs Time 4). (3) The excretion of citric acid ($p=0.002$) and urea ($p<0.0005$) were still after four hours significantly higher than prior to the intervention (Time 0 vs Time 4). Moreover, the excretion profiles of some unknown and unimportant metabolites, gave further support for the observed trajectory profiles (Figure 3-17).

The interpretation of changes in the excretion of these remaining metabolites paved the way for the construction of a metabolic profile, reflecting the consequences of detoxification of a single xenobiotic – benzoic acid as used here – as discussed below.

Information on important metabolites		Time-dependent characteristics of important metabolites following flavoured water consumption									
Metabolite (δ [ppm], multiplicity and chemical group)	Normal values Mean [Range]	Mean	Mean	WC	Mean	WC	Mean	WC	Mean	WC	
		[SD] Time 0	[SD] Time 1	p-value 0 vs 1	[SD] Time 2	p-value 0 vs 2	[SD] Time 3	p-value 0 vs 3	[SD] Time 4	p-value 0 vs 4	
Citric acid (2.59 AB - [(CH ₂) ₂])	203 [49-600] ¹	213 [115]	264 [137]	<0.0005	322 [139]	<0.0005	320 [126]	0.001	306 [117]	0.002	
Hippuric acid (3.97d - [CH ₂])	217 [28-610] ¹	210 [100]	884 [268]	<0.0005	449 [170]	<0.0005	285 [154]	0.021	228 [142]	0.689	
Glycine (3.57 s - [CH ₂])	106 [44-300] ¹	94 [63]	114 [58]	0.001	117 [37]	0.005	115 [34]	0.014	108 [32]	0.122	
Guanidinoacetic acid (3.78 s - [CH ₂])	89 [11-124] ²	97 [193]	77 [200]	0.173	0 [200]	0.028	0 [311]	0.028	0 [291]	0.028	
Urea (5.78 broad - [(NH ₂) ₂])	12285 [174-49097] ¹	479 [230]	751 [322]	<0.0005	1004 [303]	<0.0005	937 [311]	<0.0005	914 [291]	<0.0005	
Creatine (3.04 s - [CH ₃])	46 [3-448] ¹	44 [20]	57 [27]	0.015	37 [14]	0.079	35 [9]	0.011	33 [7]	0.003	
1. Bouatra <i>et al</i> , 2013											
2. Fingerhut, 2003											

Table 3-2: Quantified data on metabolically important metabolites.

Table 3-2 contains all quantified values included are from NMR-determined urine analyses; the references are all from the Human Metabolome Database, and are expressed as μ moles metabolite/mmol creatinine. WC p-values are based on the comparison of the respective metabolite concentrations from time = 1 hour to 4 hours, relative to time zero.

3.2.6 Discussion

Using the combination of important metabolites identified employing the three statistical methods (Figure 3-7), the quantified information (Table 3-2) and biochemical interpretation of the metabolomics data, we constructed a metabolite profile reflecting the primary GLYAT-catalyzed biotransformation of benzoic acid and the metabolic consequences of benzoic acid ingested via flavored water (Figure 3-8).

Humans conjugate a variety of aliphatic and aromatic monocarboxylic acids with several amino acids; the resulting peptides are excreted in the urine and bile. In characterizing two closely related GLYATs, Nandi *et al.* (1979) indicated through kinetic studies that benzoyl-CoA is the main substrate for a benzoyltransferase, with salicyl-CoA and certain aliphatic acyl-CoAs being lesser substrates for it as well. The closely related phenylacetyltransferase utilize phenylacetyl- and indoleacetyl-CoA as substrates. Acyl-CoA substrates of one transferase did not serve as substrate for the other, but act as competitive inhibitors. Glycine is the preferred acyl acceptor for both enzymes, with K_m^{App} for benzoyltransferase being less (3 mM glycine) than for the phenylacetyltransferase (20 mM glycine). Due to the high activity of glutamine N-phenylacetyltransferase, phenylacetylglutamine is almost never detected in human urine. Almost all the phenylacetic acid is excreted as N-phenylacetylglutamine. It thus may be anticipated that formation of hippuric acid in this intervention study mainly reflects the GLYAT activity of the benzoyl variant (EC 2.3.1.71).

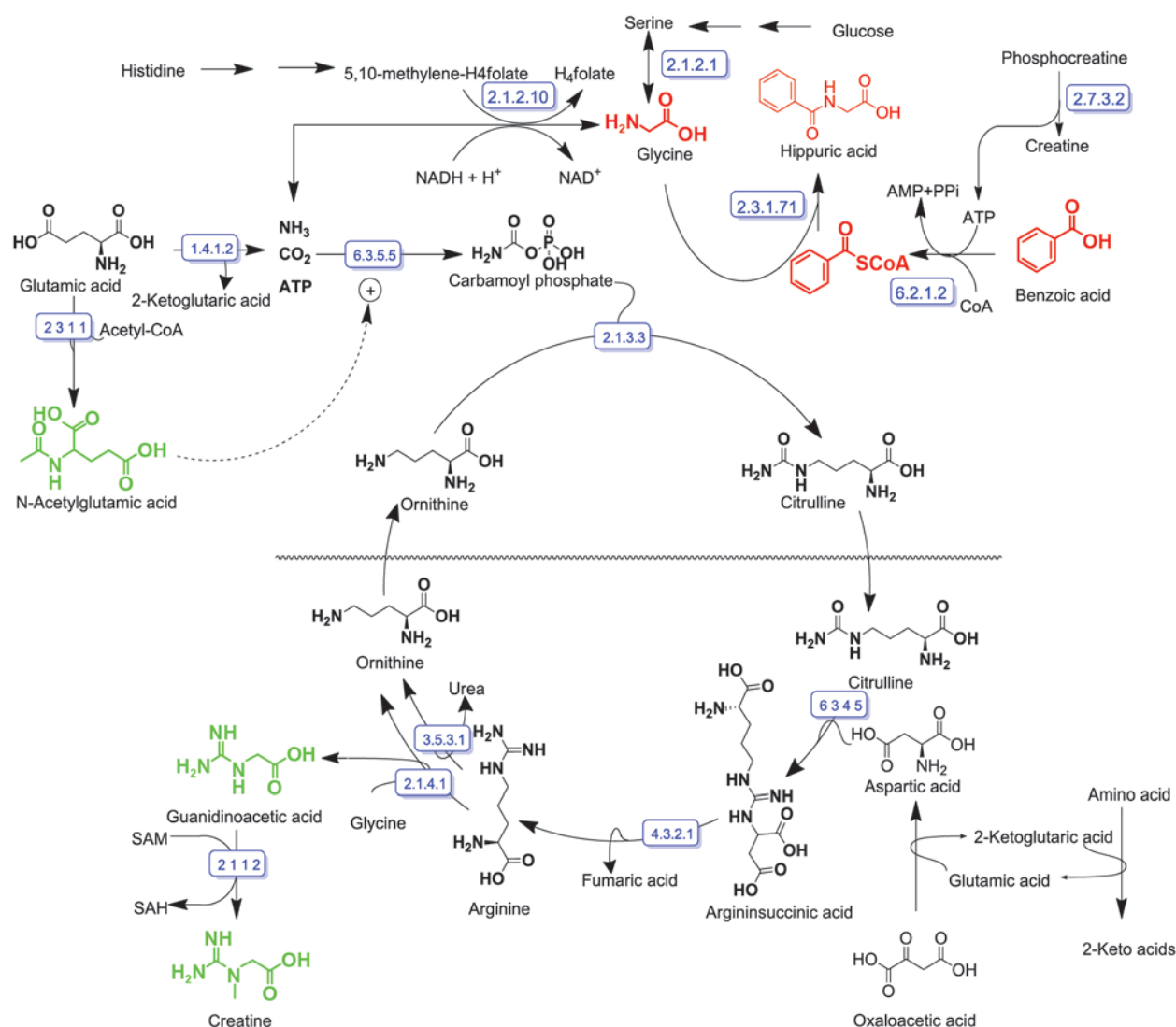


Figure 3-8: Metabolite profile of benzoic acid biotransformation with hippuric acid as outcome.

Metabolites of the primary biotransformation reactions are shown in red. Metabolites of metabolic reactions proposed as being associated with the biotransformation reactions are shown in black. Metabolites shown in green are presumed to decrease due to increased glycine demand and synthesis of glycine through the reversed glycine cleavage system. Enzyme nomenclature (names accepted by the IUBMB): EC 1.4.1.2: glutamate dehydrogenase; EC 2.1.1.2: guanidinoacetate N-methyltransferase; EC 2.1.2.1: glycine hydroxymethyltransferase (alt.: serine hydroxymethyltransferase); EC 2.1.2.10: aminomethyltransferase (a glycine synthetase); EC 2.1.3.3: ornithine carbamoyltransferase; EC 2.1.4.1: glycine amidinotransferase; EC 2.3.1.1: amino-acid N-acetyltransferase; EC 2.3.1.71: glycine N-benzoyltransferase (a GLYAT); EC 2.7.3.2: creatine kinase; EC 3.5.3.1: arginase; EC 4.3.2.1: argininosuccinate lyase; EC 6.2.1.2: butyrate-CoA ligase (a medium-chain acyl-CoA ligase); EC 6.3.4.5: argininosuccinate synthase; EC 6.3.5.5: carbamoyl-phosphate synthase (glutamine-hydrolyzing).

Early studies showed that rat liver mitochondria synthesize hippuric acid at a rate of up to 4 nmol/min per mg of protein (Gatley & Sherratt, 1977) and comparative kinetic analysis suggested that the formation of the benzoyl-CoA substrate is the rate-limiting factor. More recently, two distinct forms of xenobiotic/medium-chain fatty acid:CoA ligase (XM-ligase) were isolated from human liver mitochondria, referred to as HXM-A and HXM-B (Vessey *et al.*, 1999). Both forms had medium-chain fatty acid:CoA ligase activity but HXM-A showed 60–80% activity towards 15 different carboxylic acids relative to benzoic acid, its best xenobiotic substrate (100% activity and the highest $V_{\max}/K_m$). Hexanoic acid was the best substrate for HXM-B, although it was also active towards xenobiotic carboxylic acids. In accordance with these findings we propose that medium-chain acyl-CoA ligase (EC 6.2.1.2) is a major catalyst for production of the CoA-substrate required for hippuric acid formation in the present study.

Acyl-CoA ligases/synthetases belong to a superfamily of adenylate-forming enzymes, and catalyze the two-step activation of fatty acids or carboxylate-containing xenobiotics (Kochan *et al.*, 2009), with xenobiotic-CoA formation in parallel to endogenous fatty acid activation through the role of the ATP-dependent acid:CoA ligases (Knights & Drogemuller, 2000). We visualize the activation of benzoyl-CoA through a Bi Uni Bi Ping-Pong molecular mechanism proposed for CoA-substrate formation by a long-chain fatty acyl-CoA synthetase (Hisanaga *et al.*, 2004): The benzoyl carboxylate substrate (BA) first reacts with enzyme-bound ATP to form an acyl-adenylate intermediate (Ec:B~AMP), which then reacts with CoA to produce the activated benzoyl-CoA ester (B~CoA), having pyrophosphate (PPi) and AMP as byproducts.

Formation of hippuric acid, following GLYAT-catalyzed conjugation between benzoyl-CoA and glycine, peaks within one hour ($p < 0.0001$) following the consumption of flavored water (Table 3-2). The p -values of the Wilcoxon signed rank test (Table 3-2) for the time-dependent urinary excretion of glycine complements that of hippuric acid, and likewise peaks at one hour ($p = 0.001$) following the intervention. However, the excretion kinetics profile of glycine clearly suggests abundance of glycine during the main detoxification period (two hours after the intervention), rather than glycine depletion. Two main pathways exist for increased glycine biosynthesis (Figure 3-8) under physiological conditions of glycine demand: (1) L-serine can be converted to glycine by serine hydroxymethyltransferase (EC 2.1.2.1) in the reversible glycine biosynthesis pathway, having tetrahydrofolate as acceptor for the CH_2OH group from serine, yielding 5,10-methylene-tetrahydrofolate and water. (2) The *de novo* synthesis of glycine can occur from CO_2 and NH_3 , catalyzed by glycine synthase (EC 2.1.2.10) using 5,10-methylene-tetrahydrofolate as the source of the second carbon and yielding tetrahydrofolate – likewise in a reversible reaction.

The CO₂ and NH₃ substrates for the *de novo* synthesis of glycine are generated by mitochondrial glutamate dehydrogenase (EC 1.4.1.2) that converts glutamic acid to 2-ketoglutaric acid, with NH₃ usually as a substrate in the urea cycle. Glycine is also a substrate for glycine transaminase (EC 2.1.4.1)-catalyzed synthesis of guanidinoacetic acid. Urinary guanidinoacetic acid concentrations, notably, decreased (Table 3-2; p = 0.173) within the first hour following consumption of flavored water, towards values below the detection limit (Table 3-2; p = 0.028) two hours later. We speculate that decreased guanidinoacetic acid is caused by the preferential utilization of glycine for benzoic acid detoxification as well as lower urea cycle activity due to decreased N-acetylglutamic acid, a modulator for the up-regulation of the urea cycle. Against this background it seems that the traditional paradigm of GLYAT-catalyzed benzoic acid detoxification, supported by increased *de novo* glycine biosynthesis (Figure 3-8), prevails under the conditions of the present intervention experiment, even though up-regulation of glycine-amidinotransferase (EC 2.1.4.1) through decreased creatine (McGuire *et al.*, 1984) cannot be excluded.

Finally, we observed an increase in urinary creatine within the first hour following the intervention. The creatine/phosphocreatine system plays an important role in energy storage and energy provision, with creatine synthesis being central in cellular energy metabolism. Two main enzymes are the basis of the creatine biosynthesis pathway, namely, arginine:glycine amidinotransferase (EC 2.1.4.1) and S-adenosyl-L-methionine:N-guanidinoacetate methyltransferase (EC 2.1.1.2), as shown in Figure 3-8. Given the decrease of guanidinoacetic acid to below the detection limit (Table 3-2), it seems unlikely that the urinary creatine originates from, and excess creatine is produced by, the creatine biosynthesis pathway. Our proposal is that phosphocreatine (EC 2.7.3.2) degrades in favor of creatine and supplements the ATP reserves, required for the burst of ATP required following benzoic acid detoxification.

In summary, we have described the metabolite profile following benzoic acid intake as part of a designed intervention study. The time-dependent glycine profile supported the view of abundant glycine availability during the main detoxification period rather than that of glycine depletion. It should be noted, however, that the perturbation caused by benzoic acid consumption may be more complex than discussed above. We observed small, but significant time-dependent changes in the NMR spectra (Figure 3-17) for methylguanidine and three unknown substances that are not accounted for in the metabolic model shown in Figure 3-8.

The complex NMR spectral data, generated from cases participating in a time-dependent crossover study, could be resolved sufficiently through the application of traditional univariate and multivariate analyses combined with an ANOVA-simultaneous component analysis (ASCA), repeated measures analysis of variance (RM ANOVA) and unfolded principal component analysis (unfolded PCA) – an approach that opens a novel way for analyses and understanding of complex metabolomics data that reflect perturbations from normal or pathophysiological endogenous or exogenous origin. Furthermore the combination of the complementary statistical methods together with the transdisciplinary approach followed in metabolomics research provided a more holistic view of the data. This proved useful in elucidating the effect of the intervention and provided novel insights into benzoic acid biotransformation, which is not typically observed by more reductionist methods of analysis.

3.2.7 Additional Information

Author Contributions

This investigation required a multidisciplinary approach and the coordinated inputs of all authors were essential to produce the outcome of the study. CR and JW defined the aim of the study, developed the methodological design and supervised the planning and execution of the investigation. CR obtained research funds for the project and ethical approval from the Health Sciences Ethical Committee of North-West University (Ethical approval number: NWU-00045-12-S1), including written informed consent from the 24 participants of the crossover study. CI conducted all experimental analyses and data collection. SM gave guidance and assessment on ¹H-NMR data generation and spectral analyses; CI and SM were responsible for the data curation, peak annotation and production of the metadata. MvR and JW conceptualized the complementary techniques required to analyze the data and MvR applied selected computer programs; implemented computer codes and supporting algorithms; and performed the statistical and computational analyses of the original and curated data. LM participated throughout the investigation as expert on metabolism and in modeling of the proposed metabolic outcome of the study. All authors gave input according to personal expertise in the development of the original draft manuscript as well as reviews and editing, gave final approval of the version as submitted and agreed to the accountability requirements.

Conceptualization: CJR JAW; Data curation: CI SM MvR JAW; Formal analysis: CI SM; Funding acquisition: CJR; Investigation: CI SM; Methodology: CJR JAW; Project administration: CJR JAW; Software: JAW MvR; Supervision: CJR JAW; Writing,review & editing: CI MvR SM LJM JAW CJR.

Acknowledgment of Funders

Research funding for the analytical and computational aspects of the project was provided by the Technological Innovation Agency (TIA) of the Department of Science and Technology of South Africa. CI received a postgraduate bursary from the National Research Foundation (NRF) of South Africa. The views and findings of this investigation are those of the authors and do not reflect or represent any policies of the NRF or TIA funders. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests

The authors declare that there is no conflict of interest.

3.3 Supplementary Information: Contribution towards a Metabolite Profile of the Detoxification of Benzoic Acid through Glycine Conjugation: An Intervention Study

3.3.1 Original ^1H -NMR spectral data for intervention 1

Intervention 1 refers to consumption of flavored water only – the cause of the benzoic acid biotransformation described in the main paper. The data include urine samples from the 23 experimental cases that completed the study, yielding a total of 138 study samples. The original design was based on 24 cases (individual subjects), although one did not participate in the flavored water experiment. The original ^1H -NMR spectrum of each of the urine samples analyzed was divided into 0.02-ppm equal-sized bins between 0.5 and 10 ppm, excluding the region of the water peak (4.72–4.88 ppm). This resulted in 467 bins containing spectral data. Table 3-3 represents a small extract from the original, raw spectral data matrix of all the samples from the vehicle intervention analyzed for this study (available online as S2 File: CI_unscaled raw vehicle data.xlsx).

Sample Information					Bin no					
Batch	ID	Case	Treatment	Time	9.99	9.97	9.95	...	0.53	0.51
1	F41-1	F	Vehicle	-1	252890.2	171823.4	163215.5		384216.4	266010.8
1	F410	F	Vehicle	0	151299.1	100954.3	86908.15		262467	216768.2
1	F411	F	Vehicle	1	79268.87	105579.2	95048.77		76325.53	126937.3
1	F412	F	Vehicle	2	84927.78	101751.3	127028.2		77872.51	68765.72
1	F413	F	Vehicle	3	74747.86	144056.7	47937.61		91158.41	58079.11
1	F414	F	Vehicle	4	67488.71	70024.98	68237.62		103052	89226.37
2	E51-1	E	Vehicle	-1	120268.5	125437.8	123530.6		143314	122366.1
2	E510	E	Vehicle	0	206635.9	142868.1	112930.5		198141.4	156402.8
2	E511	E	Vehicle	1	106837.9	97280.69	132238.9		168890.4	128759.2
2	E512	E	Vehicle	2	88524.67	110966.2	92268.4		75853.71	102751.3
2	E513	E	Vehicle	3	101023.8	97107.34	59599.65		76011.03	88203.6
2	E514	E	Vehicle	4	134874.8	93287.56	67920.9		182858.7	85690.59
3	G41-1	G	Vehicle	-1	183044.6	110147.5	92701.17		368924.5	308433.9
3	G410	G	Vehicle	0	220505.4	134572.8	99752.4		451830.2	324699.9
3	G411	G	Vehicle	1	229915.6	116229.6	130475.5		951134.8	407340.4

Table 3-3: Excerpt of raw ^1H -NMR spectral data.

3.3.2 Threshold value and normalization

The use of NMR spectroscopy as a tool for metabolomics is limited by the sensitivity (limit of detection) of NMR (Krojanski *et al.*, 2008), requiring an approach in discarding noise while retaining and then measuring real peaks. The spectral intensities within the original 467 bins generated here ranged from 7.8×10^5 to 1.7×10^9 . Based on previous empirical experience with NMR spectral analysis, we defined a threshold value of 2×10^6 , being approximately the limit of detection of metabolomic substances presumed to be present in a spectral bin. To reduce the uncertainty in the data at the level of the approximate detection limit, we applied the threshold of

2×10^6 to the data matrix, setting all values below this threshold to zero. We subsequently applied a 'zero-filter' to remove bins with more than 50% zero observations across all six time points, as applied for fusion of MS-based metabolomics data (Smilde *et al.*, 2005). This resulted in a reduction in the number of bins containing spectral data from 467 to 347. To account for the dilution differences between the urine samples, the value of each bin was made relative to the CH₂ peak of creatinine. This peak is contained in the 4.05 and 4.07 ppm bins. Therefore, each bin value was divided by the sum of the value in these two bins to normalize the bin values relative to creatinine. After normalization, the three bins related to creatinine (4.05, 4.07 and 3.05 ppm) were also removed from the matrix, giving a final total of 344 bins containing spectral data. The above processing was performed in MATLAB (2012b) as outlined in the script provided as additional material.

3.3.3 Data Pre-processing

The data resulting from this intervention study potentially contain volumes of information. However, unlocking it is a challenge for various reasons, the first being the presence of three major sources of variation: (i) the four experimental treatments; (ii) the variation over time; and (iii) the variation between cases participating in the study (potentially the greatest source of variation). The second relates to the analytical platform used. NMR spectra using equidistant binning results in hundreds of bins, in this instance 467. This far exceeds the number of cases measured (23), which is known to cause more traditional statistical methods such as MANOVA (multivariate analysis of variance) to fail (Zwanenburg *et al.*, 2011). Third, the quantified values may be very low for some bins and carry large uncertainty as to the accuracy of the value reported. As a result, values below a certain threshold are converted to zeros, which again poses a challenge for many statistical methods making an assumption about distribution.

The design of the study aimed to alleviate these challenges to some extent by (i) repeatedly measuring the same individuals across both time and treatment factors; (ii) selecting cases with similar demographics, i.e. age and living conditions; and (iii) limiting the influence of dietary difference by exposing cases to each intervention while in a fasting state. Finally, the results discussed in this paper pertain only to a subset of the data, one of the four treatment interventions, making this initial analysis more palatable yet still highly informative.

Taking into account the design of the experiment as well as the limitations of the various statistical techniques, a range of methods were applied to this subset of the data, as discussed in more detail below.

Case Reduction

Reproducibility and repeatability are not known concerns in NMR analysis (Viant *et al.*, 2009). Nevertheless, three pooled quality control (QC) samples were included in each NMR batch run, in the first, middle and last positions. The role of the QC samples was not to correct for any drift or between-batch effects, as done in GC–MS for instance, as this category of confounding factors should not be present in NMR analysis. Rather, the QC samples were used to identify batches which produced unexpected and therefore untrustworthy results due to some external factor we did not or could not control or measure. Outlier detection was based on the scores from a principal component analysis (PCA) with 90% confidence regions, along with Hotelling's T^2 statistic. A two-component PCA model was constructed based on the log-scaled and centered normalized data (464 bins) from only the QC samples. Note that the 50% zero filter was not yet applied as some of the zero measurements may be due to outlying batches.

Figure 3-9 shows the results of a PCA analysis on the QC samples. The PCA scores of all three QC samples for batches 2 and 3 fall outside the 90% confidence region (Figure 3-9A). The Hotelling's T^2 statistics for these same samples flag them as outliers (Figure 3-9B). This could have been due to some analytical error, which occurred during the preparation of the samples from these two batches. This chemical shift in the NMR spectra of the samples in these two batches, however, made it difficult to compare the peaks in these samples with those of the samples from the 21 other batches. Consequently, these two batches were excluded from further analysis as batch corrections based on only three QCs may introduce more error than they remove. We could eliminate these batches with little effect as all samples from a given case represented a batch; so, even though we sacrificed in terms of sample size, the exclusions did not change the design. Two other QC samples from different batches were also flagged; however, as only one of the three QCs in these batches was flagged, these batches were not removed.

The above processing and analysis was performed in MATLAB (2012b) using PCA function provided by the PLS-Toolbox 8.2.1 (2016), as outlined in the script provided as additional material.

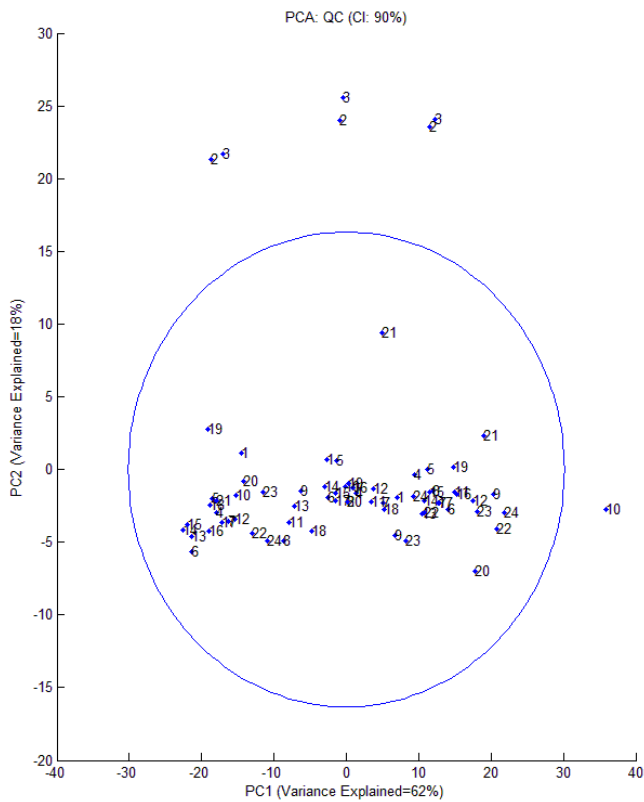
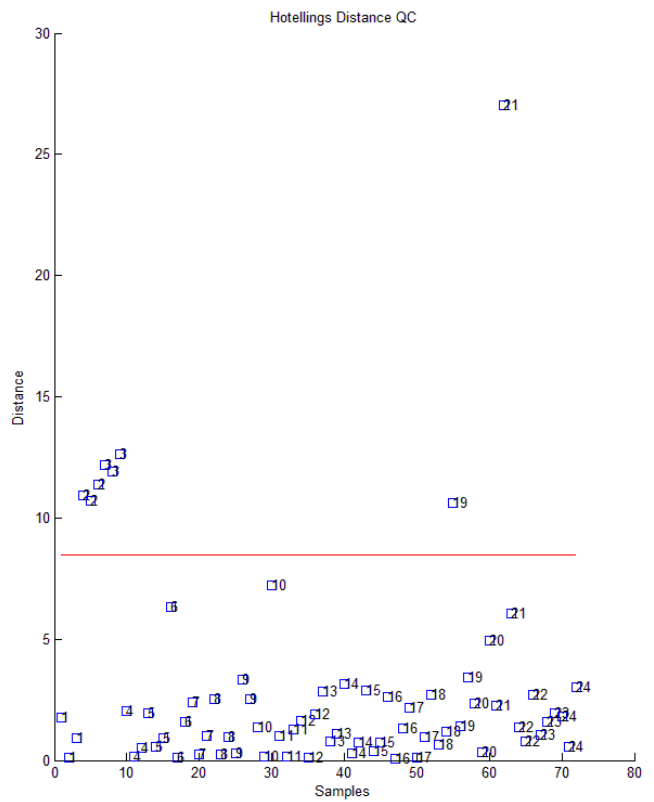
A**B**

Figure 3-9: QC Outlier Detection.

The PCA scores of all three QC samples for batches 2 and 3 fall outside the 90% confidence region (A). The Hotelling's T^2 statistics for these same samples flag them as outliers (B).

Data Transformation & Scaling

To reduce the uncertainty in the data, we applied a threshold to the data matrix followed by a 'zero-filter'. The reduced data were log transformed to ensure that the additive models used to extract information from the data were appropriate (Van den Berg *et al.*, 2006). Log transformation also makes skew distributions – often resulting when variables (i.e. bins) assume only positive values, such as here – more symmetric. Symmetry is assumed in many statistical methods, including some of the methods used here.

3.3.4 Statistical Analysis

Here we made use of various methods to compare times -1 , 1 , 2 , 3 and 4 hours with time 0 . We did this to answer the following questions:

- At what time did the effect of the vehicle first become apparent?
- At what time did the effect of the vehicle start to diminish?
- Were individuals in a fasting state more similar?
- Did their metabolism change greatly while fasting?

While the last two questions do not relate to the aims of the study, their answers influence how we interpret other results. Next, we discuss the univariate methods used. These analyses were performed in MATLAB (2012b) and included (i) univariate statistics using the statistics toolbox; (ii) PCA using the PLS-Toolbox 8.2.1 (2016); and (iii) PLS-DA using the PLS-Toolbox 8.2.1 (2016).

Univariate Analysis

The Wilcoxon signed rank test was first employed to find significant differences in the average ranks of bin values between two time points. This hypothesis test is non-parametric (i.e. makes no distributional assumptions) and assumes samples are paired. Since cases were measured repeatedly, it is important to use statistical methods that can model this dependence. Paired measures reduce some of the variation that is not of interest (i.e. between cases) and methods exploiting this are therefore better able to detect group differences. All hypothesis tests rely on random sampling, which is not the case here. Convenience sampling, as used here, can bias p-values produced by tests and they should be evaluated with caution. Further multiple testing requires controlling the family-wise error rate. However, since we did not base variable selection on hypothesis testing, these factors were less of a concern. Our aim here was simply to gain insight into the general progression of the effect of the treatment over time (i.e. whether differences became more pronounced or not).

Given the limitations of hypothesis testing listed above, other univariate measures supplemented our evaluation of the general progression. For each bin, two more intuitive measures of group

differences were also evaluated: (i) effects sizes based on the Wilcoxon signed rank z-statistic to evaluate practical significance; and (ii) fold change value.

Figure 3-10 provides the Volcano plots for all pairwise time point comparisons. The figure displays the individual Volcano plots for the pairwise comparisons of time -1 vs. time 0 (A); time 0 vs. time 1 (B); time 0 vs. time 2 (C); time 0 vs. time 3 (D); and time 0 vs. time 4 (E). A Volcano plot is essentially a scatter plot of the $\log_2$ scaled fold change values against the $-\log_{10}$ scaled p-values (in this case p-values from the Wilcoxon signed rank test). By performing scaling, the plot is able to highlight the discriminatory bins in the upper left and right corners for a given threshold. In this instance bins with p-values below 0.05 and fold change values larger than 2 were highlighted in the figures. From Figure 3-10B it is evident that individual bins start to reflect the effect of the vehicle intervention 1 hour after consumption.

Multivariate Analysis

To gain insight into the interactions between bins, i.e. to find bins which in combination differ between groups, we made use of multivariate statistical methods. Euclidean and Ward cluster analysis was used to find natural groupings between the cases based on the bin data provided, but without providing information on the known group structure (unsupervised). Cluster analysis makes use of a distance measure (Euclidean) to measure the distances or difference between samples based on the measured variables. A linkage method (Ward) was then used to group neighboring samples. The groups formed can be viewed using a dendrogram. If these groups coincide with the known group structure (as indicated by labeling the dendrogram after the fact), we can be confident that the data set contains discriminatory bins. For more detail on cluster analysis refer to Field (2009). Figure 3-11 shows the dendrograms for all pairwise time point comparisons with time 0. The interesting findings from these dendrograms are discussed in the main paper.

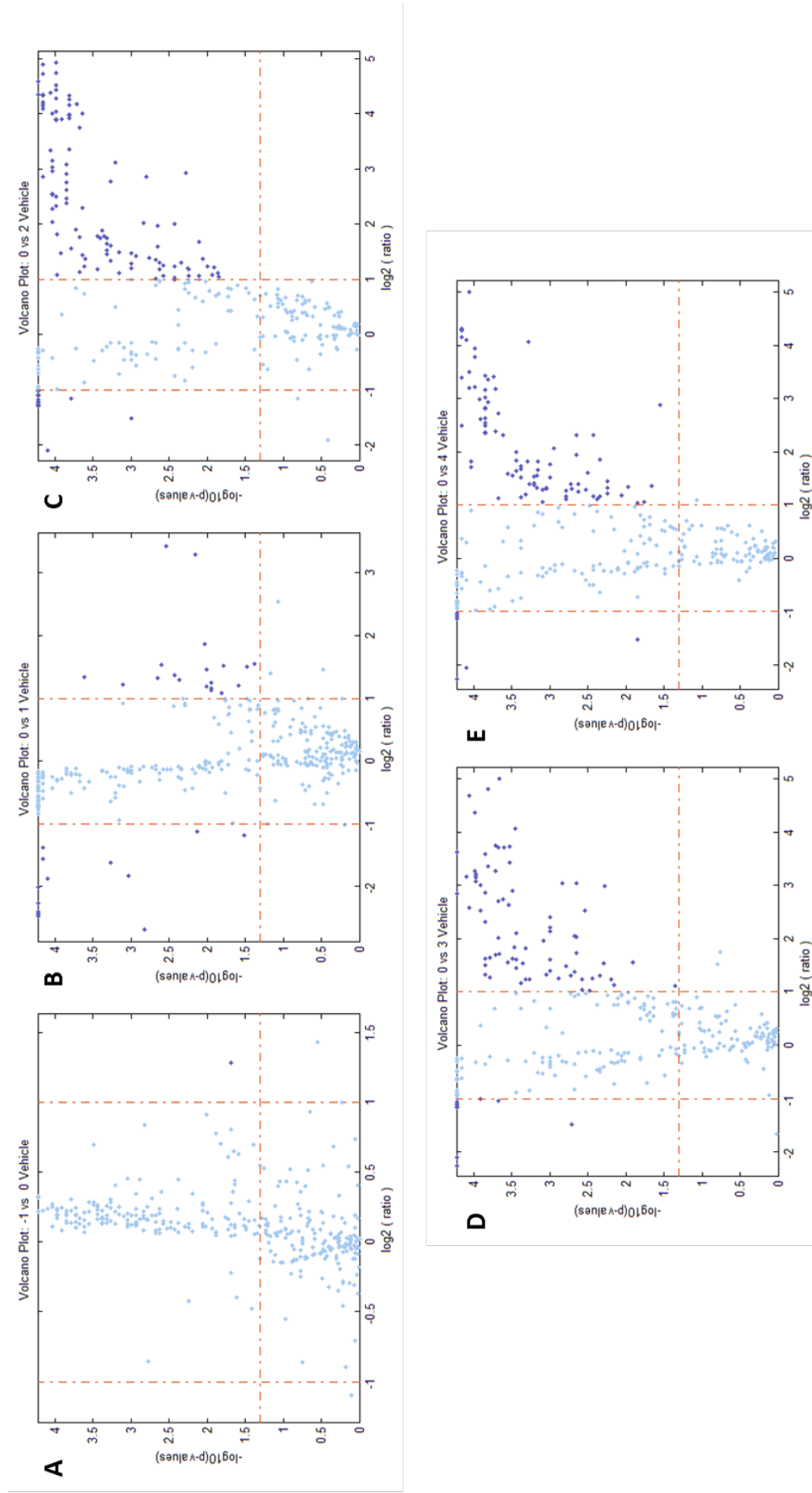


Figure 3-10: Volcano Plots of Pairwise Time Point Comparisons with time 0.

The figure displays the individual scatter plot of the log₂ scaled fold change values against the -log₁₀ scaled Wilcoxon signed rank test p-values. Each plot represents a pairwise comparison: time -1 vs. time 0 (A); time 0 vs. time 1 (B); time 0 vs. time 2 (C); time 0 vs. time 3 (D); and time 0 vs. time 4 (E).

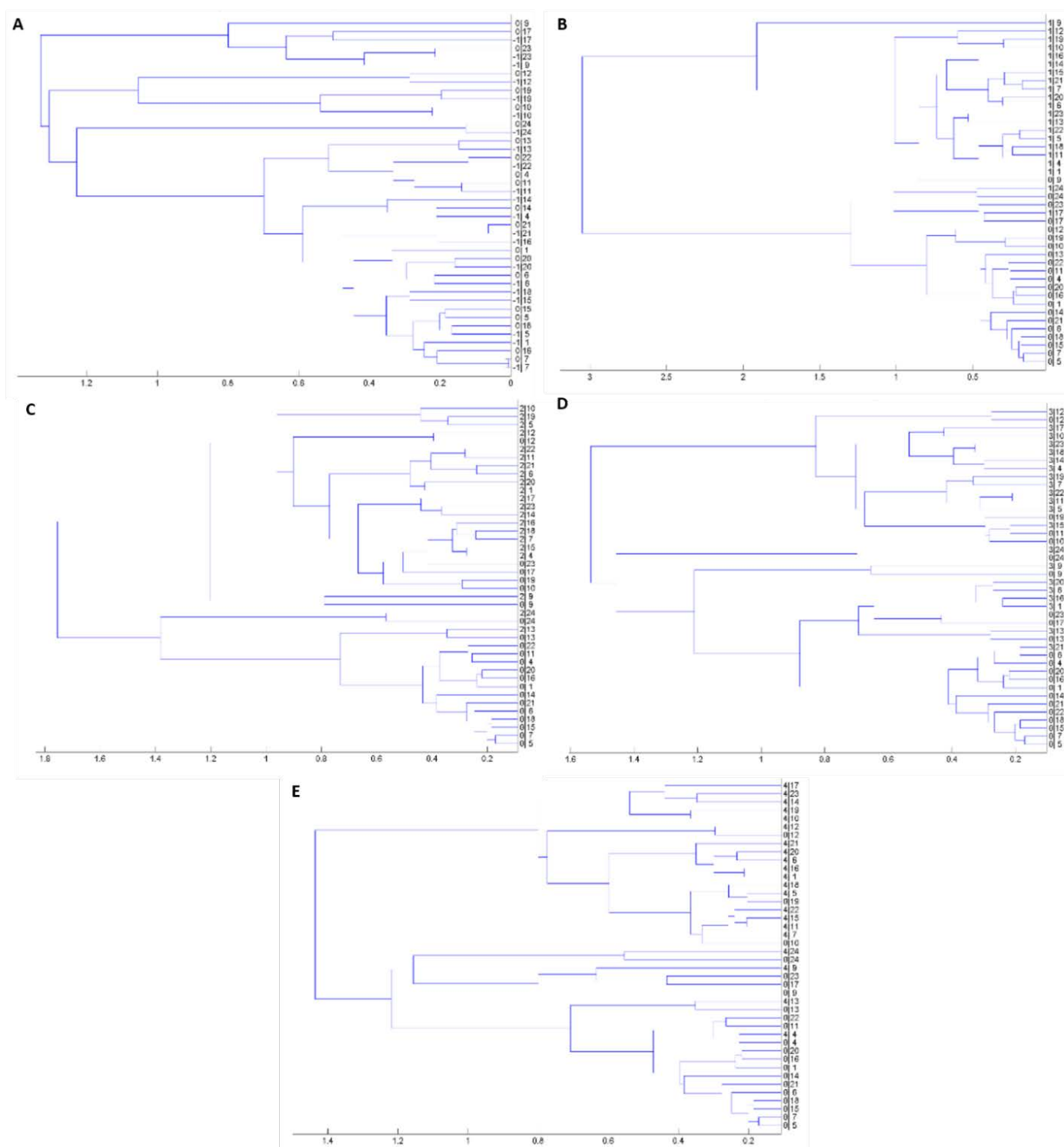


Figure 3-11: Dendrograms of Pairwise Time Point Comparisons with time 0.

The figure displays the individual dendrograms for the pairwise comparisons of time –1 vs. time 0 (A); time 0 vs. time 1 (B); time 0 vs. time 2 (C); time 0 vs. time 3 (D); and time 0 vs. time 4 (E).

Principal component analysis (PCA) was used to project the hundreds of bins onto a two-dimensional subspace using the following model:

If D represents a data matrix of N cases (i.e. samples) by K variables, where the N cases represent observations across all factor levels, then the PCA model of D can be written as: $D = TP' + E$, where T represents the scores and P' the transposed loadings matrices, respectively. Given that the dimensionality of D is reduced, the error matrix E is added to account for the variation not explained by the model (TP'). For more detail refer to (Johnson & Wichern, 2002).

Such a projection allows for a clearer view of the influence of experimental factors, specifically which factor dominates the variability in the data. PCA, like cluster analysis, is also an unsupervised method, therefore the influence of experimental factors becomes visible when labeling the scores (values for samples in the new subspace) according to associated factor level after the model has been constructed, in our case the times associated with each sample. Projecting the matrix also allows us to evaluate the influence of each variable on the projected scores. In other words, if we find that our group structure predominantly explains the covariance structure of the data, we can evaluate the loadings (values for variables or bins in the new subspace) to identify a list of variables with discriminatory ability on their own or in combination. The loadings were not used during the PCA analysis but extensively so in the ASCA model (discussed later), which is built on the PCA model.

Figure 3-12 shows the score plots for all pairwise time point comparisons with time 0. Figure 3-12A indicates how fasting state did not change over time since times -1 and 0 are very similar and the individual cases are closely comparable. Figure 3-12 B through E show how the global bin profile changes relative to the starting point of the intervention. The effect of the vehicle is already evident after the first hour, starts to diminish after two hours, and cases continue to return to normal three and four hours after the intervention.

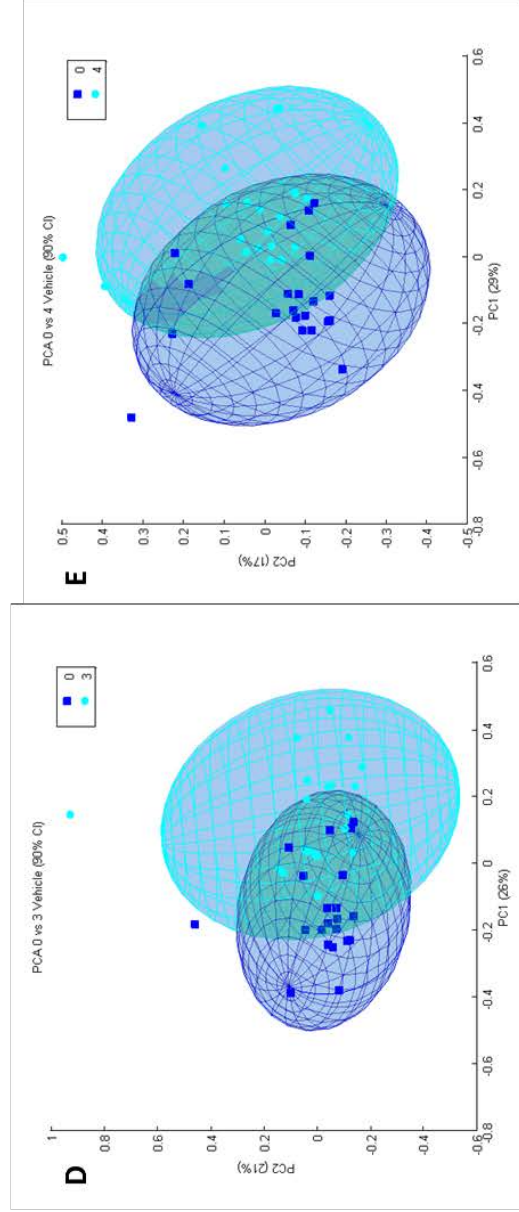
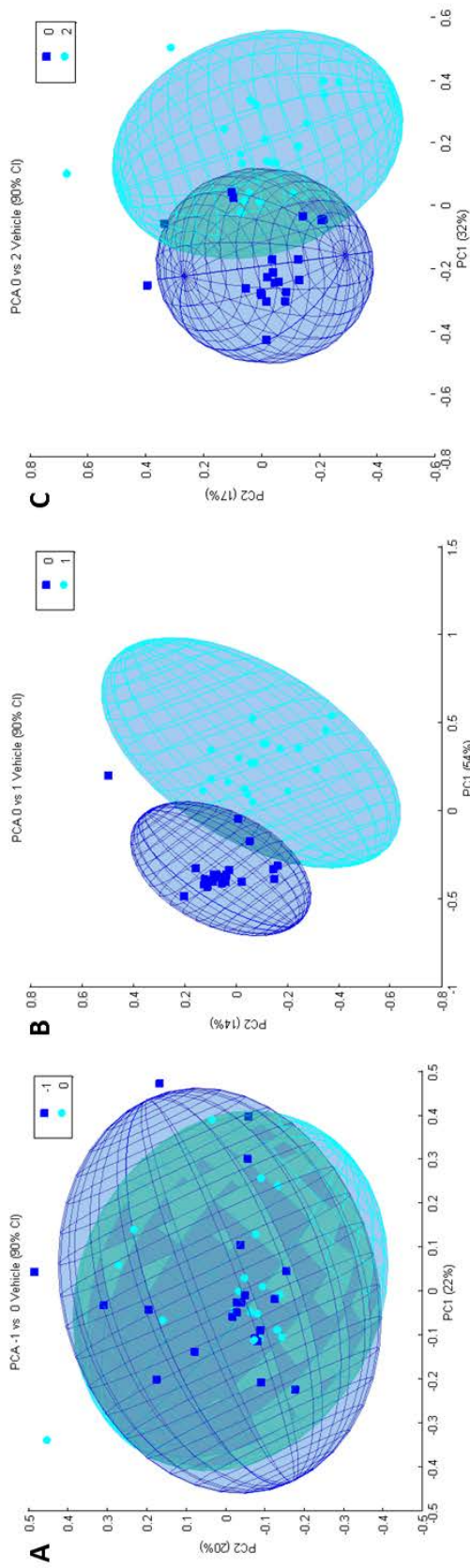


Figure 3-12: PCA Score Plots of Pairwise Time Point Comparisons with time 0.

The figure displays the individual score plots of the PCA models for the pairwise comparisons of time -1 vs. time 0 (A); time 0 vs. time 1 (B); time 0 vs. time 2 (C); time 0 vs. time 3 (D); and time 0 vs. time 4 (E).

Lastly, partial least-squares discriminant analysis (PLS-DA) was performed. This method is similar to PCA but makes use of the known group structure (i.e. it is a supervised method). The benefit of using this information is that the method is better able to identify a list of discriminatory variables. The downside is that the model itself may not generalize to other data sets as it is known to overfit (Westerhuis *et al.*, 2008). Since our aim was not to develop a classification model, we were not too concerned about this. However, we do report two measures of confidence in the bins selected: (i) R^2 to measure the goodness of the model fit; and (ii) Q^2 , the leave-one-out cross-validated R^2 , to assess generalizability or sensitivity to the individual samples used to construct the model.

Figure 3-13 shows the score plots for all pairwise time point comparisons with time 0, while Table 3-4 reports the corresponding fit statistics.

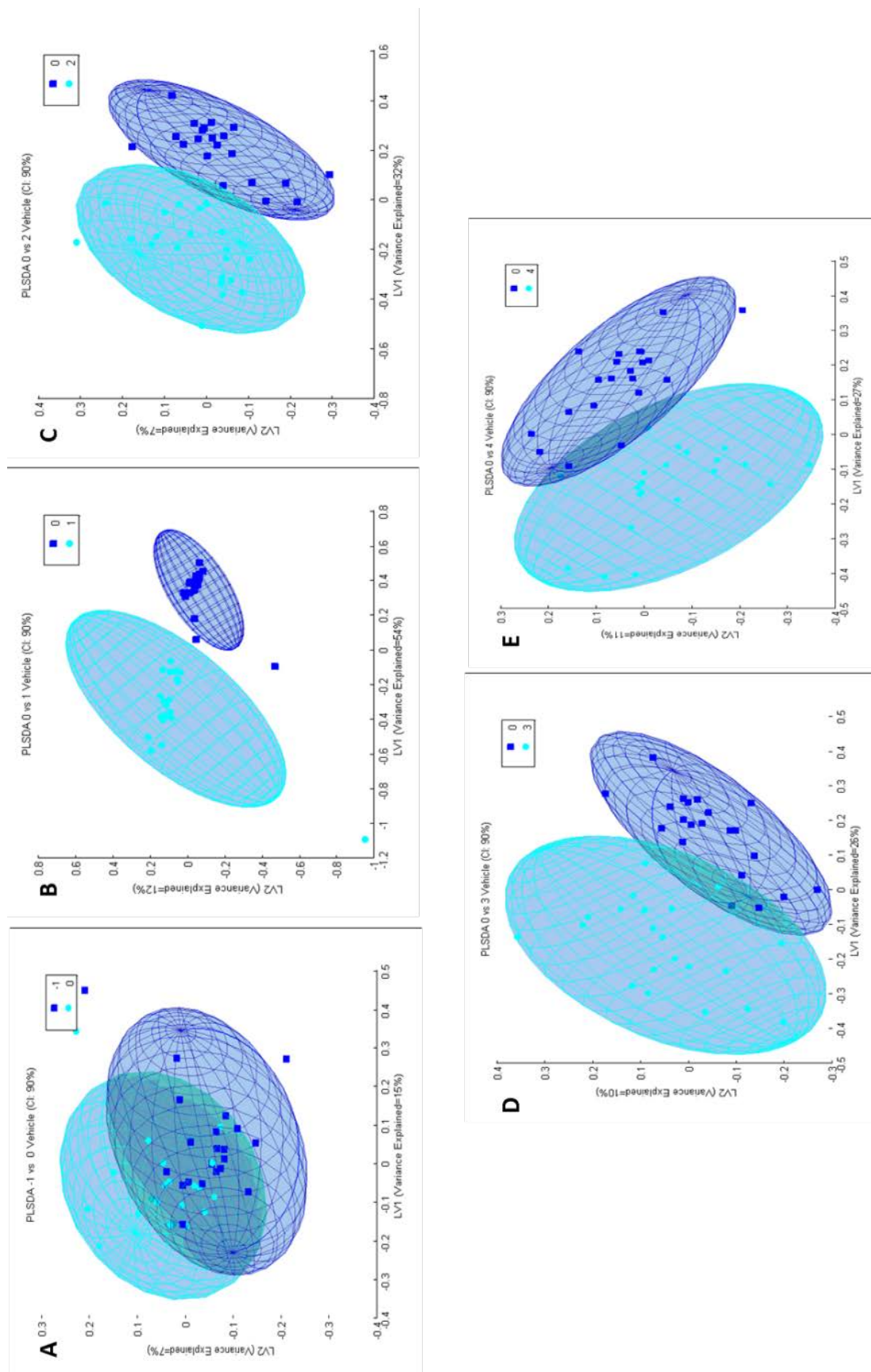


Figure 3-13: PLS-DA Score Plots of Pairwise Time Point Comparisons with time 0.

The figure displays the individual score plots of the PCA models for the pairwise comparisons of time -1 vs. time 0 (A); time 0 vs. time 1 (B); time 0 vs. time 2 (C); time 0 vs. time 3 (D); and time 0 vs. time 4 (E).

Comparison	R ²	Q ²
Time 0 vs. -1	0.171	0.002
Time 0 vs. 1	0.776	0.836
Time 0 vs. 2	0.726	0.704
Time 0 vs. 3	0.646	0.488
Time 0 vs. 4	0.628	0.501

Table 3-4: PLS-DA Fit Statistics Pairwise Time Point Comparisons with time 0.

R² to measure the goodness of model fit with respect to the group structure, while Q² represent the leave-one-out cross-validated R²

The results from the cross-sectional analysis led us to conclude that, while between-case variation is significant, we were still able to find bins that differ from their baseline measure (time 0). We also found that the fasting state was long enough to reduce between-case variation (time -1 vs. 0). We made use of the bins selected by the PLS-DA models as one of the shortlists to be included in the Venn analysis that produced the key metabolites quantified and discussed. We wanted to include bins in this list that were informative at any point in time (compared to the baseline, i.e. time 0) while accounting for the multiple PLS-DA models constructed. As such we decided to include bins with VIP values exceeding 2 in any of the models. A VIP score is calculated for each variable in the data matrix and provides an indication of the variable's importance in the projection (hence abbreviated as VIP). For more detail on this calculation refer to (Pérez-Enciso & Tenenhaus, 2003).

RM ANOVA

RM ANOVA (repeated measures analysis of variance) is a statistical method that deconstructs a data set into the known sources of variation in order to assess which of these sources significantly influenced the values recorded in the data set. This deconstruction is performed by calculating and subtracting factor level means (ANOVA) – in our case the means for the time points, as well as for each case since the same cases were measured across all time points. The primary aim of this analysis was to identify variables that changed significantly in time (across the six time points).

We constructed the ANOVA model as follows:

Let X represent all observations for a given bin and let α and β represent the Time and Case factors, respectively. We can model the change in time as follows:

$$X_{ij} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij}$$

where μ represents the overall mean of the bin

α_i represents the time effects with $i = 1 \dots 6$

β_j represents the case effects with $j = 1 \dots 21$

$\alpha\beta_{ij}$ represent the random interaction effect between case i , treatment j , i.e. the random error not explained by the model.

Therefore X_{ij} represents the observation on the bin for case j , $i - 2$ hours after consuming the vehicle, as modeled on the basis of the factor means and random noise. Bins for which the factor means play a larger role than the random noise to obtain the observed value are flagged as significant.

RM ANOVA essentially expands the cross-sectional approach to account for the complete study design of the data set (or subset since we retain only time as a factor). However, it does not account for interactions between bins as it is a univariate method. RM ANOVA can be expanded to evaluate multiple variables, but breaks down when the number of variables exceeds the number of cases, as mentioned earlier. For in-depth insight into ANOVA methods, refer to Tabachnick & Fidell (2006 & 2007). Here we used RM ANOVA to identify individual bins that changed significantly in time, as constituting the second shortlist to be included in the Venn analysis. To control the family-wise error rate resulting from applying RM ANOVA hypothesis testing multiple times, bins were selected only if their associated p-values were estimated as 0. This analysis was performed in MATLAB using the statistics toolbox (MATLAB, 2012b) and code adapted from Trujillo-Ortiz *et al.* (2004).

Unfolded PCA

Unfolded PCA was applied to gain insight into the global (i.e. multivariate) effect of the vehicle over time. This method extends the RM ANOVA idea to the multivariate setting. However, it does not explicitly take the study design into account as it does not truly decompose the data into the

known sources of variation. The data tensor (with three dimensions representing cases, times and bins) was unfolded in time as illustrated in Figure 3-14.

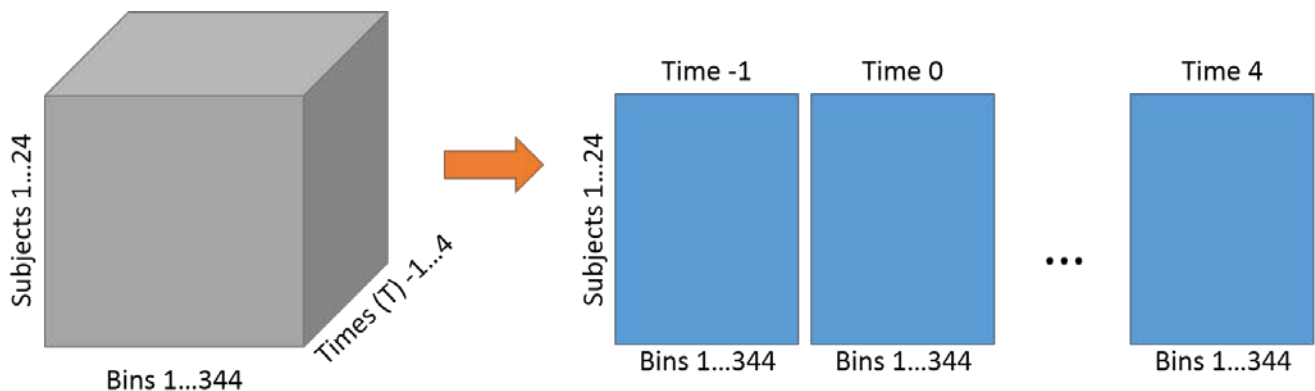


Figure 3-14: Unfolding Data.

The figure illustrates how three dimensions representing cases, times and bins are unfolded for data analysis.

Unfolding transforms a three-dimensional tensor into a two-dimensional matrix, allowing for principal component analysis. PCA of the unfolded tensor provides insight into the effect of the vehicle in time on the bins (indicated by the ellipses and centroids), as well as individual cases as indicated by the overlaid trajectories (Figure 3-5). For a more in-depth discussion on unfolding refer to Villez *et al.* (2009). This analysis was performed in MATLAB (2012b) using the PCA function of the PLS-Toolbox 8.2.1 (2016).

ASCA

ASCA (ANOVA-simultaneous component analysis) was developed for analyzing designed metabolomics data. It can not only account for the study design but also for the high-dimensionality of the data. In essence ASCA combines MANOVA and PCA (as can be deduced from the name) by first deconstructing the data as done in ANOVA and subsequently applying PCA to each subset. The model can be summarized as follows:

If D represents a data tensor then ASCA will deconstruct D similarly to RM ANOVA:

$$D = \mu + \alpha_i + \beta_j + \alpha\beta_{ij}$$

Note the adoption of boldface symbols, as the factor now represents matrices of means across all variables as opposed to just a vector of means for one variable for each factor level.

Next, PCA is performed for each matrix

$$D = 1m + T_{\alpha}P'_{\alpha} + T_{\beta}P'_{\beta} + T_{\alpha\beta}P'_{\alpha\beta}$$

For an in-depth discussion on the ASCA method, refer to Jansen *et al.* (2005) and Smilde *et al.* (2005).

As mentioned in the PCA discussion, we can evaluate the loadings (values for variables in the new subspace) to identify a list of bins with discriminatory ability on their own or in combination. The sum of the squared loadings (SSL) for the first two latent variables was used to identify perturbed bins because in combination they explained 88% of the variation. Bins were seen to have discriminatory ability if their SSL values were above 0.01, the threshold identified based on a Scree plot of all SSL scores (Figure 3-15).

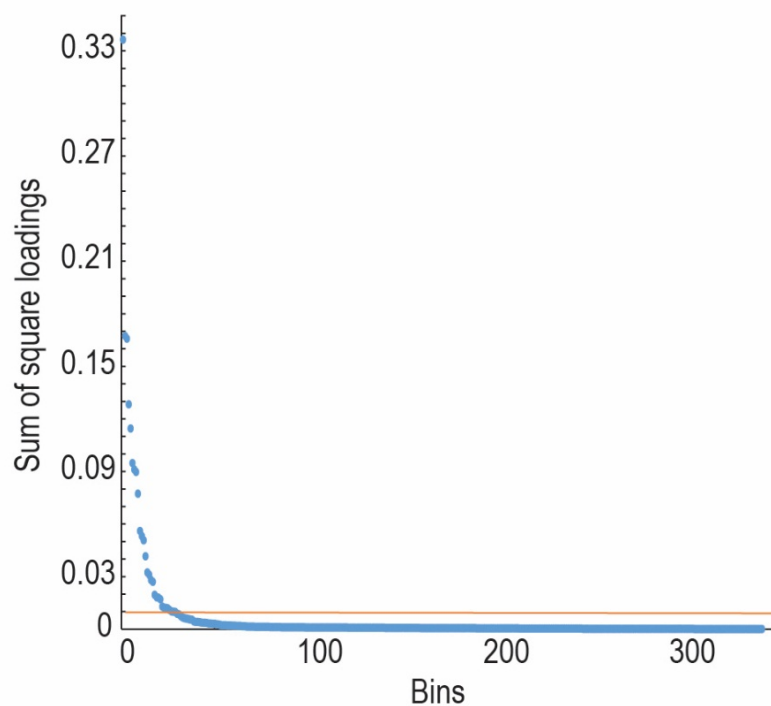


Figure 3-15: Sum of Squared Loadings of ASCA model.

The figure ranks the bins from 1 to 344 according to the sum of their squared loadings.

Though this method may have sufficed to identify key bins, as is evident from the Venn results reported in the main paper, it was used in conjunction with the cross-sectional (specifically PLS-DA) and RM ANOVA results to ensure we did not exclude potentially important bins. This analysis was performed in MATLAB using the statistics toolbox (MATLAB, 2012b) and code provided by: Gooitzen Zwanenburg (available under APACHE Licence 2.0 <http://www.apache.org/licenses/LICENSE-2.0.html>).

3.3.5 Graphs on Excretion Kinetics

The graphs shown in Figure 3-16 provide a visualization of summary statistics for the quantified data of the metabolically important metabolites summarized in Table 3-2. The raw data are scattered over a 95% confidence interval (red) and 1 standard deviation of the mean (blue). Concentration–time profiles of hippuric acid (A), creatine (B), guanidinoacetic acid (C), glycine (D), citric acid (E) and urea (F) for times 1, 2, 3 and 4 hours following consumption of 500 ml benzoic acid containing flavored water at time 0 hour.

All samples contained all metabolites, except for guanidinoacetic acid, as indicated in (C). In two cases, outliers of more than 1.5 times the interquartile range were excluded from the figures and the calculations (Hippuric acid 4250 and creatine of 2500 μ moles metabolite/mmol creatinine, both observed at 1 hour following the intervention).

Reaction kinetics are represented by concentration–time profiles of the six metabolites for the 22 participants that participated in the intervention, shown at times 1, 2, 3 and 4 hours following the intervention (time 0 hour). All quantified values of metabolites in all samples were normalized relative to the creatinine concentration recorded for each sample.

The plots in Figure 3-16 were generated using notBoxPlot.m a MATLAB function developed by Rob Campbell and can be downloaded from: <http://www.mathworks.com/matlabcentral/fileexchange/26508-raacampbell13-notboxplot>.

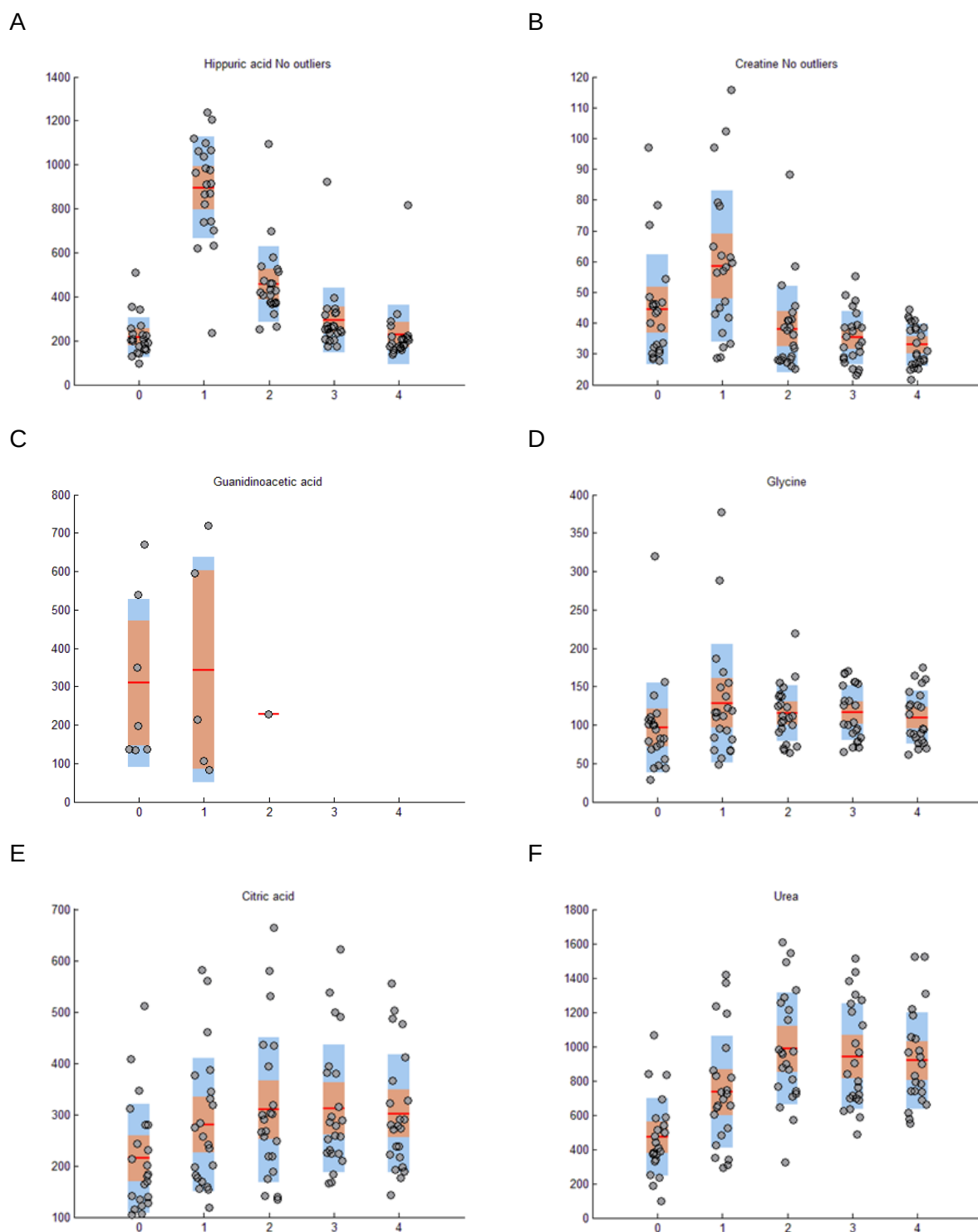


Figure 3-16: Urinary Excretion kinetics of important metabolites.

The raw data are shown as dots scattered over a 95% confidence interval (red) and 1 standard deviation of the mean (blue). Concentration–time profiles of hippuric acid (A), creatine (B), guanidinoacetic acid (C), glycine (D), citric acid (E) and urea (F) for times 1, 2, 3 and 4 hours (representing the x-axis) following consumption of 500 ml benzoic acid containing flavored water at time 0 hour.

3.3.6 NMR spectra on the excretion of six substances

The authentic spectra and power values from time dependent PCA's were used to re-analyze the outcomes of the intervention. Figure 3-17 displays the 500 MHz ^1H -NMR spectra of minor components with the top spectrum taken from a urine sample of one selected case; spectra from left to right are: (A) unknown (8.11 ppm), (B) unknown (6.485 ppm), (C) methylguanidine (2.83 ppm) and (D) unknown (1.075 ppm). The substances for times 0 (black), 1 (blue), 2 (green), 3 (red) and 4 (orange).

It appeared that three unknown substances (Figure 3-17: (A) 8.22 ppm, (B) 6.485 ppm and (D) 3.075 ppm), as well as methylguanidine (2.83 ppm - shown in Figure 3-17C) - a known endogenous metabolite, changed due to the intervention. Methylguanidine is synthesized from creatinine concomitant with the synthesis of hydrogen peroxide from endogenous substrates in peroxisomes. It will be speculative to indicate its link to the biotransformation study, but it may be linked to oxidative stress. Although none of the four substances were identified as VIP's in the PLS-DA, their concentrations changed in time following the intervention. As such, they could contribute to the centroid values returning towards the value observed before the intervention, without reaching the exact value of time 0. These observations indicate that the consequence of benzoic acid consumption is more complex than just its biotransformation to hippuric acid.

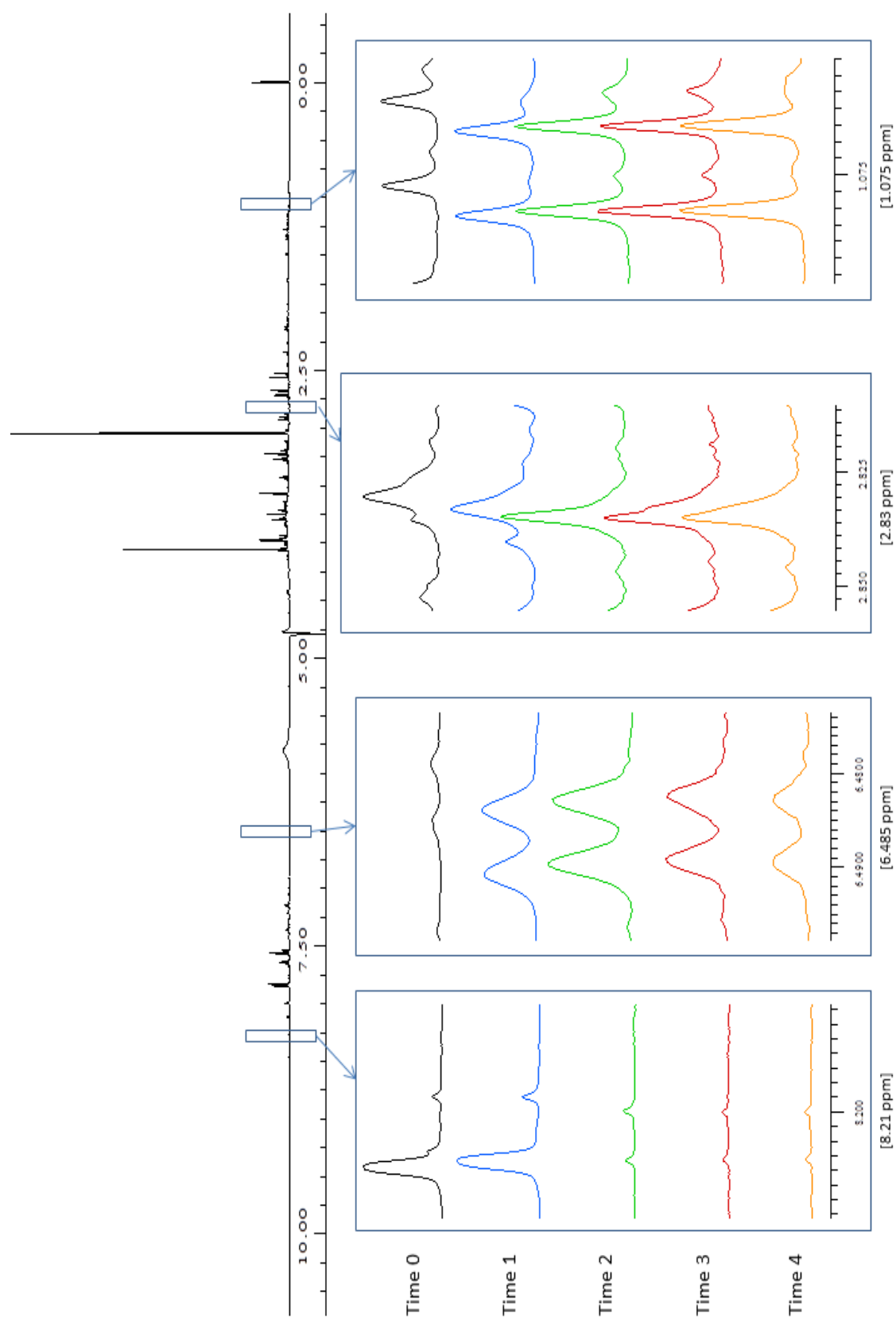


Figure 3-17: 500 MHz ^1H -NMR spectra of minor components from urine.

500 MHz ^1H -NMR spectra of minor components with the top spectrum taken from a urine sample of one selected case; spectra from left to right are: (A) unknown (8.11 ppm), (B) unknown (6.485 ppm), (C) methylguanidine (2.83 ppm) and (D) unknown (1.075 ppm). The substances for times 0 (black), 1 (blue), 2 (green), 3 (red) and 4 (orange).

3.3.7 NMR analysis on guanidinoacetic acid

Theoretically, the chemical structure of guanidinoacetic acid contains three carbon atoms of which only one has free hydrogen atoms ($-\text{CH}_2-$) that would produce a singlet in the NMR spectrum. This is observed in an example for one of our cases, as shown in the one dimensional (1D) spectrum below. Since guanidinoacetic acid does not have any correlating protons, COrrrelation Spectroscopy (COSY) two dimensional analyses is not applicable. In the spectra shown in Figure 3-18 we compared the experimental spectrum from one of our samples (in green) to that of the pure guanidinoacetic acid standard (in red) – both in 1D and 2D ^1H J-Resolved (JRES) NMR Spectroscopic analysis. The results indicate that the experimental spectrum lines up exactly to the guanidinoacetic acid singlet (3.77 ppm), confirming the identity of guanidinoacetic acid in the urinary sample. Analysis of the spectra from a urine sample shown in green and from a guanidinoacetic acid standard shown in red.

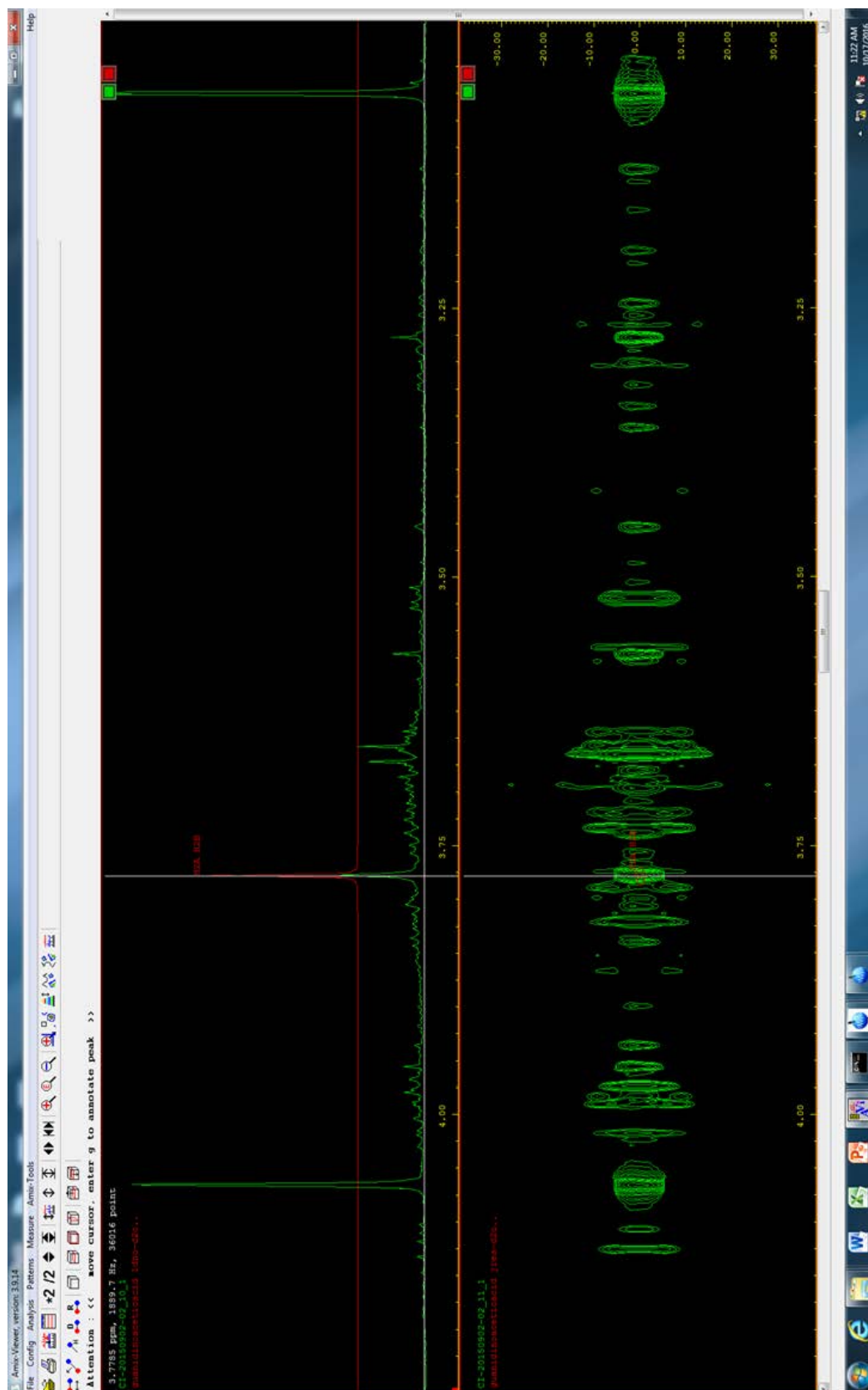


Figure 3-18: 1D and 2D JRES NMR spectra.

The figure compares the experimental spectrum from a sample (green) to that of the pure guanidinoacetic acid standard (red) in 1D (top spectra) and 2D ^1H J-Resolved or JRES (bottom spectra) NMR Spectroscopic analysis.

3.4 Accepted Publication: The ¹H-NMR-based metabolite profile of acute alcohol consumption: a metabolomics intervention study

3.4.1 Title Page

The ¹H-NMR-based metabolite profile of acute alcohol consumption: a metabolomics intervention study

Authors:

Cindy Irwin^{#1}, Mari van Reenen^{#1,2}, Shayne Mason¹, Lodewyk J. Mienie¹, Ron A. Wevers³, Johan A. Westerhuis^{2,4} and Carolus J. Reinecke^{*1}

#Set of equal first authors:

These authors contributed equally to this work:

Cindy Irwin email: cindyirwin74@gmail.com

Mari van Reenen email: 12791733@nwu.ac.za

***Corresponding author:**

Carolus J. Reinecke: email: carools.reinecke@nwu.ac.za

Addresses:

¹Centre for Human Metabolomics, Faculty of Natural Sciences and Agriculture, North-West University (Potchefstroom Campus), Potchefstroom, South Africa

²Department of Statistics, Faculty of Natural Sciences and Agriculture, North-West University (Potchefstroom Campus), Potchefstroom, South Africa

³Department of Laboratory Medicine, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands

⁴Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, The Netherlands

Other email addresses:

Shayne Mason: nmr.nwu@gmail.com

Lodewyk J. Mienie: japie.mienie@gmail.com

Ron A. Wevers: ron.wevers@radboudumc.nl

Johan A. Westerhuis: j.a.westerhuis@uva.nl

3.4.2 Abstract

Metabolomics studies of disease conditions related to chronic alcohol consumption provide compelling evidence of several perturbed metabolic pathways underlying the pathophysiology of alcoholism. The objective of the present study was to utilize proton nuclear magnetic resonance (^1H -NMR) spectroscopy metabolomics to study the holistic metabolic consequences of acute alcohol consumption in humans. The experimental design was a crossover intervention study which included a number of substances to be consumed — alcohol, a nicotinamide adenine dinucleotide (NAD) supplement, and a benzoic acid-containing flavoured water vehicle. The experimental subjects — 24 healthy, moderate-drinking young men — each provided six hourly-collected urine samples for analysis. Complete data sets were obtained from 20 of the subjects and used for data generation, analysis and interpretation. The results from the NMR approach produced complex spectral data, which could be resolved sufficiently through the application of a combination of univariate and multivariate methods of statistical analysis. The metabolite profiles resulting from acute alcohol consumption indicated that alcohol-induced NAD^+ depletion, and the production of an excessive amount of reducing equivalents, greatly perturbed the hepatocyte redox homeostasis, resulting in essentially three major metabolic disturbances — up-regulated lactic acid metabolism, down-regulated purine catabolism and osmoregulation. Of these, the urinary excretion of the osmolyte sorbitol proved to be novel, and suggests hepatocyte swelling due to ethanol influx following acute alcohol consumption. Time-dependent metabolomics investigations, using designed interventions, provide a way of interpreting the variation induced by the different factors of a designed experiment, thereby also giving methodological significance to this study. The outcomes of this approach have the potential to significantly advance our understanding of the serious impact of the pathophysiological perturbations which arise from the consumption of a single, large dose of alcohol — a simulation of a widespread, and mostly naive, social practice.

Keywords

acute alcohol, metabolomics, proton magnetic resonance spectroscopy, metabolite profile, nicotinamide adenine dinucleotide

3.4.3 Introduction

The use of metabolomics in intervention, or challenge, studies greatly enhances the holistic understanding of the effects of single, or combined, consumed substances on metabolic pathways (Gibney *et al.*, 2005; Wishart, 2008; Irwin *et al.*, 2016). In this paper, we present the experimental design and outcome of an intervention study, which included a number of interventions: (1) vehicle only — commercial flavoured water containing sodium benzoate as preservative (Irwin *et al.*, 2016); (2) a defined acute dose of alcohol, consumed alongside the vehicle; (3) an NAD supplement, taken one hour prior to the study, followed by consumption of the vehicle; and (4) the NAD supplement, followed by consumption of the vehicle and the alcohol dose.

Alcohol oxidation by alcohol dehydrogenase (ADH) and aldehyde dehydrogenase (ALDH) results in the reduction of oxidized NAD (NAD⁺) to its reduced form (NADH), thereby generating a highly reduced cytosolic environment in hepatocytes. An increased NADH:NAD⁺ ratio has been described to influence several metabolic processes, resulting in, amongst others: decreased glycolysis (Bradford & Rusyn, 2005); reduced Krebs cycle (Hawkins & Kalant, 1972; Zakhari & Li, 2007); and decreased gluconeogenesis (Zakhari & Li, 2007; Lieber, 1997). The use here of NAD as a supplement is based on a perceived view that it has a therapeutic capacity if taken prior to alcohol consumption; the theory being that it supplements alcohol-induced NAD⁺ depletion.

The data set of the present study was generated from longitudinal (urine samples, collected hourly over a five-hour period), multi-subject (several experimental participants) and multi-group (in this case four, separate interventions) measurements, and expressed as multivariate data - numerous variables generated and identified through a metabolomics approach utilizing ¹H-NMR spectroscopy. The main aim of the investigation was to determine the effect of acute (a single dose indigested in a short time) alcohol consumption, as well as the potential effect of an NAD supplement on the resulting metabolite profile.

Several metabolomics studies have been reported on the pathophysiological consequences of chronic consumption and alcoholism in humans (Nahon *et al.*, 2012; Rachakonda *et al.*, 2014; Amathieu *et al.*, 2014; Lian *et al.*, 2011), as well as on chronic intragastric alcohol feeding of rodents, followed by mass spectrometric based metabolite profiling (Gika *et al.*, 2012; Loftus *et al.*, 2010). The metabolomics technique has been applied for global metabolite profiling using GC–MS (Gao *et al.*, 2011; Kirpich *et al.*, 2016), in several LC–MS studies (Gika *et al.*, 2012; Loftus *et al.*, 2010; Manna *et al.*, 2011; Gika & Wilson, 2016), as well as in untargeted studies using ¹H-

NMR spectroscopy (examples hereof are summarized in Table 3-5). ^1H -NMR spectroscopy is a robust method with broad metabolite-class coverage (albeit with limited sensitivity), and is well suited for the diverse metabolomics studies listed in Table 3-5 using rodents (Nicholas *et al.*, 2008; Bradford *et al.*, 2008; Masuo *et al.*, 2009; Fernando *et al.*, 2010; Yoseph *et al.*, 2013), selected human volunteers at risk for cardiovascular diseases (Vázquez-Fresno *et al.*, 2012), and population-based cohorts (Würtz *et al.*, 2016). With the exception of the study by Bradford *et al.* (2008), all the studies listed in Table 3-5 used serum or tissue material as the samples for the generation of the metabolite data.

Here we report on an untargeted NMR metabolomics intervention study on acute alcohol consumption, which expands the limited data on alcohol-induced metabolic profiling in humans. This study followed a non-invasive sampling procedure (hourly-collected urine samples) which is ideal for generating the untargeted ^1H -NMR-based metabolic profiles required for the longitudinal experimental design. As a biological waste material, urine is the biofluid of choice (Bouatara *et al.*, 2013) to reflect the metabolic breakdown products following acute alcohol consumption, as well as any potential by-products derived from the gut microbiome, which has been shown to be important in several alcohol-related studies (Gao *et al.*, 2011; Manna *et al.*, 2011; Bala *et al.*, 2014). The data pre-processing applied to the spectral data in the present study are comparable to the approaches applied in the studies listed in Table 3-5, albeit with minor differences.

The statistical methods used in the intervention studies listed in Table 3-5 all use a pre vs post approach — multivariate model-based analysis (PCA, PLS-DA, OPLS and ANOVAs) of the data was performed to determine the metabolites responsible for the separation of the control ('pre') versus the alcohol-treated ('post') groups. Analysis of the data set from the present study is theoretically complex and required methods specifically designed for longitudinal, multi-subject, multi-group and multivariate data (Antti *et al.*, 2002; Smilde *et al.*, 2005). We accordingly generated a matched-sample series through a crossover study of participating subjects, collecting samples over a distinct time frame. First, data from the 'vehicle only' and 'alcohol plus vehicle' interventions were compared at each time point to determine the impact of the consumption of a single, large dose of alcohol. Next, the statistical analysis focused on the metabolic perturbations due to the acute consumption of alcohol, measured over time. A shortlist of significantly perturbed NMR spectral bins was compiled and quantified as relative concentrations for biological interpretation. Finally, the effect of NAD supplementation was evaluated for a select list of biologically relevant metabolites.

Most remarkable in the present study was the highly increased presence of urinary sorbitol, a response to alcohol consumption not observed in the studies listed in Table 3-5, nor previously reported elsewhere. The other main metabolic observations that emanated from this NMR-based intervention study was the increased levels of urinary lactic acid (indicated as a biomarker in two of the listed rodent studies) and hypoxanthine (not observed in the studies listed in Table 3-5, but observed in another metabolomics study on patients with liver cirrhosis (Lian *et al.*, 2011)). Both these observations strongly support the view that the reduced cytosolic environment is the primary metabolic consequence of acute alcohol consumption.

Taken together, the metabolic information acquired from this metabolomics study further underscores the view that alcohol consumption is associated with severe risks and remains one of the world's leading health risk factors for disability, morbidity and mortality. This is succinctly expressed in the 2014 World Health Organization global status report on alcohol (World Health Organization, 2014): *"Of all deaths worldwide, 5.9% are attributable to alcohol consumption; this is greater than, for example, the proportion of deaths from HIV/AIDS (2.8%), violence (0.9%) or tuberculosis (1.7%)"*.

Authors (year) & Summary of design	Data pre-processing	Statistical analysis	Conclusion on metabolic observations
Rodent studies			
Nicolas et al (2008) C: Sprague-Dawley rats (N=5) S: Serum and tissue E: Single dose/binge protocol A: Untargeted	Zero-filled Fourier transformed Baseline corrected	D: Spectral bins T: Mean centred G: Pre vs Post U: t-test M: PCA	A cycle of aborted gluconeogenesis stimulated by the increased NADH/NAD ratio but short-circuited by decreased alanine levels during hepatic ethanol metabolism.
Bradford et al (2008) C: Male C57Bl/6J mice (N=4) S: Urine, serum and liver E: Chronic: 7–36 g/kg/day; Fed <i>ad libitum</i> A: Untargeted	Spectra phased Baseline corrected Normalized to the sum of all integrals set to 1000	D: Spectral bins T: Pareto scaled G: Pre vs Post U: t-test M: PCA & OPLS	Energy utilization is important in understanding the pathogenesis of alcohol induced liver injury, with acetylglutamine, n-acetylglycine and taurine potential novel non-invasive markers of alcohol consumption and oxidative stress.
Masuo et al (2009) C: Female Fisher rats (N=5) S: Liver and brain E: Chronic: 10 mo 15% Sake; Fed <i>ad libitum</i> A: Untargeted	Spectra normalize to total integral area Referenced to Na-TMS-tetra- ¹ H-propionate	D: Spectral bins T: Pareto scaled G: Pre vs Post U: t-test M: PCA & PLS-DA	In liver: An attenuation of mitochondrial function; In brain: (1) Perturbed amino acids (Increased and decreased); and (2) Decreased N-acetyl-aspartate, taurine and GABA. Chronic Sake intake may cause alterations in the intoxicated body but also in the next generation.
Fernando et al (2010) C: Male Fischer rats (N=344) S: Plasma and liver E: Chronic: 5% alcohol diet A: Lipidome	Spectra manually phased Referenced to TMS Spectra divided into equal bins of 0.01 ppm width Baseline corrected	D: Spectral bins T: Auto scaled G: Pre vs Post U: t-test M: Hierarchical clustering & PCA	Alcohol consumption alters metabolism of cholesterol, triglycerides and phospholipids that could contribute to the development of fatty liver and indicates that fatty liver precedes oxidative stress and inflammation.
Yoseph et al (2013) C: Male FVB/N mice (N:8–9) S: Pancreatic tissue E: Chronic; Fed <i>ad libitum</i> with increased interventions A: Untargeted/Focus: Pancreatic metabolome	Spectra phased; Baseline corrected Spiking experiments to validate suspected metabolites identified in the spectra.	D: 52 Metabolites (mM) T: Variance scaled G: Pre vs Post U: t-test M: PLS-DA & One-way ANOVA (Tukey post-test)	Pancreatic metabolome following chronic alcohol intake indicates (1) increased acetate, adenosine, xanthine, acetoacetate, 3-hydroxybutyrate and betaine; and (2) decreased cytidine, uracil, fumarate, creatine phosphate creatine, and choline. Mice with chronic alcohol ingestion have increased mortality when encountered with sepsis.
Human studies			
Vázquez-Fresno et al (2012) C: Males volunteers at risk (Age≥55; N=61) S: 24-hour urine E: Nutritional wine A: Untargeted and selected metabolites	Untargeted: Spectra divided into equal bins Profiling integration; Spectra phased Baseline corrected; Calibrated Selected metabolites: Chenomx NMR Suite 7.0 profiler compared on:HMDB/BMRB/MMCD and literature	D: Spectral bins T: Pareto scaled G: Pre vs Post U: t-test; Pearson's correlation M: ANOVA & Fisher's LSD	(1) Food metabolome: Mannitol (diet); tartrate (wine intake). (2) Endogenous modifications after wine consumption indicated by branch-chain amino acids. (3) Gut metabolites, 4-hydroxyphenylacetate and hippurate
Wurtz et al (2016) C: 3 population-based cohorts, (meta-analysis N=9778) S: Serum E: Habitual consumption A: Targeted lipids and 86 metabolic measures	Automated NMR metabolite profiling: Robotics-controlled and fully automated with a capacity of about 150–180 samples in 24 h. Integrated computational methods for the data-driven systems biology approach to biomedical research	D: Computational generated T: Log _e transformed & SD Scaled G: Cross-sectional (i.e. 2 group comparisons) U: Bonferroni method; t-test & Mann-Whitney test M: Linear regression (R-squared)	Prominent metabolic associations with alcohol consumption include monounsaturated fatty acids, omega-6 fatty acids, glutamine, citrate and lipoprotein particle size. Many of these cardiometabolic biomarkers strongly associated with alcohol intake as HDL cholesterol.
Irwin et al (the present study) C: Healthy males (Age=22; N=24) S: Hourly urine (5 times) E: Acute (1.5 g vodka/kg) A: Untargeted and quantified metabolites	Normalized relative to the creatinine at 4.05 ppm Baseline corrected – 50% zero-filter Batch comparisons using QC samples	D: Spectral bins (excluding H ₂ O region) T: log _e transformed; Auto-scaled G: Cross-over, time dependent matched series U: WR-test; Effect size; Fold change values M: Hierarchical cluster; Multi-level PCA NAD-effect; Two-way RM ANOVA	Aims: (1) Effect of vehicle consumption: Irwin <i>et al</i> (2016) PLOS ONE. 11:e0167309 (2) Impact of consumption of a single large dose of alcohol (time dependence) (3) Statistical analysis of quantified metabolites following consumption (4) Effect of NAD supplementation (5) Model of the metabolite profile following acute alcohol consumption
Abbreviations used in columns 1 and 3: C: Cases studied (experimental animals or humans); S: Samples used for metabolite identification; E: Ethanol exposure; A: Approach for the generation of the variable or metabolite data; N: number of experimental cases studied; D: Data used for statistical analyses; T: Transformation and scaling approach; G: Groups compared; U: Univariate analyses applied; M: Multivariate analyses applied.			

Table 3-5: A summary of the experimental designs, data analysis approaches and main metabolic conclusions of some NMR-based ethanol administration studies.

3.4.4 Materials and methods

Chemicals and reagents

The substances used for the interventions reported here were: commercial flavoured water as vehicle (aQuellé lemon-flavoured sparkling water, containing carbonated natural spring water; fructose; citric acid; flavouring; sodium benzoate preservative; sodium cyclamate, aspartame, acesulfame K non-nutritive sweeteners; and vitamin C – www.aquelle.co.za, product of South Africa); commercial alcohol for consumption (Smirnoff No. 21 triple-distilled vodka: 43% alcohol – product of South Africa); NAD (*NAD ASSIST* – product of Future Health, South Africa); and commercial bottled water (Valpré still spring water, inorganic contents specified – product of South Africa). The internal standard for the ^1H -NMR analysis was trimethyl-2,2,3,3-tetradeuteriopropionic acid (TSP, sodium salt; Sigma Aldrich).

Experimental subjects and protocol

The participants in the intervention study consisted of a group of 24 medically confirmed healthy males, of various ethnicities, between 20 and 24 years of age. All of the participants were in a healthy, athletic condition, with weights in the 61–92 kg range. They were neither alcohol addicts nor total abstainers, but confirmed their use of alcohol at a moderate social level. No participants took any medication, all were asked to refrain from vitamins, minerals, and other supplementation, and were requested to follow a similar dietary and lifestyle pattern for the duration of the study. The complete protocol, described previously (Irwin *et al.*, 2016), was approved by the Health Sciences Ethical Committee of North-West University (Ethical approval number: NWU-00045-12-S1) and conducted in accordance with guidelines for good clinical practice. The study was performed at the Health Clinic of the university, under the supervision of a medical doctor as well as a nurse, and all participants could leave the premises only after approval by the doctor.

The experiments were conducted on Saturday mornings between 08:00 and 12:00. All participants had to abstain from breakfast and had to provide an early morning urine sample, collected one hour before the start of the experiment (time –1). The participants were randomly assigned to an intervention group until all 24 had participated in all four interventions ('vehicle only', 'alcohol plus vehicle', 'vehicle plus NAD' and 'vehicle plus alcohol and NAD'). Owing to commitments of some participants, the experiments were performed over a period of 7 consecutive Saturdays. However, three of the participants failed to participate in all four interventions, and their samples were therefore excluded from the data generation process.

Through data analysis, one further participant was marked as an outlier in one of the interventions (see Figure 3-21 below), and was therefore also excluded from further analyses. Thus, the data generated were based on the complete sets of information obtained from 20 participants.

The four interventions consisted of the consumption of: (1) 500 mL lemon-flavoured water only, to measure the baseline effect of the vehicle; (2) 1.5 g vodka per kg body mass; (3) one tablet containing 50 mg NAD; and (4) 1.5 g vodka per kg body mass plus 50 mg NAD. In all three of the latter interventions, the substances were consumed with 500 mL flavoured water as vehicle. NAD (where applicable) was taken one hour before the start of the experiment, directly after collection of the first (time -1) urine sample. The dose and time schedule for consuming the alcohol (maximally five minutes) comply with established criteria to result in a tolerable but moderately severe level of acute alcohol intoxication (Numminen *et al.*, 2000). All participants were provided with 1.5 L pure spring water, which was the only substance that could be consumed over the four-hour period of sample collection. Urine samples were collected at time 0, just prior to consumption of the substances, followed by four further samples at 1, 2, 3 and 4 hours thereafter, providing six samples in total from each participant for each of the four interventions. All samples were treated, stored, prepared and analysed according to the protocol described previously (Irwin *et al.*, 2016) and included in Section 3.5.1 of the Supporting Information (refer to Section 3.5). Although repeatability and reproducibility are not major concerns in NMR analyses (Viant *et al.*, 2009), the measurement design included the use of pooled quality control (QC) samples to estimate any batch effect or other interfering analytical aspect. The collected samples were analyzed in 24 separate batches, each batch containing the 24 samples of a single subject and three QC samples. The batch analysis order is given in Section 3.5.1. The quality assurance results (given in 3.5.2) indicated that the QC samples clustered close together and therefore no batch correction was required.

Uric acid does not have NMR-detectable protons at physiological pH, making it essentially “*NMR invisible*” (Bouatra *et al.*, 2013). Uric acid concentrations were therefore determined in each sample individually on a ThermoFisher Scientific Clinica Chemistry Analyzer (type 863), using ThermoFisher Scientific Uric Acid Reagents (described in greater detail in Section 3.5.3).

Data analysis

The original NMR spectral data (referred to in Section 3.5.4) were subjected to pre-processing — data were normalized relative to the CH₂ creatinine peak (at 4.05 ppm), and very low values were replaced with zero, before performing a 50% zero-filter (details in Section 3.5.2). Further bin

reduction, based on batch comparisons of the coefficient of variation of the bins from the QC samples, yielded 289 bins for further analyses. The resulting data were analysed along all dimensions in order to understand the metabolic effect of the different interventions in time, as well as the differences between the participants, even when considering a relatively homogeneous group such as the subjects involved in this study. The analysis was performed on different cross-sections and blocks of the data tensor. Univariate as well as multivariate methods were applied. Data were log transformed (natural log with shift parameter equal to 1) and auto-scaled (per bin) prior to univariate (parametric tests only) and multivariate analysis.

The first analysis aimed to shed light on the impact of the consumption of a single, large dose of alcohol. To illustrate this effect, data from the 'vehicle only' and 'alcohol plus vehicle' interventions were compared at each point in time. Cross-sectional analysis made use of multivariate techniques, including hierarchical cluster analysis (Euclidian distance and Ward linkage) and multi-level principal component analysis (ML-PCA) (Westerhuis *et al.*, 2010). Univariate approaches included determining Wilcoxon signed-rank test (WRT) p-values and associated effect size, as well as fold change (FC) values.

Next, statistical analysis focused on the metabolic perturbations that result from the consumption of alcohol as measured in time. Through the application of a combination of statistical methods, and using time 0 as a reference point, or control group, a shortlist of significantly perturbed NMR spectral bins was compiled and quantified from the 'alcohol plus vehicle' intervention data for biological interpretation. This list of important bins was extended through the inclusion of metabolites that were well presented in the spectra, although they did not contribute significantly to the statistical differentiation within the 'alcohol plus vehicle' consumption profiles.

Lastly, the effect of NAD supplementation was of interest. In order to evaluate this intervention, a select list of biologically relevant metabolites was modelled across the time period and across all four interventions using a two-way repeated measures analysis of variance (2-way RM ANOVA) model. The significance of the differences was calculated using Greenhouse–Geisser-corrected p-values. In addition, Wilcoxon signed-rank tests were performed to calculate the significance of the differences between selected pairs of measurements.

3.4.5 Results

An example of a ^1H -NMR spectrum (10.00 to 0.00 ppm), generated from the analysis of a representative experimental sample collected one hour following the 'alcohol plus vehicle' intervention, is shown in Figure 3-19a, together with enlargements of four sections of the spectrum to further illustrate some detail (Figure 3-19b–e): EtOH = ethanol (1.18 t, 3.64 q); LA = lactic acid (1.33 d, 4.12 q); SA = succinic acid (2.41 s); TMAO = trimethylamine-N-oxide (3.27 s); CA = citric acid (2.61 AB); MA = methylamine (2.61 s); TMA = trimethylamine (2.90 s); DMG = N,N-dimethylglycine (2.93 s); DMF = N,N-dimethylformamide (2.87 s, 3.02 s); CT = creatine (3.04 s, 3.93 s); GLY = glycine (3.57 s); SO = sorbitol (3.60–3.69 m, 3.73 d, 3.74–3.80 m, 3.82 d, 3.85 m); IS = indoxyl sulphate (7.51 d, 7.70 d); HX = hypoxanthine (8.20 d); HA = hippuric acid (3.97 d, 7.56 tt, 7.64 tt, 7.84 dd); Cr = creatinine (3.05 s, 4.06 s). The following were not observed in the present spectrum: fumaric acid (6.52 s) and 3-hydroxybutyric acid (1.20 d, 2.36 m, 4.15 m).

The dominating peaks in Figure 3-19a were from ethanol (due to the alcohol consumption), hippuric acid (the biotransformation product of benzoic acid, derived from the gut microbiome and present in the vehicle used in all the interventions, as well as observed in a moderate red wine nutritional study) (Vázquez-Fresno *et al.*, 2012), creatinine (a normal constituent of urine), and trimethylamine-N-oxide (TMAO), a known osmolyte and protein stabilizer.

Most notable was the presence of an exceedingly complex area, approximately between 3.60 and 3.90 ppm (Figure 3-19b). Interpretation of this region is particularly difficult, since it may contain overlapping resonances from several metabolites.

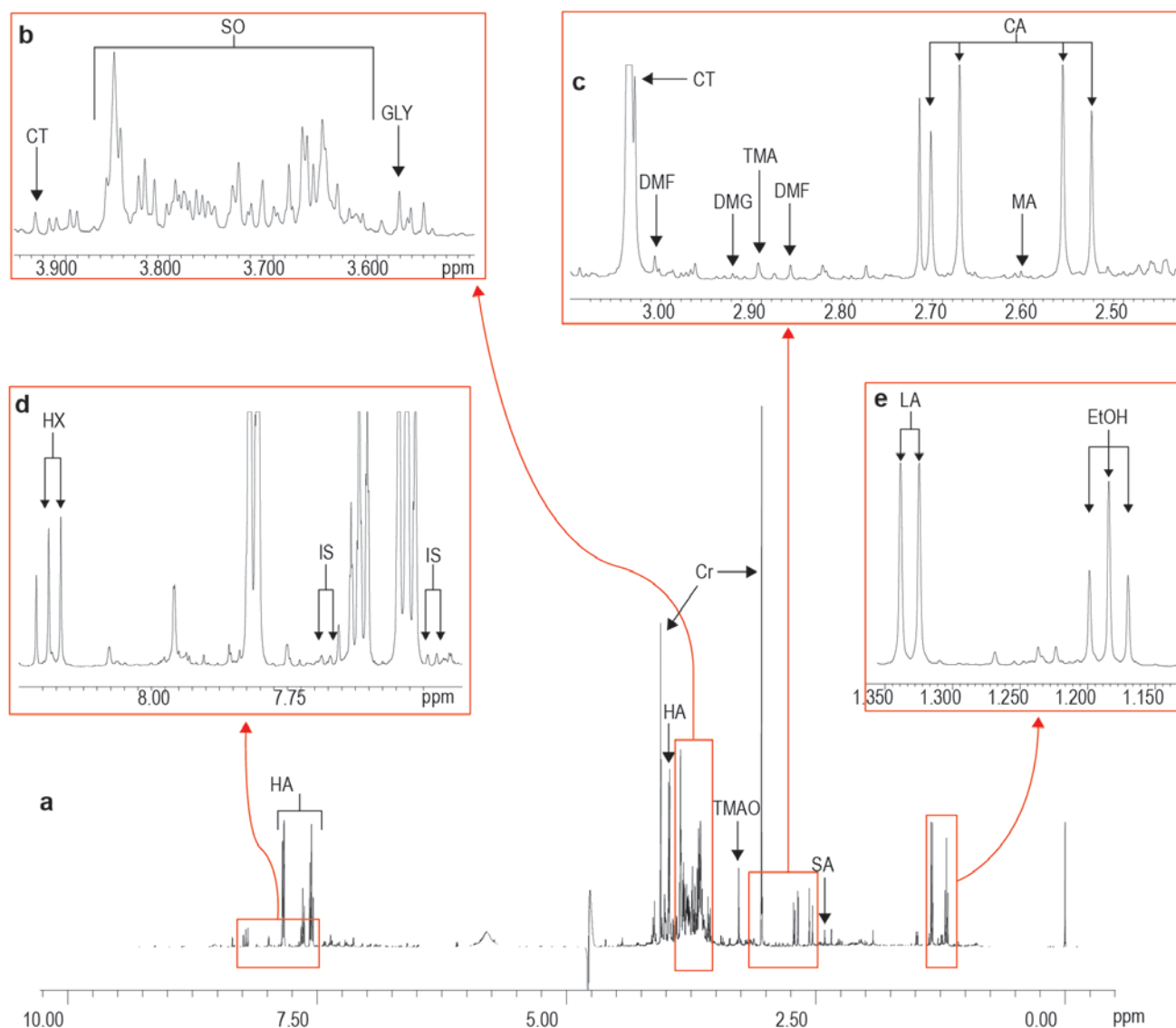


Figure 3-19: Representative ^1H -NMR spectrum of urine collected one hour following the 'alcohol plus vehicle' intervention.

A ^1H -NMR spectrum (10.00 to 0.00 ppm) with enlargements of four sections (b–e) highlighting the following compounds: EtOH = ethanol (1.18 t, 3.64 q); LA=lactic acid (1.33 d, 4.12 q); SA=succinic acid (2.41 s); TMAO=trimethylamine-N-oxide (3.27 s); CA=citric acid (2.61 AB); MA=methylamine (2.61 s); TMA=trimethylamine (2.90 s); DMG=N,N-dimethylglycine (2.93 s); DMF=N,N-dimethylformamide (2.87 s, 3.02 s); CT=creatine (3.04 s, 3.93 s); GLY=glycine (3.57 s); SO=sorbitol (3.60–3.69 m, 3.73 d, 3.74–3.80 m, 3.82 d, 3.85 m); IS=indoxyl sulphate (7.51 d, 7.70 d); HX=hypoxanthine (8.20 d); HA=hippuric acid (3.97 d, 7.56 tt, 7.64 tt, 7.84 dd); Cr=creatinine (3.05 s, 4.06 s).

The strong signals observed at 3.60–3.69, 3.73, 3.74–3.80 and 3.85 ppm were suggestive of sorbitol in the urine, following alcohol consumption. This suggestion was confirmed by comparing the spectra obtained through two-dimensional (2D) correlation spectroscopy (COSY) analysis of an experimental sample, obtained one hour after alcohol consumption, and a sample containing sorbitol as a standard. The confirmatory two-dimensional spectra is shown in Figure 3-20: (a) 1D ¹H-NMR of a representative urine sample collected one hour after alcohol consumption, zoomed into the 3.50–3.90 ppm region (black), compared to the pure compound spectrum of sorbitol (blue). (b) Correlating 2D ¹H-NMR COSY, confirming the sorbitol annotation based upon proton correlation.

Next, an area with several related amines or amine derivatives is shown in Figure 3-19c, while Figure 3-19d indicates the presence of hypoxanthine (an intermediate in the catabolism of purines towards uric acid) and indoxyl sulphate (a biotransformation product of tryptophan). Finally, Figure 3-19e indicates the presence of ethanol itself, as well as the lactic acid doublet (1.33 ppm), the main known marker of lactic acidosis, and one of the primary effects of acute alcohol consumption.

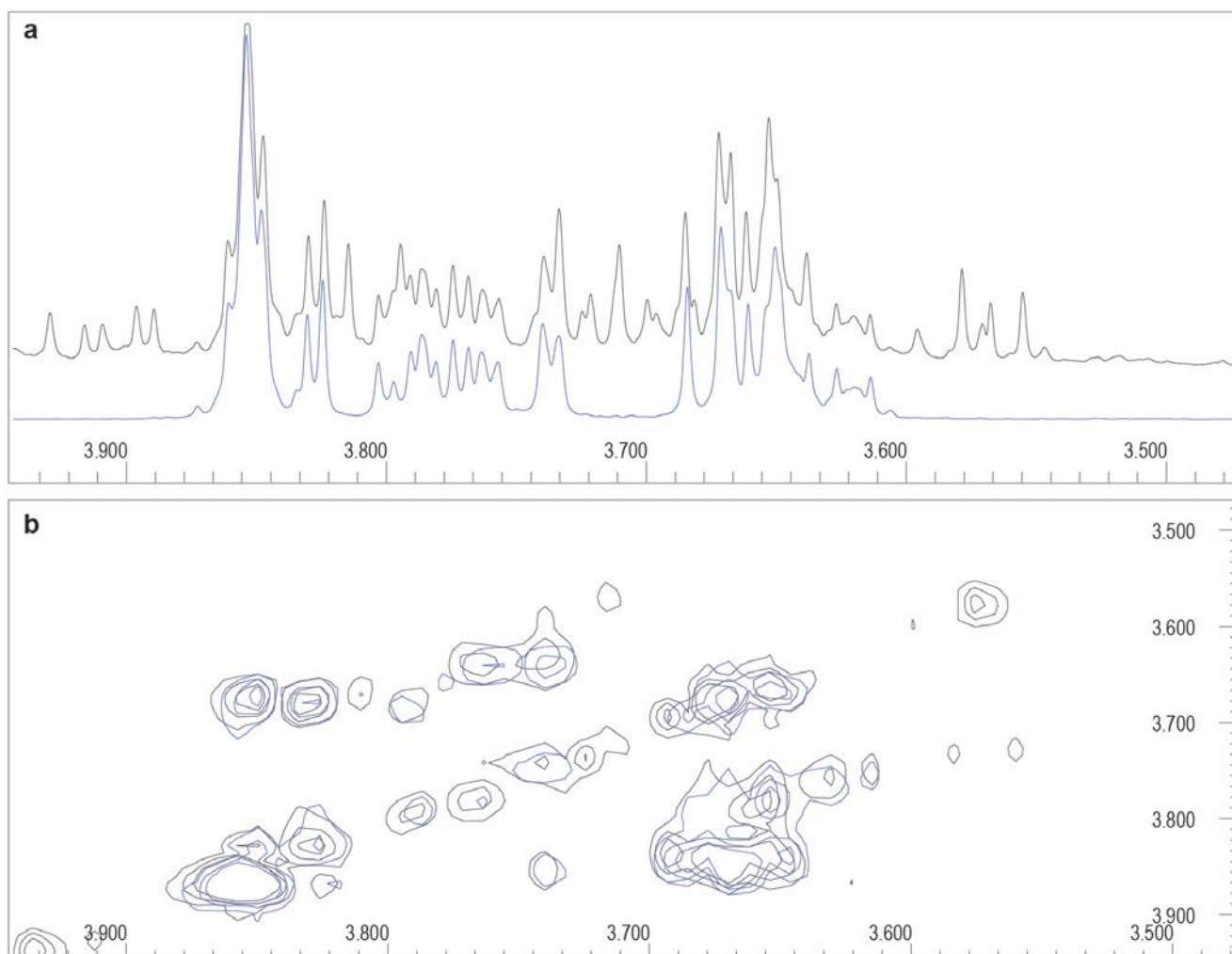


Figure 3-20: Confirmation of sorbitol annotation.

The top image (a) displays 1D ¹H-NMR 3.50–3.90 ppm region of a urine sample collected one hour after alcohol consumption (black), compared to the pure compound spectrum of sorbitol (blue). The bottom image (b) displays the correlating 2D ¹H-NMR COSY, confirming the sorbitol annotation based upon proton correlation.

Spectral analysis of the representative sample gives a clear indication of the important metabolites present in urine following alcohol consumption (Figure 3-19), but no information on the underlying dynamic effects following the consumption. The next step was therefore to determine the time-dependent impact of the consumption of a single, large dose of alcohol. Statistical analyses of the data obtained from the experimental samples from the ‘vehicle only’ and ‘alcohol plus vehicle’ interventions provide such information, and are illustrated in Figure 3-21.

The respective analyses were constructed on subsets of the data representing the same three time points — time 0 (a, d and g), 2 hours (b, e and h) and 4 hours (c, f and i) following the two interventions. Data from the 21 participants in the dendrograms and ML-PCA plots are shown as blue dots/areas for the ‘vehicle only’ intervention and pink dots/areas for the ‘alcohol plus vehicle’ intervention. The single outlier is shown as a red square in the dendrograms. All data from this participant were excluded from further analyses, resulting in the analysis of the data from a total of 20 participants.

Figure 3-21 a and d (time 0) showed no group separation prior to vehicle or alcohol consumption. It did, however, become apparent that one individual may have used an exogenous substance on the day of his ‘vehicle only’ intervention experiment, which completely separated this person from all others in the group (indicated as a red square in Figure 3-21a). This rendered the data from this individual unfit for analysis, producing the final number of 20 cases used for further analysis. The bin profiles obtained two hours after the ‘vehicle only’ and ‘alcohol plus vehicle’ interventions clearly indicated group differences in the unsupervised analyses due to the addition of alcohol (Figure 3-21 b and e). The Volcano plot from data collected one hour after alcohol consumption revealed a larger number of bins with significant up- and down-regulated values ($p \leq 0.05$ and $|FC| \geq 2$) relative to the number of bins from data collected one hour after vehicle consumption (data not shown). This number of significant bins increased to at least 58 out of the 289 bins (20%) two hours following consumption (Figure 3-21h). The number of these bins progressively decreased 3 and 4 hours after consumption, becoming only 13 bins after 4 hours that differed significantly between the two interventions (Figure 3-21i).

Note: The NMR spectra generated were dominated by the presence of hippuric acid (see Irwin et al., 2016). Exclusion of hippuric acid from the data set, however, did not significantly affect the relevant information on alcohol consumption.

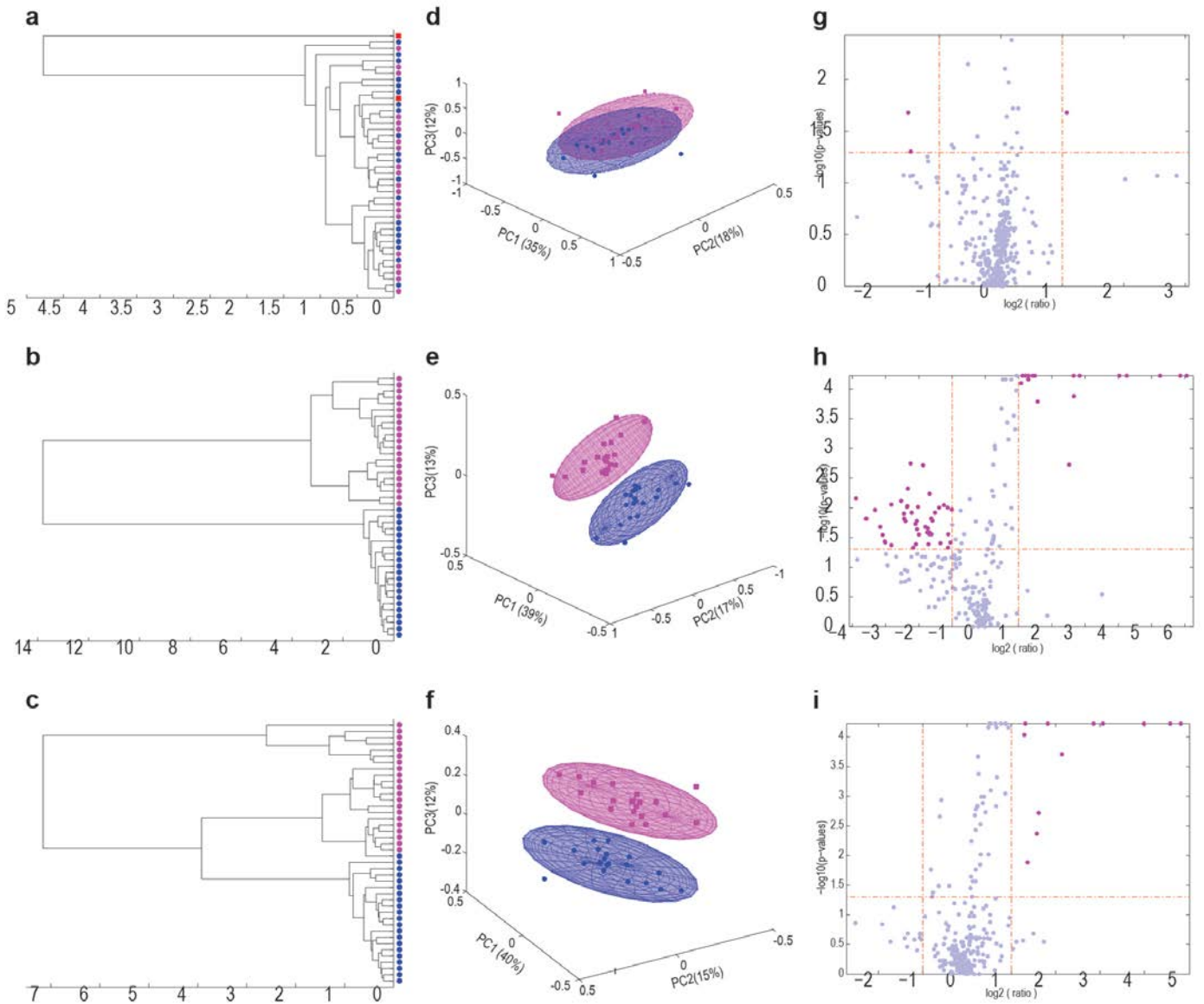


Figure 3-21: Group separation between participants, based on bins data from the ‘vehicle only’ and the ‘alcohol plus vehicle’ interventions.

The results shown represent three time points — time 0 (a, d and g), 2 hours (b, e and h) and 4 hours (c, f and i) following the two interventions. In figures a through f dots represent 21 participants, blue for the ‘vehicle only’ intervention and pink for the ‘alcohol plus vehicle’ intervention, a single outlier is shown as a red square in figure a. Dendrograms a to c were generated through hierarchical cluster analysis, while figures d to f represent ML-PCA scores plots. Figures g to i are volcano plots which show significant bins in pink, defined as such if $p \leq 0.05$ (inverted log scaled on the y-axis) and $|FC| \geq 2$ (log scaled on the x-axis), and non-significant bins in blue.

Given this marked effect, the next set of analyses were performed to identify and rank the bins most affected by the consumption of alcohol using time 0 as a point of reference. Bins were shortlisted if they differed significantly within the 'alcohol plus vehicle' intervention at any point in time (relative to time 0) based on a significant WRT p-value ≤ 0.05 and $|FC| \geq 2$. This shortlist of bins was linked to a set of metabolites and quantified. This list, including only the most perturbed metabolites, was then extended through the inclusion of metabolites which were well presented in the spectra, but did not contribute significantly to the statistical differentiation within the 'alcohol plus vehicle' consumption profiles. Quantified uric acid data (not identified by NMR analysis, but determined individually for each urine sample) was also added to this final shortlist of 13 important metabolites. The concentrations of these 14 metabolites, at all five time points related to the 'alcohol plus vehicle' intervention, are given in Table 3-6, together with the relevant summary statistics, determined one (early effect) and four (late effect) hours after alcohol consumption. Time -1 is not included in Table 3-6 since it reflects the past several hours prior to the intervention, and is not related to the 'alcohol plus vehicle' intake. The purpose of Table 3-6 was then to rank metabolites for biological interpretation of the consequences of the intervention.

Variable	Time 0 vs Time 1 (early effect)				Time 0				Time 1				Time 2				Time 3				Time 4				Time 0 vs Time 4 (late effect)			
	WRT p-value		WRT Effect Size		Mean		SD		Mean		SD		Mean		SD		Mean		SD		Mean		SD		WRT p-value		WRT Effect Size	
	[BH Adjusted p-value]		[FC]																						[BH Adjusted p-value]		[FC]	
DMF (N,N-dimethylformamide)	0.057 [0.079]		0.301 [+1.038]		15.98 [2.292]		16.58 [2.613]		15.64 [2.795]		16.01 [2.308]		16.58 [2.613]		15.64 [2.795]		16.01 [2.308]		16.58 [2.613]		17.12 [2.699]		17.12 [2.699]		0.017 [0.047]		0.378 [+1.071]	
DMG (N,N-dimethylglycine)	0.052 [0.079]		0.307 [+1.140]		8.472 [7.965]		9.656 [10.02]		9.190 [9.230]		8.168 [7.640]		9.656 [10.02]		9.190 [9.230]		8.168 [7.640]		9.656 [10.02]		8.438 [7.274]		8.438 [7.274]		1.000 [1.000]		0.000 [−1.004]	
Ethanol	<0.0001 [0.0004]		0.620 [> +100]		0.000 [0.000]		1389 [1556]		6448 [2673]		3440 [1984]		1389 [1556]		6448 [2673]		3440 [1984]		1389 [1556]		1912 [1250]		1912 [1250]		<0.0001 [0.001]		0.620 [> +100]	
Glycine	0.044 [0.077]		0.319 [+1.145]		160.2 [84.90]		183.4 [86.26]		151.4 [44.68]		146.6 [50.45]		183.4 [86.26]		151.4 [44.68]		146.6 [50.45]		183.4 [86.26]		158.3 [56.11]		158.3 [56.11]		0.601 [0.701]		0.083 [−1.012]	
Hippuric acid	0.313 [0.313]		0.159 [+1.725]		220.4 [217.2]		380.2 [393.7]		87.66 [124.4]		138.1 [344.7]		380.2 [393.7]		87.66 [124.4]		138.1 [344.7]		380.2 [393.7]		178.9 [333.7]		178.9 [333.7]		0.100 [0.156]		0.260 [−1.232]	
Hypoxanthine	<0.0001 [0.0004]		0.620 [+8.187]		9.604 [7.029]		78.63 [45.97]		29.82 [14.99]		14.44 [5.716]		78.63 [45.97]		29.82 [14.99]		14.44 [5.716]		78.63 [45.97]		12.48 [5.191]		12.48 [5.191]		0.044 [0.077]		0.319 [+1.299]	
Indoxyl sulphate	0.002 [0.005]		0.490 [+1.147]		24.30 [10.29]		27.86 [13.42]		24.87 [16.04]		21.51 [14.25]		27.86 [13.42]		24.87 [16.04]		21.51 [14.25]		27.86 [13.42]		22.87 [11.19]		22.87 [11.19]		0.794 [0.855]		0.041 [−1.063]	
Lactic acid	0.001 [0.004]		0.508 [+2.537]		70.90 [61.39]		179.9 [274.8]		92.74 [19.59]		75.96 [14.04]		179.9 [274.8]		92.74 [19.59]		75.96 [14.04]		179.9 [274.8]		72.79 [16.26]		72.79 [16.26]		0.006 [0.022]		0.431 [+1.027]	
Methylamine	0.006 [0.013]		0.431 [+1.159]		4.556 [2.146]		5.282 [2.477]		5.937 [2.256]		5.186 [2.380]		5.282 [2.477]		5.937 [2.256]		5.186 [2.380]		5.282 [2.477]		5.059 [2.263]		5.059 [2.263]		0.033 [0.073]		0.336 [+1.110]	
Sorbitol	<0.0001 [0.0004]		0.620 [> +100]		0.000 [0.000]		653.2 [482.0]		852.8 [625.7]		426.2 [331.6]		653.2 [482.0]		852.8 [625.7]		426.2 [331.6]		653.2 [482.0]		289.0 [253.3]		289.0 [253.3]		0.000 [0.001]		0.614 [> +100]	
Taurine	0.062 [0.079]		0.295 [+1.076]		100.6 [29.71]		108.2 [32.54]		109.7 [29.86]		113.5 [37.45]		108.2 [32.54]		109.7 [29.86]		113.5 [37.45]		108.2 [32.54]		118.7 [36.49]		118.7 [36.49]		0.006 [0.022]		0.437 [+1.180]	
TMAO (trimethylamine N-oxide)	0.002 [0.004]		0.502 [+1.139]		51.18 [23.90]		58.32 [25.48]		60.31 [29.26]		62.45 [35.68]		58.32 [25.48]		60.31 [29.26]		62.45 [35.68]		58.32 [25.48]		62.00 [32.48]		62.00 [32.48]		0.037 [0.073]		0.331 [+1.211]	
Trimethylamine	0.100 [0.117]		0.260 [+1.133]		1.753 [1.229]		1.987 [1.657]		1.763 [1.680]		1.654 [1.578]		1.987 [1.657]		1.763 [1.680]		1.654 [1.578]		1.987 [1.657]		1.615 [1.436]		1.615 [1.436]		0.167 [0.234]		0.218 [−1.085]	
Uric acid	0.247 [0.266]		0.183 [−1.351]		0.963 [0.760]		0.713 [0.279]		0.336 [0.168]		0.487 [0.243]		0.713 [0.279]		0.336 [0.168]		0.487 [0.243]		0.713 [0.279]		0.743 [0.566]		0.743 [0.566]		0.211 [0.269]		0.198 [−1.296]	

- All quantified values, except those for uric acid (expressed as mmol/L), are from NMR-determined urine analyses, and are expressed as $\mu\text{mol metabolite}/\text{mmol creatinine}$.
- WRT p-values are based on the comparison of the respective metabolite concentrations relative to time 0 for time 1 and time 4 of the "alcohol plus vehicle" intervention.
- P-values adjusted for multiple testing (14 tests in total) are also reported based on the Benjamini & Hochberg (BH) approach for controlling the rate of false discoveries.
- Positive and negative fold change values indicate up- and down-regulation of metabolites, relative to time 0, respectively

Table 3-6: Quantified data of important metabolites following alcohol.consumption.

The data summarized in Table 3-6 indicate that seven metabolites (ethanol, hypoxanthine, indoxyl sulphate, lactic acid, methylamine, sorbitol and TMAO) were significantly up-regulated ($p \leq 0.05$) in the first hour following alcohol consumption. Five of these metabolites (ethanol, hypoxanthine, lactic acid, sorbitol and TMAO) remained up-regulated at every time point up to time 4 after alcohol consumption, although less significantly so ($p \leq 0.05$) than after the first hour. This observation illustrates a general characteristic of metabolic profiling, well known in the area of inherited diseases: serial examinations of urinary metabolites show that the amounts of these acids excreted varies greatly in time following a perturbation, mostly due to different metabolic consequences related to the perturbation. The uric acid excretion profile observed in this study further emphasizes this characteristic. The profile of urinary uric acid differed distinctly from those of the other metabolites mentioned — its concentration decreased from time 0, and became significantly reduced 2 and 3 hours after alcohol consumption ($p \leq 0.05$ relative to time 0). Thereafter, its concentration steadily increased to near its initial level after 4 hours. Three other metabolites, DMF, taurine and glycine, became significantly up-regulated at some stage following alcohol consumption. The up-regulation of two metabolites, DMG and trimethylamine, was not significant at any time point relative to time 0 following the intervention. The excretion profile of hippuric acid followed the same trend as previously reported for the ‘vehicle only’ intervention (Irwin *et al.*, 2016).

These mean values of the metabolites over time indicate the dynamic aspect following alcohol consumption, but do not reflect individual responses towards the intervention. Inter-individual variation has already been well established even for consumption of only the vehicle used in this study (Irwin *et al.*, 2016), and therefore, this aspect is not discussed in detail here. Instead, focus was placed on the individual, as well as the group, concentration changes of the five important metabolites that were significantly up-regulated during the whole time period following alcohol consumption, as well as on the concentration changes of uric acid, owing to its unique excretion profile.

Figure 3-22 displays subplots as a set of solid lines (and one dotted line) representing the observations from the 20 individuals across 5 time points for each of the six metabolites shown to be significantly up-regulated in Table 3-6. Each subplot displays a set of lines (coloured solid lines for 19 of the individuals and a red dotted line for one highlighted individual) representing the observations from the 20 individuals across 5 time points for a given metabolite: (a) ethanol; (b) sorbitol; (c) TMAO; (d) lactic acid; (e) hypoxanthine; and (f) uric acid. The black dashed lines

represent one potential LOWESS (locally weighted scatterplot smoothing) regression for each metabolite against time using a bi-weight kernel.

Most individuals excreted a limited amount of ethanol in their urine one hour after alcohol consumption, and, generally, ethanol excretion peaked two hours following alcohol consumption (Figure 3-22a). Likewise, most individuals showed a distinct excretion of the osmolyte sorbitol one hour after alcohol consumption (Figure 3-22b), albeit at different concentrations. The LOWESS regression curve for sorbitol, however, peaked at 2 hours for the group as a whole. The concentration of TMAO, another osmolyte, increased significantly from time 0 (Figure 3-22c), reaching a plateau at 3 hours. Its final concentration was, however, still higher ($p = 0.037$) than the baseline value at time 0.

The next observations from Figure 3-22 are closely associated with the known alcohol-induced disturbance of the $\text{NAD}^+:\text{NADH}$ ratio in hepatocytes. The excretion of lactic acid for the group peaked at 2 hours, but 4 individuals showed an early and excessive response one hour after alcohol consumption (Figure 3-22d). The profile of the LOWESS regression curve for hypoxanthine (Figure 3-22e) distinctly peaked one hour following alcohol consumption. Two of the early high lactic acid responders showed similar, high hypoxanthine excretion profiles. The general profile of uric acid excretion (Figure 3-22f) was a mirror image of that of hypoxanthine, indicating the NAD^+ -dependence of its formation through dehydrogenation of hypoxanthine (catalysed by xanthine dehydrogenase; EC 1.17.1.4).

Most individuals followed the general trend indicated by the LOWESS regression line, although one showed an early excretion of ethanol following alcohol consumption (shown as a red dotted line in Figure 3-22). Compared to the group, this individual also presented with a high excretion of sorbitol and TMAO, peaking at one hour, an early and extremely high lactic acid excretion, high excretion of hypoxanthine, but comparable excretion of uric acid. Since this person did not present as an outlier, we attribute this variation to the unique response of this individual to alcohol consumption. Diversity in the excretion profiles of the individuals within the group, emphasized by the one highlighted person, clearly illustrates the importance of inter-individuality, previously described for the consumption of only the vehicle (Irwin *et al.*, 2016).

Taken together, the combination of observations summarized in Table 3-6 and illustrated in Figure 3-22 indicate that the cases themselves are a noteworthy source of variation, while the group response provided the opportunity for the development of a holistic model (illustrated in Figure 3-24) of the effect of acute alcohol consumption.

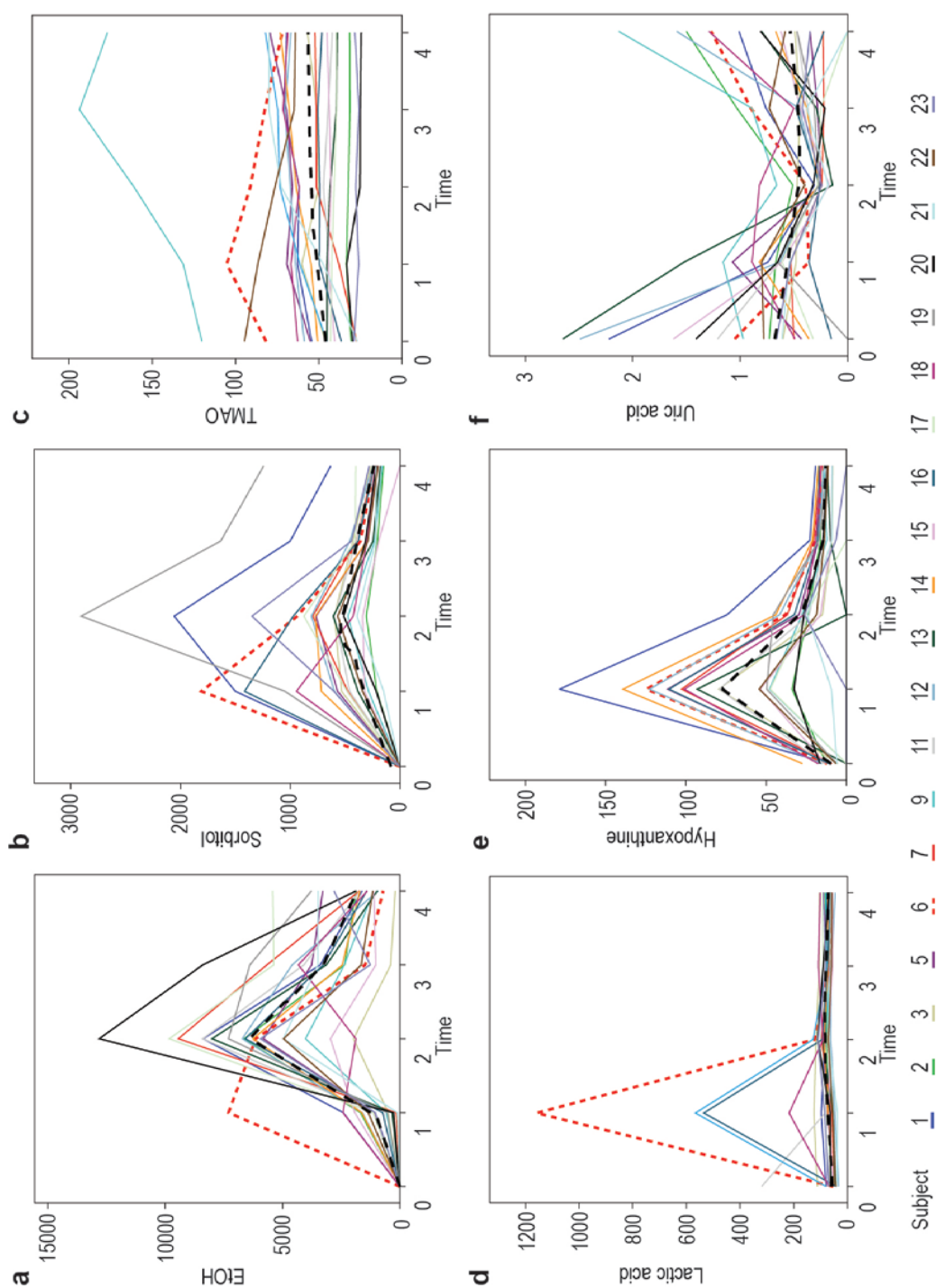


Figure 3-22: Changes in the concentrations of the six up-regulated metabolites from time 0 to time 4 following alcohol consumption.

Each subplot displays a set of lines (colored solid lines for 19 of the individuals and a red dotted line for one highlighted individual) representing the observations from the 20 individuals across 5 time points for a given metabolite: (a) ethanol; (b) sorbitol; (c) TMAO; (d) lactic acid; (e) hypoxanthine; and (f) uric acid. The black dashed lines represent one potential LOWESS (locally weighted scatterplot smoothing) regression for each metabolite against time using a bi-weight kernel.

Finally, the ingestion of one tablet of an NAD-containing supplement, proposed to potentially counteract the effects of alcohol consumption, did not show a notable effect if taken one hour before the alcohol dose. To illustrate this, the complete design of this study (that is, four interventions measured over 6 time points) was modelled for quantified hypoxanthine and sorbitol, using a 2-way RM ANOVA model based on log transformed data to improve normality. Reported in the graphs are the Greenhouse–Geisser-corrected p-values for the main effects (where significant), as well as specific comparisons based on the Wilcoxon signed-rank tests (Section 3.5.2 provides more details on this analysis). These two metabolites were selected based on the NAD⁺-dependence of their catabolism (hypoxanthine → xanthine → uric acid, both reactions catalysed by xanthine dehydrogenase; and sorbitol → fructose, catalysed by sorbitol dehydrogenase (SD; EC 1.1.1.14), respectively).

Figure 3-23 provides a brief overview of the information extracted from this model: (a) Differences in the average levels of hypoxanthine from time –1 to time 4 between the four interventions; (b) Differences in the average change in hypoxanthine levels, measured from time 0, following the four interventions; (c and d) The comparative results for sorbitol. Significant differences were based on the Greenhouse–Geisser-corrected p-values from the RM ANOVA model or the Wilcoxon signed-rank tests, assessing differences between the sets of means, and are indicated by the arrows.

Figure 3-23 a and c indicate a significant difference in the average levels of hypoxanthine and sorbitol, respectively, between the interventions including and excluding alcohol, as well as in time between the interventions including alcohol. Figure 3-23 b and d, likewise, indicate a significant difference in the average change in the levels of hypoxanthine and sorbitol across the time period between the four interventions.

All the results from this approach therefore arrive at the same conclusion - the observed differences were mainly attributable to the consumption of alcohol. For sorbitol, the differences were only significant when comparing the two interventions involving alcohol with the two not involving alcohol. No significant differences were observed between the two interventions not involving alcohol, nor between the two interventions involving alcohol. For hypoxanthine, significant differences were observed between the 'alcohol plus vehicle' and the 'vehicle plus alcohol and NAD' interventions at 2 and 3 hours after consumption, with significance on the average levels at 2 and 4 hours, respectively. Although alcohol consumption caused increased hypoxanthine excretion, it appears that the addition of NAD resulted in even higher amounts of

hypoxanthine being excreted. These results indicate that NAD, as used in this study, might potentially have an effect in the later phases following alcohol consumption, but, as a whole, do not justify its application as a supplement to counteract the effects of alcohol consumption.

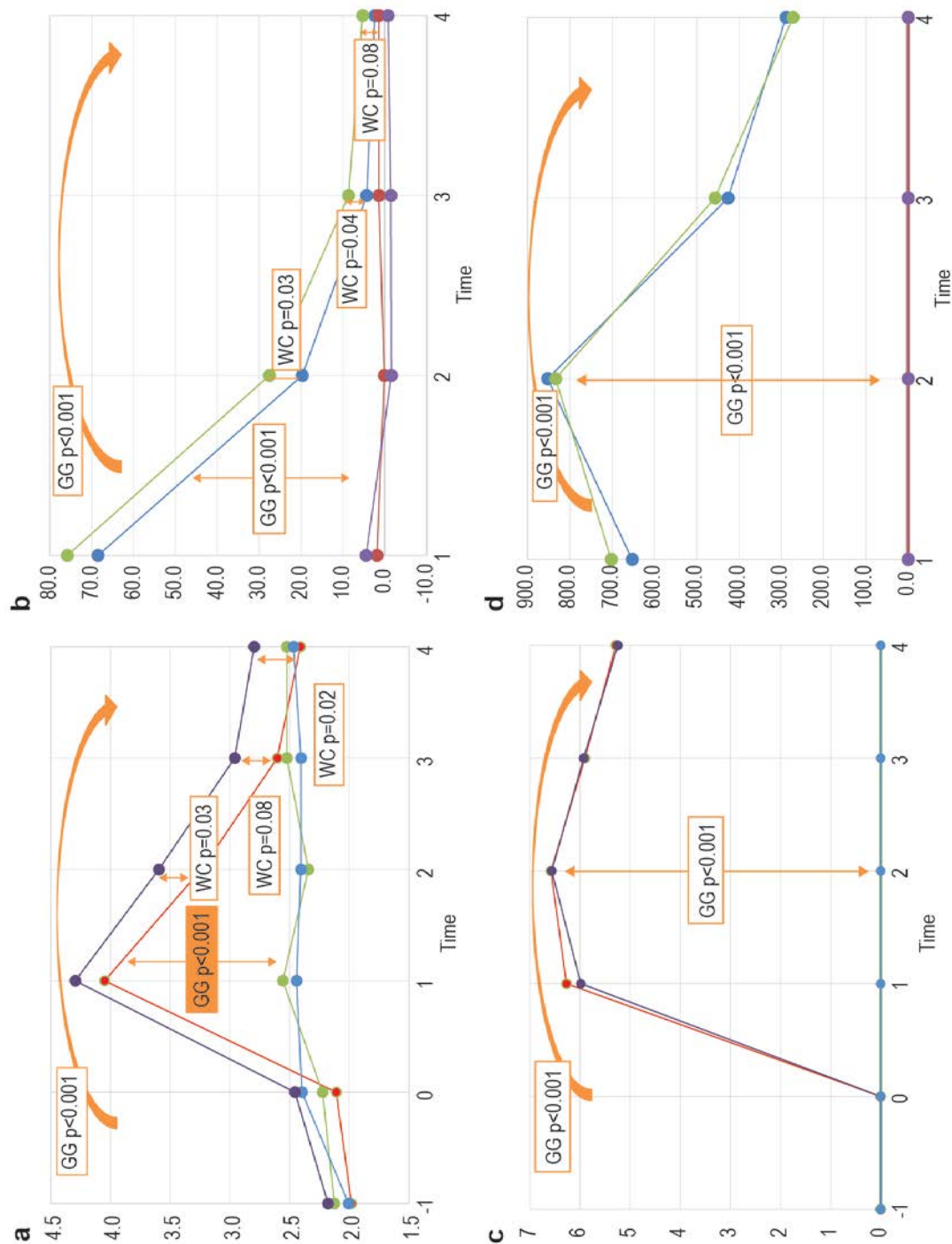


Figure 3-23: Indications of differences in the average levels of hypoxanthine and sorbitol across the four interventions and six time points.

(a) Differences in the average levels of hypoxanthine from time -1 to time 4 between the four interventions; (b) Differences in the average change in hypoxanthine levels, measured from time 0, following the four interventions; (c and d) Significant p-values were based on the Greenhouse-Geisser-corrected p-values from the RM ANOVA model (GG) or the Wilcoxon signed-rank tests (WC), assessing differences between the sets of means, and are indicated by the arrows.

3.4.6 Discussion

It is known that three mechanisms for ethanol metabolism may be operative in the liver: (1) the ADH-pathway in the cytosol; (2) the microsomal ethanol oxidizing system (MEOS) of the endoplasmic reticulum; and (3) the catalase mechanism, located in peroxisomes (Lieber, 1994). In the ADH-pathway, ethanol is converted to acetaldehyde and acetate in two consecutive dehydrogenase reactions, both dependent on NAD^+ . During both of these reactions NAD^+ is reduced, generating excess amounts of NADH. Following acute alcohol consumption, NAD^+ depletion, and the subsequent excessive production of reducing equivalents, greatly perturb the hepatocyte redox homeostasis. This perturbation is known, but, through the present NMR metabolomics study, new insights were revealed — illustrated in Figure 3-24.

The main organs involved are the gut and the liver. Ethanol absorption is indicated in the upper region of the gut, and benzoic acid (from the vehicle and microbiome metabolism) in the lower gut. The main metabolism of ethanol, and the associated consequences of the disturbed $\text{NAD}^+:\text{NADH}$ ratio, occur in the liver. The structural formulas and names of the six important metabolites are shown in red (uric acid is indicated in brackets because it was not measured by NMR). Blue arrows are not related to enzyme kinetic reactions, but are used as indicators of the proposed flow directions following alcohol consumption. Osmoregulation is proposed as efflux in hepatocytes in the early phases following alcohol consumption, and potential influx (shown in brackets) in the later phases. Abbreviations: ATP, adenosine triphosphate; ADP, adenosine diphosphate; AMP, adenosine monophosphate; IMP, inosine monophosphate; G-3-P, glyceraldehyde-3-phosphate; GLYAT, glycine-N-acyltransferase; SD, sorbitol dehydrogenase; AR, aldose reductase; OXPHOS, oxidative phosphorylation.

Figure 3-24 proposes a model that outlines three essential disturbances that occur following acute alcohol consumption — up-regulated lactic acid metabolism, down-regulated purine catabolism and osmoregulation.

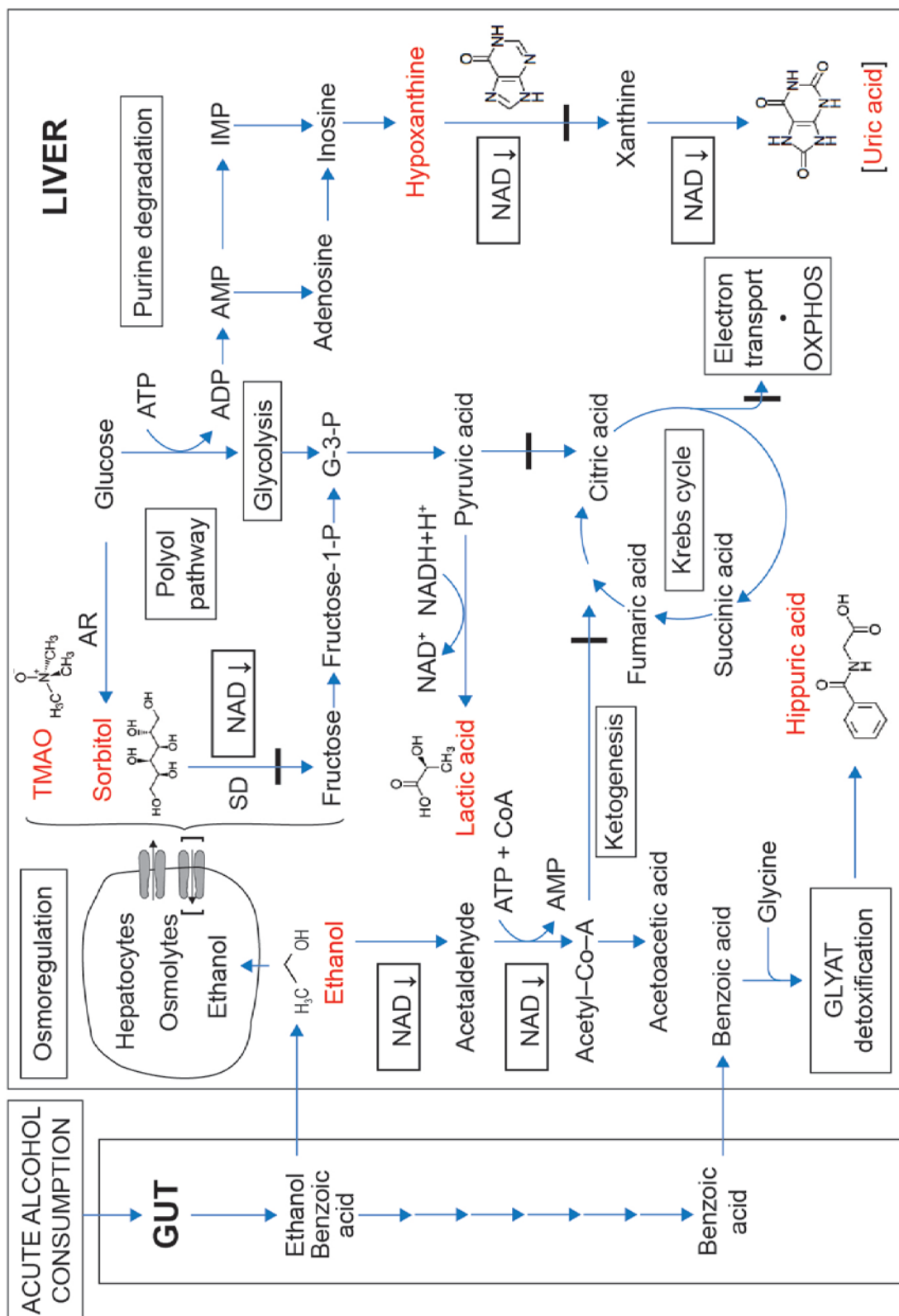


Figure 3-24: Model of the metabolite profile based on the important metabolites up-regulated following alcohol consumption.

First, up-regulation of lactic acid is coupled to the high hepatic NADH:NAD⁺ ratio (due to ADH- and ALDH-catalysed ethanol catabolism), which diverts pyruvic acid metabolism towards lactic acid, and subsequently inhibits gluconeogenesis. During low or chronic alcohol consumption, excess lactic acid is exported from the liver to peripheral tissues, where NADH levels are lower, and lactic acid may be reconverted to pyruvic acid for metabolic needs. During acute alcohol consumption, however, aerobic oxidation (that is, the Krebs cycle, the respiratory chain and oxidative phosphorylation) is known to be inhibited (Hawkins & Kalant, 1972; Zakhari & Li, 2007), which was indicated in our study by the increased levels of fumaric acid (not observed in all experimental subjects) and succinic acid.

Similarly, due to the high hepatic NADH:NAD⁺ ratio and inhibition of the Krebs cycle, ethanol-derived acetyl-CoA may be converted to acetoacetic acid and 3-ketobutyric acid, as seen in the urine samples from some of our experimental subjects. The values of these metabolites were, however, generally too low to enable their quantification. A noteworthy observation was that a few of the experimental subjects showed excessive urinary lactic acid excretion shortly after alcohol intake (Figure 3-22d), whereas the excretion of the majority of the subjects peaked at two hours following alcohol consumption. These observations are yet a further example of the individual differences in coping with the consumed alcohol. The common paradigm is that variation in response to alcohol consumption is genetically controlled, and is suspected to cause a predisposition towards the development of alcohol-induced liver disease and alcoholism (Day *et al.*, 1991). Furthermore, we observed that the mean value of urinary lactic acid peaked one hour after alcohol consumption (increasing roughly 2.5 times from 70.90 mM to 179.9 mM ($p = 0.001$)) and declined to a near normal value three hours after consumption (75.90 mM). This is comparable to the decreased lactic acid levels measured in liver and serum samples from rats, decapitated three hours following treatment with a single intragastric dose of ethanol (Nicholas *et al.*, 2008).

Second, hypoxanthine, an intermediate in purine catabolism and a precursor of uric acid, appeared to be an important indicator of acute alcohol consumption (Figure 3-22e). This is in agreement with the increased hypoxanthine observed in patients suffering from alcohol- and hepatitis B-induced cirrhosis (Lian *et al.*, 2011), and in those listed with stearamide as biomarker for hepatic cirrhosis (Gika & Wilson, 2016). In addition, several intravenous ethanol infusion, and related, studies on purine metabolism in humans (reviewed in Yamamoto *et al.*, 2005) indicated that many factors affect this metabolic pathway - daily drinking habits; the type of alcoholic beverages; exercise; and, ultimately, ALDH polymorphisms, are all important contributing factors.

Increased AMP following alcohol consumption seems to be the important metabolic departure point in reflecting on purine catabolism - alcohol-induced diminished ATP production via glycolysis (Yuki & Thurman, 1980) and increased adenine nucleotide turnover (Faller & Fox, 1982), contribute towards increased ADP levels, and its increased conversion to AMP. Additionally, during the ADH- and ALDH-catalysed degradation of ethanol, two equivalents of ATP are consumed and two equivalents of AMP are produced for each equivalent of ethanol converted to acetyl-CoA (Puig & Fox, 1984). All these perturbations lead to a considerable alcohol-induced increase in AMP, which is then catabolised and reflected in the hypoxanthine and, ultimately, uric acid profiles. Hyperuricaemia, gout and increased urinary uric acid excretion, accompanied by raised urinary hypoxanthine and xanthine concentrations, are accepted indicators of chronic alcohol consumption, although excretion of hypoxanthine was shown to be lower than that of xanthine in regular drinkers (Nishimura *et al.*, 1994).

In our study, we observed a highly significant increase (Table 3-6: $p \leq 0.0001$; FC = +8.187) in hypoxanthine excretion in all 20 cases studied, which peaked at one hour after alcohol consumption in 18 of the cases, and at 2 hours in the other two cases (Figure 3-22e). The general pattern of the urinary excretion of uric acid (Figure 3-22f) presents as a mirror image to that of hypoxanthine, initially decreasing and then returning to near its starting concentration. We relate these observations to the alcohol-induced perturbed hepatic redox state — depleted NAD^+ , the cofactor for xanthine dehydrogenase, interrupts the purine catabolic pathway, decreasing the conversions of hypoxanthine to xanthine, and xanthine to uric acid. However, as the $\text{NADH}:\text{NAD}^+$ ratio normalizes, these reactions can once again take place, resulting in the catabolism of hypoxanthine to produce uric acid. It should, however, be noted that lactic acid, which is highly increased following alcohol consumption, competitively inhibits the clearance of uric acid through the renal proximal tubule (Yü *et al.*, 1957), which further supports the observed initial decrease in uric acid excretion following alcohol consumption. Thus, while hyperuricaemia is an accepted marker of regular and chronic alcohol consumption, increased excretion of hypoxanthine may possibly act as an indicator of acute alcohol consumption.

Third, of notable interest, was the observation of the increased urinary excretion of sorbitol following ethanol consumption, which is not listed in the review on global metabolic profiling studies on alcohol-related disorders, covering NMR, GC–MS and LC–MS approaches (Gika & Wilson, 2016). Sorbitol was not detected in any of the urine samples collected prior to any of the interventions (normal reference value: $3.506 \pm 2.24 \mu\text{mol}/\text{mmol creatinine}$) (Bouatara *et al.*, 2013). Its excretion, however, increased rapidly in most individuals following alcohol consumption

(Figure 3-22b). Moreover, the sorbitol excretion profile did not seem to be affected by the consumption of the NAD supplement, taken prior to alcohol consumption (Figure 3-23 c and d). Sorbitol is an organic osmolyte present in all human cells, and, together with other osmolytes, reaches very high concentrations (in the millimolar range) in the cytosol. In human cells, the osmolytes are classified into three groups (Strange, 2004): (1) amino acids and their derivatives (including taurine); (2) methylamines (including TMAO); and (3) polyols (including sorbitol). Osmolytes play key roles as cytoprotectants, and in maintaining cell volume homeostasis (Lang *et al.*, 1998). They function as nonperturbing solutes, which permits their accumulation to high levels and large shifts in their concentrations without having deleterious effects on cellular structure and function (Strange *et al.*, 2004).

These unique characteristics of osmolytes opened several lines of thought regarding the perturbation of sorbitol following acute alcohol consumption. Of these postulates, sorbitol synthesis and catabolism (occurring in the polyol pathway), as well as its function in osmoregulation, are pertinent:

(1) The NADPH-dependent enzyme aldose reductase (AR; EC 1.1.1.21) catalyses the synthesis of sorbitol from glucose. This reaction is highly operative under hyperglycaemic conditions, such as in diabetes mellitus (Tang *et al.*, 2012), when up to 30% of glucose is channelled into the AR-catalysed polyol pathway. However, given the fasting state of the experimental subjects before the 'alcohol plus vehicle' intervention, and the increased anaerobic oxidation of glucose towards lactic acid, it seems unlikely that sorbitol accumulation was due to the activation of sorbitol synthesis.

(2) The next step in the polyol pathway is the degradation of sorbitol to fructose, a reaction catalysed by the NAD⁺-dependent enzyme sorbitol dehydrogenase (SD). In this study, sorbitol accumulation may be attributed to the alcohol-induced inhibition of its catabolism - due to the metabolism of ethanol, the NAD⁺ required for the dehydrogenation of sorbitol becomes depleted. Furthermore, significant down-regulation of SD has been observed in a study on the changes of the cytoplasmic proteome in response to alcoholic hepatotoxicity in rats (Kim *et al.*, 2015), which lends support to this viewpoint.

(3) Increased sorbitol as a consequence of its role in osmoregulation, however, seems to be the preferred explanation for interpreting the observed urinary excretion profile of sorbitol. Ethanol is both water and lipid soluble, which renders it a membrane-permeable substance. Abnormal cell volume regulation significantly contributes to the pathophysiology of several disorders, and cells

respond to these changes by importing, exporting, or synthesizing osmolytes to maintain volume homeostasis (Diamond & Gordon, 1997). On an experimental level, oedema/cell swelling could be induced by binge-simulated ethanol exposure in slice cultures of the developing rat brain (Moon *et al.*, 2014). In a cell-to-medium flux study, hyperosmotically induced intracellular accumulation of sorbitol in renal epithelial cells showed a greater than 150-fold increased efflux within five minutes after exposure to an isosmotic medium.

We thus speculate that the increased sorbitol excretion in our alcohol-exposed subjects relates to ethanol-induced hepatocyte swelling, which is compensated for by sorbitol release in order to maintain cellular volume homeostasis. Although less pronounced in this study than that of sorbitol, perturbations of other osmolytes, such as TMAO and taurine (observed in this study, see Table 3-6), should also be considered in the examination of osmolyte responses.

In summary, we have described the urinary metabolite profile of healthy, young males following acute alcohol consumption as part of a designed intervention study. The complex NMR spectral data, generated from individuals participating in a time-dependent crossover study, could be resolved sufficiently through the application of univariate and multivariate statistical analyses. This approach provided a novel method for the analysis and understanding of the complex metabolomics data that were produced due to acute alcohol consumption. We indicate that NAD⁺ depletion, and the production of an excessive amount of reducing equivalents, greatly perturb the hepatocyte redox homeostasis, resulting in metabolic disturbances, of which urinary excretion of sorbitol is novel. We postulate that sorbitol is a marker of a cell volume regulatory response to ethanol-induced hepatocyte swelling. This may have a wider significance related to brain oedema-induced neurodegenerative damage following chronic binge alcohol exposure (Moon *et al.*, 2014; Collins *et al.*, 1998).

3.4.7 Additional Information

Acknowledgements

Research funding for the analytical and computational aspects of the project was provided by the Technological Innovation Agency (TIA) of the Department of Science and Technology of South Africa. The views and findings of this investigation are those of the authors and do not reflect or represent any policies of TIA. The funder was not involved in the writing of the manuscript, neither in its editing nor in the decision for it to be distributed for publication.

3.5 Supplementary Information: The ^1H -NMR-based metabolite profile of acute alcohol consumption: a metabolomics intervention study

3.5.1 Methods for sample treatment, storage, preparation and ^1H -NMR analysis

Sample collection, characterization and storage

All the subjects were randomly assigned to one of the four intervention groups until all 24 had participated in all four interventions. Six urine samples were collected from each subject for each intervention, giving a total of 24 samples from each subject over the course of the study.

One 5 mL and two 1 mL vials were used to provide aliquots of each of the urine samples; these aliquots, together with the remainder of the bulk urine samples, were stored at -80°C . Once all the urine samples were collected, one 1 mL aliquot of each was thawed and combined to prepare a pooled quality control (QC) sample for the experiment as a whole. This QC sample was then divided into 15 mL aliquots and once again stored at -80°C .

The collected samples were analysed in 24 separate batches, each batch containing the 24 samples of a single subject and three QC samples. The 27 samples from each of the 24 batches were analysed in the following order:

$QC_1 [S_{-1}S_0S_1S_2S_3S_4]_{Vehicle} [S_{-1}S_0S_1S_2S_3S_4]_{Alcohol} QC_2 [S_{-1}S_0S_1S_2S_3S_4]_{NAD} [S_{-1}S_0S_1S_2S_3S_4]_{NAD+Alcohol} QC_3$

where S_{-1} represents the sample collected at time -1 , S_0 represents the sample collected at time 0, and so on.

Sample preparation and ^1H -NMR analysis

Spectral analyses were conducted at the NMR facility of the Centre for Human Metabolomics at North-West University. Prior to analysis, an aqueous 1.5 M KH_2PO_4 deuterated buffer solution at pH 7.4 was prepared (Dona *et al.*, 2014). This solution served to lock the signal during analysis, maintained a stable pH in the sample and contained trimethyl-2,2,3,3-tetradeuteriopropionic acid (TSP, sodium salt; Sigma Aldrich) as the internal standard to provide a chemical shift reference of $\delta = 0.00$. The urine samples, stored at -80°C , were thawed at room temperature for analysis. A 600 μL volume of each sample was centrifuged at $12\,000 \times g$ for 5 min to remove any sediments or debris. A 60 μL volume of buffer solution was added to 540 μL of the supernatant, vortexed and transferred to a 5-mm NMR tube.

Each sample so prepared was analysed on a Bruker Avance III HD 500 MHz NMR spectrometer equipped with a triple-resonance inverse (TXI) $^1\text{H}\{^{15}\text{N},^{13}\text{C}\}$ probe head and x, y, z gradient coils. ^1H spectra were acquired as 128 transients in 32K data points with a spectral width of 6002 Hz. The sample temperature was maintained at 300 K and the H_2O resonance was pre-saturated by single-frequency irradiation during a relaxation delay of 4 s, with a 90° excitation pulse fixed at 8 μs . Shimming of the sample was performed automatically on the deuterium signal. The resonance line widths for TSP and metabolites were <1 Hz (measurements at half the height of the peak). Fourier transformation and phase and baseline correction were done automatically. The software used was Bruker Topspin (V3.2) and Bruker AMIX (V3.9.9) (Ellinger *et al.*, 2013).

All urine samples were normalized with reference to the creatinine CH_2 peak at 4.05 ppm. We employed two methods of spectral analysis: (1) Applying equidistant binning (using a bin width of 0.02 ppm) to the spectral region between 0.5 and 10 ppm, excluding the region of the water peak (4.72–4.88 ppm). This gave a total of 467 integrated units per NMR spectrum for each sample for statistical analysis — spectral data. (2) Using information from the statistical outputs to identify discriminatory metabolites (based upon pure compound spectral libraries), and then accurately integrating the selected metabolites — providing quantified concentration data (μmol metabolite/mmol creatinine) for univariate analysis (see Table 3-6).

3.5.2 Data Processing

Pre-processing

The use of NMR spectroscopy as a tool for metabolomics is limited by the sensitivity (limit of detection) of NMR (Krojanski *et al.*, 2008), requiring an approach in discarding noise while retaining and then measuring real peaks. The spectral intensities within the original 467 bins generated here ranged from 7.8×10^5 to 1.7×10^9 . Based on previous empirical experience with NMR spectral analysis, we defined a threshold value of 2×10^6 , being approximately the limit of detection of metabolomic substances presumed to be present in a spectral bin. To reduce the uncertainty in the data at the level of the approximate detection limit, we applied the threshold of 2×10^6 to the data matrix, setting all values below this threshold to zero. We subsequently applied a ‘zero-filter’ to remove bins with more than 50% zero observations across all six time points, as applied for fusion of MS-based metabolomics data (Smilde *et al.*, 2005). This resulted in a reduction in the number of bins containing spectral data from 467 to 347.

The remaining zero observations were replaced by random numbers below the minimum non-zero observation for each bin, after which the data were log scaled and auto scaled.

To account for the dilution differences between the urine samples, the value of each bin was normalized relative to the CH₂ peak of creatinine. This peak is contained in the 4.05 and 4.07 ppm bins. Therefore, each bin value was divided by the sum of the value in these two bins to normalize the bin values relative to creatinine. After normalization, the three bins related to creatinine (4.05, 4.07 and 3.05 ppm) were also removed from the matrix, giving a final total of 344 bins containing spectral data.

Quality assurance

Quality control (QC) samples were included in each batch at the start, middle and end of the analytical run. The variation and correlation within and between bins for QC and experimental samples were compared using principal component analysis (PCA) to attain a birds-eye view of any batch effects. NMR data are not known typically to present with such variation structures (that is, within or between batch effects), and this was also the case here.

Figure 3-25 is a PCA scores plot illustrating the variation and correlation within and between bins for QC and experimental samples. The close clustering of the QC samples compared to the experimental samples indicates that the variation in the QC samples is much less than the variation in the experimental samples. From these results it was assumed that no batch corrections were required.

The interaction effect of NAD and alcohol

As explained in the main text, the complete design of the study (that is, four intervention measured over 6 time points across all individuals) was modelled for quantified hypoxanthine and sorbitol using a two-way repeated measures analysis of variance (2-way RM ANOVA) model. A brief overview of the information extracted from the model is presented in Figure 3-23 of the main text.

The absence of these metabolites prior to the intervention and sudden increase in their levels shortly thereafter, resulted in a high frequency of zero observations, thus introducing a spike in the distribution of the data. Therefore, the nature of the experiment and metabolites under investigation resulted in the data not being normally distributed. To account for this, log transformed data were used to build the RM ANOVA model, while Greenhouse–Geisser-

corrected p-values were used to assess the significance of the main effects (that is, intervention and time).

Unfortunately, inspection of the residuals indicated that the transformation was not able to produce normality across all effect levels. The Wilcoxon signed-rank test was used as a confirmation test on the untransformed data to test for significantly different pairs of means as indicated in Figure 3-23 a and c in the main text. However, since the performance of the Wilcoxon signed-rank test may also be compromised due to the spike in the data, difference data were assessed as a final confirmation to provide additional proof of the differences observed. The difference data were created by subtracting time 0 data from the remaining time points, while ignoring time -1. Doing so significantly improved the distribution of the data. The results based on the difference data are summarised in Figure 3-23 b and 5d in the main text.

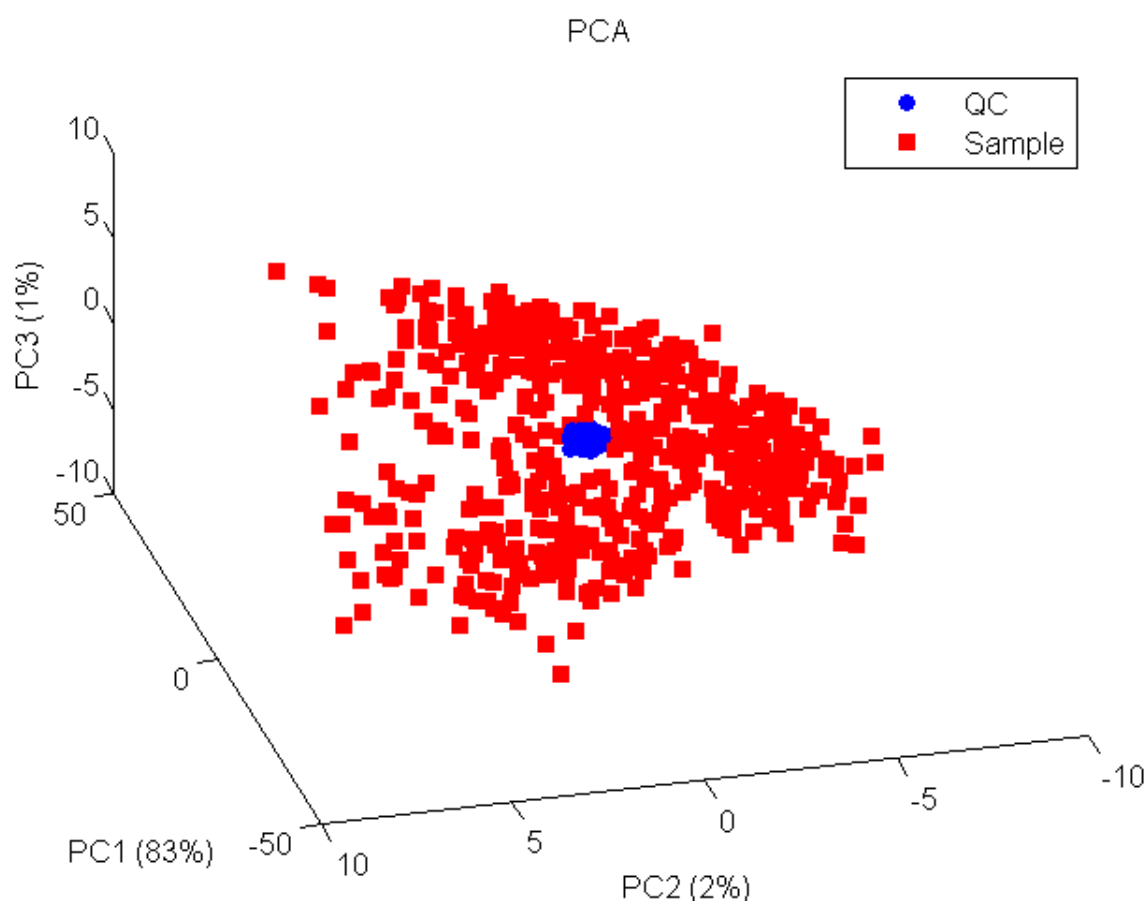


Figure 3-25: PCA scores plot illustrating the variation and correlation within and between bins for QC and experimental samples.

3.5.3 Uric acid analysis

The manual of the protocol detailing the uric acid analysis is available online as a downloadable PDF at <https://www.thermofisher.com/order/catalog/product/TR24321> (ThermoFisher Scientific, 2012).

Infinity™ Uric Acid Liquid Stable Reagent

PRODUCT SUMMARY

Stability	:	Until Expiry at 2 - 8°C
Linear Range	:	0.03-1.5 mmol/L (0.5-25.2 mg/dL)
Specimen Type	:	Serum or Urine
Method	:	Enzymatic Endpoint
Reagent Preparation	:	Supplied ready to use.



INTENDED USE

This reagent is intended for the in vitro quantitative determination of Uric Acid in human serum or urine.

CLINICAL SIGNIFICANCE

Uric acid is a metabolite of purines, nucleic acids and nucleoproteins; consequently, abnormal levels may be indicative of a disorder in the metabolism of these substances. Hyperuricaemia may be observed in renal dysfunction, gout, leukemia, polycythaemia, atherosclerosis, diabetes, hypothyroidism, or in some genetic diseases. Decreased levels are present in patients with Wilson's Disease.^{1,2,3}

METHODOLOGY

This reagent is based upon the methods of Trivedi and Kabasakalian^{4,5} with a modified Trinder⁶ peroxidase assay using 2,4,6-Tribromo-3-hydroxy benzoic acid (TBHB).

The series of reactions involved in the assay system is as follows:



1. Uric Acid is oxidised to allantoin by uricase with the production of H_2O_2 .
2. The peroxide reacts with 4-aminoantipyrine (4-AAP) and TBHB in the presence of peroxidase to yield a quinoneimine dye. The resulting change in absorbance at 520nm (520-550nm) is proportional to uric acid concentration in the sample.

REAGENT COMPOSITION

Active Ingredients	Concentration
4-Aminoantipyrine	0.5 mmol/L
TBHB	1.75 mmol/L
Uricase (Bacillus Sp.)	> 120 U/L
Peroxidase (Horseradish)	> 500 U/L
Tris Buffer	50 mmol/L
pH 8.25 ± 0.1 at 20°C.	

WARNING: Do not ingest. Avoid contact with skin and eyes. If spilt thoroughly wash affected areas with water. Reagent contains Sodium Azide which may react with copper or lead plumbing. Flush with plenty of water when disposing. For further information consult the Infinity Uric Acid Liquid Stable Reagent Material Safety Data Sheet.

CAUTION: This product contains animal source material. Handle and dispose of this product as if it were potentially infectious.

REAGENT PREPARATION

Reagent is supplied ready to use.

STABILITY AND STORAGE

When stored refrigerated at 2-8°C the reagent is stable until the expiry date stated on the bottle and kit box label.

Indications of Reagent Deterioration:

- Turbidity;
- Reagent Absorbance >0.5 AU at 520nm; and/or
- Failure to recover control values within the assigned range.

SPECIMEN COLLECTION AND HANDLING

Collection: No special preparation of the patient is required.

SYMBOLS IN PRODUCT LABELLING

	Authorized Representative		Temperature Limitation
	For in vitro diagnostic use		Use by/Expiration Date
	Batch code/Lot number		CAUTION: CONSULT INSTRUCTIONS FOR USE.
	Catalogue number		Manufactured by
	Consult instructions for use		

Serum: Use non-haemolysed serum.

Urine: It is recommended that 15 mL of 2 mol/L NaOH be added to the collection vessel. Upon receipt of the urine sample, pH should be checked. If the pH is less than 8.0 it should be adjusted with NaOH. A 1:10 dilution of urine is typically required prior to analysis.⁷

Storage: Serum samples are stable for at least 3 days at room temperature (18-25°C) and for at least 6 months frozen.² Stabilized urine may be stored at room temperature for 5 days.⁷

ADDITIONAL EQUIPMENT REQUIRED BUT NOT PROVIDED

- If required, pipettes for accurately dispensing measured volumes.
- A clinical chemistry analyzer capable of maintaining constant temperature (37°C) and measuring absorbance at 520 nm.
- Analyzer specific consumables, eg: sample cups.
- Normal and abnormal assayed control material.
- Calibrator or a suitable aqueous standard (see calibration section).

ASSAY PROCEDURE

The following system parameters are recommended. Individual instrument applications are available upon request from the Technical Support Group.

SYSTEM PARAMETERS

Temperature	37°C
Primary Wavelength	520 nm (520-550 nm)
Secondary Wavelength	600 - 660 nm
Assay Type	Endpoint
Direction	Increase
Sample :	Reagent Ratio
eg: Sample Vol	1 : 50
Reagent Vol	3 µL
Incubation Time	150 µL
Reagent Blank Limits	300 Seconds
(520nm, 1cm light path)	Low 0.0 AU
Linearity	High 0.5 AU
Analytical Sensitivity	0.03-1.50 mmol/L
(520nm, 1cm light path)	(0.5-25.2 mg/dL)
	0.42 ΔA per mmol/L
	0.025ΔA per mg/dL

CALCULATIONS

Results are calculated, usually automatically by the instrument, as follows:

$$\text{Uric Acid} = \frac{\text{Absorbance of Unknown}}{\text{Absorbance of Calibrator}} \times \text{Calibrator Value}$$

Example:

$$\begin{aligned} \text{Absorbance of calibrator} &= 0.302 \\ \text{Absorbance of unknown} &= 0.071 \\ \text{Value of calibrator} &= 0.720 \text{ mmol/L (12.1 mg/dL)} \end{aligned}$$

$$\text{Uric Acid} = \frac{0.071}{0.302} \times 0.720 = 0.16 \text{ mmol/L}$$

$$\text{Uric Acid} = \frac{0.071}{0.302} \times 12.1 = 2.8 \text{ mg/dL}$$

NOTES

1. The reagent and sample volumes may be altered proportionally to accommodate different spectrophotometer requirements.
2. The color development is stable for 15 minutes.
3. Specimens with Uric Acid concentrations greater than 1.50 mmol/L (25.2 mg/dL) should be diluted with saline and reassayed. Multiply results by the dilution factor.

Thermo
SCIENTIFIC

4. S.I. unit conversion factor: mmol/L x 16.8 = mg/dL.

CALIBRATION

Calibration is required. An aqueous standard or serum based calibrator, with an assigned value traceable to a primary standard (eg NIST or IRMM) is recommended. Standards should not contain formaldehyde or enzyme inhibitors as preservatives. For calibration frequency on automated instruments, refer to the instrument manufacturers specifications.

However, calibration stability is contingent upon optimum instrument performance and the use of reagents which have been stored as recommended in the stability and storage section of this package insert. Recalibration is recommended at anytime if one of the following events occurs:-

- The lot number of reagent changes.
- Preventative maintenance is performed or a critical component is replaced.
- Control values have shifted or are out of range and a new vial of control does not rectify the problem.

QUALITY CONTROL

To ensure adequate quality control, normal and abnormal control with assayed values should be run as unknown samples:-

- At least once per day or as established by the laboratory.
- When a new bottle of reagent is used.
- After preventative maintenance is performed or a critical component is replaced.
- With every calibration.

Control results falling outside the upper or lower limits of the established ranges indicate the assay may be out of control.

The following corrective actions are recommended in such situations:-

- Repeat the same controls.
- If repeated control results are outside the limits, prepare fresh control serum and repeat the test.
- If results are still out of control, recalibrate with fresh calibrator, then repeat the test.
- If results are still out of control, perform a calibration with freshly prepared reagent, then repeat the test.
- If results are still out of control, contact Technical Services or the local distributor.

LIMITATIONS

1. Studies to determine the level of interference from haemoglobin, bilirubin (free and conjugated) and lipaemia were carried out. The following results were obtained:

Haemoglobin: No interference from haemoglobin up to 424 mg/dL.

Free Bilirubin: No interference from free bilirubin up to 212 µmol/L (12 mg/dL).

Conjugated Bilirubin: No interference from conjugated bilirubin up to 212 µmol/L (12 mg/dL).

Lipaemia: No interference from lipaemia, measured as absorbance at 630nm, up to 1.68 AU.

2. Young DS[®] has published a comprehensive list of drugs and substances which may interfere with this assay.

EXPECTED VALUES⁹

Child:	0.12 - 0.32 mmol/L	2.0 - 5.0 mg/dL
Adult Male:	0.21 - 0.42 mmol/L	3.5 - 7.2 mg/dL
Adult Female:	0.15 - 0.35 mmol/L	2.6 - 6.0 mg/dL
Urine:	1.48 - 4.43 mmol/day	250 - 750 mg/day

The quoted values should serve as a guide only. It is recommended that each Laboratory verify this range or derives a reference interval for the population it serves.¹⁰

PERFORMANCE DATA

The following data was obtained using the Infinity Uric Acid Liquid Stable Reagent on a well maintained automated clinical chemistry analyzer. Users should establish product performance on their specific analyzer used.

IMPRECISION

Imprecision was evaluated over a period of 20 days using two levels of commercial

control and following the NCCLS EP5-T procedure.¹¹

	LEVEL I	LEVEL II
Number of data points	80	80
Mean (mmol/L / mg/dL)	0.279 / 4.69	0.603 / 10.13
Within run: SD (mmol/L / mg/dL)	0.007 / 0.12	0.009 / 0.15
CV (%)	2.3	1.5
Total: SD (mmol/L / mg/dL)	0.019 / 0.32	0.021 / 0.35
CV (%)	6.8	3.4

METHOD COMPARISON

Comparison studies were carried out using a similar commercially available reagent as a reference. Serum and urine samples were assayed in parallel and the results compared by least squares regression. The following statistics were obtained.

Serum:	
Number of sample pairs	60
Range of sample results	0.11-0.61 mmol/L (1.85-10.25 mg/dL)
Mean of reference method results	0.315 mmol/L (5.29 mg/dL)
Mean of Uric Acid results	0.336 mmol/L (5.65 mg/dL)
Slope	0.931
Intercept	0.042 mmol/L (0.71 mg/dL)
Correlation coefficient	0.987
Urine:	
Number of sample pairs	41
Range of sample results	0.48 - 11.7 mmol/L (8.0 - 196 mg/dL)
Mean of reference method results	3.0 mmol/L (49.6 mg/dL)
Mean of Uric Acid results	2.5 mmol/L (42.7 mg/dL)
Slope	0.967
Intercept	-0.32 mmol/L (-5.3 mg/dL)
Correlation coefficient	0.990

LINEARITY

When run as recommended the assay is linear between 0.03 and 1.50 mmol/L (0.5-25.2 mg/dL).

ANALYTICAL SENSITIVITY

When run as recommended the sensitivity of this assay is 0.42 ΔAbs per mmol/L or 0.025 ΔAbs per mg/dL (1cm light path, 520nm).

REFERENCES

1. Searcy R.L., Diagnostic Biochemistry. McGraw-Hill, New York, NY, 1969.
2. Henry R.J., Common C. and Winkelman J.W. (eds), Clinical Chemistry: Principles and Techniques. Harper & Row, Hagerstown, MD, 1974.
3. Balis M. E., Adv. Clin. Chem. 18(213) 1976.
4. Trivedi R., Berta E. and Rebar L., Clin. Chem. 22(1223), 1976.
5. Kabasakalian P. Kalliney S. and Wescott A. Clin. Chem. 19(522) 1973.
6. Trinder P., J. Clin. Pathol. 22(246) 1949.
7. Shephard MD, Mezzachi RD. Clin Biochem Revs 1983; 4:61-7.
8. Young DS. Effects of Drugs on Clinical Laboratory Tests. Third Edition. 1990; 3:360-370.
9. Tietz Textbook of Clinical Chemistry and Molecular Diagnosis (4th Ed.) Burtis, Ashwood & Bruns (Eds), Elsevier Saunders, 2005; 2301.
10. Wachtel M et al, Creation and Verification of Reference Intervals. Laboratory Medicine 1995; 26:593-7.
11. National Committee for Clinical Laboratory Standards. User evaluation of Precision Performance of Clinical Chemistry Devices. NCCLS; 1984, NCCLS Publication EP5-T.

 Fisher Diagnostics
a division of Fisher Scientific Company, LLC
a part of Thermo Fisher Scientific Inc.
Middletown, VA 22645-1905 USA
Phone: 800-528-0494
540-869-3200
Fax: 540-869-8132

 WMDE
Bergerweg 18
6085 AT Horn
The Netherlands
JL840768-en (R1)



© 2012 Thermo Fisher Scientific Inc. All rights reserved. Hitachi is a registered trademark of Roche Diagnostics, Indianapolis, IN 46250. iLab 600 is a registered trademark of Instrumentation Laboratory Company, Lexington, MA 02421. All other trademarks are the property of Thermo Fisher Scientific Inc and its subsidiaries.

REF	Reorder Information	
	Catalogue No.	Configuration
	TR24321	2 x 125 mL

3.5.4 Original ^1H -NMR spectral data

The data include urine samples from all 24 experimental subjects and all four interventions (576 samples), as well as 72 QC samples, yielding a total of 648 study samples. Data from four of the subjects were, however, removed, as explained in the main text. The spectral region between 0.5 and 10 ppm, excluding the region of the water peak (4.72–4.88 ppm), of the original ^1H -NMR spectrum of each of the urine samples analysed was divided into 0.02-ppm equal-sized bins. This resulted in 467 bins containing spectral data. Table 3-7 represents a small extract from the original, raw spectral data matrix of all the samples analysed for this study.

Batch	Sample Name	Case	Treatment	Time	Bin no:					
					9.99	9.969999	9.95	...	0.53	0.51
1	QC1	F	QC		144185.2	85175.12	97222.12		162878.1	164512.9
1	F41-1	F	Vehicle	–1	252890.2	171823.4	163215.5		384216.4	266010.8
1	F410	F	Vehicle	0	151299.1	100954.3	86908.15		262467	216768.2
1	F411	F	Vehicle	1	79268.87	105579.2	95048.77		76325.53	126937.3
1	F412	F	Vehicle	2	84927.78	101751.3	127028.2		77872.51	68765.72
1	F413	F	Vehicle	3	74747.86	144056.7	47937.61		91158.41	58079.11
1	F414	F	Vehicle	4	67488.71	70024.98	68237.62		103052	89226.37
1	F32-1	F	Alcohol	–1	140557.8	103222.9	98290.31		297911.4	217084.5
1	F320	F	Alcohol	0	186432.5	118033	86594.67		224899.6	205953.7
1	F321	F	Alcohol	1	138804.5	89211.63	82459.67		137392	87124.63
1	F322	F	Alcohol	2	74946.59	76699.26	70328.33		125527.2	139514.5
1	F323	F	Alcohol	3	70183.16	75978.23	71975.92		90170.89	86251.59
1	F324	F	Alcohol	4	170387.9	119688.5	133812.7		125650	97010.52
1	QC2	F	QC		148854.6	107287.3	109676.4		194299.3	116989.8
1	F22-1	F	NAD	–1	86274.71	155479	145948.6		176155.9	165445.7
1	F220	F	NAD	0	137788.9	91139.97	80084.29		169180.1	168179.4
1	F221	F	NAD	1	101979	64201.66	84790.44		92116.46	112283.7
1	F222	F	NAD	2	81776.94	62372.04	73383.53		94053.74	71139.18
1	F223	F	NAD	3	103830	99023.64	83557.65		87499.72	105244.2
1	F224	F	NAD	4	81614.46	71494.57	85544.25		99941.91	85364.31
1	F14-1	F	NAD + Alcohol	–1	167690.9	147427.5	99771.65		171520.8	176960.6
1	F140	F	NAD + Alcohol	0	79105.6	75316.07	71888.47		111565.1	74134.41
1	F141	F	NAD + Alcohol	1	67927.37	97794.58	50990.82		76890.9	114086.5
1	F142	F	NAD + Alcohol	2	70637.47	54988.69	86134.48		69929.7	94239.57
					70637.47	54988.69	86134.48		69929.7	94239.57

Table 3-7: A small extract from the file containing the raw ^1H -NMR spectral data.

3.6 Accepted Publication: GC–MS-based urinary organic acid profiling reveals multiple dysregulated metabolic pathways following experimental acute alcohol consumption

3.6.1 Title Page

GC–MS-based urinary organic acid profiling reveals multiple dysregulated metabolic pathways following experimental acute alcohol consumption

Authors:

Cindy Irwin¹, Lodewyk J. Mienie¹, Ron A. Wevers², Shayne Mason¹, Johan A. Westerhuis^{1,3}, Mari van Reenen⁴, *Carolus J. Reinecke¹

***Corresponding author:**

Carolus J. Reinecke email: carools.reinecke@nwu.ac.za

Addresses:

¹Centre for Human Metabolomics, Faculty of Natural Sciences and Agriculture, North-West University (Potchefstroom Campus), Private Bag X6001, Potchefstroom, South Africa.

²Radboud University Nijmegen Medical Centre, Translational Metabolic Laboratory, Department of Laboratory Medicine, PO Box 9101, 6500 HB Nijmegen, The Netherlands.

³Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam, The Netherlands.

⁴Department of Statistics, Faculty of Natural Sciences and Agriculture, North-West University (Potchefstroom Campus), Private Bag X6001, Potchefstroom, South Africa.

Email addresses:

Cindy Irwin: cindyirwin74@gmail.com

Lodewyk J. Mienie: 10061533@g.nwu.ac.za

Ron A. Wevers: ron.wevers@radboudumc.nl

Shayne Mason: nmr.nwu@gmail.com

Johan A. Westerhuis: j.a.westerhuis@uva.nl

Mari van Reenen: 12791733@nwu.ac.za

Carolus J. Reinecke: carools.reinecke@nwu.ac.za

3.6.2 Abstract

Metabolomics studies of diseases associated with chronic alcohol consumption provide compelling evidence of several perturbed metabolic pathways. Moreover, the holistic approach of such studies gives insights into the pathophysiological risk factors associated with chronic alcohol-induced disability, morbidity and mortality. Here, we report on a GC–MS-based organic acid profiling study on acute alcohol consumption. Our investigation — involving 12 healthy, moderate-drinking young men — simulated a single binge drinking event, and indicated its metabolic consequences. We generated time-dependent data that predicted the metabolic pathophysiology of the alcohol intervention. Multivariate statistical modelling was applied to the longitudinal data of 120 biologically relevant organic acids, of which 13 provided statistical evidence of the alcohol effect. The known alcohol-induced increased NADH:NAD⁺ ratio in the cytosol of hepatocytes contributed to the global dysregulation of several metabolic reactions of glycolysis, ketogenesis, the Krebs cycle and gluconeogenesis. The significant presence of 2-hydroxyisobutyric acid supports the emerging paradigm that this compound is an important endogenous metabolite. Its metabolic origin remains elusive, but recent evidence indicated 2-hydroxyisobutyrylation as a novel regulatory modifier of histones. Metabolomics has thus opened an avenue for further research on the reprogramming of metabolic pathways and epigenetic networks in relation to the severe effects of alcohol consumption.

3.6.3 Introduction

Notwithstanding the encyclopaedic information on alcoholism, the WHO asserts that alcohol remains one of the world's leading risk factors for disability, morbidity and mortality — 5.9% of all deaths worldwide are attributable to alcohol consumption, exceeding those from HIV/AIDS (2.8%), violence (0.9%) or tuberculosis (1.7%) (Who, 2014). The tenth edition of the *International Classification of Diseases* lists at least 25 chronic conditions that are entirely attributable to alcohol; alcohol is also a risk factor in certain cancers, some tumours, numerous cardiovascular and digestive diseases, and many neuropsychiatric conditions (Shield *et al.*, 2013). Brain image studies have revealed changes in brain structure during the progression from adolescence to adulthood (Blakemore & Choudhury, 2008; Lenroot & Giedd, 2006), a critical period characterized by increased brain connectivity and maturation of brain neural circuits. These changes are highly susceptible to the effects of exogenous substances, which likely include alcohol (White *et al.*, 2000; Fein *et al.*, 2013; Guerrini *et al.*, 2014), making children and adolescents especially vulnerable to alcohol-related harm (Mäkelä & Mustonen, 2000). It was reported from a recent

survey that approximately 50% of children (aged 11–14 years) in the UK have consumed alcohol, and 33% of adolescents (15–16 years) admitted to having experienced at least one episode of acute alcohol intoxication in the month preceding the survey (O’Neil *et al.*, 2012)

Human genome-wide association studies (GWAS) have identified polymorphisms and candidate genes associated with individuals’ innate risks for alcohol dependence (Spanagel *et al.*, 2010; Rietschel & Treutlein, 2013). With metabolomics involving large-scale molecular studies of metabolic systems, it provides novel and complementary approaches to GWAS studies (Nahon *et al.*, 2012; Rachakonda *et al.*, 2014; Amathieu *et al.*, 2014). Additionally, metabolomics has revealed the extensive perturbations of various metabolic pathways in response to chronic alcohol consumption, which underpin alcohol-induced disease states. Such studies show much promise for disease profiling and biomarker identification of the conditions associated with and arising from chronic alcohol consumption (Manna *et al.*, 2011). The burden attributable to acute alcohol consumption (rapid ingestion of alcoholic beverages) or binge drinking (drinking too much too fast) has been shown to be similarly high (Gill, 2002; Mathurin & Deltenre, 2009; Askgaard *et al.*, 2015). Impediments associated with acute and binge drinking include perturbed metabolism (e.g. glycogen depletion, acidosis and hypoglycaemia) (Zakhari & Li, 2007), damage to intestinal epithelial cells (Szabo, 2015), neurobiological diseases (Petit *et al.*, 2013), and ultimately even abrupt premature death due to physical or mental disabilities (Rehm *et al.*, 2006; Taylor *et al.*, 2010).

Several important observations on acute alcohol consumption have recently been obtained from animal-based studies:

- A rodent “*intragastric feeding model*” was investigated with ultra-high performance liquid chromatography-time of flight mass spectrometry (UHPLC-TOFMS) technology to determine changes in global metabolite profiles for plasma and urine from alcohol treated rats and mice compared to control animals (Gika *et al.*, 2012). Apart from several other observations, the researchers reported changes in the concentrations of 5-hydroxytryptophan and xanthurenic acid, both of which are intermediates in tryptophan metabolism. These observations provide further insight into the association of liver metabolism in response to ethanol exposure.
- A study on the immune response of healthy individuals (11 males and 14 females, aged 21–56 years) with no history of alcohol use disorder, revealed that acute binge drinking resulted in a rapid increase in serum Gram-negative bacterial endotoxin lipopolysaccharide (LPS) and bacterial 16S rDNA. LPS is a potent trigger of the inflammatory cascade via activation of the

Toll-like receptor 4 (TLR4) complex and increase in the portal and/or systemic circulation in several types of chronic liver diseases (Szabo *et al.*, 2010).. The elevation of bacterial 16S rDNA levels after acute binge drinking indicated transient gut-derived microbial translocation as a likely mechanism for the serum LPS increase. The authors suggest that the increased serum levels of bacterial products following acute consumption might contribute to innate immune responses and potentially to the behavioural effects associated with alcohol binge drinking. The study also draws attention to potential perturbations in metabolite profiles which may be due to the role of the gut and the microbiome in binge drinking. This point is substantiated by GC–MS urinary metabolomics results from another study, which revealed metabolite differences between Sprague–Dawley and Wistar rats following different perturbations, including consecutive acute ethanol interventions (Gao *et al.*, 2011). These results directed to different metabolic pathways and differences in the intrinsic metabolism and symbiotic gut microflora between these animal strains.

- A review of the analytical technologies used in profiling studies of animal or human serum, plasma, urine and tissue samples, obtained following exposure to alcohol, summarizes a range of endogenous metabolites that have been proposed as potential ethanol consumption-related biomarkers (Gika & Wilson, 2014). This range of biomarkers provides biochemical insights that are essential for understanding the effects and mechanisms of ethanol toxicity.

Notwithstanding the potential of metabolic profiling, a recent review noted that few studies on acute alcohol consumption have been undertaken in humans (Gika & Wilson, 2016), recognizing that these models could provide a basis for studying the biochemical effects of prolonged ethanol exposure, as well as to potentially identify biomarkers for monitoring the progression of alcoholism in man (Gika *et al.*, 2012).

We thus postulated that our metabolomics approach could provide further insights into the metabolic signature which arises from a single excessive dose of alcohol. However, the response to experimental alcohol consumption is complex, as shown by the diverse observations noted from the study on two related animal strains (Gao *et al.*, 2011). Alcohol studies in humans, likewise, vary greatly according to the extent and method of usage (chronic or acute), individual variability (genetic and behavioural), and environmental factors. Also, individuals' attitude to their consumption habits is a private issue, and important ethical considerations (such as restricted case participation and policy guidelines) have to be taken into account during the selection of sufficient and appropriate participants for the laboratory assessment of alcohol use (Plant, Plant

& Vernon, 1996). Therefore, recording the influence of alcohol in laboratory studies involving a select group of moderate-drinking young men, as we present here, is complex from several points of view.

Despite these limitations and qualifications, the application of metabolomics to intervention or challenge studies is a preferred practical approach towards a holistic understanding of the effects of consumed substances on metabolism (Gibney *et al.*, 2005; Wishart, 2008). Intervention studies produce extensive data sets due to longitudinal (time-dependent), multi-subject (several experimental participants), multi-group (number of interventions), and multivariate data (numerous metabolic variables) inputs (Antti *et al.*, 2002; Smilde *et al.*, 2005). Our previous intervention study on consuming commercial flavoured water with a benzoic acid preservative (Irwin *et al.*, 2016) indicated that time-dependent metabolomics investigations, using designed interventions, provide a way of interpreting the variation induced by the different factors of a designed experiment. This approach has the potential to significantly further our understanding of normal and pathophysiological perturbations of endogenous or exogenous origin.

Here we report on a metabolomics study which used commercial flavoured water as a vehicle (Irwin *et al.*, 2016) for vodka consumption, to simulate acute but controlled alcohol consumption (resembling a single incident of acute or binge drinking) in young male participants. All metabolomics studies are inherently hampered by an analytical inadequacy to provide a comprehensive coverage of the metabolome. Given the known main pathways of alcohol metabolism, we selected a gas chromatographic–mass spectrometric (GC–MS) approach to analyse the urinary samples collected during this time-dependent crossover study. These analyses generated an extensive data set, from which (by using conventional and extended statistical methods) we could reveal a multitude of disturbances in the urinary organic acid profiles over time due to the alcohol intervention.

3.6.4 Results

Effect of the acute alcohol intervention across all cases

Observations of the behaviour of the participants during the intervention stage indicated a moderate degree of acute alcohol consumption characteristic of the early phase of a binge-drinking event by young male students (Roberts, 2015). To determine the effect of the alcohol intervention, the data sets of 120 organic acid metabolites, generated from the samples collected at times 1, 2, 3 and 4 hours following the intervention, were compared with the data set from the

samples collected prior to the intervention (time 0), and presented for unsupervised principal component analysis (PCA). Subsequently, a supervised partial least squares discriminant analysis (PLS-DA) was performed to maximize the discrimination between the controls (time 0) and the subsequent hourly-collected data. The PCA and PLS-DA results are shown in Figure 3-26: Input data were from the 120 quantified metabolites for time 0 vs time 1 (a & e), time 0 vs time 2 (b & f), time 0 vs time 3 (c & g), and time 0 vs time 4 (d & h) following alcohol consumption. The samples for time 0 were collected just prior to the intervention, and are therefore regarded as the control samples. The figures were based on the non-paired method of analysis. Results from the paired method are shown in the Supplementary Information (SI) Figure 3-33.

The PCA scores plots showed some differentiation 1 hour after the alcohol consumption (Figure 3-26a), followed by complete separation after 2 hours, and a progressive return to the time 0 profile after 3 and 4 hours. The PLS-DA scores plots (Figure 3-26 e through h) showed a complete separation for all four times following alcohol consumption relative to time 0.

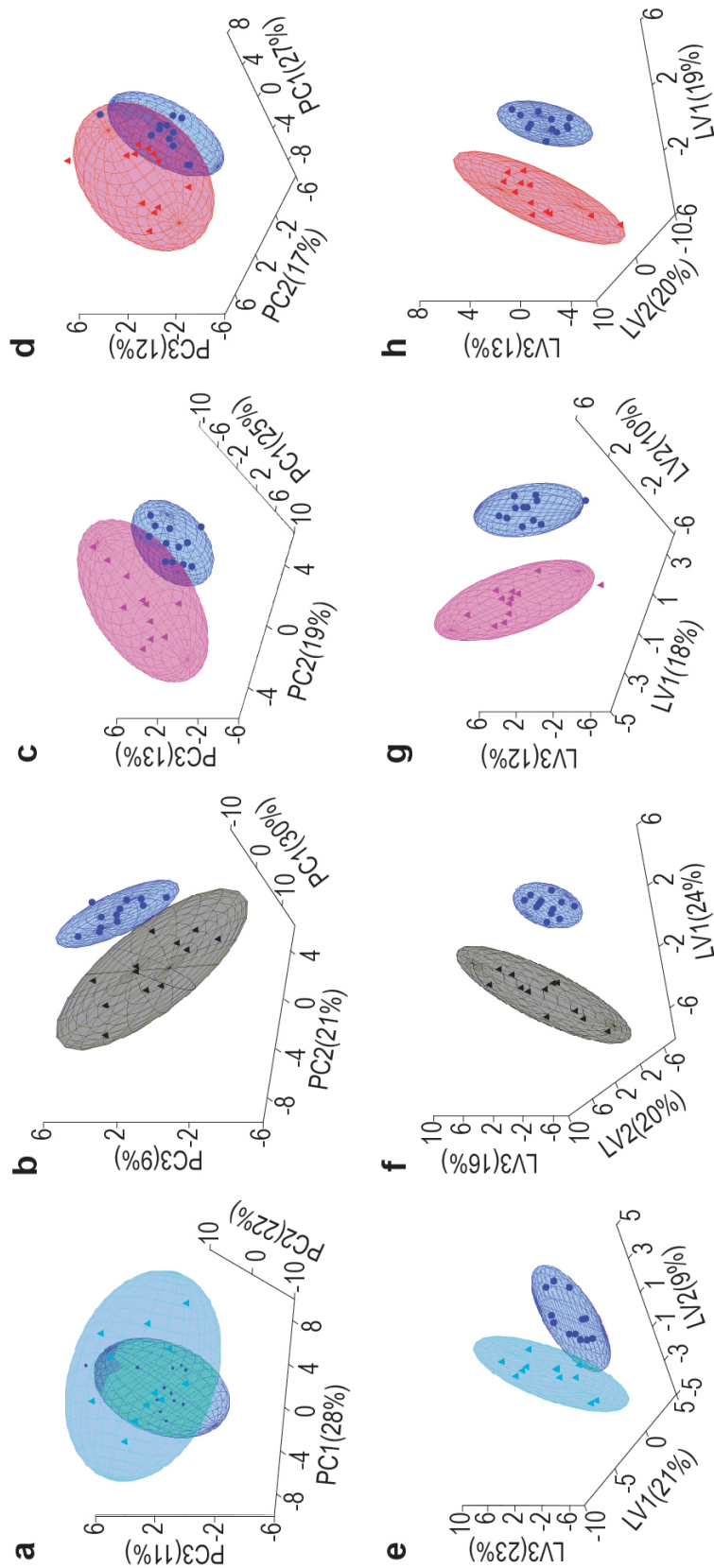


Figure 3-26: PCA and PLS-DA scores plots following alcohol consumption.

The PCA and PLS-DA results are shown in the top and bottom rows, respectively. Input data were from the 120 quantified metabolites for time 0 vs time 1 (a & e), time 0 vs time 2 (b & f), time 0 vs time 3 (c & g), and time 0 vs time 4 (d & h) following alcohol consumption.

For better insight into the time effect of the intervention, two complementary approaches were used, both of which are founded on Principal Component Analysis (PCA). The first, ANOVA Simultaneous Component Analysis (or ASCA), enabled us to partition variation similar to an analysis of variance (or ANOVA) approach using the experimental design, and summarize variation observed across many variables through projection to a lower dimensional space.

The PCA and ASCA scores plots are shown in Figure 3-27: (a) ASCA scores on the first three latent variables (LV1 to LV3), along with 95% confidence ellipsoids for the centroids, are shown and colour-coded according to time following the intervention (0, 1, 2, 3 and 4 hours shown in black, red, pink, dark blue and light blue, respectively), with the arrow showing the time-dependent trend, using the same discriminating colour sequence. (b) Unfolded PCA scores centroids for each time point based on the first two principle components (PC1 and PC2) are shown. The colours of the centroids are the same as for the ASCA and the direction of the trajectory is indicated by the red arrows, starting from time 0.

The scores plot resulting from the ASCA model (Figure 3-27a) indicates a marked change one hour after the intervention. After 3 hours a state of homeostasis is reached as indicated by the similarity between times 3 and 4. This is, however, a new state of homeostasis, since it is not comparable to the state prior to the intervention, as will be discussed below. Though time 2 appears to show a larger perturbation compared to time 1, following the intervention, time 2 also shows a progression towards the new state of homeostasis, making it less indicative of the initial alcohol effect. The ASCA model partitioned variation from various sources in the design (that is, time and participant) by averaging out individual effects. Secondly, performing PCA on the dataset unfolded in time provides another view on the variation observed, which ties repeated observations together into a single profile. The scores from such a model provide a complimentary and, in a sense, confirmatory summary of the effect of the alcohol consumption in time, as indicated by the score centroids shown in Figure 3-27b. It is evident that there was a change in the organic acid profile (i.e. across all cases) over time, with the most extensive change visible 1 hour (time 1) after the intervention. The centroids plotted for times 2 to 4 complement the interpretation made from the ASCA.

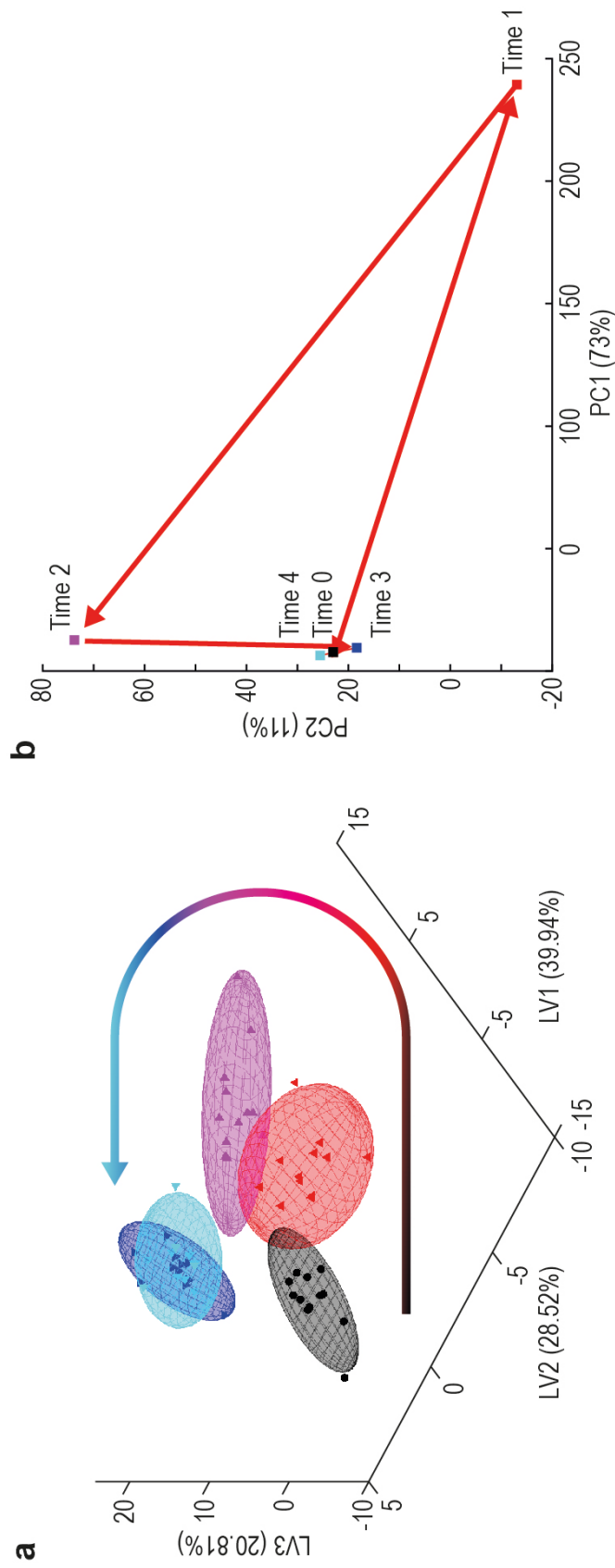


Figure 3-27: Multivariate approaches to indicate the time effect.

(a) ASCA scores on the first three latent variables (LV1 to LV3), along with 95% confidence ellipsoids for the centroids, are shown and colour-coded according to time following the intervention (0, 1, 2, 3 and 4 hours shown in black, red, pink, dark blue and light blue, respectively), with the arrow showing the time-dependent trend, using the same discriminating colour sequence. (b) Unfolded PCA scores centroids for each time point based on the first two principle components (PC1 and PC2) are shown. The colours of the centroids are the same as for the ASCA and the direction of the trajectory is indicated by the red arrows, starting from time 0.

Based on the statistical criteria, a total of 13 variables (Table 3-8) were identified as those responsible for the separation between the control samples and those collected one hour after alcohol consumption. Shown in Table 3-8 are the VIP values, *p*-values and FC values at time 1 relative to time 0, as well as the mean and standard deviation values for all times. The FC and *p*-values for time 4 relative to time 0 are also included to show the eventual diminishing effect of the alcohol on the 13 metabolites. In addition, since ethanol data is essential in the context of this intervention study, these values are included for reference purposes. However, ethanol cannot be reliably detected with the GC–MS method used here, and therefore the quoted urinary ethanol values were measured through a targeted proton nuclear magnetic resonance [¹H-NMR] spectroscopy analysis. The full metabolite profile that became apparent through this NMR metabolomics study will be reported in a subsequent paper.

Table 3-8 contains descriptive statistics as well as univariate and multivariate information on perturbed metabolites following alcohol consumption. A total of 13 organic acid metabolites were identified as important discriminatory variables one hour after alcohol consumption (VIP ≥ 1.0, *p* ≤ 0.05 and |FC| ≥ 1.5). Abbreviations used: [Std. Dev.], standard deviation; Cr, urinary creatinine value; HMDB, Human Metabolome Database; [WRT]0-1 and [WRT]0-4, Wilcoxon signed-rank test *p*-values showing the significance of a metabolite at times 1 and 4 hours relative to time 0; [FC]0-1 and [FC]0-4, fold change values for a metabolite at times 1 and 4 hours relative to time 0; n/a, not applicable for the purpose of this table.

Consumed ethanol is known to be dispersed through exhaling (not measured), metabolic conversions and urinary excretion, which peaked at two hours following its consumption (Table 3-8). However, the results from the ASCA and unfolded PCA (Figure 3-27) — which indicated that the most dramatic metabolite changes occurred one hour following the intervention — encouraged us to preferably study in detail the metabolic effects seen one hour after alcohol consumption.

Metabolites ¹	Controls vs 1 hr		Mean concentrations of metabolites*					Controls vs 4 hr		HMDB Ref. values**
	PLS-DA VIP≥1.0	[WRT] ₀₋₁ [FC] ₀₋₁	t = 0	t = 1	[Std. Dev.] t = 2	t = 3	t = 4 hr	[WRT] ₀₋₄ [FC] ₀₋₄		
Hippuric acid	19.6	0.002 +3.2	948 [737]	3037 [1357]	2775 [2051]	1157 [1111]	726 [764]	0.239 -1.3		27.9–932.7
Lactic acid	14.4	0.015 +27.4	8.73 [3.65]	239 [412]	28.1 [19.7]	15.5 [14.7]	12.1 [8.51]	0.209 +1.4		3.9–9.8
Fumaric acid	7.24	0.003 +6.4	0.78 [0.47]	4.96 [4.53]	5.88 [6.08]	2.62 [1.94]	2.8 [2.15]	0.006 +3.6		0.75–1.2
Vanillylmandelic acid	5.03	0.034 -1.6	15.4 [3.61]	9.87 [7.39]	2.96 [2.32]	1.87 [1.11]	1.75 [0.99]	0.002 -8.8		1.1–1.7
2-Hydroxybutyric acid	4.06	0.034 +4.3	2.68 [1.54]	11.4 [16.0]	3.85 [2.68]	2.26 [1.16]	2.52 [2.15]	0.695 -1.1		1.2–6.9
Succinic acid	3.88	0.010 +2.6	3.81 [3.81]	9.83 [7.90]	20.2 [24.6]	12.8 [12.8]	12.7 [12.3]	0.004 +3.3		4.9–14.9
3-Hydroxybutyric acid	2.76	0.041 +5.9	0.40 [0.54]	2.39 [2.96]	0.59 [0.55]	0.47 [0.40]	0.77 [0.68]	0.182 +1.9		1.4–2.2
2-Ethylhydracrylic acid	1.66	0.010 +2.1	1.09 [1.00]	2.30 [1.73]	2.53 [2.78]	1.76 [1.61]	1.86 [1.47]	0.015 +1.7		1.3–2.9
3-Hydroxyisobutyric acid	1.56	0.034 +1.9	3.01 [1.27]	5.76 [3.86]	3.31 [2.05]	2.74 [1.25]	3.44 [1.95]	0.239 +1.1		4.1–19.0
2-Hydroxyisobutyric acid	1.39	0.034 +1.8	11.1 [4.03]	20.5 [14.2]	20.5 [13.4]	12.8 [6.93]	16.3 [13.1]	0.209 +1.5		4.4–7.6
Malic acid	1.32	0.023 +4.9	0.27 [0.16]	1.30 [1.60]	1.81 [1.59]	0.86 [0.67]	1.19 [1.37]	0.006 +4.5		0.7–5.3
N-Tiglylglycine	1.22	0.034 +1.7	1.55 [1.15]	2.65 [1.39]	10.9 [9.15]	10.9 [9.35]	10.8 [8.64]	0.002 +7.0		0.78–1.2
2-Hydroxyglutaric acid	1.06	0.023 +1.7	2.88 [0.99]	4.91 [3.39]	10.0 [5.90]	7.22 [3.62]	8.50 [5.73]	0.003 +3.0		0.8–52.0
Ethanol ²	n/a	n/a	0.0	1368	5173	2852	1810	n/a	n/a	n/a

*: μmol metabolite/ mmol Cr | **: μmol metabolite/ mmol Cr | 1: 13 organic acid metabolites (concentrations expressed as $\mu\text{mol}/\text{mmol}$ Cr), identified through GC–MS analysis. | 2: Ethanol (concentrations expressed as $\mu\text{mol}/\text{mmol}$ Cr), identified through an independent, targeted ^1H -NMR analysis of the same urine samples as used for the GC–MS analyses

Table 3-8: Univariate, multivariate and descriptive statistics for the most perturbed metabolites following alcohol consumption.

Abbreviations used: [Std. Dev.], standard deviation; Cr, urinary creatinine value; HMDB, Human Metabolome Database; [WRT]0-1 and [WRT]0-4, Wilcoxon signed-rank test p-values showing the significance of a metabolite at times 1 and 4 hours relative to time 0; [FC]0-1 and [FC]0-4, fold change values for a metabolite at times 1 and 4 hours relative to time 0; n/a, not applicable for the purpose of this table.

Of the important metabolites, hippuric acid, the phase II biotransformation product of benzoic acid (Gao *et al.*, 2011; Irwin *et al.*, 2016; Roberts, 2015; Conti & Bickel, 1977; Badenhorst *et al.*, 2014), had the highest VIP value (Table 3-8, VIP = 19.6). The concentration of hippuric acid was already high at time 0 (948 $\mu\text{mol}/\text{mmol Cr}$), due to its being the normal excretion product of benzoic acid derived from the gut microbiome. This concentration increased to 3037 $\mu\text{mol}/\text{mmol Cr}$ one hour after the alcohol consumption, owing to the benzoic acid preservative in the flavoured water vehicle consumed with the alcohol. Hippuric acid likewise appeared to be the most important discriminating variable in the previously described effects of the vehicle-only intervention (Irwin *et al.*, 2016), as well as in the consecutive acute ethanol intervention study on rats, where the change in urinary hippuric acid is suggested to be due to a metabolic dysfunction of damaged liver tissue (Gao *et al.*, 2011). The presence of the high concentration of hippuric acid does not influence the outcome of the analysis, since the same list of important metabolites (VIP > 1.0) is obtained after excluding hippuric acid from the data (see Section 3.7.7 of the SI). It is also worth noting that four other gut-related urinary metabolites observed in the rat study (Gao *et al.*, 2011) were also present in the urine samples from our study (see Table 3-10 in the SI), but did not appear to be discriminatory metabolites due to the alcohol intervention. Other important biotransformation products from the rat and mouse intervention studies, such as ethyl glucuronide and ethyl sulphate (Gika *et al.*, 2012), were not detectable by the GC–MS methods used in the present study.

The metabolic interrelations among the remaining 12 metabolites in Table 3-8 correspond to several metabolic consequences of alcohol consumption. The oxidation of ethanol by alcohol dehydrogenase (ADH) creates a highly reduced cytosolic environment in hepatocytes, and favours the production of lactic acid (second-highest VIP) from pyruvic acid, resulting in downstream metabolic consequences due to pyruvic acid depletion. The reduced environment also accounts for the perturbations observed for vanillylmandelic (FC = –1.6), 2-hydroxybutyric (FC = +4.3) and 3-hydroxybutyric (FC = +5.9) acids, and increased urinary excretion of succinic, fumaric, malic and 2-hydroxyglutaric (derived from 2-ketoglutaric acid) acids that implicates a dysfunctional Krebs cycle. Increased excretion of N-tiglylglycine (the phase II biotransformation product of tiglyl-CoA) and 2-ethylhydracrylic acid (intermediates in the R- and S-pathways of isoleucine catabolism, respectively), as well as 3-hydroxyisobutyric acid (produced from valine catabolism) points to amino acid mobilization but inhibition of gluconeogenesis (Costantini *et al.*, 1996). The reason for the increased excretion of 2-hydroxyisobutyric acid (2-HIBA) remains speculative and will be discussed below.

To characterize the metabolic relationship between the 13 important variables further, we calculated Spearman's rho correlation coefficients over the full period of the alcohol intervention, as described in Section 3.7.6 of the SI, and shown in Figure 3-28. Correlations extend from high positive (red; $r \geq 0.6$) through neutral (green; $-0.2 < r < 0.2$) to high negative (blue; $r \leq -0.6$). The high correlation between lactic, 2-hydroxybutyric and 3-hydroxybutyric acids and the high correlation between the Krebs cycle-associated metabolites (together with their respective time slots) are blocked in green and orange, respectively.

Important observations from the correlation analysis correspond to the metabolic interrelations: (1) the profile for vanillylmandelic acid was unique — it was the only metabolite that showed a negative correlation with time ($r = -0.7$), indicating that its urinary concentration constantly decreased over the study period, and remained very low ($1.75 \mu\text{mol}/\text{mmol Cr}$; $p = 0.034$; $\text{FC} = -8.8$) towards the end (at time 4 hours). (2) A good correlation ($r = 0.4$ to 0.7) was observed between the indicators of ketosis (lactic, 2-hydroxybutyric and 3-hydroxybutyric acids), but only in the initial phase following alcohol consumption, as indicated by their neutral correlations with time ($r = 0$ to -0.2 with time). (3) A high to very high correlation ($r = 0.6$ to 0.8) was observed between the Krebs cycle-associated metabolites, all of which also show a reasonable correlation ($r = 0.3$ to 0.5) over the full period following the intervention. (4) Likewise, N-tiglylglycine showed a high correlation with the Krebs cycle intermediates ($r = 0.6$ to 0.8), which also extended over the full study period ($r = 0.7$).

The high correlation values shown in Figure 3-28 collectively substantiate a strong interrelationship between the 13 organic acid metabolites that were regarded as important due to acute alcohol consumption. The mean concentrations, and the time 4 relative to time 0 p -values ($[\text{WRT}]_{0-4}$) and fold changes ($[\text{FC}]_{0-4}$) for the 13 variables (listed in Table 3-8), support the conclusions about the individual and group correlations observed in Figure 3-28; they also provide an alternative and additional support for the observations from the correlation study.

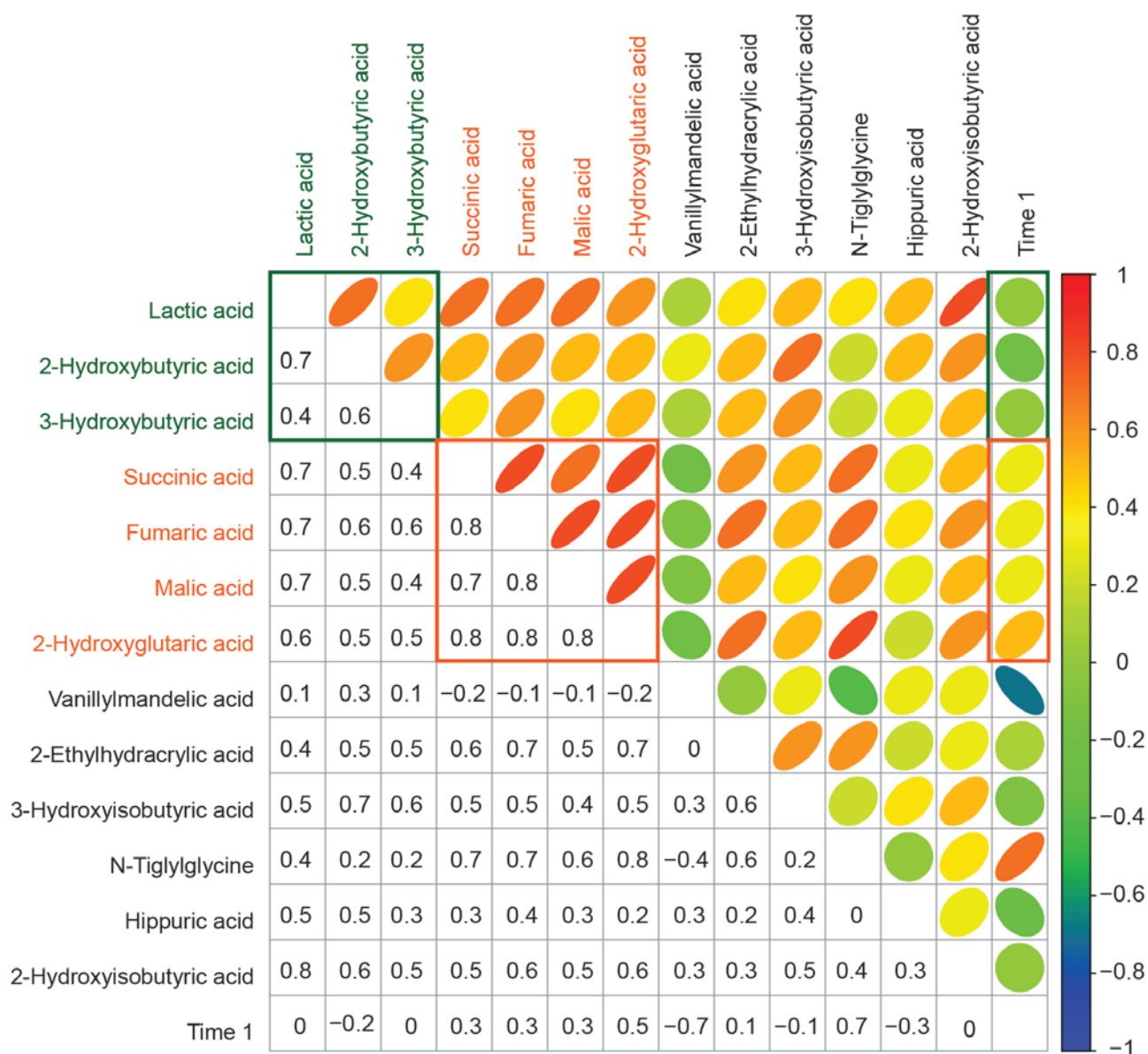


Figure 3-28: Correlation matrix over the full time period for the 13 metabolites listed in Table 3-8.

Correlations are colour-coded according to the scale to the right of the matrix. The high correlation between lactic, 2-hydroxybutyric and 3-hydroxybutyric acids and the high correlation between the Krebs cycle-associated metabolites are blocked in green and orange, respectively.

Inter-individual variation following acute alcohol consumption

In order to illustrate the inter-individual variation between the participants, as well as the longitudinal effect of the alcohol consumption, PCA of the data unfolded in time was performed as described in Section 3.7.6 of the SI. The unfolding transformed the three-dimensional data (a tensor of cases, interventions and time) into a two-dimensional matrix, and thus allowed PCA to account for the longitudinally repeated measures.

This analysis of the unfolded data provides insight into the effect of the acute alcohol consumption over time on the 120 metabolites. Figure 3-29(a-c) illustrates the scores based on the first two principal components (PC1 and PC2) of the PCA model. The centroids of the five time-dependent scores are shown in red, along with the 90% confidence ellipsoids for each centroid. Figure 3-29 b and c show the trajectories from time 0 onwards for two individuals, illustrating the variation in profiles in response to the alcohol intervention. Figure 3-29 d to show bi-plots for the seven most important metabolites when ranking was based on the sum of the squares of loadings of the first two components. These metabolites are more primarily responsible for the pattern in the plot, that is, causing separation in the unfolding of the PCAs, as applicable for times 0 (d), 1 hour (e) and 4 hours (f). Encircled numbers identify the metabolites: 1, hippuric acid; 2, phenylacetylglutamine; 3, 4-hydroxyphenylacetic acid; 4, indole-3-acetic acid; 5, pyrrole-2-carboxylic acid; 6, pyroglutamic acid; 7, lactic acid; 8, 2-hydroxybutyric acid; 9, fumaric acid; 10, N-cinnamoylglycine; 11, glucuronic acid.

The centroids of the PCA scores for each time (Figure 3-29a), as well as those of two individual cases (Figure 3-29 b and c), illustrate the inter-individual variation in response to the intervention. The averaged and individual trajectories showed similarities in the individual responses to the alcohol consumption over time, since they all indicated a biphasic response pattern (phase 1: times 0 to 1 to 2, and phase 2: times 2 to 3 to 4). However, distinct differences were also noted, indicated by the unique orientation and biphasic profiles of the trajectories of the individuals (Figure 3-29 b and c). The metabolic basis for the latter observation is illustrated, for example, by a comparison of the lactic acid excretion observed in the two cases, which progressed from 8.2 to 6.0, 74.5, 14.7 and 16.0 $\mu\text{mol}/\text{mmol Cr}$ for case 1 (Figure 3-29b), but from 10.6 to 31.2, 19.8, 8.7 and 11.2 $\mu\text{mol}/\text{mmol Cr}$ for case 2 (Figure 3-29c). Similar differences in the individual responses to the alcohol consumption were observed for several other metabolites as well, which cumulatively result in the unique trajectory of each individual subject — an observation that agrees with the contemporary view on genetic and metabolic individuality.

Figure 3-29 d to f illustrates the bi-plots for the top most influential metabolites responsible for the separation in the unfolding of the PCAs. The dominant metabolites observed at time 0 were phase II biotransformation products (hippuric acid and phenylacetylglutamine) and organic acids originating from the gut microbiome (4-hydroxyphenylacetic, indole-3-acetic, pyrrole-2-carboxylic and pyroglutamic acids). This gut-derived organic acid profile dominates at all the times measured, as also indicated in Figure 3-29 e and f for times 1 and 4 hours. Notably, one hour after alcohol consumption, lactic, 2-hydroxybutyric and fumaric acids appear as additional important metabolites, in accordance with the results in Table 3-8 and Figure 3-28. At time 3, aconitic acid is indicated as an important metabolite (figure not shown), whereas N-cinnamoylglycine and glucuronic acid (both associated with phase II biotransformation) appear to be important 4 hours after alcohol consumption. It thus appears that the endogenous detoxification mechanisms through biotransformation remained functional despite the acute alcohol consumption.

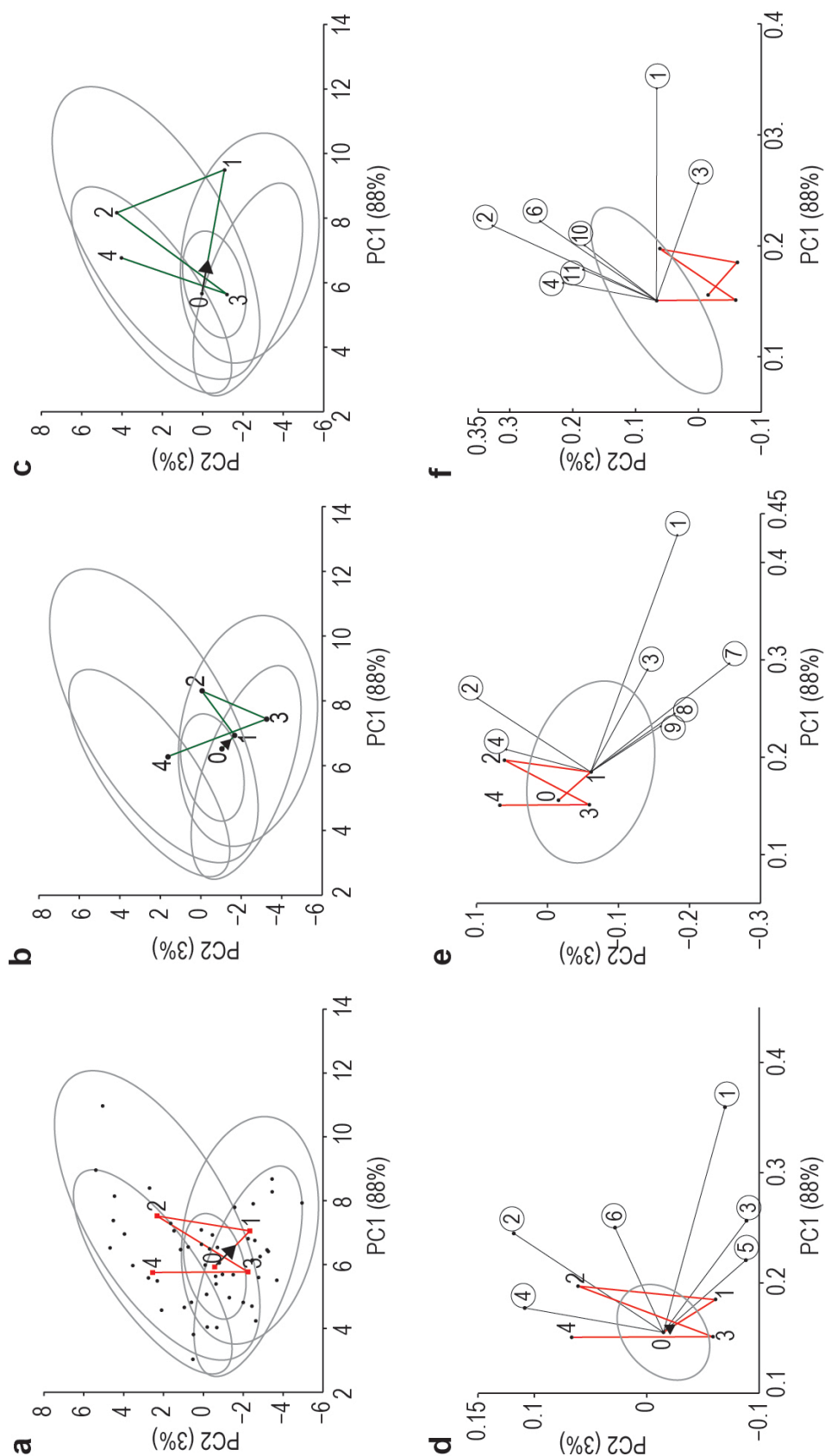


Figure 3-29: Unfolded PCA scores and selected bi-plots.

Figures a to c show the scores based on the first two principal components (PC1 and PC2). The centroids of the scores at each time are shown in red, with the 90% confidence ellipsoids. Figures b and c show the trajectories from time 0 onwards for two individuals. Figure d to f is a bi-plot showing the seven metabolites with the highest sum of the squares of loadings of the first two components for times 0 (d), 1 hour (e) and 4 hours (f). Encircled numbers identify the metabolites: 1, hippuric acid; 2, phenylacetylglutamine; 3, 4-hydroxyphenylacetic acid; 4, indole-3-acetic acid; 5, pyrrole-2-carboxylic acid; 6, pyroglutamic acid; 7, lactic acid; 8, 2-hydroxybutyric acid; 9, fumaric acid; 10, N-cinnamoylglycine; 11, glucuronic acid.

3.6.5 Discussion

In this study, the metabolomics organic acid profile revealed significant metabolic effects of a single dose of alcohol, consumed in a well-defined vehicle by healthy, moderate-drinking males. Although their analytical constraints limit the scope of profiling studies, we concur with the view that carefully conducted studies in humans are warranted, and would provide valuable new insights into the short and long term effects of alcohol exposure, alcoholic liver disease and alcoholism, in man (Gika & Wilson, 2014).. The method followed here offered a convenient and sensitive approach to uncover perturbed metabolic pathways, of which we modelled the main ones as illustrated in Figure 3-30.

Several distinct insights deserve special attention. The first insight deals with hepatic robustness. Glycine N-acyltransferase (GLYAT)-based biotransformation was shown to play a vital background role during the intervention study. The scores and bi-plots from the unfolded PCAs (Figure 3-29d–f) revealed the reasons for the observed differentiation between the groups shown in the scores plots of the unsupervised PCAs (Figure 3-26a–d). The observed differentiation was caused primarily by detoxification (gut metabolites) and biotransformation (phase II conjugates) processes. Apart from the alcohol load from the intervention, the liver is simultaneously confronted with the benzoic acid load from the vehicle, which induced a significant hepatic biotransformation response (Irwin *et al.*, 2016). Two closely related acyl-CoA:amino acid N-acyltransferases have been characterized as operative in liver mitochondria (Nandi *et al.*, 1979). One transferase was specific for benzoyl-CoA, salicyl-CoA, and certain short straight- and branched-chain fatty acyl-CoA esters as substrates, whereas the other enzyme specifically used either phenylacetyl-CoA or indoleacetyl-CoA. It thus seems probable that the benzoyl-CoA:glycine N-acyltransferase (using glycine as the preferred acyl acceptor) was functional in the biotransformation of the vehicle-induced accumulated benzoyl-CoA, but notably also of tiglyl-CoA and 2-methylenebutyryl-CoA, which accumulated from perturbations due to the alcohol consumption. This biotransformation resulted in the formation and increased excretion of hippuric acid, N-tiglylglycine and N2-methylenebutyrylglycine (a minor metabolite increased due to the alcohol consumption -data not shown, but included in Figure 3-30), respectively. The metabolomics approach thus provides insight into a background process that underscores the robustness of hepatic function in healthy individuals, notwithstanding the impact of the acute alcohol consumption - an observation that would not hold true for the liver affected by chronic alcoholism.

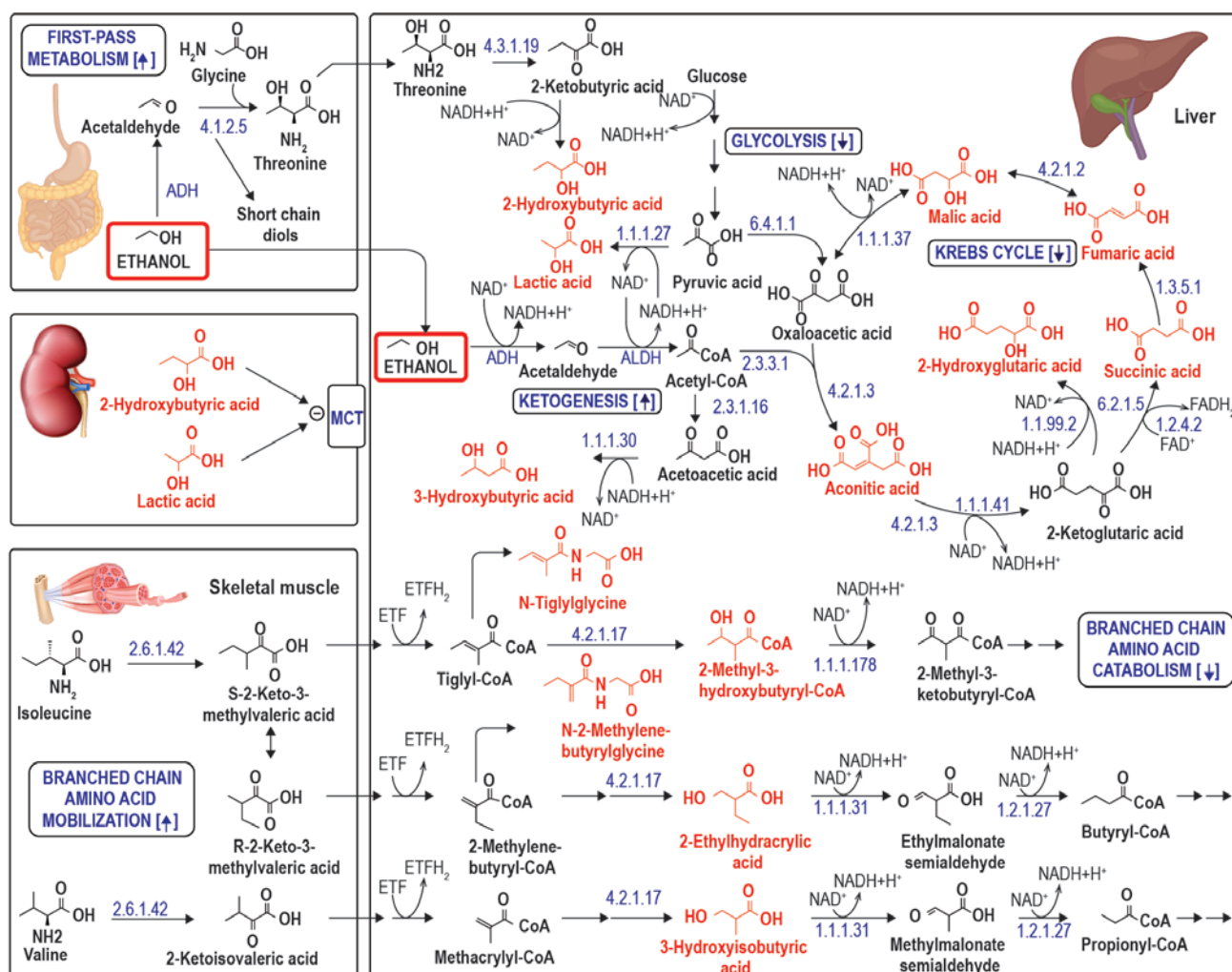


Figure 3-30: Proposed model indicating some important metabolic pathways affected by acute alcohol consumption.

Up- or down-regulated metabolites are shown in red. Abbreviations used: MCT, monocarboxylic acid transporter; four-figure numbers (e.g. 2.6.1.42) refer to the IUBMB enzyme nomenclature (EC number); S, main catabolic pathway of isoleucine; R, minor catabolic pathway of isoleucine, which may act as a safety valve for overflow of accumulating metabolites from the S-pathway. Credit for images used: Kidney – ID 20446327 © Natis76 | Dreamstime.com; Muscle – ID 69693231 © Tigatelu | Dreamstime.com; Liver & Digestive system – ID 43552467 © Bluezace | Dreamstime.com

The second insight deals with the comprehensive metabolic perturbations caused by the acute alcohol consumption, as disclosed by the supervised PLS–DA. The metabolic reactions following alcohol consumption showed a phased response profile, to which the gut is known to contribute following alcohol consumption. In the early phase, a fraction of the alcohol consumed orally may be immediately metabolized by the gut alcohol dehydrogenase (ADH) isoforms in a process

known as first-pass metabolism (DiPadova *et al.*, 1987). A study on acute alcohol binge drinking showed increased serum levels of bacterial products (endotoxin and bacterial 16S rDNA) derived from the gut microbiome (Bala *et al.*, 2014). The serum endotoxin levels rapidly increased within 30 minutes following consumption, remained elevated for 3 hours and returned to lower than baseline levels by 24 hours after alcohol intake. The early increase in 2-hydroxybutyric acid following alcohol consumption observed in the present study (Table 3-8, Figure 3-28, and Figure 3-29e — metabolite 8) is an indication of a limited contribution from first-pass metabolism in the present experimental group: threonine is anticipated to be produced from the increased intestinal acetaldehyde (the product of ethanol metabolism), and is subsequently catabolised in the liver to 2-hydroxybutyric acid, as proposed in the representation shown in Figure 3-30. However, since the liver is important in alcohol metabolism, and considering its wide spectrum of lesions (Louvet & Mathurin, 2015), we linked most of the metabolomic perturbations observed in the present study to the liver metabolism, as shown in the conceptual model in Figure 3-30. Alcohol oxidation by hepatic ADH results in the reduction of NAD^+ to NADH, thereby generating a highly reduced cytosolic environment in hepatocytes. The increased NADH:NAD^+ ratio influences the acetaldehyde–lactate dehydrogenase coupled lactic acid accumulation (ALDH–LDH = EC1.2.1.10–EC1.1.1.27). Greatly increased lactic acid excretion (Table 3-8, Figure 3-28, and Figure 3-29e— metabolite 7) is proposed as the next early hallmark of acute alcohol consumption and, together with the concomitant early increase in 3-hydroxybutyric acid (Table 3-8, Figure 3-28), is linked to ketogenesis. Acetoacetic acid, the precursor of 3-hydroxybutyric acid, showed a similar early increase (0.012, 0.016 and 0.044 $\mu\text{mol}/\text{mmol Cr}$ at times 0, 1 and 2 hours, respectively), and peaked at two hours after alcohol consumption (VIP = 0.004; FC = 3.64; $p = 0.02$). Taken together, first-pass metabolism and ketogenesis seem to be the immediate salvage responses to disturbed homeostasis caused by the acute alcohol consumption.

Further, an increased NADH:NAD^+ ratio (depleted NAD^+) has been described to influence the direction of several metabolic processes, including decreased glycolysis (Bradford & Rusyn, 2005), decreased Krebs cycle (Zakhari & Li, 2007; Hawkins & Kalant, 1972), and decreased gluconeogenesis (Zakhari & Li, 2007; Lieber, 19997). The early increase in fumaric acid (Table 3-8, Figure 3-28, and Figure 3-29e— metabolite 9) indicates reduced Krebs cycle functionality, which persisted to the end of the 4-hour period of the present intervention study (Table 3-8, Figure 3-28). The notable and persistently reduced Krebs cycle functionality is further supported by the inhibition of the catabolism of all three branched-chain amino acids, indicated by the profiles of N-tiglylglycine, 2-ethylhydracrylic acid and 3-hydroxyisobutyric acid (Table 3-8, Figure 3-28). It is worth noting that the perturbed urinary metabolite profile had not normalized 4 hours after alcohol

consumption, although the supervising physician regarded the participants sufficiently sober to permit them to leave the clinic at this time.

The third insight highlights the observations on vanillylmandelic acid. This metabolite was significantly down-regulated by depleted NAD^+ , and remained so over the full 4-hour period (Table 3-8, Figure 3-28). The down-regulation of vanillylmandelic acid reflects the upstream inhibition in the catabolism of the catecholamines noradrenaline and adrenaline, which occurs due to the inhibition of the NAD^+ dependent aldehyde dehydrogenase-catalysed reactions (EC1.2.1.3 and EC1.2.1.5, respectively). The known relative rise in adrenaline and the delayed increase in noradrenaline concentrations after alcohol consumption is a physiological effect of the inhibition of these enzymes (Ireland *et al.*, 1984), which, together with the corticotrophin-releasing factor, may contribute to negative affective states and relapse vulnerability during alcohol abstinence (Becker, 2012). The observations on vanillylmandelic acid may also be important for other alcohol related studies. In a study on alcoholic patients, higher levels of urinary homovanillic acid (an upstream metabolite of the catecholamine pathway) were observed in patients with a A1 allele in their genotype, when compared to patients homozygous for the *TaqIA2* allele (Ponce *et al.*, 2004). We observed that homovanillic acid peaked two hours after acute alcohol consumption (VIP = 0.129; FC = 1.13; $p = 0.44$). These findings suggest that determination of urinary levels of catecholamine end-products should be included in alcohol-related studies, since variance in their levels could relate to a polymorphism, and could affect addictive behaviour.

As a last insight, 2-hydroxyisobutyric acid (2-HIBA) was indicated in this study to be an important and significant organic acid indicator regarding the metabolic consequences of acute alcohol consumption. 2-Hydroxybutyric acid and 2-HIBA are normally reabsorbed in the kidney by a monocarboxylic acid transporter (MCT). This transporter is, however, also responsible for transporting lactic acid between the urine and the blood. After alcohol consumption, the increased lactic acid contributes to the saturation of these transporters. Consequently, not all of the 2-hydroxybutyric acid and 2-HIBA is reabsorbed, leading to their increased excretion in the urine. However, the conventional view is that urinary 2-HIBA is a non-metabolite, obtained from environmental exposure to tertiary-butylacetate (a commercial solvent used in industrial coatings and cleaners) and certain gasoline additives (Bus *et al.*, 2015; Dekant *et al.*, 2001; Amberg *et al.*, 2001). But things change. New research casts a valuable light on 2-HIBA, suggesting that it appears to be an important metabolite with potential value for health and disease — 2-HIBA is reported to be up-regulated in known human disorders such as chronic kidney disease (Mutsaers *et al.*, 2013), diabetes mellitus (Zhang *et al.*, 2014; Li *et al.*, 2009), human gastric cancer (Wang

et al., 2016), fibromyalgia (Malatji *et al.*, 2017), myocardial infarction (Lee *et al.*, 2015; Surekha *et al.*, 2007; Ali & Konishi, 1998), and a number of inborn errors of metabolism (Goodman *et al.*, 1980; Tanaka *et al.*, 1980). In the present study, alcohol induced significant up-regulation of 2-HIBA in the first two hours after consumption (Table 3-8: VIP = 1.39; FC = +1.8; $p = 0.034$). The very high correlation of 2-HIBA with lactic acid ($r = 0.8$), and high correlation with 2-hydroxybutyric and fumaric acids ($r = 0.6$), strongly supports the view that 2-HIBA is an endogenous metabolite. However, considering the diversity of the conditions associated with the up-regulation of 2-HIBA, a common metabolic origin remains elusive. Most notably, MS analyses of peptide fragments from human histones recently identified a new type of histone mark, namely lysine 2-hydroxyisobutyrylation, which is conserved, widely distributed, has a high stoichiometry, and induces large structural change in histones (Dai *et al.*, 2014). Histone acetylation was first identified more than 50 years ago (Allfrey *et al.*, 1964), and has since become known as wide-ranging posttranslational modifications, linked to a variety of processes, including transcription, DNA replication, and DNA damage. A comprehensive catalogue of histone modifications and their proposed functional consequences (Zhao & Garcia, 2015) includes 2-hydroxyisobutyrylation of lysine side chains, amongst others, for which a rather limited function has been determined. Histone modifications are proposed to present a large opportunity in the years to come to gain insights into chromatin biology, epigenetic events and their biological consequences, which should include the proposed alcoholism-related involvement of 2-HIBA (You *et al.*, 2014) in the modification of histones (Volmar & Wahlestedt, 2015; Li *et al.*, 2017).

Metabolomics studies, however, have limitations:

(1) GC–MS analysis, our method of choice, produces data that are noisier than other metabolomics data types, such as NMR or triple quadrupole liquid chromatography–mass spectrometry (LC–QQQ–MS). This may be overcome by using more experimental subjects (but their number was limited in this study due to ethical considerations), or by using replicates of all samples (rather than only random repeat samples as used in our measurement design).

(2) Although it would be useful, quantifying metabolomics data absolutely is often difficult to execute accurately, is relatively expensive, and is rarely done in diagnostic or intervention study designs as presented here. As an acceptable alternative, therefore, we expressed our metabolite concentrations as values relative to an internal standard.

(3) Concomitant plasma or serum analyses did not form part of this experimental design.

Therefore, the metabolomic variables considered in this investigation were generated from only organic acid analyses of urine samples, and were modelled on theoretical grounds on metabolic pathways of the gut, muscle and liver (Figure 3-30). Given the intoxicating influence of alcohol, metabolite levels in the brain are also important, but could also not be considered here due to the biofluid used. However, as baseline, the perturbations of urinary organic acid profiles provide a prime and feasible way for clinical tests to screen for individual response types, as well as for other relevant aspects of acute alcohol exposure.

3.6.6 Conclusions

Taken together, our findings indicate that metabolomics provided a systematic and standardized method for detecting a range of metabolic responses over time, not previously described comprehensively for acute alcohol consumption. These findings open avenues for potentially important future investigations in alcohol research: (1) genotype-based selection of individuals in follow-up alcohol intervention studies is advised, given the clear inter-individual responses to alcohol consumption; (2) low values of urinary vanillylmandelic acid may be an indicator of a binge drinking or acute alcohol consumption episode in seemingly non-intoxicated individuals; and (3) the striking presence of 2-HIBA supports the emerging new paradigm of 2-HIBA being an important endogenous metabolite. Moreover, detailed studies on the biological origin of 2-HIBA, as well as on its perceived gene-modification role through lysine 2-hydroxyisobutyrylation of histones may take us one step closer to understanding the personalized responses to acute alcohol consumption and the perceived epigenetic changes that are induced. All in all, we concur that acute alcohol consumption studies broaden insights on significant adverse health effects of alcohol even in healthy individuals (Bala *et al.*, 2014). These insights will help researchers to define novel approaches to treat or ameliorate alcohol-induced disability, organ damage and morbidity.

3.6.7 Materials & Methods

Intervention study design

The experimental group consisted of 12 clinically selected healthy males (aged 20–24 years), who admitted to consuming alcohol at a moderate, social level (baseline alcohol consumption was defined by the participants' declared levels of drinking). None of the participants used any medication. They were asked to refrain from vitamins, minerals, and other supplementation, and to follow a similar dietary and lifestyle pattern for the duration of the study. The experimental

interventions were performed at the Health Clinic of North-West University under controlled conditions. A medical doctor and nurse were present during the period of intervention, and all participants could leave the premises only after approval by the medical doctor. The protocol was approved by the Health Sciences Ethical Committee of North-West University (ethical approval number: NWU-00045-12-S1), conducted in accordance with guidelines for good clinical practice, and all participants provided informed written consent to the research protocol (an example of the informed consent form is included in Section 3.7.8 of the SI). The protocol was registered as a clinical trial on 3 November 2017 under the Pan African Clinical Trial Registry (registration number: PACTR201711002748255), under the title: A metabolomics investigation on experimental interventions of acute alcohol consumption.

The experiments were conducted on two Saturday mornings between 08:00 and 12:00. All participants were required to abstain from alcohol consumption for at least 48 hours preceding the experiment, and to abstain from breakfast on the days of the experiments (that is, to remain in an overnight fasted state).

On the first Saturday, half of the participants (randomly selected) were given 500 mL lemon-flavoured sparkling water as vehicle only (contents: fructose and citric acid flavouring; sodium benzoate preservative; sodium cyclamate, aspartame, acesulfame K sweeteners; vitamin C). The other half of the participants received the same quantity of the vehicle, as well as a predefined quantity of alcohol — 1.5 mL alcohol per kilogram body mass (Numminen *et al.*, 2000). The alcohol used was triple-distilled vodka: 43% alcohol. On the second Saturday, the participants received the alternate intervention to the one received on the first Saturday.

On both days the participants were also provided with 1.5 L bottled water, which was the only substance that could be consumed over the 4-hour period of the experiment. Differences in the concentrations of urinary metabolites due to variation in water consumption between the participants were accounted for by determining the creatinine concentration of each sample, and expressing the concentrations of all the quantified metabolites as $\mu\text{mol metabolite}/\text{mmol creatinine}$.

Initial (time 0) urine samples were collected just prior to the intervention. Subsequent urine samples were collected at 1, 2, 3 and 4 hours after the start of the experiment. This gave a total of five urine samples from each participant for each intervention. Time 0 urine samples served as controls for each participant, with longitudinal data being compared accordingly.

Sample handling

After collection, each urine sample was divided into aliquots and stored at -80°C . Once all the urine samples had been collected, a 1 mL aliquot of each sample was thawed and combined to prepare a pooled quality control (QC) sample. This QC sample was divided into aliquots and once again stored at -80°C . Another 1 mL aliquot of each urine sample was used for creatinine determination, performed by an external pathology service.

Metabolomics workflow

The workflow of the intervention study started with the generation of time-dependent quantitative metabolomics data, progressed to the application of various models of statistical analysis, which eventually led to the biological interpretation of the effect of the interventions on the group as well as on individual cases. The samples collected prior to alcohol (or vehicle) consumption were used as controls for the subsequent hourly samples collected after the interventions.

Figure 3-31 graphically depicts this workflow: (a) Longitudinal data were generated through the GC–MS analyses for 120 metabolites across five hourly intervals. (b) Two metabolic states — before alcohol consumption (time 0) and after alcohol consumption (times 1 to 4 hours) — were compared through multivariate, as well as univariate, analyses to establish the levels of significance of the observed differences. (c) Correlation analysis was used to indicate the relationship between important metabolites. (d) Unfolded PCA was applied to the data that emerged in time to summarize the subsequent variability, and, more importantly, between individuals, due to the alcohol effect, as well as for the identification of the metabolites responsible for this variability. (e) Construction of a global metabolite profile from the combined results provided the framework for the discussion on the effects of acute alcohol consumption on the subjects' metabolism.

The effect of the vehicle-only intervention was previously described (Irwin *et al.*, 2016), and will not be discussed here.

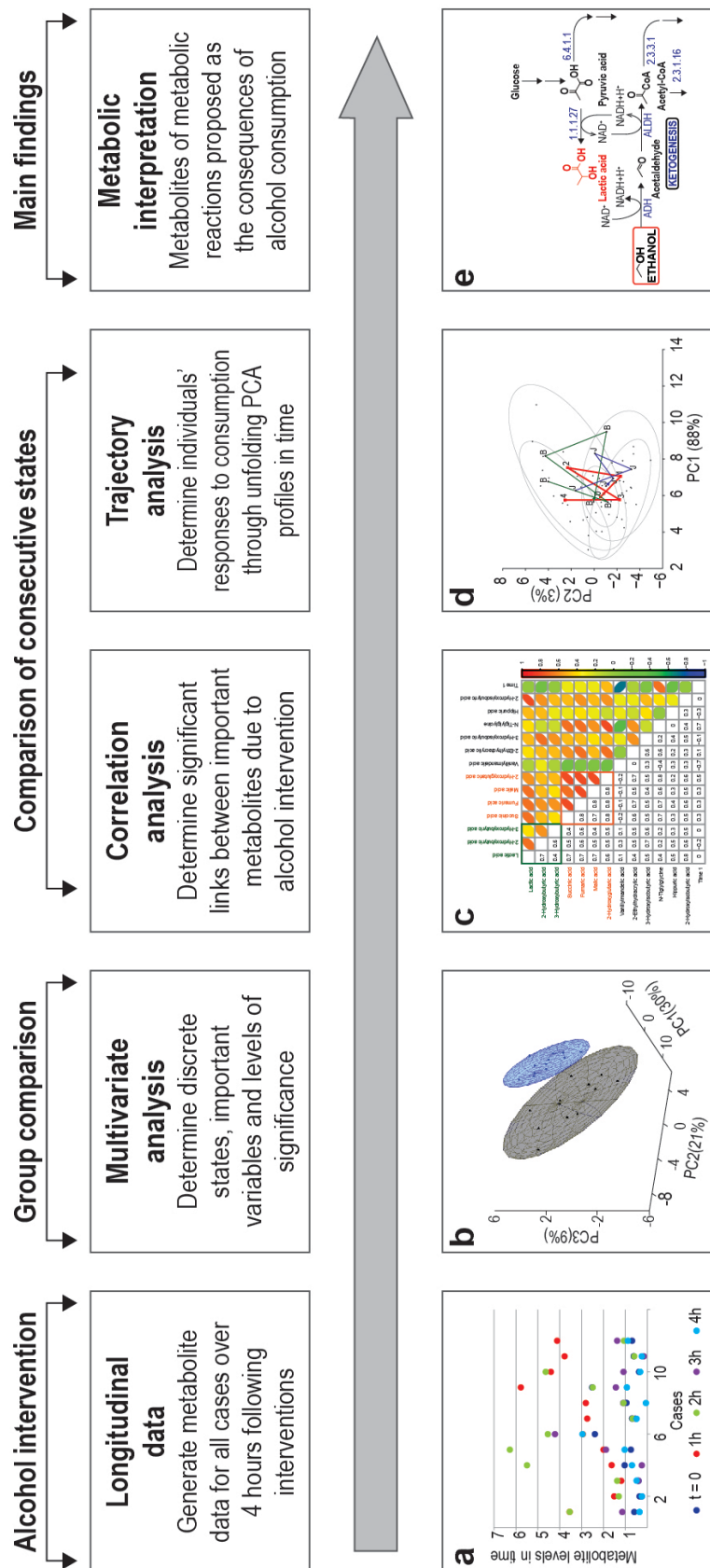


Figure 3-31: Representation of the metabolomics workflow to investigate the effect of acute alcohol consumption.

(a) Longitudinal data were generated through the GC-MS analyses for 120 metabolites across five hourly intervals. (b) Two metabolic states — before alcohol consumption (time 0) and after alcohol consumption (times 1 to 4 hours) — were compared through multivariate, as well as univariate, analyses to establish the levels of significance of the observed differences. (c) Correlation analysis was used to indicate the relationship between important metabolites. (d) Unfolded PCA was applied to the data that emerged in time to summarize the subsequent variability, and, more importantly, between individuals, due to the alcohol effect, as well as for the identification of the metabolites responsible for this variability. (e) Construction of a global metabolite profile from the combined results provided the framework for the discussion on the effects of acute alcohol consumption on the subjects' metabolism.

Measurement design

For the generation of data through GC–MS analysis, each participant's samples were analysed in a separate batch. Each batch included the participant's 10 urine samples (S), repeat samples (R) and quality control samples (Q), and each batch was constructed and analysed as follows:

QQQQQSS_VS_VS_VR_AQS_VS_VR_AQS_AS_AS_AR_VQS_AS_AR_VQQ

where: Q	= Quality control sample	[Total = 11]
S _V	= Randomly selected vehicle sample	[Total = 5]
S _A	= Randomly selected alcohol sample	[Total = 5]
R _A	= Randomly selected alcohol repeat sample	[Total = 2]
R _V	= Randomly selected vehicle repeat sample	[Total = 2].

QC samples were included to estimate any batch effect or other interfering analytical aspects. The data from the five QC samples at the start of each batch were used to condition the column and were excluded from further analysis. Repeat samples were included to determine their repeatability across the batch — the first two repeat samples in each batch were repeats of experimental samples from later in the batch, and the last two repeat samples were repeats of experimental samples from earlier in the batch. The total GC–MS running time for each batch was approximately 21 hours.

Organic acid extraction and GC–MS analysis

A 5 mL aliquot of each urine sample was used for organic acid extraction before GC–MS analysis, as described previously (Mason *et al.*, 2016). All experimental, QC and repeat samples were prepared and derivatized individually and in the same way. Full details of the method are presented in Section 3.7.1 of the SI. An excerpt of the data used for the alcohol intervention is shown in the Supplementary Information (Table 3-9). Metabolite concentrations are expressed as $\mu\text{mol}/\text{mmol}$ creatinine relative to an internal standard (4-phenylbutyric acid).

Quantitative urinary alcohol excretion

Urinary ethanol is not detectable by the GC–MS method used here due to its high volatility, but was quantified in all samples obtained from the vehicle and alcohol interventions by means of NMR spectroscopy. The data indicated that ethanol excretion peaked at 2 hours following alcohol consumption (data not shown).

Variable identification, classification and reduction

The QC samples were used to identify and classify a list of representative variables in all the urine samples. Following untargeted GC–MS data generation, a total of 172 variables were detected through AMDIS, excluding the 2 internal standards used. Several of the variables were present in concentrations just above the detection limit, and were not observed in all QC samples. The Human Metabolome Database (www.hmdb.ca) was used as the reference for the biological description of each feature, and as the basis for classifying the variables. Information on variables not included in the HMDB (e.g. exogenous substances or artifacts formed during the derivatization reactions) was obtained from other established chemical databases or from the literature; failure with this resulted in the classification of 7 variables as ‘no annotation’. Details of the 172 variables are summarized in the Supplementary Information (Table 3-10), and exclusion criteria for the determination of metabolite relevance are described in Section 3.7.3 of the SI. From this protocol, 120 metabolites were identified and used for the statistical evaluation of the effect of alcohol consumption.

3.6.8 Statistical Analysis

Identification of important organic acid metabolites

Metabolites causing the separations were regarded as important if they varied substantially between the samples collected prior to alcohol consumption and those collected one hour after alcohol consumption. Changes in metabolite levels were ranked based on their multivariate VIP values (Variable Importance in Projection), fold change (FC) values and non-parametric Wilcoxon signed-rank test (WRT) p -values. The selection criteria were: $VIP \geq 1.0$, $WRT\ p \leq 0.05$ and $|FC| \geq 1.5$. The aim with the selection was for a deeper understanding of the dominant biological changes rather than to model the observed data.

Multivariate statistical analysis

Two modes of multivariate statistical analyses were applied to the metabolomics data generated during this intervention experimental design: (1) cross-sectional analysis of time points using traditional multivariate methods to compare two groups (PCA and PLS-DA); and (2) longitudinal analysis performed across all times (ASCA, Smilde *et al.*, 2005; Zwanenburg *et al.*, 2011 and unfolded PCA, Villegz *et al.*, 2009). Details of these methods were previously described (Irwin *et al.*, 2016).

The statistical analyses indicated that we should achieve acceptable power (0.8) for a large effect size (0.9) given 12 paired observations and a 5% significance level. We should therefore be able to identify large differences between two factor levels (that is, between two points in time or two interventions) for a single variable (see SI Section 3.7.4 for details). Prior to statistical analysis, the data were pre-processed by: (i) treatment of zero-valued observation; and (ii) transformation and scaling. This is explained in greater detail in Section 3.7.5 of the SI. PCA was used to project the observed data to new spaces that maximizes the variation along fewer hyperplanes while not taking the group membership into consideration. PLS–DA was applied to build models to predict group membership, by projecting the variance in the observed data measured and the membership to new spaces. PLS–DA was used as a supervised method to rank and select the metabolites most changed by the intervention. The significance of these changes was established through univariate analysis using the Wilcoxon signed-rank test and fold change ratios. The figures shown in Figure 3-26 do not take into account the paired nature of the data. However, multi-level PCA and PLS–DA were also performed, which cater specifically for repeated measures. The multi-level results are closely related to those reported here, and are included in Section 3.7.6 of the SI.

3.6.9 Additional Information

Data availability

The full data set is given in Excel format as part of the SI (available online).

Acknowledgements

Research funding for the analytical and computational aspects of the project was provided by the Technological Innovation Agency (TIA) of the Department of Science and Technology of South Africa. The views and findings of this investigation are those of the authors and do not reflect or represent any policies of TIA. The funder had no role in designing the study, data collection and analysis, decision to publish, or preparation, writing or editing of the manuscript.

Author Contributions

Conceptualization: CJR JAW.

Data curation: CI SM MvR JAW.

Formal analysis: CI.

Funding acquisition: CJR.

Investigation: CI SM.

Methodology: CJR JAW.

Project administration: CJR JAW.

Software: JAW MvR.

Supervision: CJR JAW.

Writing – original draft: CI MvR SM LJM JAW CJR RAW.

Writing – review & editing: CI MvR SM LJM JAW CJR RAW.

Additional Information

Supplementary information accompanies this manuscript.

Competing Interests

The authors declare that they have no competing interests.

3.7 Supplementary Information: GC–MS-based urinary organic acid profiling reveals multiple dysregulated metabolic pathways following experimental acute alcohol consumption

3.7.1 Sample Preparation, Organic Acid Extraction, and GC–MS Analysis

Extraction and derivatization of organic acids

The 5 mL aliquots used for the analyses were thawed at room temperature and vortexed before use.

The volume of urine used for the organic acid analysis was based on the creatinine value of each of the samples:

- For creatinine values higher than 100 mg% use 0.5 ml urine
- For creatinine values less than 100 mg% but higher than 5 mg% use 0.5 ml urine
- For creatinine values less than 5 mg% but higher than 2 mg% use 2 ml urine
- For creatinine values less than 2 mg% use 3 ml urine.

The correct volume of urine was transferred to a 6 ml Kimax tube and 6 drops of 5M HCl was added to adjust the pH of the urine to below 2.

The volume of internal standard (IS) added was determined by multiplying the creatinine value (in mg %) with five times the volume of urine used. A mix of malonic acid (RT ~14 minutes) and 4-phenylbutyric acid (RT ~20 minutes) was used as internal standard. The IS was prepared by dissolving 52.5 mg of each in 50 ml milli-Q water. The solution was sonicated for 30 minutes to ensure that all compounds were properly dissolved. 4-Phenylbutyric acid was used owing to its absence in normal urine and in known pathological conditions. In addition, it elutes almost in the middle of the organic acid profile and theoretically co-elutes with very few, if any, other organic acids. Malonic acid was added as a second IS reference peak in the GC chromatogram.

The first extraction was done by adding 6 ml of ethylacetate to each of the samples. The tubes were capped and checked for leakage by inverting the tubes before being mixed on a rotary wheel for 30 minutes. The samples were then centrifuged at 3000 rpm for 3 minutes and the organic phase (top phase) was aspirated into a clean 6 ml Kimax tube.

The second extraction was done by adding 3 ml of diethylether to the aqueous phase (lower phase) and the tubes were once again capped and checked for leakage. The samples were then mixed on the rotary wheel for 10 minutes and centrifuged at 3000 rpm for 3 minutes. The organic phase was aspirated and added to the ethylacetate phase and the aqueous phase was discarded.

Approximately 2 ml of anhydrous sodium sulphate (Na_2SO_4) was added to each sample and the tubes were capped and vortexed to mix. Proper dispersion of the Na_2SO_4 ensures that all the water is removed from the organic phase. The samples were then centrifuged at 3000 rpm for 1 minute and the organic phase was decanted into a clean 3 ml Kimax tube. The samples were finally evaporated to dryness in a heating block at 37°C under nitrogen gas for approximately 45 minutes.

Using a Hamilton syringe, the derivatization reagents O-bis(trimethylsilyl)trifluoroacetamide (BSTFA), trimethylchlorosilane (TMCS) and pyridine were added to the dried samples. The volume of BSTFA added was determined by multiplying the creatinine value (in mg %) with three times the volume of urine used. The volume of TMCS and pyridine added was determined by multiplying the creatinine value (in mg %) with 0.6 times the volume of urine used. The tubes were capped and incubated at 70°C for 45 minutes.

The derivatized samples were then transferred to 1.5 ml GC–MS vials with inserts, capped, and placed in the GC–MS autosampler for GC–MS analysis. Blank (hexane), QC and repeat samples were included in the injection sequence, as described in the main text.

The internal standards 4-phenylbutyric acid and malonic acid as well as the anhydrous Na_2SO_4 were purchased from BDH. Ethylacetate and diethylether were obtained from Merck Chemicals. The derivatization reagents O-bis(trimethylsilyl)trifluoroacetamide (BSTFA), trimethylchlorosilane (TMCS) and pyridine were purchased from Sigma Chemical Company.

GC–MS analysis

An Agilent GC–MS system was used in this study and consisted of a model 7890A gas chromatograph, a model 5975C mass selective detector, a Hewlett Packard 5970C quadrupole mass spectrometer and Agilent Chemstation. GC separation was achieved on a WCOT fused silica capillary column [30 m x 0.32 mm (i.d.)] coated with SE30 CB, film thickness 0.30 μm (Machery Nagel).

Samples were introduced into the column via a splitless injector at 280°C. The initial oven temperature was kept at 60°C for 2 minutes and then programmed to rise 6°C/min to a final temperature of 280°C. This temperature was maintained for 5 minutes. Helium was used as carrier gas at a pressure of 60 kPa and at a constant flow rate of 1 ml/min. The column was inserted directly into the ion source with an interface temperature of 280°C. The mass spectra of all GC peaks were generated by a mass spectrometer operated in the electron impact (EI) mode, with electron energy of 70 eV. The MS source and quadrupole temperatures were 250°C and 150°C respectively.

Qualitative analysis of metabolites in QC samples (metabolite identification)

After GC separation, each peak at an indicated RT was analysed in EI mode in order to investigate the fragmentation pattern of each compound. The m/z values selected for feature identification were either the base peak ion or one of the more abundant characteristic fragment ions.

Deconvolution and data analyses were conducted using AMDIS software (Version 2.71) linked to NIST Mass Spectral Search Program for the NIST/EPA/NIH Mass Spectral Library (Version 2.0F, built Oct. 8, 2008). Where authentic standards were available, their respective response factors were used, and for those compounds where no authentic standards were available, a response factor of 1 was assumed. The analytical setting of the AMDIS software was as follows: minimum factor – 60%, threshold – ‘off’, Scan Direction – ‘high to low’, and type of analysis – ‘Use of an internal standard for RI’. The deconvolution settings were: component width – 12, adjacent peak subtraction – 1, resolution – medium, sensitivity – low, and shape requirements – low. The first hit of identified compounds and integrated area of the peaks were exported to Microsoft Excel.

3.7.2 Alcohol Intervention Study Data

Table 3-9 shows an excerpt of the data used for the alcohol intervention study (the complete file is available online as part of the supplementary material).

Batch	Time	1	2	3	4	5	6	7	8	9	...	119	120
		Glyoxylic-1,2-Dihyd	1,2-Dihyd	Lactic-acid	Hexanoic-	2-Hydroxy	Glycolic-a	Oxalic-acid	2-Hydroxy			Octadecar	Eicosanoic
1	0	0.181	0.825	6.424	8.188	0.198	8.959	7.566	2.127	3.362		1.918	0.030
1	1	0.004	0.863	0.586	6.011	0.194	7.354	3.867	1.043	1.983		1.086	0.035
1	2	0.000	3.348	23.576	73.513	1.017	37.804	11.328	1.472	5.640		0.125	0.014
1	3	0.004	1.715	2.798	14.757	0.219	11.194	1.895	1.719	2.231		5.534	0.057
1	4	0.004	1.544	3.411	16.042	0.525	12.337	3.209	0.507	1.175		3.472	0.012
2	0	0.380	6.090	3.299	10.635	0.213	13.432	5.542	1.656	2.546		1.380	0.006
2	1	0.577	6.087	5.495	31.238	0.207	51.693	12.475	3.070	8.968		7.982	0.015
2	2	2.431	10.583	6.763	19.799	0.040	24.936	3.507	4.564	4.293		16.834	0.064
2	3	0.165	6.804	2.695	8.701	0.031	13.519	1.426	2.046	2.385		3.687	0.016
2	4	0.392	5.628	0.717	11.225	0.016	22.899	4.268	1.535	3.195		2.861	0.018
3	0	0.057	6.082	1.117	5.128	0.017	7.560	2.057	0.249	1.832		0.587	0.002
3	1	0.263	6.834	1.312	6.123	0.005	11.984	2.682	0.708	2.046		0.579	0.008
3	2	0.074	7.555	3.482	11.516	0.041	10.888	1.091	0.437	0.051		4.187	0.009
3	3	0.009	7.164	0.762	2.578	0.011	4.549	0.295	0.566	0.173		1.701	0.000
3	4	0.142	7.788	2.470	7.058	0.023	10.243	1.352	2.480	0.027		5.931	0.012
4	0	1.260	5.450	14.115	13.301	0.008	19.194	3.939	1.727	3.589		1.668	0.004
4	1	0.786	6.998	4.797	7.539	0.005	11.048	1.319	0.234	0.600		2.648	0.019
4	2	5.384	8.793	7.255	22.786	0.074	34.901	1.477	1.401	2.113		20.830	1.148
4	3	0.452	6.635	2.458	3.065	0.002	6.041	0.249	0.109	0.308		1.754	0.022
4	4	10.047	5.740	14.662	20.996	0.176	32.883	3.912	0.401	2.180		5.673	0.038
5	0	0.190	5.314	1.160	13.823	0.059	18.080	6.760	1.977	3.394		1.605	0.040
5	1	0.348	5.662	2.762	16.770	0.036	20.638	8.213	2.311	4.511		3.696	0.830
5	2	0.004	13.903	16.727	58.947	0.143	44.790	9.347	11.309	9.393		20.223	0.187
5	3	0.041	9.541	9.511	52.548	0.150	29.792	5.183	1.630	3.260		21.992	0.056
5	4	3.452	6.023	1.227	12.596	0.041	19.668	4.064	1.310	1.860		5.391	0.045

Table 3-9: An excerpt of the data used for the alcohol intervention study.

3.7.3 Identified Features

Supplementary Table 3-10 gives the details of each of the 172 identified and classified variables. A reference value for the normal concentration range in urine for adults of each identified variable is also given.

Tentative variable name based on AMDIS and the library used	RT (min)	Variable no.	Index reference	Reference range (μmol/mmol Cr)	Excluded variable and criteria
Glyoxylic acid	7.769	1	HMDB00119	1.27 (0.46–3.26)	
2-Undecenoic acid	8.029		CS 21171839	n/r	Not a recognized biological substance
1,2-Dihydroxyethane	8.693	2	KEGG C01380	n/r	
1,2-Dihydroxypropane	9.211	3	ChEBI 16997	n/r	
Boric acid	9.24		HMDB35731	n/r	Not a recognized biological substance
Phenol	10.158		HMDB00228	4.8 (0.6–12.8)	Not a recognized biological substance
O-acetylphenol	10.323		HMDB32568	n/r	Not a recognized biological substance
Lactic acid	10.652	4	HMDB00190	3.9–9.8	
Hexanoic acid	10.75	5	HMDB00535	n/q	
2-Hydroxyisobutyric acid	10.881	6	HMDB00729	4.4–7.6	
Glycolic acid	1.952	7	HMDB00115	64.0 (21.0–107.0)	
Oxalic acid	12.229	8	HMDB02329	27.00 (0.00–54.00)	
2-Hydroxybutyric acid	12.549	9	HMDB00008	2.8 (1.2–6.9)	
4-Ketovaleric acid	12.603	10	HMDB00720	1.3 (0.4–2.0)	
4-Cresol	12.834		HMDB01858	46.0 (1.2–118.9)	Questionable biological substance/origin
3-Hydroxypropionic acid	12.882	11	HMDB00700	8.1 (3.1–11.8)	
2-Methyl-2-hydroxybutyric acid	12.921	12	HMDB01987	0.8 (0.4–1.5)	
Sulphate	12.954		HMDB01448	2407.89 (2171.05–2697.36)	Not a recognized biological substance
2-Hexenoic acid	13.057	13	HMDB10719	n/r	
3-Hydroxybutyric acid	13.421	14	HMDB00357	1.4–2.2	
3-Hydroxyisobutyric acid	13.442	15	HMDB00023	11 (4.1–19)	
2-Hydroxyisovaleric acid	13.665	16	HMDB00407	0.23–0.42	
N-Crotonylglycine	13.82	17	CAS71428-89-2	n/r	Glycine conjugation product of crotonic acid
Malonic acid	14.402	IS-2			
2-Methyl-3-hydroxybutyric acid	14.518	18	HMDB00354	4.2 (1.6–6.7)	
Methylmalonic acid	14.74	19	HMDB00202	1.8 (0.00–3.6)	
3-Hydroxyisovaleric acid	14.802	20	HMDB00754	11 (6.9–25)	
Benzoic acid	15.208		HMDB01870	4.2 (1.9–6.5)	Preservative of the vehicle
2-Ethyl-3-hydroxypropionic acid (2-ethylhydracrylic acid)	15.277	21	HMDB00396	2.1 (1.3–2.9)	
3-Hydroxyvaleric acid	15.462	22	HMDB00531	1 (0–2)	
Acetoacetic acid	15.532	23	HMDB00060	0.15 (0.01–0.58)	
2-Methyl-3-ketobutyric acid	15.601	24	HMDB03771	1.0 (0.0–2.0)	
4-Pyridinecarboxylic acid	15.741		HMDB60665	n/r	Derivatization artifact
Octanoic acid	15.899	25	HMDB00482	n/q	
2,2-Dihydroxyacetic acid	15.929		CS 2005843	n/r	Not a recognized biological substance
2-Ketoisovaleric acid	15.6	26	HMDB00019	0.13 (0.00–0.54)	
N-Acetyl glycine	15.66	27	HMDB00532	n/r	

Tentative variable name based on AMDIS and the library used	RT (min)	Variable no.	Index reference	Reference range (μmol/mmol Cr)	Excluded variable and criteria
Phosphoric acid	16.288	28	HMDB02142	784 (425–1170)	
Octanoic acid	15.899	29	HMDB00482	n/q	
3-Ketovaleric acid	16.198	30	CS 388751	n/r	Can be of endogenous origin in certain metabolic disorders
Phenylacetic acid	16.37	31	HMDB00209	0.3 / 1.9	
Ethylmalonic acid	16.379	32	HMDB00622	1.16–2.99	
Glycerol	16.687	33	HMDB00131	13 (4–19)	
2-Octenoic acid	16.954	34	HMDB00392	n/r	
Succinic acid	17.024	35	HMDB00254	9.9 +/- 5.0	
2-Hydroxyglutarylactone	17.141		n/a	n/r	From 2-hydroxyglutaric acid
3-Hydroxycaproic acid	17.169	36	HMDB02203	n/r	
1,2-Dihydroxybenzene	17.263	37	HMDB00957	4.06 +/- 1.80	
Methylsuccinic acid	17.382	38	HMDB01844	2.3 (0.8–10.8)	
Pyrimidinedione	17.48		n/a	n/a	Derivatization artifact
Uracil	17.555	39	HMDB00300	9.5 (2.6–22.8)	
5-Hydroxyvaleric acid	17.52	40	HMDB61927	n/r	
Citraconic acid	17.81		HMDB00634	1.6 (0.9–2.1)	Not a recognized biological substance
Glyceric acid	17.815	41	HMDB00139	1.7 (0.2–6.0)	
Glutaconic acid	17.839	42	HMDB00620	3.1 (1.2–3.1)	
Fumaric acid	17.889	43	HMDB00134	0.75–1.2	
Nonanoic acid	18.115	44	HMDB00847	n/r	
5-Hydroxyhexanoic acid	18.224	45	HMDB00525	2.7 (0.8–5.7)	
N-Isobutyrylglycine	18.297	46	HMDB00730	n/q	
2,3-Dihydroxybutanoic acid	18.391	47	HMDB00498	5.0 +/- 3.0	
Lactyllactate	18.762		CS 91775	n/r	From lactic acid
1,3-Dihydroxybenzene	18.62		HMDB32037	n/r	Questionable biological substance
Hydroxymalonic acid	18.913		HMDB35227	n/r	Derivative from IS-2
Glutaric acid	18.975	48	HMDB00661	1.3 (0.6–2.6)	
Phenoxyacetic acid	19.083		HMDB31609	n/r	Questionable biological substance
Benzamide	19.153		HMDB04461	n/r	Not a recognized biological substance
2,3,4-Trihydroxybutyric acid-lactone	19.32		n/a	n/r	From 2,3,4-trihydroxybutyric acid
N-Butyrylglycine	19.46	49	HMDB00808	0.24 +/- 0.36 [infant]	
3-Methylglutaric acid	19.442	50	HMDB00752	1.0–6.5	
N-Isobutyrylglycine(2)	19.46		HMDB00730	n/q	Second peak (see no. 58)
Malonic acid-triTMS	19.602		Agilent G1676AA	n/r	Derivative of IS-2
4-phenylbutyric acid	19.636	IS-1			
3-Methylglutaconic acid	19.698	51	HMDB00522	6.2 (2.8–8.3)	
3-Hydroxyadipylactone	19.748	52	n/a	n/r	From adipic acid (see no. 86)
N-Acetylleucine	19.768	53	HMDB11756	n/r	
3,4-Dihydroxybutyric acid	19.952		HMDB06118	34.9 (12.5–57.2)	Questionable biological substance

Tentative variable name based on AMDIS and the library used	RT (min)	Variable no.	Index reference	Reference range (μmol/mmol Cr)	Excluded variable and criteria
Decanoic acid	20.1	54	HMDB00511	n/q	
N-Isovalerylglycine	20.22	55	HMDB00678	2.0 (0.4–4.0)	
Parabanic acid	20.308		CID 67126	n/r	Medication/Artifact
2,2-Dihydroxymalonic acid	20.473		CS 61695	n/r	Derivative from IS-2
3-Ethylglutaric acid	20.54	56	n/a	n/a	Potential ethanol derivative
5-Methyldihydropyrimidine-2,4(1H,3H)-dione	20.572		CS 84456	n/r	Not a recognized biological substance
3,5-Dihydroxy-3-methylvaleric acid	20.697	57	HMDB00227	0.14 (0.06–0.22)	
N-Isovalerylglycine(2)	20.823		HMDB00678	2.0 (0.4–4.0)	Second peak (see no. 76)
Pyroglutamic acid	21.128	58	HMDB00267	14.0 +/- 7.2	
3-Methylglutaconic acid	20.626	59	HMDB00522	6.2 (2.8–8.3)	Second peak (see no. 71)
Citramalic acid	20.717		HMDB00426	2.4 (1.0–4.8) [NMR]	Questionable biological substance
Adipic acid	20.94	60	HMDB00448	5.1 (0.8–35)	Can be of endogenous origin in certain metabolic disorders
Malic acid	20.942	61	HMDB00156	2.0 (0.7–5.3)	
2-Deoxy-3,5-dihydroxypentonic acid-g-lactone	20.971	62	n/a	n/a	
2-Piperidinecarboxylic acid	21.14		HMDB00070	0.28 (0.0–0.56)	Questionable biological origin
4-Hydroxycyclohexylcarboxylic acid	21.243		HMDB01988	6.1 (1.1–28.1) [Adolescent]	Questionable biological substance
5-Hydroxyhydantoin	21.323		CID 4157426	n/r	Derivatization artifact
N-Acetylthreonine	21.356	63	CS 3841173	n/r	
3-Methyladipic acid	21.518	64	HMDB00555	2.9 (0.7–10.5)	
5-(Hydroxymethyl)furan-2-carboxylic acid	21.651		HMDB02432	1.7	Questionable biological substance
2,4-Hexadienedioic acid	21.65		HMDB29581	n/r	Questionable biological substance
2-Hydroxyphenylacetic acid	22.081	65	HMDB00669	2.0 (0.9–4.5)	
N-Tiglylglycine	21.736	66	HMDB00959	0.78–1.2	
Threitol	21.744	67	HMDB04136	19.3 (5.3–32.7)	
2-Hydroxy-2-methylglutaric acid	21.901	68	Yancey <i>et al.</i> , 1986	n/r	
2-Hydroxyphenylacetic acid	22.081	69	HMDB00669	2.0 (0.9–4.5)	
N-3-Methylcrotonylglycine	22.128	70	HMDB00459	1.0 (0.0–2.0)	
2,3,4-Trihydroxybutyric acid	22.275		HMDB00613	20.0 +/- 8.0	Questionable biological substance
N-2-Pentenoylglycine	22.358	71	n/a	n/a	
2-Hydroxyglutaric acid	22.453	72	HMDB00694	0.8–52	
3-Hydroxyglutaric acid	22.49	73	HMDB00428	<11.51 [children]	
2,3,4-Trihydroxybutyric acid(2)	22.52	74	HMDB00613	20.0 +/- 8.0	Second peak (see no. 102)
2,4,6-Trihydroxypyrimidine	22.665		HMDB41833	n/r	Derivatization artifact
Pimelic acid	22.678	75	HMDB00857	2.2 (0.7–4.0)	
Hexahydropyrimidine-2,4,5-trione	22.696		n/a	n/a	Not a recognized biological substance
3-Ketoglutaric acid	22.841	76	HMDB13701	0–0.11	

Tentative variable name based on AMDIS and the library used	RT (min)	Variable no.	Index reference	Reference range (μmol/mmol Cr)	Excluded variable and criteria
N-Hexanoylglycine	22.869	77	HMDB00701	1.00 (0.0–2.0)	
2-Ketoglutaric acid	22.949	78	HMDB00208	2.87 (0.18–14.3)	
3-Hydroxyphenylacetic acid	22.753	79	HMDB00440	0.6 (0.4–0.9)	
3-Hydroxy-3-methylglutaric acid	22.992	80	HMDB00355	3.2 (1.1–5.2)	
4-Hydroxycyclohexylacetic acid	23.047		HMDB00909 (trans)	n/r	Not a recognized biological substance. Alternative reference: CS 13628091 (cis)
4-Hydroxybenzoic acid	23.143	81	HMDB00500	1.8 (0.7–29)	
4-Hydroxyphenylacetic acid	23.249	82	HMDB00020	6.0 (2.4–9.7)	
2,5-Furandicarboxylic acid	23.255		HMDB04812	1.9 (0.1–5.4) [Adolescent]	Derivatives found in food prepared by strong heating. Not a recognized biological substance
N-Acetylaspartic acid	22.18	83	HMDB00812	4.66 +/- 1.14	
2,3,4,5-Tetrahydroxypentanoic acid-1,4-lactone	23.496	84	CAS 179091-67-9	n/r	From ascorbic acid (see no. 151)
Dodecanoic acid	23.807	85	HMDB00638	0.03 +/- 0.02	
2-Furanylcarbonylglycine	23.639		HMDB00439	9.95 (2.00–18.66)	Not a recognized biological substance (see no. 118)
N-Cyclohexylsulfamic acid	23.849		HMDB31340	n/r	Not a recognized biological substance
N-Acetylanthranilic acid	23.908	86	CID 6971	n/r	
Octenedioic acid	23.897	87	HMDB00341	n/r	
Isocitric acid-lactone	24.048		CID 98259	n/r	From isocitric acid (see no. 145)
Suberic acid	24.255	88	HMDB00837	0.5 (0.0–2.9)	
Erythro-pentonic acid	24.561	89	n/a	n/r	Possible CAS 74742-30-6
Acetyl-4-phenol	24.595		CS 7189	n/r	Not a recognized biological substance
Tricarballic acid	25.195		HMDB31193	n/r	Derivatization artifact
Citric acid-diethylester	25.195	90	CAS 19958-02-2	n/r	
Aconitic acid	25.195	91	HMDB00461	13 (2.7–44)	
Vanillic acid	25.283	92	HMDB00484	1.0 (0.0–2.5)	
Homovanillic acid	25.347	93	HMDB00118	5.6 (2.1–47.3)	
Pyrrole-2-carboxylic acid	25.551	94	HMDB03094	0.03 +/- 0.02	
2,5-Dihydroxybenzoic acid	25.686	95	HMDB00152	1.35 +/- 0.79	
4-Hydroxymandelic acid	25.737	96	HMDB00822	0.98–1.5	
Azelaic acid	25.819		HMDB00784	4.8 (1.3–15)	Questionable biological substance
Hippuric acid-diTMS	26.055		CS 454967	n/r	Derivative of hippuric acid
Citric acid-monoethylester	25.923	97	CID 57345948	n/r	
Hippuric acid-TMS	26.058	98	HMDB00714	27.92–932.66	
3,5-Dihydroxybenzoic acid	26.281	99	HMDB13677	0.218 +/- 0.124	
3,4-Dihydroxybenzoic acid	26.34		HMDB01856	0.048 +/- 0.008	Questionable biological substance
3,4-Dihydroxyphenylacetic acid	26.454	100	HMDB01336	0.9 (0.6–1.3)	
Isocitric acid	26.552	101	HMDB00193	56.8 (19.4–119.1)	

Tentative variable name based on AMDIS and the library used	RT (min)	Variable no.	Index reference	Reference range (μmol/mmol Cr)	Excluded variable and criteria
Citric acid	26.584	102	HMDB00094	172.8 +/- 87.4	
Tetradecanoic acid	26.65	103	HMDB00806	0.06 +/- 0.03	
3-Hydroxyphenylhydracrylic acid	26.844		HMDB02643	5.9 (1.4–22.1)	Questionable biological substance
Methylcitric acid	27.008	104	HMDB03610	0.59 (0.05–1.15)	
Vanillylmandelic acid	27.168	105	HMDB00291	1.1–1.7	
Ascorbic acid	27.303		HMDB00044	32.5 (4.6–78)	Vehicle additive
Hydantoinpropionic acid	27.255		HMDB01212	n/r	Questionable biological substance
4-Hydroxyphenyllactic acid	27.571	106	HMDB00755	1.1 (0.2–2.6)	
1H-Indole-3-acetic acid	28.028	107	HMDB00197	1.5–2.6	
3,4-Dihydroxyphenylpropionic acid	28.131		HMDB00423	0.27 +/- 0.079	Questionable biological substance
3-(4-Hydroxy-2,5-dioxoimidazolidin-4-yl) propanoic acid	28.155		n/a	n/a	Not a recognized biological substance
3-Methoxy-4-hydroxyphenylhydracrylic acid	28.761		Smith, 1969	n/r	Not a recognized biological substance
2,4,6-Trihydroxybenzoic acid	28.771		CID 66520	n/r	Not a recognized biological substance
Palmitic acid	29.416	108	HMDB00220	11 (6.0–23)	
2-Hydroxyhippuric acid	29.501	109	HMDB00840	0.5	
Glucuronic acid	29.734	110	HMDB00127	9.7 (3.7–20.6)	
4-Methoxy-3-hydroxycinnamic acid (Isoferulic acid)	29.886	111	HMDB00955	0.394 +/- 0.075	
N-Acetyltyrosine	29.886	112	HMDB00866	<10 [children]	
4-Hydroxyphenylpyruvic acid	29.738	113	HMDB00707	1.65 (0.15–8.74)	
Heptadecanoic acid	30.693	114	HMDB02259	n/r	
N-Cinnamoylglycine	30.723	115	HMDB11621	n/r	
4-Hydroxyhippuric acid	31.15	116	HMDB13678	0–14	
Phenylacetylglutamine	30.71	117	HMDB06344	47.03 (3.84–85.51)	
4-Hydroxyhippuric acid(2)	31.423		HMDB13678	0–14	Second peak (see no. 167)
Oleic acid	31.583	118	HMDB00207	5.2 (0.3–13)	
Octadecanoic acid	31.904	119	HMDB00827	2.9 (1.6–6.6)	
Eicosanoic Acid	33.85	120	HMDB02212	n/r	

Table 3-10: Identification, classification and reference ranges of 172 identified variables.

The original list of 172 variables contained substances that were excluded from the metabolomics analysis to assess the influence of acute alcohol consumption. Only the 120 included substances are numbered in column 4 of the table. The reasons for the exclusion of the other 52 substances are given in the final column of the table.

Categories for exclusion of variables were:

- Exogenous contaminants that are not recognized biological substances.
- Substances of a questionable biological function or origin, based on literature assessments.
- Artifacts formed from chemicals (e.g. urea) present in the urine.
- Artifacts due to the reaction conditions (e.g. formation of lactones, used for correction of the parent substance).
- Artifacts due to formation of an additional TMS derivative (used for correction of the parent substance).
- Substances showing multiple peaks in the GC-profile (used for correction to only one substance).

3.7.4 Statistical Power

Only a limited number of subjects could be included in an intervention study of this nature, i.e. exposure to a substance (alcohol) with known health risks. It was therefore necessary to ensure that the small sample size would not prevent the quantification of the effect of the intervention. A power calculation was performed assuming the following:

- The effect of the intervention would be large, with an expected effect size of at least 0.8 when calculated based on Cohen's d--value (Cohen, 1988). This was considered a valid assumption given the design of the study, i.e. consuming a relatively large volume of alcohol on an empty stomach.
- The dependence between samples from the same individual would be taken into account by using the dependent samples, or paired t-test, to assess the significance of the effect of the intervention. Though this method is a parametric univariate method comparing only two groups, it was deemed appropriate as we could transform the data to achieve normality, we did not know beforehand how many metabolites would be extracted, and the most dramatic effect of the intervention was expected to be between time 0 (before consumption of the alcohol) and 1 hour after consumption based on a pilot study.
- A significance level of 5% for a two-tailed test and a sample size of 12. We did not want to speculate on the metabolites that would be extracted and hence could not speculate on the direction of the expected change. The two-tailed hypothesis was therefore selected.

The power under these conditions is plotted against the effect size in Figure 3-32 (assuming a two-sided (a) and one-sided (b) alternative hypothesis, from which it is evident that power values above 0.7 are achieved for effect sizes of 0.8 or higher. This was acceptable in the current context. The power curve given a one-sided hypothesis is also included for reference purposes to show the increased power.

3.7.5 Data Pre-processing

The collection and analysis of biological samples culminated in 12 subjects observed across 5 time points and measured over 120 metabolites. A 50% zero filter was applied, taking time into account — that is, variables were excluded if they contained more than 50% zero-values in all 5 times. Two variables were removed based on these criteria. The remaining zero-values were imputed for each variable by randomly generating numbers from the tail of beta-distribution fitted to the non-zero observations. The resulting random numbers were all smaller than half of the minimum observed value. The data were then log transformed and centred prior to further statistical analysis.

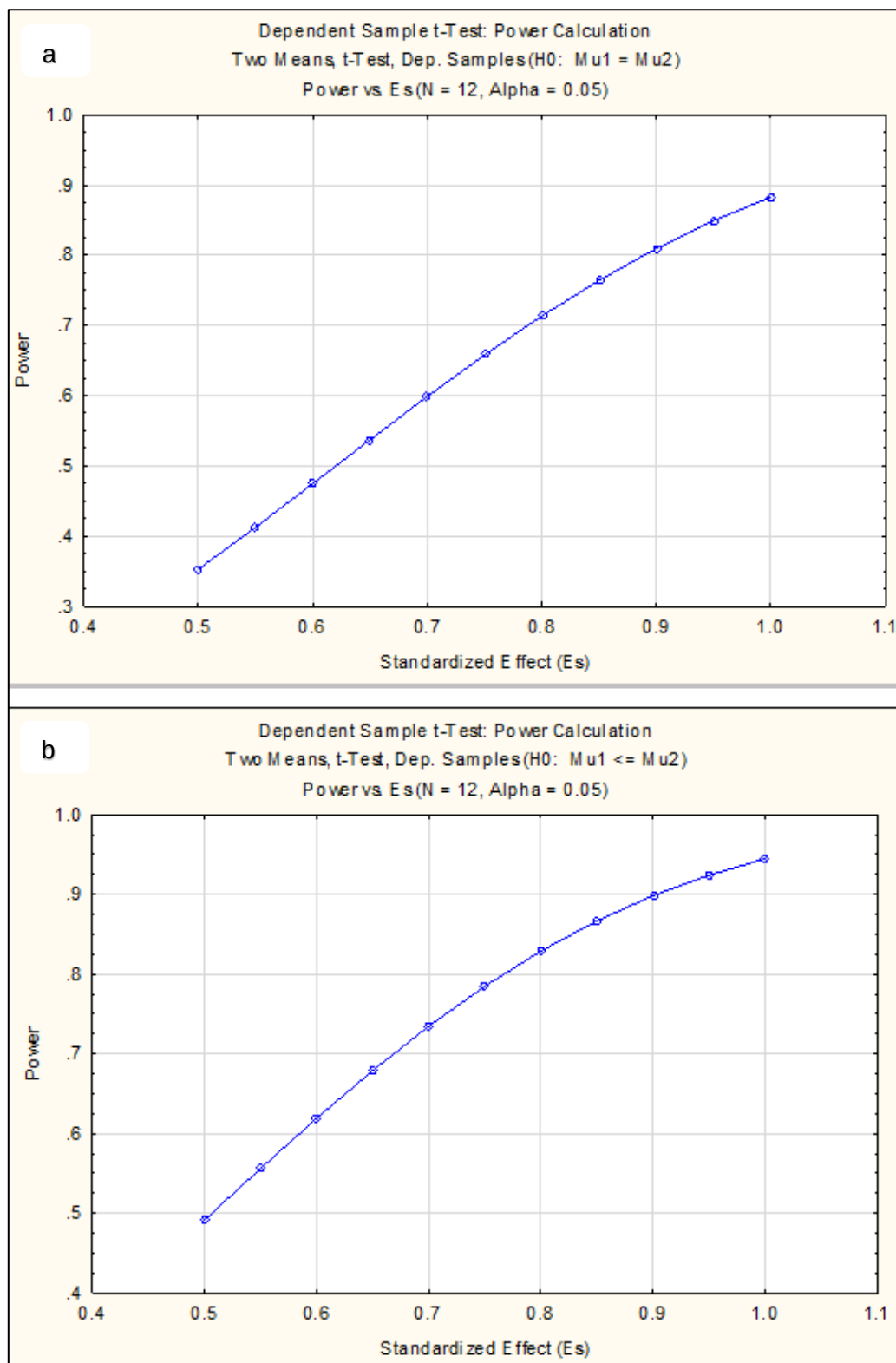


Figure 3-32: Power curves.

The graphs plot power (y-axis) against the effect size (x-axis) assuming a two-sided (a) and one-sided (b) alternative hypothesis

3.7.6 Statistical Methods & Additional Results

It was decided to deconstruct the three-dimensional data tensor into less complex cross-sections in order to select a subset of statistically important metabolites to aid biological interpretation of the observed perturbations. The assumption was made that the most biologically acute effect of the intervention would peak one hour after alcohol consumption. The cross-section between time 0 (before consumption of the alcohol) and one hour after consumption was therefore analysed using univariate and multivariate statistical methods.

Univariate analysis of paired data

The univariate nonparametric Wilcoxon signed-rank test was used to assign significance levels (p -values) to changes in metabolite levels from time 0 to time 1. The dependent t-test was also applied, after data transformation to ensure normality, but not used as current research revealed that the Wilcoxon signed-rank test to be a suitable alternative with marginally more power for a non-normally distributed data set of smaller sample size (Imam *et al.*, 2014).

Multivariate analysis of unpaired data

Multivariate partial least squares-discriminant analysis (PLS-DA) is a method that constructs a regression model to predict group membership by projecting variance in the metabolites measured, as well as in the observed group membership, to new spaces. PLS-DA models can easily overfit and produce models with inflated predictive ability unless extensively validated through test data and permutation testing. For this reason, the PLS-DA results are used cautiously as the small sample size did not allow for proper validation. The results used are limited to the VIP values, which are produced by the PLS-DA model for each metabolite as an indication of its predictive strength. Given our concern here, it is again important to emphasize that the aim of this selection was not to model the observed data, but rather to gain a deeper understanding of the predominant biological changes occurring due to the intervention. Ranking and selecting metabolites according to VIP values enabled us to identify metabolite combinations which dominated the observed change in metabolomic state. Ultimately, we revert back to the raw data (means and standard deviations) to interpret the resulting list against established metabolic pathways, and including proposals for extension or modification of them, resulting in a representation of a global metabolite profile reflecting the metabolic consequences of the alcohol consumption.

PCA is similar to PLS–DA in that observed data are also projected, but to new spaces that maximize variation along fewer hyperplanes, not taking the group membership into consideration.

Variable selection

A subset of metabolites was subsequently identified to gain a deeper understanding of the predominant biological changes occurring due to the intervention. The selection was not made for the purpose of modelling the effect of the intervention in time nor to predict group membership. Changes in metabolites levels were then ranked based on their multivariate VIP values. The significance of changes in high-ranking variables was then established through fold change (FC) values and non-parametric Wilcoxon signed-rank test (WRT) p -values. The selection criteria were then as follows: $VIP \geq 1.0$, $WRT\ p \leq 0.05$ and $|FC| \geq 1.5$.

Correlation analysis

Associations between metabolites shortlisted for biological interpretation were assessed using Spearman's rho or rank order correlation coefficient. This method is nonparametric as this takes the ranks of the data as inputs. The correlation coefficients (r -values) produced are seen as biologically relevant if $|r| \geq 0.5$ (Cohen, 1988).

Multivariate analysis of time-dependent data

The data tensor evaluated here has a very specifically designed structure with the same individuals being measured over and over, hence a repeated measures design over 5 points in time. Although the Wilcoxon signed-rank test accounts for groups of data not being independent but paired, standard PCA and PLS–DA applications do not. To understand the importance of this information for these models, a multi-level PCA and PLS–DA on the within-subject variation was also performed.

Supplementary Figure 3-33 provides the scores plots corresponding to the cross-sections of the data used in the main text, i.e. data from the 120 quantified metabolites for time 0 vs time 1 (a & e), time 0 vs time 2 (b & f), time 0 vs time 3 (c & g), and time 0 vs time 4 (d & h) following alcohol intake. The samples for time 0 were collected just prior to the interventions and are therefore regarded as the control samples.

The PCA scores plots are comparable with those included in the main text, showing some differentiation one hour after the alcohol consumption (Figure 3-33a), followed by complete

separation after 2 and 3 hours. In contrast, the separation is retained after 4 hours in the multi-level PC model. That said, a slight yet progressive return after 3 and 4 hours to the profile at time 0 is still evident when considering the decreasing variance explained by the first principal component. The PLS–DA plots again showed a complete separation for all four times following alcohol consumption relative to time 0.

Finally, the entire data tensor was modelled using PCA. This was achieved by unfolding the data along the time dimension (Blakemore & Choudhury, 2006). This approach allows for the projection of data to fewer dimensions, while retaining the maximum amount of variation and understanding of the changes observed in time. This change can now be observed and compared over all individuals, by averaging scores, or for each individual to allow for comparison. In addition, the contribution of each metabolite to the projection can be established based on their loadings, but more noteworthy, based on the directional loadings vectors as reflected in the bi-plots (Figure 3-29).

3.7.7 The Hippuric Acid Effect

The consumption of the flavoured water as vehicle is an intervention in its own right, given the relatively high concentration of sodium benzoate, used as a preservative. We have previously described the effect of the vehicle consumption (Irwin *et al.*, 2016). From this investigation it became clear that hippuric acid dominates the urine profile, which we took in consideration when using flavoured water as vehicle for alcohol consumption - we applied a paired PLS–DA on the original data, excluding the information on hippuric acid, and compared the VIP values from this analysis to those from the unpaired PLS–DA which included hippuric acid. Both analyses produced 27 metabolites with a VIP > 1.0, with only the 12 metabolites defined as important indicators, and listed in Table 3-8, being common to both analyses when applying the additional criteria of WRT $p \leq 0.05$ and $|FC| \geq 1.5$. These results imply that the presence of hippuric acid in the data set did not influence the metabolites listed in Table 3-8, and, per implication, did not affect the outcome of the study.

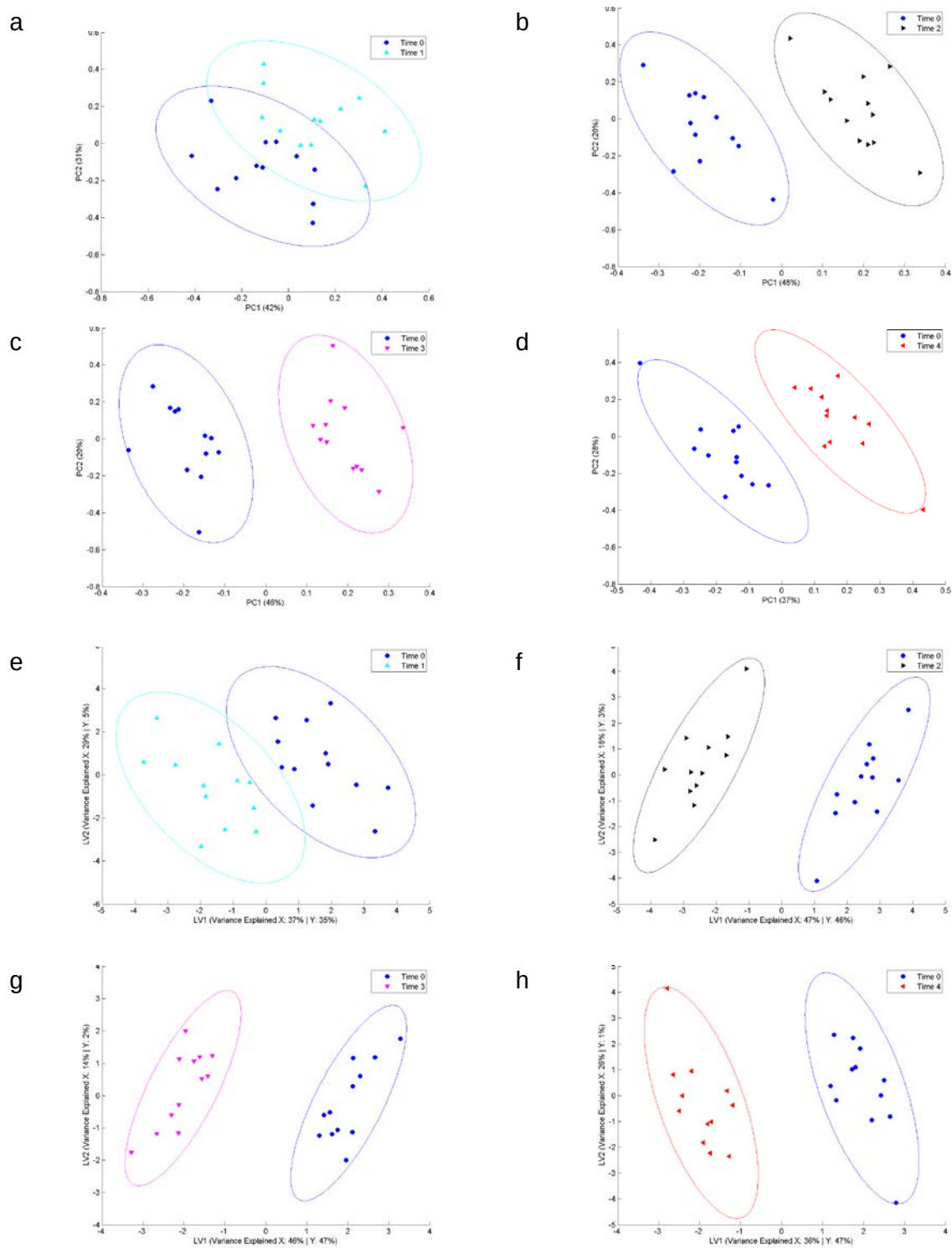


Figure 3-33: PCA and PLS-DA plots following alcohol consumption.

Scores plots of ML-PCA (left column) and PLS-DA on paired data (right column) for time 0 vs time 1 (a & e), time 0 vs time 2 (b & f), time 0 vs time 3 (c & g), and time 0 vs time 4 (d & h) following alcohol intake.

3.7.8 Example of Informed Consent Form

INFORMED CONSENT TO PARTICIPATE IN EXPERIMENT WITH ETHICAL APPROVAL FROM THE NORTH-WEST UNIVERSITY

TITLE:

An investigation into the metabolic responses to acute alcohol consumption

AIM OF EXPERIMENT:

To determine the metabolic perturbations in young males in a fasted state, due to the consumption of a fixed acute dose of alcohol and/or NAD, as investigated by a metabolomics methodology.

INVESTIGATOR'S NAME: Cindy Irwin

SUPERVISOR: Prof. C.J. Reinecke

INVESTIGATOR SITE NAME & ADDRESS: Centre for Human Metabonomics
NWU, Potchefstroom Campus
Private Bag X6001
POTCHEFSTROOM 2520
South Africa
Tel: 018 299 2309
Fax: 018 293 5248

INTRODUCTION

The North-West University's Centre for Human Metabolomics aims to investigate perturbations associated with human metabolism by means of a metabolomics approach. Metabolomics is a biochemical technique, involving a comprehensive study of low molecular weight biomolecules, commonly known as metabolites. Biochemical analysis of biological samples (such as urine or blood) provides a large comprehensive list of metabolites. The data are analysed by means of bioinformatics, a field of science incorporating statistical multivariate analysis, providing information used to determine/distinguish any potential irregularities within the metabolic profile. The focus is the identification of very specific metabolites that can statistically discriminate between normal and abnormal metabolic situations. These metabolites are known as biomarkers.

PURPOSE OF THE EXPERIMENT

This experiment constitutes part of the investigator's M.Sc. thesis involving the study into the metabolic perturbations associated with alcohol consumption. This experiment is aimed at determining the changes in the metabolite profile that occur after the consumption of an acute alcohol dose and what, if any, effect the administration of NAD together with the alcohol has to prevent these metabolic changes. To minimize variation, a homogeneous, defined experimental group, namely young males between the ages of 20 and 30 years in an overnight fasted state, will be used. The biological specimens used in this experiment will be a blood sample taken before the start of the experiment, as well as urine samples taken at defined intervals of time, followed by a metabolomics analysis.

PROTOCOL OF EXPERIMENT

Acute Alcohol and/or NAD Dose Study:

- Participants are required to abstain from alcohol consumption at least 48 hours preceding this experiment; as well as abstaining from consumption of food prior to initiation of the experiment (i.e. overnight fasted state).
- A questionnaire involving the basic clinical profile, including history of alcohol consumption, of the participant needs to be completed.
- An initial urine sample will be taken and labelled "0 hour" before alcohol and/or NAD dose is consumed.
- Group 1 participants will consume only the vehicle with which the alcohol will be mixed (i.e. lemon flavoured water) over a 15-minute period.
- Group 2 participants will consume a fixed dose of alcohol (Smirnoff vodka 43%, v/v, mixed with lemon flavoured water) of 1.5 g/kg body weight amount over a 15-minute period.
- Group 3 participants will consume the vehicle (i.e. lemon flavoured water), to which 50 mg of NAD has been added, over a 15-minute period.
- Group 4 participants will consume a fixed dose of alcohol (Smirnoff vodka 43%, v/v, mixed with lemon flavoured water) of 1.5 g/kg body weight, to which 50 mg of NAD has been added, over a 15-minute period.
- Urine and saliva for metabolomic investigations will be collected at an agreed time sequence over a period of 5 hours after alcohol and/or NAD dose first consumed, yielding 7 data points (consumption of water 15 minutes before a voiding is needed to facilitate obtaining urine samples).

- The experiment will be done in an environment compatible with clinical and ethical requirements
- A general practitioner will attend the early phases of the experiment (first 2 hours) and will be on call for the remaining part of the experiment.

NOTE: Urine samples must be labelled clearly and precisely and stored in a refrigerated environment and no additional alcohol should be consumed during the experimental time frame.

INFORMED CONSENT PROCEDURES

Participation in the project is fully voluntary. You are free to enquire about the experiment through the investigator and/or supervisor and, if agreeing to participate, the participant will be asked to sign this informed consent form. Should any participant request feedback on the outcomes of this study, such information can be made available to them.

It is required from all participants in this study to complete a questionnaire which provides information which is essential for the project. The participating physician will evaluate the information of all participants, and will approve their participation based on the information given in the questionnaire.

BENEFITS ASSOCIATED WITH THE STUDY

The Master's study, to which this experiment will contribute, is a research project aimed at better understanding the biological relationship between chronic/acute alcohol consumption and metabolic perturbations within humans. As in all cases, improved knowledge of the normal physiology will eventually yield to a better understanding and treatment of any deviation from normal physiology. The outcome of this research will be used by the researcher for a M.Sc. thesis and no reference will be included in the thesis regarding any individual who participated in the study.

PAYMENT OR REIMBURSEMENT

Participants will not be paid for their participation and do not contribute to the costs of the study. An amount of R100 will however be paid to each of the participants for travel expenses and other inconveniences that resulted from participation in the study as well as R50 for the light meal of choice after completion of the experiment.

CONFIDENTIALITY

All research records are confidential unless the law requires disclosure. No name or other personal identifying information of the participants will be used in any reports or publications resulting from this study. Data from this study will be used in an anonymous statistical analysis and reported as such by the NWU. No patient's identification details will be reported or made known to other parties.

VOLUNTARY PARTICIPATION AND CONDITIONS OF WITHDRAWAL

Your participation in this study is completely voluntary. You may choose not to participate in this study to which you are otherwise entitled.

CONSENT

I, _____, have read and understood the preceding information describing this research study and my questions have been answered to my satisfaction. I voluntarily consent to participate in this research study. I do not waive my legal rights by signing this consent form. I will receive a signed and dated copy of this consent form.

PARTICIPANT:

_____	_____	_____
Printed name	Signature	Date

INVESTIGATOR:

_____	_____	_____
Printed name	Signature	Date

PHYSICIAN:

_____	_____	_____
Printed name	Signature	Date

3.8 Statistical Software

Table 3-11 lists the data analysis software used to perform the different statistical analyses discussed throughout this chapter.

Analysis	Software
Univariate statistics	MATLAB and Statistics Toolbox (2012b)
Boxplots	notBoxPlot.m a MATLAB function developed by Rob Campbell and can be downloaded from: http://www.mathworks.com/matlabcentral/fileexchange/26508-raacampbell13-notboxplot .
Multivariate statistics	PLS-Toolbox 8.2.1 (2016). Eigenvector Research. Software available at http://www.eigenvector.com .
Correlation analysis	Corrplot by Wei & Simko (2016)
Power curves	Dell Statistica (2016): www.software.dell.com .
ML-PCA	MATLAB m files created by Biosystems Data Analysis Group, Universiteit van Amsterdam. www.bdagroup.nl/content/Downloads/software . Copyright 2008
ASCA	MATLAB using the statistics toolbox and code provided by Gooitzen Zwanenburg (available under APACHE Licence 2.0 http://www.apache.org/licenses/LICENSE-2.0.html).

Table 3-11: List of statistical software used throuout this chapter.

3.9 Comments on: GC–MS-based urinary organic acid profiling reveals multiple dysregulated metabolic pathways following experimental acute alcohol consumption.

Following the publication of the GC–MS manuscript (3.6), entitled, GC–MS-based urinary organic acid profiling reveals multiple dysregulated metabolic pathways following experimental acute alcohol consumption (Irwin *et al.*, 2018), some points of interest on the relation between the methods used and uncovered information were raised by the readership. Considering the validity of these comments, I felt it beneficial to include responses in the current thesis on these points.

(1) The use of PCA-based methods

The difference between PCA, multilevel PCA, unfolded PCA and ACSA is not adequately addressed in the paper. For example: How were scores derived for time points when using unfolded PCA?

Response:

PCA is a projection method, projecting observations measured across many variables (columns) to fewer linear combinations of those variables than explain as much of the variation across observations (rows) as possible. A scores plot is then a scatter plot for observations based on their values for the linear combinations on display. This remains the essence of all PCA-based methods. The difference between the methods applied here, relates to the structure of the data presented, i.e. the content of rows and columns. PCA assumes columns and rows are unique, so to link observations relating to repeated measures on the same subject or individual, call for the restructuring of the data tensor.

Multilevel PCA restructures the data by replacing observations by the distances between those observations and their respective averages for a given individual. The difference data or data set of distances to individual averages, is then projected, rather than the original observations. This tends to visually inflate whatever differences there may be as is evident when comparing the scores plots from the standard PCA (Figure 3-26) to that of the multilevel PCA (Figure 3-33).

Unfolded PCA attempts to capture the repeated nature of observation by splitting data tensors into two-dimensional subsets representing a given factor level and tiling each set as new columns representing additional variables. Therefore, all observations relating to a given individual are now located in a single row. The restructured data are then projected and the scores and loadings can

again be folded into the original structure to provide scores for each measurement, e.g. point in time, for each individual. This is achieved by recalculating the scores for each tiled subset based only on the data and loadings for that specific subset. The average scores across individuals can also be calculated and projected for each time point, as done here. The variance captured will remain the same for different visualizations as a single projection was done. The loadings can now be used as communalities to rank variables in order of importance (i.e. their contribution to the components of interest) by squaring and summing these loadings.

The ASCA is presently the most advanced PCA-based method as it restructures the data by partitioning them into subsets of difference data associated with a specific design aspect such as groups and times. That is, data now represent factor level means, while distance between observations and factor level means become residuals which are used in scores plots and can be explored further. Subsets of difference data are projected separately, which is useful as this informs on the variance contribution of different design factors, as well as their interactions. However, this partitioning can make the results difficult to interpret in that it cannot be folded, or in this case summed, back to its original structure. Partitioning can make it difficult to identify where significant perturbation occur, more specifically, it is difficult to pinpoint which variables are perturbed at which factor level or point in time. There are no post hoc test such as those known for ANOVA and so unfolded PCA is complementary to ASCA: ASCA is better equipped to detect significant factors, while unfolded PCA is better equipped to pin-point which factor levels are more extreme with respect to others.

(2) Projection plot dimensionality

A practical question was raised on why only the unfolded PCA plots were shown in two dimensions, instead of three, like the other PCA-based and PLS-DA plots?

Response:

As with all projection methods, deciding which components to include in the final model should be well-considered. We selected the highest ranking components with respect to the proportions of variation explained, i.e. the first two or three dimensions. Subsequently, the most visually informative three-dimensional rotation was selected. For the unfolded PCA scores plots the two-dimensional plots were much clearer compared to the three-dimensional plots. The scatter overlay made the pattern in time difficult to distinguish for Figure 3-29, while the distances between time point means were difficult to gauge in three dimensions for Figure 3-27.

Visualization were then to some extent subjective, but were only used to provide a graphical overview of the data and not to model any biological process or even for variable selection.

(3) The selection of Time 1 for the primary comparison

Most of the visualization methods and numerical values of metabolite concentrations included in the paper, show Time 2 as the most perturbed compared to Time 0, with the exception of Figure 3-27. Two comments were made: (i) Figure 3-27 and Figure 3-29 made use of the same method, but show different and somewhat odd results, i.e. the first component does not appear to be centred; (ii) Time 2 is more consistently the most perturbed compared to Time 0 across the statistical methods applied, explain then why Time 1 was selected as the primary comparison.

Response:

The decision to make Time 0 vs 1 the primary comparison was primarily motivated by biological reasoning and experience drawn from the pilot study (Mason, 2010). Wanting to sequentially mine the data to allow for comparison with previous findings was a secondary motive. However, the reviewers of the manuscript specifically requested statistical verification of our result that the predominant effect of alcohol consumption was most evident 1 hour after consumption. For this purpose unfolded PCA was selected to identify the point in time with the largest deviation from baseline (i.e. Time 0). The primary reason for this selection was the way unfolded PCA restructures the data prior to projection, as already discussed in the response to the previous set of questions. This specific unfolding allows loading values to be different for every time point. The data was not mean-centred such that the average time profiles would not be lost. ASCA could easily have been viewed as an alternative and so the resulting scores plot was also included. That said, ASCA projects deviations of the factor levels from the factor mean, while, from a biological point of view, we were primarily interested in deviations from baseline. ASCA loadings are forced equal for all time points, which might not always provide the best interpretation.

The unfolding presented in Figure 3-29 is different to the unfolding presented in Figure 3-27 as it did not include hippuric acid, which dominates the profile, as is evident from Table 3-8. This was purposefully done to detect the time point most perturbed from baseline, across all variables and not a profile dominated by hippuric acid.

Unfortunately, a mistake was made in the data processing leading up to the unfolded PCA – the data were correctly not scaled. The data for Figure 3-29 were only log transformed and not centred, while data for Figure 3-27 should have been scaled relative to baseline (Time 0) for the

interpretation regarding Time 1 to be valid. I am uncertain at the present time what a proper interpretation of the current scaling would be. For example, given that the scores were recalculated as described, I am unsure if the first component describes the distance from the origin or overall centroid (Cadima & Jolliffe, 2009) or neither. The impact of my mistake will be investigated further to judge what corrections are necessary.

Fortunately, as noted earlier, Figure 3-27 and Figure 3-29 were only included to visualize the effect detected by other statistical methods which were performed as presented. Biological interpretations were drawn from Table 3-8, which is correct. That said, it is now evident that, from a statistical point of view, Time 2 showed a more pronounced deviation from baseline. Nonetheless, from a biological point of view, Time 1 remains of chief importance and as noted previously, future work will compare Time 2 to baseline.

(4) PCA bi-plots

It is not well described how the bi-plots were generated. Specifically, why do the scores in the top row of Figure 3-29 differ from those in the bottom row?

Response:

The scores were rescaled to the maximum sum of absolute loadings for better visualization of the loadings vectors relative to the scores. This rescaling will then also reflect on the axes of the plot, however the variation explain will remain the same as the model did not change in any way. The bi-plots in the manuscript were constructed to reflect the direction and size of the seven highest ranking variables (in terms of their sum of squared loadings of the top two components when ranked by variance explained) for each level of the time factor.

3.10 References

- Ali, M.A. & Konishi, T. 1998. Enhancement of hydroxyl radical generation in the Fenton reaction by alpha-hydroxy acid. *Biochemistry and Molecular Biology International*, 46(1):137–145.
- Allfrey, V.G., Faulkner, R. & Mirsky, A.E. 1964. Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. *Proceedings of the National Academy of Science*, 51:786–794.
- Amathieu, R., Triba, M.N., Nahon, P., Bouchemal, N., Kamoun, W., Haouache, H., Trinchet, J.C., Savarin, P., Le Moyec, L. & Dhonneur, G. 2014. Serum ¹H-NMR metabolomic fingerprints of acute-on-chronic liver failure in intensive care unit patients with alcoholic cirrhosis. *PloS one*, 9(2), p.e89230.
- Amberg, A., Rosner, E. & Dekant, W. 2001. Toxicokinetics of methyl tert-butyl ether and its metabolites in humans after oral exposure. *Toxicological Sciences*, 61:62–67.
- Antti, H., Bollard, M.E., Ebbels, T., Keun, H., Lindon, J.C., Nicholson, J.K. & Holmes, E. 2002. Batch statistical processing of ¹H NMR-derived urinary spectral data. *Journal of Chemometrics*, 16:461–468.
- Aprison, M.H. & Werman, R. 1965. The distribution of glycine in cat spinal cord and roots. *Life Sciences*, 4:2075-2083.
- Askgaard, G., Grønbaek, M., Kjær, M. S., Tjønneland, A. & Tolstrup, J. S. 2015. Alcohol drinking pattern and risk of alcoholic liver cirrhosis: a prospective cohort study. *Journal of Hepatology*, 62(5):1061–1067.
- Badenhorst, C.P.S., Erasmus, E., Van der Sluis, R., Nortjé, C. & Van Dijk, A.A. 2014. A new perspective on the importance of glycine conjugation in the metabolism of aromatic acids. *Drug Metabolism Reviews*, 46(3):343–361.
- Badenhorst, C.P.S., Van der Sluis, R., Erasmus, E. Van Dijk, A.A. 2013. Glycine conjugation: importance in metabolism, the role of glycine N-acyltransferase, and factors that influence interindividual variation. *Expert Opinion on Drug Metabolism & Toxicology*, 9:1139-1153.
- Bala S, Marcos M, Gattu A, Catalano D, Szabo G. 2014. Acute binge drinking increases serum endotoxin and bacterial DNA levels in healthy individuals. *PloS one*, 9(5):e96864. Becker, H.C.

2012. Effects of alcohol dependence and withdrawal on stress responsiveness and alcohol consumption. *Alcohol Research*, 34(4):448–458.

Beyoğlu, D. & Idle, J.R. 2012. The glycine deportation system and its pharmacological consequences. *Pharmacology & Therapeutics*, 135:151-167.

Beyoğlu, D., Smith, R.L. & Idle, J.R. 2011. Dog bites man or man bites dog? The enigma of the amino acid conjugations. *Biochemical Pharmacology*, 83:1331-1339.

Blakemore, S-J. & Choudhury, S. 2006. Development of the adolescent brain: implications for executive function and social cognition. *Journal of Child Psychology & Psychiatry*, 47:296–312.

Bouatra, S., Aziat, F., Mandal, R., Guo, A.C., Wilson, M.R., Knox, C., Bjorndahl, T.C., Krishnamurthy, R., Saleem, F., Liu, P. & Dame, Z.T. 2013. The human urine metabolome. *PloS one*, 8(9):e73076.

Bradford, B.U., O'Connell, T.M., Han, J., Kosyk, O., Shymonyak, S., Ross, P.K., Winnike, J., Kono, H. & Rusyn, I. 2008. Metabolomic profiling of a modified alcohol liquid diet model for liver injury in the mouse uncovers new markers of disease. *Toxicology and Applied Pharmacology*, 232:236–243.

Bradford, B. U. & Rusyn, I. 2005. Swift increase in alcohol metabolism (SIAM): understanding the phenomenon of hypermetabolism in liver. *Alcohol* 35:13–17.

Bus, J.S. Banton, M.I., Faber, W.D., Kirman, C.R., McGregor, D.B. & Pourreau, D.B. 2015. Human health screening level risk assessments of tertiary-butyl acetate (TBAC): calculated acute and chronic reference concentration (RfC) and hazard quotient (HQ) values based on toxicity and exposure scenario evaluations. *Critical Reviews in Toxicology*, 45(2):142–171.

Cadima, J. & Jolliffe, I., 2009. On relationships between uncentered and column-centered principal component analysis. *Pakistan Journal of Statistics*, 25(4).

Cohen, J. 1988. Statistical power analysis for behavioural sciences. 2nd ed. Hillsdale, NJ: Erlbaum.

Collins, M,A., Zou, J.Y. & Neafsey, E.J. 1998.Brain damage due to episodic alcohol exposure in vivo and in vitro: furosemide neuroprotection implicates edema-based mechanism. *FASEB Journal*, 12:221–230.

Conti, A. & Bickel, M.H. 1977. History of drug metabolism: discoveries of the major pathways in the 19th century. *Drug Metabolism Review*, 6(1):1–50.

Costantini, P., Chernyak, B.V., Petronilli, V. & Bernardi, P. 1996. Modulation of the mitochondrial permeability transition pore by pyridine nucleotides and dithiol oxidation at two separate sites. *Journal of Biological Chemistry*, 271(12):6746–6751.

Dai, L., Peng, C., Montellier, E., Lu, Z., Chen, Y., Ishii, H., Debernardi, A., Buchou, T., Rousseaux, S., Jin, F. & Sabari, B.R. 2014. Lysine 2-hydroxyisobutyrylation is a widely distributed active histone mark. *Nature chemical biology*, 10(5):365–370.

Day, C.P., Bashir, R., James, O.F., Bassendine, M.F., Crabb, D.W., Thomasson, H.R., Li, T.K. & Edenberg, H.J. 1991. Investigation of the role of polymorphisms at the alcohol and aldehyde dehydrogenase loci in genetic predisposition to alcohol-related end-organ damage. *Hepatology*, 14(5):798–801.

Dekant, W., Bernauer, U., Rosner, E. & Amberg, A. 2001. Toxicokinetics of ethers used as fuel oxygenates. *Toxicology Letters*, 124:37–45.

Dell Statistica. 2016. Dell Inc. Version 13.

Diamond, I. & Gordon, A.S. 1997. Cellular and molecular neuroscience of alcoholism. *Physiological Reviews*, 77(1):1–20.

DiPadova, C., Worner, T.M., Julkunen, R.J. & Lieber, C.S. 1987. Effects of fasting and chronic alcohol consumption on the first-pass metabolism of ethanol. *Gastroenterology*, 92(5):1169–1173.

Dona, A.C., Jiménez, B., Schäfer, H., Humpfer, E., Spraul, M., Lewis, M.R., Pearce, J.T., Holmes, E., Lindon, J.C. & Nicholson, J.K. 2014. Precision high-throughput proton NMR spectroscopy of human urine, serum, and plasma for large-scale metabolic phenotyping. *Analytical Chemistry*, 86(19):9887–9894.

Ellinger, J.J., Chylla, R.A., Ulrich, E.L. & Markley, J.L. 2013. Databases and Software for NMR-based metabolomics. *Current Metabolomics*, 1:28–40.

- Faller, J. & Fox, I.H. 1982. Ethanol-induced hyperuricemia: evidence for increased urate production by activation of adenine nucleotide turnover. *New England Journal of Medicine*, 307(26):1598–1602.
- Favé, G., Beckmann, M., Lloyd, A.J., Zhou, S., Harold, G., Lin, W., Taillart, K., Xie, L., Draper, J. & Mathers, J.C. 2011. Development and validation of a standardized protocol to monitor human dietary exposure by metabolite fingerprinting of urine samples. *Metabolomics*, 7:469-484.
- Fein, G. Greenstein, D., Cardenas, V.A., Cuzen, N.L., Fouche, J.P., Ferrett, H., Thomas, K. and Stein, D.J. 2013. Cortical and subcortical volumes in adolescents with alcohol dependence but without substance or psychiatric comorbidities. *Psychiatry Research: Neuroimaging*, 214(1):1–8.
- Fernando, H., Kondraganti, S., Bhopale, K.K., Volk, D.E., Neerathilingam, M., Kaphalia, B.S., Luxon, B.A., Boor, P.J. & Shakeel Ansari, G.A. 2010. ¹H and ³¹P NMR lipidome of ethanol-induced fatty liver. *Alcoholism: Clinical and Experimental Research*, 34(11):1937–1947.
- Field, A.P. 2009. Discovering statistics using IBM SPSS Statistics: and sex and drugs and rock 'n' roll. 3rd ed. London: Sage Publications.
- Fingerhut, R. 2003. Stable isotope dilution method for the determination of guanidinoacetic acid by gas chromatography/mass spectrometry. *Rapid communications in mass spectrometry*, 17(7):717-722.
- Gao, X., Zhao, A., Zhou, M., Lin, J., Qiu, Y., Su, M., & Jia, W. 2011. GC/MS-based urinary metabolomics reveals systematic differences in metabolism and ethanol response between Sprague–Dawley and Wistar rats. *Metabolomics*, 7(3):363–374.
- Gatley, S.J. & Sherratt, H.A.S. 1977. The synthesis of hippurate from benzoate and glycine by rat liver mitochondria. *Biochemical Journal*, 166:39-47.
- Gibney, M.J., Walsh, M., Brennan, L., Roche, H.M., German, B. & Van Ommen, B. 2005. Metabolomics in human nutrition: opportunities and challenges. *American Journal of Clinical Nutrition*, 82:497–503.
- Gika, H.G. & Wilson, I.D. 2014. Global metabolic profiling for the study of alcohol-related disorders. *Bioanalysis*, 6(1):59–77.

- Gika, H.G. Ji, C., Theodoridis, G.A., Michopoulos, F., Kaplowitz, N. & Wilson, I.D. 2012. Investigation of chronic alcohol consumption in rodents via UHPLC–MS based metabolite profiling. *Journal of Chromatography A*, 1259:128–137.
- Gika, H.G. & Wilson, I.D. 2016. Metabolic profiling approaches for biomarkers of ethanol intake. (In Patel, V.B., ed. *Molecular aspects of alcohol and nutrition*. Elsevier: Academic Press. pp. 213–222).
- Gill, J.S. 2002. Reported levels of alcohol consumption and binge drinking within the UK undergraduate student population over the last 25 years. *Alcohol and Alcoholism*, 37(2):109–120.
- Goodman, S.I., McCabe, E.R., Fennessey, P.V. & Mace, J.W. 1980. Multiple acyl-CoA dehydrogenase deficiency (glutaric aciduria type II) with transient hypersarcosinemia and sarcosinuria; possible inherited deficiency of an electron transfer flavoprotein. *Pediatric Research*, 14(1):12–17.
- Guerrini, I., Quadri, G. & Thomson, A.D. 2014. Genetic and environmental interplay in risky drinking in adolescents: a literature review. *Alcohol and Alcoholism*, 49:138–142.
- Hawkins, R.D. & Kalant, H. 1972. The metabolism of ethanol and its metabolic effects. *Pharmacological Reviews*, 24(1):67–157.
- Hisanaga, Y., Ago, H., Nakagawa, N., Hamada, K., Ida, K., Yamamoto, M., Hori, T., Arai, Y., Sugahara, M., Kuramitsu, S. & Yokoyama, S. 2004. Structural basis of the substrate-specific two-step catalysis of long chain fatty acyl-CoA synthetase dimer. *Journal of Biological Chemistry*, 279:31717–31726.
- Holtzclaw, W.D., Dinkova-Kostova, A.T. & Talalay, P. 2004. Protection against electrophile and oxidative stress by induction of phase 2 genes: the quest for the elusive sensor that responds to inducers. *Advances in Enzyme Regulations*, 44:335–367.
- Imam, A., Mohammed, U. & Abanyam, C.M. 2014. On consistency and limitation of paired t-test, sign and Wilcoxon sign rank test. *IOSR Journal of Mathematics*, 10(1):1–6.
- Ireland, M.A., Van Dongen, R., Davidon, L., Beilin, L.J. & Rouse, I.L. 1984. Acute effects of moderate alcohol consumption on blood pressure and plasma catecholamines. *Clinical Science*, 66:643–648.

- Irwin, C., Mienie, L.J., Wevers, R.A., Mason, S., Westerhuis, J.A., van Reenen, M. & Reinecke, C.J. 2018. GC–MS metabolomics reveals multiple dysregulated metabolic pathways following experimental acute alcohol consumption. *Scientific reports*, 8:5775
- Irwin, C., van Reenen, M., Mason, S., Mienie, L.J., Westerhuis, J.A. & Reinecke, C.J. 2016. Contribution towards a metabolite profile of the detoxification of benzoic acid through glycine conjugation: an intervention study. *PloS one*, 11(12):e0167309.
- Irwin, C., van Reenen, M., Mason, S., Mienie, L.J., Wevers, R.A., Westerhuis, J.A. & Reinecke, C.J., 2018. The ¹H-NMR-based metabolite profile of acute alcohol consumption: a metabolomics intervention study. *PloS one*, 13(5):e0196850.
- Jana, S. & Mandlekar, S. 2009. Role of phase II drug metabolizing enzymes in cancer chemoprevention. *Current Drug Metabolism*, 10:595-616.
- Jansen, J.J., Hoefsloot, H.C.J., Van der Greef, J., Timmerman, M.E., Westerhuis, J.A. & Smilde, A.K. 2005. ASCA: analysis of multivariate data obtained from an experimental design. *Journal of Chemometrics*, 19:469-481.
- Johnson, R.A. & Wichern, D.W. Applied multivariate statistical analysis. 2002. 5th ed. Englewood Cliffs, New Jersey: Prentice Hall.
- Kim, D.H., Lee, E-M., Do, S-H., Jeong, D-H. & Jeong, K.S. 2015. Changes of the Cytoplasmic Proteome in Response to Alcoholic Hepatotoxicity in Rats. *International Journal of Molecular Sciences*, 16:18664–18682.
- Kirpich, I.A., Petrosino, J., Ajami, N., Feng, W., Wang, Y., Liu, Y., Beier, J.I., Barve, S.S., Yin, X., Wei, X. & Zhang, X. 2016. Saturated and unsaturated dietary fats differentially modulate ethanol-induced changes in gut microbiome and metabolome in a mouse model of alcoholic liver disease. *American Journal of Pathology*, 186(4):765–776.
- Knights, K.M. & Drogemuller, C.J. 2000. Xenobiotic-CoA Ligases: Kinetic and Molecular Characterization. *Current Drug Metabolism*, 1:49-66.
- Kochan, G., Pilka, E.S., Von Delft, F., Oppermann, U. & Yue, W.W. 2009. Structural snapshots for the conformation-dependent catalysis by human medium-chain acyl-coenzyme A synthetase ACSM2A. *Journal of Molecular Biology*, 388:997-1008.

- Krojanski, H.G., Lambert, J., Gerikalan, Y., Suter, D. & Hergenröder, R. 2008. Microslot NMR probe for metabolomics studies. *Analytical Chemistry*, 80:8668-8672.
- Lang, F., Busch, G.L., Ritter, M., Völkl, H., Waldegger, S., Gulbins, E. & Haussinger, D. 1998. Functional significance of cell volume regulatory mechanisms. *Physiological Reviews*, 78(1):247–306.
- Lee, J., Jung, Y., Park, J.Y., Lee, S.H. & Hwang, G.S. 2015. LC/MS-based polar metabolite profiling reveals gender differences in serum from patients with myocardial infarction. *Journal of pharmaceutical and biomedical analysis*, 115, 475–486.
- Lenroot, R. K. & Giedd, J.N. 2006. Brain development in children and adolescents: insights from anatomical magnetic resonance imaging. *Neuroscience and Biobehavioral Reviews*, 30:718–729.
- Li, X., Xu, Z., Lu, X., Yang, X., Yin, P., Kong, H., Yu, Y. & Xu, G. 2009. Comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry for metabonomics: biomarker discovery for diabetes mellitus. *Analytical Chimica Acta*, 633(2):257–262.
- Li, Z., Bridges, B., Olson, J. & Weinman, S.A. 2017. The interaction between acetylation and serine-574 phosphorylation regulates the apoptotic function of FOXO3. *Oncogene*, 36(13):1887–1898.
- Lian, J.S., Liu, W., Hao, S.R., Guo, Y.Z., Huang, H.J., Chen, D.Y., Xie, Q., Pan, X.P., Xu, W., Yuan, W.X. & Li, L.J. 2011. A serum metabonomic study on the difference between alcohol- and HBV-induced liver cirrhosis by ultraperformance liquid chromatography coupled to mass spectrometry plus quadrupole time-of-flight mass spectrometry. *Chinese Medical Journal*, 124(9):1367–1373.
- Lieber, C.S. 1994. Alcohol and the liver: 1994 update. *Gastroenterology*, 106(4):1085–1105.
- Lieber, C.S. 1997. Ethanol metabolism, cirrhosis and alcoholism. *Clinica Chimica Acta*, 257(1):59–84.
- Lindon, J.C., Holmes, E. & Nicholson, J.K. 2001. Pattern recognition methods and applications in biomedical magnetic resonance. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 39:1-40.

- Loftus, N., Barnes, A., Ashton, S., Michopoulos, F., Theodoridis, G., Wilson, I., Ji, C. & Kaplowitz, N. 2010. Metabonomic investigation of liver profiles of nonpolar metabolites obtained from alcohol-dosed rats and mice using high mass accuracy MSn analysis. *Journal of Proteome Research*, 10(2):705–713.
- Louvet, A. & Mathurin, P. 2015. Alcoholic liver disease: mechanisms of injury and targeted treatment. *Nature Reviews Gastroenterology & Hepatology*, 12(4):231–242.
- Mäkelä, K. & Mustonen, H. 2000. Relationships of drinking behaviour, gender and age with reported negative and positive experiences related to drinking. *Addiction*, 95:727–736.
- Malatji, B.G., Meyer, H., Mason, S., Engelke, U.F., Wevers, R.A., van Reenen, M. & Reinecke, C.J. 2017. A diagnostic biomarker profile for fibromyalgia syndrome based on an NMR metabolomics study of selected patients and controls. *BMC Neurology*, 17(1):88–102.
- Manna, S.K., Patterson, A.D., Yang, Q., Krausz, K.W., Idle, J.R., Fornace, Jr. A.J. & Gonzalez, F.J. 2011. UPLC–MS-based urine metabolomics reveals indole-3-lactic acid and phenyllactic acid as conserved biomarkers for alcohol-induced liver disease in the *Ppara*-null mouse model. *Journal of Proteome Research*, 10(9):4120–4133.
- Mason, S. 2010. The metabolomics of acute alcohol abuse. Potchefstroom: North-West University (MSc: Dissertation). <http://dspace.nwu.ac.za/handle/10394/5034> Date accessed:
- Mason, S., van Furth, A.M.T., Solomons, R., Wevers, R.A., van Reenen, M. & Reinecke, C.J. 2016. A putative urinary biosignature for diagnosis and follow-up of tuberculous meningitis in children: outcome of a metabolomics study disclosing host–pathogen responses. *Metabolomics*, 12(7):110.
- Masuo, Y., Imai, T., Shibato, J., Hirano, M., Jones, O.A.H., Maguire, M.L., Satoh, K., Kikuchi, S. & Rakwal, R. 2009. Omic analyses unravels global molecular changes in the brain and liver of a rat model for chronic Sake (Japanese alcoholic beverage) intake. *Electrophoresis*, 30(8):1259–1275.
- Mathurin, P. & Deltenre, P. 2009. Effect of binge drinking on the liver: an alarming public health issue? *Gut* 58:613–617.
- MATLAB with Statistics Toolbox. 2012b. The MathWorks Inc., Natick, MA, USA

- McGuire, D.M., Gross, M.D., Van Pilsum, J.F. & Towle, H.C. 1984. Repression of rat kidney L-arginine: glycine amidinotransferase synthesis by creatine at a pretranslational level. *Journal of Biological Chemistry*, 259:12034-12038.
- Moon, K.H., Tajuddin, N., Brown, J., Neafsey, E.J., Kim, H.Y. & Collins, M.A. 2014. Phospholipase A2, Oxidative Stress, and Neurodegeneration in Binge Ethanol-Treated Organotypic Slice Cultures of Developing Rat Brain. *Alcoholism: Clinical and Experimental Research*, 38(1):161–169.
- Morozova, T.V., Mackay, T.F. & Anholt, R.R. 2014. Genetics and genomics of alcohol sensitivity. *Molecular Genetics and Genomics*, 289(3):253–269.
- Mutsaers, H.A., Engelke, U.F., Wilmer, M.J., Wetzels, J.F., Wevers, R.A., van den Heuvel, L.P., Hoenderop, J.G. & Masereeuw, R. 2013. Optimized metabolomic approach to identify uremic solutes in plasma of stage 3–4 chronic kidney disease patients. *PloS one*, 8(8):e71199.
- Nahon, P., Amathieu, R., Triba, M.N., Bouchemal, N., Nault, J.C., Zioli, M., Seror, O., Dhonneur, G., Trinchet, J.C., Beaugrand, M. & Le Moyec, L. 2012. Identification of serum proton NMR metabolomic fingerprints associated with hepatocellular carcinoma in patients with alcoholic cirrhosis. *Clinical Cancer Research*, 18(24):6714–6722.
- Nandi, D.L., Lucas, S.V. & Webster Jr., L.T. 1979. Benzoyl-coenzyme A:glycine N-acyltransferase and phenylacetyl-coenzyme A:glycine N-acyltransferase from bovine liver mitochondria. *Journal of Biological Chemistry*, 254:7230–7237.
- Nicholas, P.C., Kim, D., Crews, F.T. & Macdonald, J.M. 2008. ¹H NMR-based metabolomic analysis of liver, serum, and brain following ethanol administration in rats. *Chemical Research in Toxicology*, 21:408–420.
- Nicholson, J.K., Holmes, E. & Wilson, I.D. 2005. Gut microorganisms, mammalian metabolism and personalized health care. *Nature Reviews Microbiology*, 3:431-438.
- Nishimura, T., Shimizu, T., Mineo, I., Kawachi, M., Ono, A., Nakajima, H., Kuwajima, M., Kono, N. & Matsuzawa, Y. 1994. Influence of daily drinking habits on ethanol-induced hyperuricemia. *Metabolism-Clinical and Experimental*, 43(6):745–748.

- Numminen, H., Syrjälä, M., Benthin, G., Kaste, M. & Hillborn, M. 2000. The effect of acute ingestion of a large dose of alcohol on the hemostatic system and its circadian variation. *Stroke*, 31(6):1269–1273.
- O’Neil, S., S., Coulton, S., Deluca, P., Deverill, M., Drummond, C., Gilvarry, E., Graybill, E., Harle, C., Howel, D., Kaner, E. & McArdle, P. 2012. Brief intervention to prevent hazardous drinking in young people aged 14–15 in a high school setting (SIPS JR-HIGH): study protocol for a randomized controlled trial. *Trials*, 13(1):166.
- Park, H.W., Park, E.H., Yun, H.M. & Rhim, H. Date? Sodium benzoate mediated cytotoxicity in mammalian cells. *Journal of Food Biochemistry*, 35:1034-1046.
- Pérez-Enciso, M. & Tenenhaus, M. 2003. Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach. *Human Genetics*, 112:581-592.
- Petit, G., Maurage, P., Kornreich, C., Verbanck, P. & Campanella, S. 2013. Binge drinking in adolescents: a review of neurophysiological and neuroimaging research. *Alcohol and Alcoholism*, 49(2):198–206.
- Plant, M., Plant, M. & Vernon, B. 1996. Ethics, funding and alcohol research. *Alcohol and Alcoholism*, 31(1):17–25.
- PLS-Toolbox 8.2.1. 2016. Eigenvector Research, Inc., Manson, WA, USA 98831
- Ponce, G., Hoenicka, J., Rodriguez-Jimenez, R., Gozalo, A., Jimenez, M., Monasor, R., Aragüés, M., Rubio, G., Jimenez-Arriero, M.A., Ramos, J.A. & Palomo, T. 2004. DRD2 TaqIA polymorphism is associated with urinary homovanillic acid levels in a sample of Spanish male alcoholic patients. *Neurotoxicity Research*, 6(5):373–377.
- Powers, R. 2009. NMR Metabolomics and drug discovery. *Magnetic Resonance in Chemistry*, 47, S2–S11.
- Puig, J.G. & Fox, I.H. 1984. Ethanol-induced activation of adenine nucleotide turnover: Evidence for the role of acetate. *Journal of Clinical Investigation*, 74(3):936–941.
- Rachakonda, V., Gabbert, C., Raina, A., Bell, L.N., Cooper, S., Malik, S. & Behari, J. 2014. Serum metabolomic profiling in acute alcoholic hepatitis identifies multiple dysregulated pathways. *PloS one*, 9(12):e113860.

- Rehm, J., Taylor, B. & Room, R. 2006. Global burden of disease from alcohol, illicit drugs and tobacco. *Drug Alcohol Reviews*, 25:503–513.
- Rietschel, M. & Treutlein, J. 2013. The genetics of alcohol dependence. *Annals of The New York Academy of Science*, 1282(1):39–70.
- Roberts, M. 2015. 'A big night out': young people's drinking, social practice and spatial experience in the 'liminoid' zones of English night-time cities. *Urban Studies*, 52(3):571–588.
- Schachter, D. & Taggart, J.V. 1953. Benzoyl-coenzyme A and hippurate synthesis. *The Journal of Biological Chemistry*, 203:925-934.
- Schachter, D. & Taggart, J.V. 1954. Glycine N-acylase: purification and properties. *The Journal of Biological Chemistry*, 208:263-275.
- Shield, K.D., Parry, C. & Rehm, J. 2013. Chronic diseases and conditions related to alcohol use. *Alcohol Research*, 35(2):155–173.
- Smilde, A.K., Jansen, J.J., Hoefsloot, H.C.J., Lamers, R-JAN., Van der Greef, J. & Timmerman, M.E. 2005. ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. *Bioinformatics*, 21:3043–3048.
- Smilde, A.K.M., Van der Werf, M., Bijlsma, S., Van der Werff-van der Vat, B.J.C. & Jellema, R.H. 2005. Fusion of mass spectrometry-based metabolomics data. *Analytical Chemistry*, 77:6729–6736.
- Spanagel, R, Bartsch, D., Brors, B., Dahmen, N., Deussing, J., Eils, R., Ende, G., Gallinat, J., Gebicke-Haerter, P., Heinz, A. & Kiefer, F. 2010. An integrated genome research network for studying the genetics of alcohol addiction. *Addiction Biology*, 15(4):369–379.
- Strange, K. 2004. Cellular volume homeostasis. *Advances in Physiology Education*, 28:155–159.
- Surekha, R.H., Srikanth, B.B.M.V., Jharna, P., Ramachandra, R.V., Dayasagar, R.V. & Jyothy, A. 2007. Oxidative stress and total anti oxidant status in myocardial infarction. *Singapore Medical Journal*, 48(2):137–142.
- Szabo, G. 2015. Gut–liver axis in alcoholic liver disease. *Gastroenterology*, 148(1):30–36.

- Szabo, G., Bala, S., Petrasek, J. & Gattu, A. 2010. Gut-liver axis and sensing microbes. *Digestive Diseases*, 28(6):737–744.
- Tabachnick, B.G. & Fidell, L.S. 2007. Experimental designs using ANOVA. Belmont: Thomson Brooks/Cole.
- Tabachnick, B.G. & Fidell, L.S. 2006. Using multivariate statistics. 5th ed. Boston: Allyn & Bacon.
- Tanaka, I. & Ezure, K. 2004. Overall distribution of GLYT2 mRNA-containing versus GAD67 mRNA-containing neurons and colocalization of both mRNAs in midbrain, pons, and cerebellum in rats. *Neuroscience Research*, 49:1874-1882.
- Tanaka, K., Hine, D. G., West-Dull, A. & Lynn, T. B. 1980. Gas-chromatographic method of analysis for urinary organic acids. I. Retention indices of 155 metabolically important compounds. *Clinical Chemistry*, 26(13):1839–1846.
- Tang, W.H., Martin, K.A. & Hwa, J. 2012. Aldose reductase, oxidative stress, and diabetic mellitus. *Frontiers in Pharmacology*, 3:87.
- Taylor, B., Irving, H.M., Kanteres, F., Room, R., Borges, G., Cherpitel, C., Greenfield, T. & Rehm, J. 2010. The more you drink, the harder you fall: a systematic review and meta-analysis of how acute alcohol consumption and injury or collision risk increase together. *Drug and Alcohol Dependence*, 110(1):108–116.
- ThermoFisher Scientific. 2012. Infinity™ Uric Acid Liquid Stable Reagent. 2012:1–2. <https://www.thermofisher.com/order/catalog/product/TR24321>.
- Trujillo-Ortiz, A., Hernandez-Walls, R. & Trujillo-Perez, R.A. 2004. RMAOV1: one-way repeated measures ANOVA. A MATLAB file. <http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=5576> Date accessed: 19 May 2009.
- Van den Berg, R.A., Hoefsloot, H.C.J., Westerhuis, J.A., Smilde, K.A. & Van der Werf, M.J. 2006. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*, 7:1-15.
- Vázquez-Fresno, R., Llorach, R., Alcaro, F., Rodríguez, M.A., Vinaixa, M., Chiva-Blanch, G., Estruch, R., Correig, X. & Andrés-Lacueva, C. 2012. (1)H-NMR-based metabolomic analysis of

the effect of moderate wine consumption on subjects with cardiovascular risk factors.

Electrophoresis, 33(15):2345–2354.

Vessey, D.A., Kelley, M. & Warren, R.S. 1999. Characterization of the CoA ligases of human liver mitochondria catalyzing the activation of short and medium-chain fatty acids and xenobiotic carboxylic acids. *Biochimica et Biophysica Acta*, 1428:455-462.

Viant, M., Bearden, D.W., Bundy, J.G., Burton, I.W., Colette, T.W., Ekman, D.R., Ezernieks, V., Karakach, T.K., Lin, C.Y., Rochfort, S. & Ropp, J.S.D. 2009. International NMR-based environmental metabolomics intercomparison exercise. *Environmental Science and Technology*, 43:219–225.

Villez, K., Steppe, K. & De Pauw, D.J.W. 2009. Use of unfold PCA for on-line plant stress monitoring and sensor failure detection. *Biosystems Engineering*, 103:23–24.

Volmar, C.H. & Wahlestedt, C. 2015. Histone deacetylases (HDACs) and brain function. *Neuroepigenetics*, 1:20–27.

Wang, H., Zhang, H., Deng, P., Liu, C., Li, D., Jie, H., Zhou, Z. and Zhao, Y.L. 2016. Tissue metabolic profiling of human gastric cancer assessed by ¹H NMR. *BMC Cancer*, 16:371.

Wei, T. & Simko, V. 2016. corrplot: Visualization of a Correlation Matrix. R package version 0.77. <https://CRAN.R-project.org/package=corrplot>

Wen, H., Yang, H.J., An, Y.J., Kim, J.M., Lee, D.H., Jin, X., Park, S.W., Min, K.J. & Park, S. 2013. Enhanced phase II detoxification contributes to beneficial effects of dietary restriction as revealed by multi-platform metabolomics studies. *Molecular & Cellular Proteomics*, 12:575-586.

Westerhuis, J.A., Van Velzen, E.J.J., Hoefsloot, H.C.J. & Smilde, A.K. 2010. Multivariate paired data analysis: multilevel PLSDA versus OPLSDA. *Metabolomics*, 6(1):119–128.

Westerhuis, J.A., Hoefsloot, H.C.J., Smit, S., Vis, D.J., Smilde, A.K., Van Velzen, E.J.J., Van Duinhoven, J.P. & Van Dorsten, F.A. 2008. Assessment of PLSDA cross validation. *Metabolomics*, 4(1):81-89.

White, A., Matthews, D. & Best, P. 2000. Ethanol, memory, and hippocampal function: a review of recent findings. *Hippocampus*, 10:88–93.

Wishart, D.S. 2008. Metabolomics: applications to food science and nutrition research. *Trends in Food Science and Technology*, 19:482–493.

World Health Organization. 2014. Global status report on alcohol and health, 2014. www.who.int/substance_abuse/publications/global_alcohol_report/en/ Date accessed: July 16, 2017.

Würtz, P., Cook, S., Wang, Q., Tiainen, M., Tynkkynen, T., Kangas, A.J., Soininen, P., Laitinen, J., Viikari, J., Kähönen, M. & Lehtimäki, T. 2016. Metabolic profiling of alcohol consumption in 9778 young adults. *International Journal of Epidemiology*, 45(5):1493–1506.

Yamamoto, T., Moriwaki, Y. & Takahashi, S. 2005. Effect of ethanol on metabolism of purine bases (hypoxanthine, xanthine, and uric acid). *Clinica Chimica Acta*, 356(1):35–57.

Yoseph, B.P., Breed, E., Overgaard, C.E., Ward, C.J., Liang, Z., Wagener, M.E., Lexcen, D.R., Lusczek, E.R., Beilman, G.J., Burd, E.M. & Farris, A.B. 2013. Chronic alcohol ingestion increases mortality and organ injury in a murine model of septic peritonitis. *PloS one*, 8(5):e62792.

You, C., Zhang, H., Sakharkar, A. J., Teppen, T. & Pandey, S.C. 2014. Reversal of deficits in dendritic spines, BDNF and Arc expression in the amygdala during alcohol dependence by HDAC inhibitor treatment. *International Journal Of Neuropsychopharmacology*, 17:313–322.

Yü, T.F., Sirota, J.H., Berger, L., Halpern, M. & Gutman, A.B. 1957. Effect of Sodium Lactate Infusion on Urate Clearance in Man. *Proceedings of the Society for Experimental Biology and Medicine*, 96(3):809–813.

Yuki, T. & Thurman, R.G. 1980. The swift increase in alcohol metabolism: Time course for the increase in hepatic oxygen uptake and the involvement of glycolysis. *Biochemical Journal*, 186(1):119–126.

Zakhari, S. & Li, T-K. 2007. Determinants of alcohol use and abuse: impact of quantity and frequency patterns on liver disease. *Hepatology*, 46(6):2032–2039.

Zhang, A-H., Qiu, S., Xu, H-Y., Sun, H. & Wang, X-J. 2014. Metabolomics in diabetes. *Clinica Chimica Acta*, 429:106–110.

- Zhao, Y. & Garcia, B. A. 2015. Comprehensive catalog of currently documented histone modifications. *Cold Spring Harbor Perspectives in Biology*, 7(9):a025064.
- Zimniak, P. 2008. Detoxification reactions: relevance to aging. *Ageing Research Reviews*, 7:281-300.
- Zwanenburg, G., Hoefsloot, H.C.J., Westerhuis, J.A., Jansen, J.J., & Smilde, A.K. 2011. ANOVA-principal component analysis and ANOVA-simultaneous component analysis: a comparison. *Journal of Chemometrics*, 25:561–567.

CHAPTER 4 — NOVEL STATISTICAL APPROACHES FOR THE ANALYSIS OF METABOLOMICS DATA

4.1 Background

In Chapter 2 the characteristics of metabolomics data are described. One of these characteristics is a large proportion of missing data, specifically when generating data through the GC–MS (Gas chromatography–mass spectrometry) analysis of samples. In metabolomics, missing values are often automatically replaced by zeroes by the platform software. This is done as missing values are believed to primarily result from observations below the limit of detection (LOD) of the platform and metabolites simply not being present in a given biological sample (truly zero). This implies that a zero value may be more than just a zero, in turn making zero-value processing a much more involved process than most researchers currently consider. The third publication in Chapter 3 illustrated the mainstream approach to the zero-value problem, which advocates the exclusion of variables with high proportions of zeros not related to any group structure, followed by the imputation of the remaining zero values. Given the potential impact of imputing a large proportion of a data set, this was prioritized with regard to the development of statistical approaches for metabolomics data, i.e. the primary aim of this thesis.

Studies comparing two groups dominate metabolomics research even where designs are more complex, as illustrated in Chapter 3. The primary reason, I experienced and therefore infer, is interpretability, i.e. more complex designs still require two-group comparisons to aid biological interpretation. The aim is then to identify metabolites that differ between groups and rank them according to their discriminatory ability. Ranks are then compared to a threshold to select a subset of variables for metabolic mapping and/or for constructing classification models. It therefore seemed reasonable to develop a method that worked well or was custom-built for comparing two groups. This was then the first part of the development (Section 4.2): Accepted Publication: Variable selection for binary classification using error rate p -values applied to metabolomics data (van Reenen *et al.*, 2016).

The initial publication laid a foundation upon which many more approaches could be built. Accordingly, the second publication strived to extend the foundational approach, error rate p -value or ER p , by enabling it to more accurately reflect the nature of zero-valued observations, which was the original aim. The endeavour turned out to be more fruitful as the development of ER p did more than introduce the theory upon which its extension (extended ER p or XER p) could

be staged, ERp can already address many other characteristics of metabolomics data: (i) not normality distributed; (ii) groups of different size; (iii) measuring many times more variables than subjects; and (iv) the impact of classification error, that is, correctly classifying ill individual is usually of greater consequence than correctly classifying healthy individuals. In addition, the resulting approach is easy to implement and even more so to move to a clinical setting. ERp is easy to comprehend and interpret, not only making user interaction possible, but also robust. For example, users are able to: (i) rank variables in an interpretable order of importance; (ii) relax or make stricter the selection criteria of variables based on a sound understanding of the selection criteria and underlying biology.

The second publication presented in Section 4.4 - Accepted Publication: Metabolomics variable selection and classification in the presence of observations below the detection limit using an extension of ERp (van Reenen *et al.*, 2017) - built on the first and aimed to answer the question of zero-value processing by extending the functionality of ERp. XERp is able to achieve this based on the concept of a jump discontinuity in distributions, at least for the primary sources of zero-values, i.e. absent from the sample (truly zero) and the LOD, without requiring any differentiation in source or knowledge of the LOD. XERp also further broadened the horizon of future prospects to accommodate even more of the characteristics of metabolomics data (as discussed in Chapter 5).

Note that the publications presented in this chapter were altered somewhat to conform to the template used for this thesis.

4.2 Accepted Publication: Variable selection for binary classification using error rate p -values applied to metabolomics data

4.2.1 Title Page

Variable selection for binary classification using error rate p -values applied to metabolomics data

Authors:

Mari van Reenen^{3,4,§}, Carolus J Reinecke³, Johan A Westerhuis^{1,3}, J Hendrik Venter²

§Corresponding author:

Mari van Reenen email: 12791733@nwu.ac.za

Addresses:

¹Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands

²Centre for Business Mathematics and Informatics, Faculty of Natural Sciences, North-West University (Potchefstroom Campus), Private Bag X6001, Potchefstroom, South Africa.

³Centre for Human Metabolomics, Faculty of Natural Sciences, North-West University (Potchefstroom Campus), Private Bag X6001, Potchefstroom, South Africa

⁴Department of Statistics, Faculty of Natural Sciences, North-West University (Potchefstroom Campus), Private Bag X6001, Potchefstroom, South Africa

Email addresses:

Mari van Reenen: 12791733@nwu.ac.za

Carolus J Reinecke: carools.reinecke@nwu.ac.za

Johan A Westerhuis: j.a.westerhuis@uva.nl

J Hendrik Venter: hennie.venter@nwu.ac.za

4.2.2 Abstract

Background

Metabolomics datasets are often high-dimensional though only a limited number of variables are expected to be informative given a specific research question. The important task of selecting informative variables can therefore become complex. In this paper we look at discriminating between two groups. Two tasks need to be performed: (i) finding variables which differ between the two groups; and (ii) determining how the selected variables can be used to classify new subjects. We introduce an approach using minimum classification error rates as test statistics to find discriminatory and therefore informative variables. The thresholds resulting in the minimum error rates can be used to classify new subjects. This approach transforms error rates into p -values and is referred to as ERp.

Results

We show that non-parametric hypothesis testing, based on minimum classification error rates as test statistics, can find statistically significantly shifted variables. The discriminatory ability of variables becomes more apparent when error rates are evaluated based on their corresponding p -values, as relatively high error rates can still be statistically significant. ERp can handle unequal and small group sizes, as well as account for the cost of misclassification. ERp retains (if known) or reveals (if unknown) the shift direction, aiding in biological interpretation. The threshold resulting in the minimum error rate can immediately be used to classify new subjects.

We use NMR generated metabolomics data to illustrate how ERp is able to discriminate subjects diagnosed with *Mycobacterium tuberculosis* infected meningitis from a control group. The list of discriminatory variables produced by ERp contains all biologically relevant variables with appropriate shift directions discussed in the original paper from which this data is taken.

Conclusions

ERp performs variable selection and classification, is non-parametric and aids biological interpretation while handling unequal group sizes and misclassification costs. All this is achieved by a single approach which is easy to perform and interpret. ERp has the potential to address many other characteristics of metabolomics data. Future research aims to extend ERp to account for a large proportion of observations below the detection limit, as well as expand on interactions between variables.

Keywords

Variable selection; significance testing; non-parametric; binary classification; metabolomics.

4.2.3 Background

A major aim of metabolomics studies is to find metabolites that distinguish a control group of reference or 'normal' subjects from a group of experimental or 'abnormal' subjects which differ from the control group subjects as a result of disease, treatment with drugs, toxicity, environmental, genetic or physiological effects (Madsen *et al.*, 2010; Bundy *et al.*, 2009; Kaddurah-Daouk *et al.*, 2008). The interpretation of those metabolites in terms of the underlying biological phenomena and the development of discriminating biomarkers are important goals (Xia *et al.*, 2013). Traditional statistical methods often make assumptions which make the validity of results questionable in the case of metabolomics. Metabolite concentrations are non-negative, requiring suitable transformation to accommodate the distributional assumptions of parametric statistical methods. While non-parametric methods such as the Mann-Whitney test make no distributional assumptions, they do not produce classification rules for new subjects. Metabolomics datasets generated through spectroscopic or spectrometric methods can consist of hundreds and even thousands of variables, making the selection of discriminatory variables important, yet complex (Gromski *et al.*, 2014). In addition, it is often difficult or expensive to obtain sufficient subjects or sample material to make group sizes large and equal. However, methods such as logistic regression require large sample sizes (N), especially when confronted with a large number of explanatory variables (V) (Bewick *et al.*, 2005). Variable selection prior to developing a classification model is often performed on large V small N data. However, inference after variable selection is not advisable without correcting for the uncertainty associated with the selection step (Berk *et al.*, 2013). These methods also require model specifications such as linearity in the variables of the regression function. This linearity may well be misspecified. Methods such as PLS-DA (Partial least squares regression for discriminant analysis) are often used, but are biased and more likely to classify new subjects as belonging to the smaller of the two groups (Brereton & Lloyds, 2014) whereas the cost of misclassification may be opposite to this. Furthermore, variable selection based on a PLS-DA model is still problematic (Kvalheim *et al.*, 2014). Such projection based statistical methods are generally rather sophisticated, limiting their practicality (Xia *et al.*, 2013; Brereton & Lloyd, 2014). This is likely the reason why metabolomics researchers often combine the results of a variety of statistical and even machine learning methods to select a subset of variables. Doing so can become cumbersome if they do not reach

the same decision and again requires estimation of “*post-selection error*” (Steel *et al.*, 2004) when used in further model development.

The notions of sensitivity (*se*) and specificity (*sp*) of statistical methods are often used to evaluate the classification ability of models. These can be combined into the Youden Index (*J*) and the area under the receiver operating characteristic curve (AUC) (Yin & Tian, 2014; Lai *et al.*, 2012). In this paper we also combine sensitivity and specificity but in the form of a weighted sum of misclassification error rates depending on a threshold. The choice of threshold resulting in the minimum error rate provides us with a rule to classify new subjects. There are some parallels with CART (classification and regression trees) in this respect. However, we show that non-parametric hypothesis testing can be based on the minimised error rates. This enables us to convert the error rates into *p*-values which in turn lead to selecting variables that contain discriminatory information. These *p*-values provide a ranking of the variables and the notion of multiple testing corrections can be used to decide how far up in this ranking the variables are considered to contain significant discriminatory information. This approach is referred to as ERp below.

ERp takes unequal and small group sizes explicitly into account and allows for the specification of the relative cost of misclassification, which is desirable when selecting an appropriate threshold. ERp provides information regarding the direction of the shift, i.e. whether metabolite levels are higher or lower in the experimental group, thus aiding biological interpretation of the results. ERp simultaneously provides us with a classification rule which can be applied immediately to classify new subjects. That is, once statistically significantly shifted variables have been selected, final classification can be based on a majority vote (as used here) or a more complex weighting structure taking *p*-values into account.

The paper is structured as follows: First, we review the notion of classification error rates and their dependence on thresholds, as well as how they can be estimated from available data. Secondly, we show that the testing used in ERp is non-parametric and how to calculate the relevant *p*-values. Finally, we illustrate the application of ERp on ¹H nuclear magnetic resonance (NMR) spectroscopy data from cerebrospinal fluid (CSF) samples to discriminate subjects diagnosed with *Mycobacterium tuberculosis* infected meningitis from a control group and compare our results with those obtained using traditional methods (Mason *et al.*, 2015)..

4.2.4 Methods

Introduction

If there is a shift in the concentration of a metabolite from the control to the experimental group, the shift is either upwards or downwards. For such a metabolite there is a concentration threshold which can be used to discriminate between the groups and classify new subjects. The combined error rate is associated with a choice of threshold and if it can be chosen to make this combined error rate small, the metabolite is important as a discriminator between the groups. This raises the question: How small must this error rate be for the associated variable (e.g. metabolite) to be a good discriminator? ERp makes use of significance testing and transformation of this error rate to a p -value to answer this question.

Classification error rates

Consider a single variable X and let F_0 and F_1 denote the population cumulative distribution functions (CDFs) for control and experimental subjects (represented by the variable Y) respectively. We assume that experimental subjects ($Y = 1$) tend to have either upwardly or downwardly shifted values of X when compared to control subjects ($Y = 0$). It is important to treat these shift directions separately in order to properly determine the role of X in discriminating between the groups.

The **upward rule** is to choose a threshold c and classify a subject as experimental group if $X > c$ and as control if $X \leq c$. The rate of misclassification of control subjects is $1 - F_0(c)$ and of experimental subjects is $F_1(c)$. Let weights w_0 and w_1 ($w_0 + w_1 = 1$) represent the relative costs of misclassification of control and experimental subjects respectively, so that the weighted combined error rate is

$$er_{up}(c) = w_0(1 - F_0(c)) + w_1F_1(c) \quad (4.2-1)$$

Choosing equal weights $w_0 = w_1 = \frac{1}{2}$ implies that it is equally important to keep both rates of misclassification low. In other applications the weights could be selected differently, e.g. if experimental subjects are ill individuals, then misclassifying an individual as an experimental subject may imply costly or invasive treatment, while not identifying an ill individual as an experimental subject may have fatal consequences. The incidence rates of the two groups may also need to be taken into account in the choice of the weights.

For the **downward rule** a subject is classified as experimental if $X \leq c$ and as control otherwise. The weighted combined error rate is then

$$er_{down}(c) = w_0 F_0(c) + w_1 (1 - F_1(c)) \quad (4.2-2)$$

Both error rates are functions of the threshold c . Let c_{up} and c_{down} represent the choices of c that minimize 4.2-1 and 4.2-2 with minimized error rates er_{up}^* and er_{down}^* respectively. If er_{up}^* is small, the variable X can be used to classify subjects following the upward rule with threshold c_{up} . Alternatively, if er_{down}^* is small, the downward rule with threshold c_{down} can be applied.

As mentioned above a shift in distribution is either upward or the downward and this is reflected in only one of the pair er_{up}^* and er_{down}^* being small. With this in mind, we introduce also a **minimum error rate** together with a threshold and direction by letting

$$er_{min}^* = er_{up}^*, c_{min} = c_{up} \text{ and } d_{min} = \text{"up"} \text{ if } er_{up}^* < er_{down}^* \quad (4.2-3)$$

and

$$er_{min}^* = er_{down}^*, c_{min} = c_{down} \text{ and } d_{min} = \text{"down"} \text{ if } er_{up}^* \geq er_{down}^* \quad (4.2-4)$$

In other words, er_{min}^* is the minimum of the up and down error rates while c_{min} and d_{min} are the threshold and shift direction associated with this minimum. Sometimes subject matter reasons may dictate to only consider the upward or downward shift, but in the absence of such reasons, we choose the smaller of the two error rates (er_{min}^*). We can then classify a new subject using the rule specified by d_{min} .

Estimating error rates from data

All the quantities introduced above depend on the population CDFs F_0 and F_1 which are usually unknown and can only be estimated from the data at hand. Notation is needed for this purpose. Denote the number of subjects observed in the control and experimental groups by N_0 and N_1 respectively. Let $N = N_0 + N_1$, and for $n = 1, 2, \dots, N$, let x_n represent the value of X for the n -th subject and y_n its group indicator taking the value 0 for the control group and 1 for the experimental group.

The empirical estimates of $F_0(c)$ and $F_1(c)$ are given by $\frac{1}{N_0} \sum_{n=1}^N (1 - y_n) I(x_n \leq c)$ and $\frac{1}{N_1} \sum_{n=1}^N y_n I(x_n \leq c)$ respectively, where $I(A)$ is the indicator function of the event A .

Replacing $F_0(c)$ and $F_1(c)$ in 4.2-1 by their estimates, the estimated combined error rate is

$$\widehat{er}_{up}(c) = \frac{w_0}{N_0} \sum_{n=1}^N (1 - y_n) I(x_n > c) + \frac{w_1}{N_1} \sum_{n=1}^N y_n I(x_n \leq c) \quad (4.2-5)$$

Let $\hat{c}_{up}$, denote the value of c which minimizes 4.2-5 and let the corresponding minimized error rate be $\widehat{er}_{up}(\hat{c}_{up}) = \min_c \{\widehat{er}_{up}(c)\} = \widehat{er}_{up}^*$.

This minimization can be performed by ranking the x_n 's increasingly. As c is varied, $\widehat{er}_{up}(c)$ remains constant between the successively ranked x_n values. Hence it is sufficient to compute $\widehat{er}_{up}(c)$ only at the midpoints of the intervals formed by the ranked x_n values and then to choose $\hat{c}_{up}$ as the value which minimizes $\widehat{er}_{up}(c)$ (Fluss *et al.*, 2005).

Thus $\hat{c}_{up}$ and $\widehat{er}_{up}^*$ provide estimates of c_{up} and er_{up}^* respectively when using the upward rule. If an upward shift in the values of the variable X is of interest and $\widehat{er}_{up}^*$ turns out to be small, X can be used to classify subjects by applying the upward rule with threshold $\hat{c}_{up}$. Similar statements hold when specifying a downward shift or specifying no shift direction (see Section 4.3.1).

Using classification error rates as test statistics

Clearly, the discriminating ability of X is related to the size of $\widehat{er}_{up}^*$, but it is not obvious how small $\widehat{er}_{up}^*$ should be for X to be a good discriminator. Furthermore, the true but unknown error rate er_{up}^* may differ from $\widehat{er}_{up}^*$, making it inadvisable to judge the importance of the variable X solely on the value of $\widehat{er}_{up}^*$ without taking into account the extent of such differences. One possible way to do this is to calculate a standard error or confidence interval for er_{up}^* , as done when using the Youden index (Lai *et al.*, 2012). We propose to use $\widehat{er}_{up}^*$ as a test statistic to formally test the null hypothesis that there is no shift in the distribution of X for the experimental group compared to the control group, against the alternative of an upward shift in distribution. This enables us to judge the discriminatory importance of X in terms of the familiar concept of a p -value.

For testing purposes, the distributions of $\widehat{er}_{up}^*$, $\widehat{er}_{down}^*$ and $\widehat{er}_{min}^*$ under the null hypothesis $H_0: F(x) = F_0(x) = F_1(x)$ are required.

Assume: (i) $F(x)$ is a continuous and strictly increasing function in x starting at 0 for some sufficiently small value of x ; and (ii) the x_n 's are independent and identically distributed (*iid*) according to F .

Putting $u_n = F(x_n)$ and $b = F(c)$, equation 4.2-5 becomes

$$\begin{aligned}\widehat{er}_{up}(c) &= \frac{w_0}{N_0} \sum_{n=1}^N (1 - y_n) I(F(x_n) > F(c)) + \frac{w_1}{N_1} \sum_{n=1}^N y_n I(F(x_n) \leq F(c)) \\ &= \frac{w_0}{N_0} \sum_{n=1}^N (1 - y_n) I(u_n > b) + \frac{w_1}{N_1} \sum_{n=1}^N y_n I(u_n \leq b) = \widetilde{er}_{up}(b)\end{aligned}$$

Since this expression is only a function of b , minimising over c is equivalent to minimising over b , giving $\widehat{er}_{up}^* = \min_c \{\widehat{er}_{up}(c)\} = \min_b \{\widetilde{er}_{up}(b)\}$. This expresses $\widehat{er}_{up}^*$ as a function of the u_n 's only. The probability integral transform states that the u_n 's are independent and identically uniformly distributed on $[0, 1]$, henceforth abbreviated as *IUD* $[0, 1]$. Therefore, the null distribution of $\widehat{er}_{up}^*$ does not depend on F , i.e. $\widehat{er}_{up}^*$ is a non-parametric test statistic. Section 4.3.2 shows that this is also true for $\widehat{er}_{down}^*$ and $\widehat{er}_{min}^*$ and that the null distribution of $\widehat{er}_{down}^*$ is the same as for $\widehat{er}_{up}^*$. Moreover, the null distributions depend on the group sizes (N_0 and N_1) and the weights (w_0 and w_1), so that one does not need to be concerned about unequal sample sizes, sampling variability and biases when judging the resulting p -values as indicators of the discriminatory importance of X (see Section 4.3.3).

The actual calculation of the null distribution by analytic means seems impossible in view of the complicated expressions involved in the definitions of the error rates. However, simulation offers a solution. Figure 4-1 provides an algorithm to convert error rate test statistics into their corresponding p -values. As an alternative to simulating the null distribution, asymptotic approximation can be used, as discussed in Section 4.3.3. We illustrate the benefit of converting error rates into their corresponding p -values in the Results & Discussion below.

- Generate N $IIUD$ $[0, 1] u_n$'s
- Assign the first $N_0 y_n$'s as 0 and the remainder as 1
- Minimize $\frac{w_o}{N_o} \sum_n^N (1 - y_n) I(u_n > b) + \frac{w_l}{N_l} \sum_{n=1}^N y_n I(u_n \leq b)$ by varying b over the midpoints of the increasingly ordered u_n 's to obtain $\widehat{er}_{up}^*$
- Repeat these steps M times to build up a file of *iid* copies of $\widehat{er}_{up}^*$, say $\widehat{er}_{up}^*(m)$, $m = 1, \dots, M$, providing a simulation approximation of the null CDF
- If T of the $\widehat{er}_{up}^*(m)$'s fall below an actually observed $\widehat{er}_{up}^*$ its associated p -value is approximately T / M . Approximations are more accurate for large M .

Figure 4-1: Algorithm to simulate the null cumulative distribution functions.

ERp applied to more than one variable

Up to this point we have considered only one variable X . As mentioned previously, metabolomics studies deal with multiple variables, therefore we need to find a short list of discriminatory and biologically relevant variables which are preferably easy to detect in clinical practice (Xia *et al.*, 2013). ERp calculates the error rate for each variable, converts it into a p -values and ranks the variables by increasing p -values. The family wise error rate (FWER) or false discovery rate can be used to decide how far up the ranking variables are still considered to contain significant information. We use the Bonferroni-Holm (BH) method (Holm, 1979) to control the FWER for multiple testing, making it the only parameter involved in our approach. However, a slight adjustment is required when applying stepwise methods such as BH, since more than one variable can have the same error rate and therefore p -value. The stepwise nature of BH may indicate that some variables are significant while others, with the same error rate, are not. In such instances we prefer to be on the conservative side regarding control of the FWER. That is, if a

variable is not significant then all variables with the same error rate should also be treated as not significant. Instead of using the Bonferroni-Holm method the user may opt for any of the many other available correction methods - see for example Romano & Wolf (2005) and Manso & Machado (2015).

ERp Software

The MATLAB functions to perform ERp, as well as an example application, are provided as part of the supplementary material. Section 4.3.6 gives details of this software together with the user inputs required and the output delivered.

4.2.5 Results & Discussion

In this section we discuss two examples to illustrate the benefit of converting error rates into their corresponding p -value. The sample sizes $N_0 = 21$ and $N_1 = 12$ were selected to correspond with those of the dataset used to illustrate ERp in the application subsection below. Two weight sets, referred to as 1 ($w_0 = w_1 = \frac{1}{2}$) and 2 ($w_0 = \frac{1}{3}$ and $w_1 = \frac{2}{3}$) are used. These examples are used throughout the remainder of the paper.

Converting error rates to p -values

We calculate the null CDFs based on one million simulation repetitions using the algorithm in Figure 4-1. Since we are only interested in significantly small error rates, the left tails of the null distributions are relevant. For clarity purposes, Figure 4-2 shows the left tails of the logarithms of these null CDFs.

As is to be expected the null distribution of $\widehat{er}_{min}^*$ is shifted to the left relative to that of $\widehat{er}_{up}^*$. This is because $\widehat{er}_{min}^*$ has slightly less power being a two-sided test compared to the one-sided test of $\widehat{er}_{up}^*$. The customary 5% significance level ($\alpha = 0.05 = 10^{-1.3}$) is attained for error rates as high as 0.3 and 0.25 (light blue lines). An error rate as large as 0.3 or 0.25 would likely not lead one to conclude that X has discriminatory ability. This becomes more apparent when evaluating the corresponding p -values. Metabolomics studies have many variables requiring correction for multiple testing. Therefore, a lower significance level such as $\alpha = 0.001 = 10^{-3}$ may be relevant. Even for such a low α the observed error rates are above 0.15 (dark blue lines). Section 4.3.7 discusses similar results for other sample sizes and weight combinations.

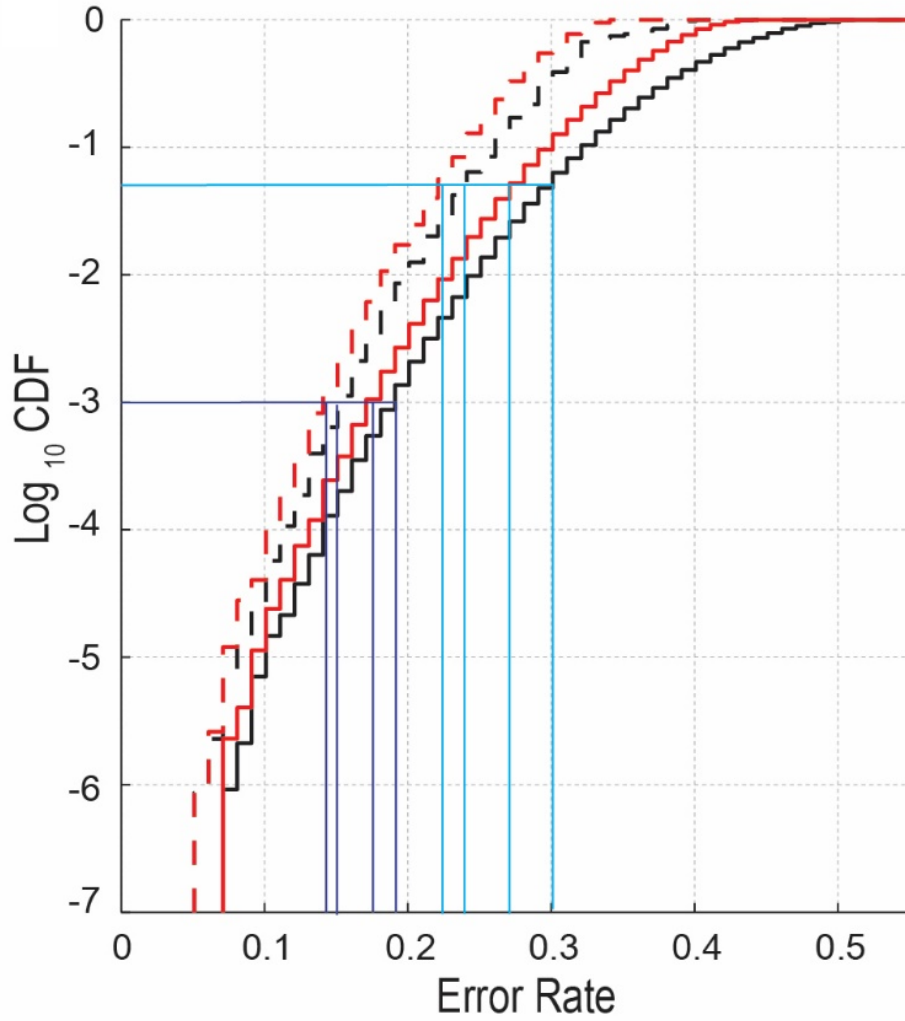


Figure 4-2: The null cumulative distribution functions.

The graphs show the $\log_{10}$ transformed null CDFs of $\widehat{er}_{up}^*$ (black lines) and $\widehat{er}_{min}^*$ (red line), for group sizes $N_0 = 21$ and $N_1 = 12$ using weight sets 1 (solid lines) and 2 (dashed lines). The dark ($\alpha=0.001$) and light ($\alpha=0.05$) blue lines represent points of reference discussed in the text.

Power comparison of error rate test statistics

ERp operates on two levels, namely it performs a hypothesis test and also delivers a classification threshold. It may be anticipated that delivering two outputs comes at a cost, i.e. less power in the hypothesis testing part. Here we briefly report a simulation study comparing the power of $\widehat{er}_{up}^*$ and $\widehat{er}_{min}^*$ as test statistics against that of the well-known non-parametric Mann-Whitney (MW) test. The results presented here assume that the control group follows a standard log-normal

distribution, while the experimental group follows an upwardly shifted log-normal distribution. That is, if $y_n = 0$, then $X_n = \exp(Z_n)$ with Z_n normally distributed with mean 0 and variance 1, while if $y_n = 1$, then $X_n = \exp(Z_n + \mu)$ where μ varies over a grid of positive values. Ten thousand simulated data sets were generated for each grid point and the p -values were calculated for the test statistics $\widehat{er}_{up}^*$, $\widehat{er}_{min}^*$ and MW, at each shift magnitude and for each repetition. These simulations were performed in MATLAB (2012b) making use of the one-sided MW test ensuring sound comparison to $\widehat{er}_{up}^*$.

As a first measure of comparison the resulting p -values were averaged over the repetitions to measure the expected power of the test statistics. Figure 4-3 compares the average p -values (A - weight set 1 and B - weight set 2) associated with $\widehat{er}_{up}^*$ (red lines), $\widehat{er}_{min}^*$ (blue lines) and the MW test statistic (black lines), as well the proportions of repetitions in which the p -values (C - weight set 1 and D - weight set 2), of $\widehat{er}_{min}^*$ (blue lines) and MW (black lines) were below the p -values of $\widehat{er}_{up}^*$. The dotted red line represents the 50% cut-off. The dashed blue lines represent points of reference to be discussed later in the text.

As is to be expected, having a priori information regarding which shift direction to evaluate does improve the discriminatory ability of ERp, since $\widehat{er}_{up}^*$ is more powerful than $\widehat{er}_{min}^*$. It is also evident that the MW test statistic delivers more power on average, but this diminishes at larger shift, i.e. at shifts of most practical relevance.

As a second measure of comparison, Figure 4-3 C and D show the proportion of p -values for each of the test statistics that are smaller than the p -values of $\widehat{er}_{up}^*$. If this proportion is above (below) 50%, then the corresponding test is frequently better (poorer) than $\widehat{er}_{up}^*$ on this measure. These figures show that the MW is better at smaller shifts, i.e. below 1.6 and 2.0, while for larger shifts, ERp is consistently better.

In summary, ERp's ability to also classify new subjects does not seem to result in a lack of power as compared to the MW test. Section 4.3.8 supplies more results on this matter.

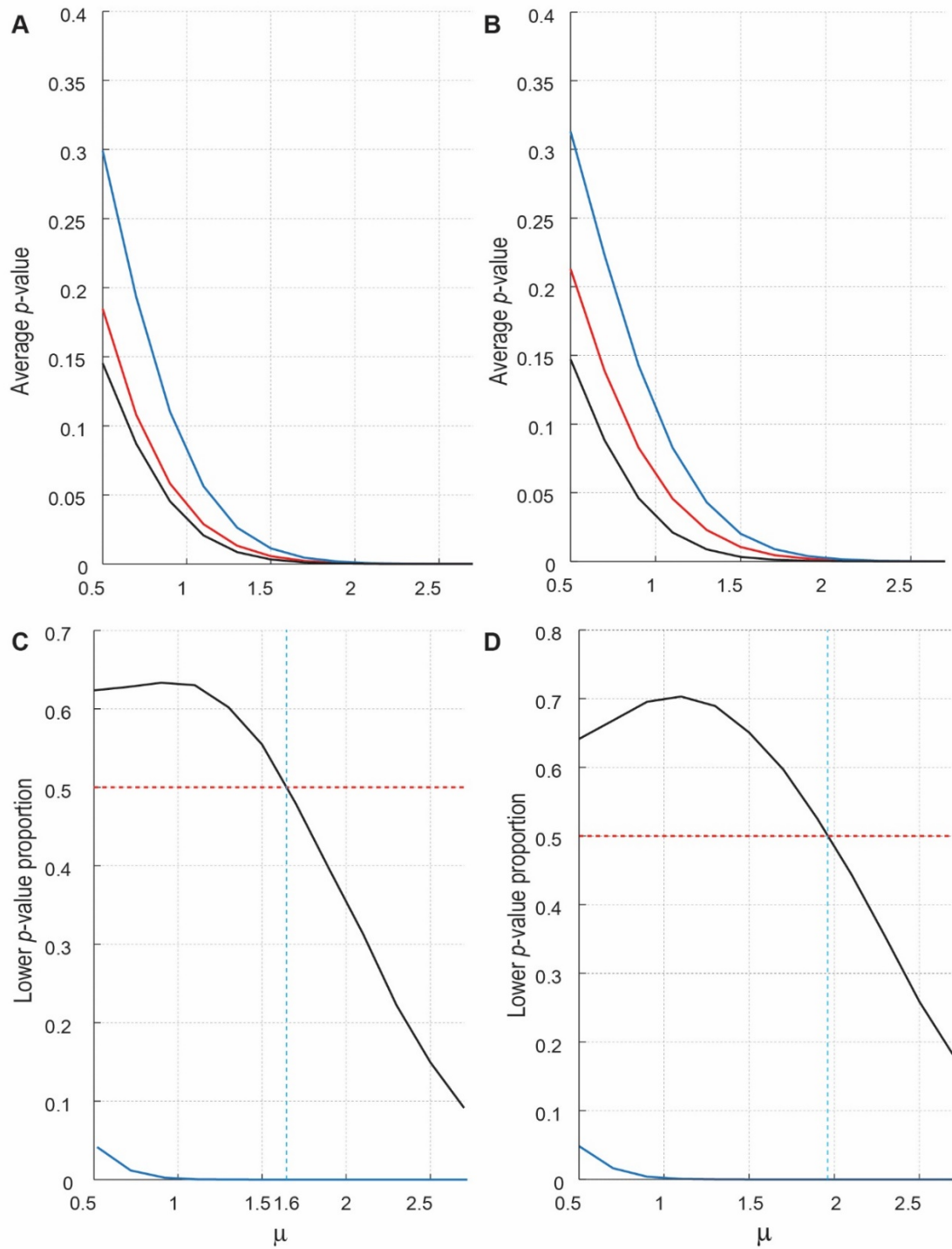


Figure 4-3: Simulation comparison of the different error rate test statistics.

A (weight set 1) and B (weight set 2) depict the average p -values associated with $\hat{e}r_{up}^*$ (red lines), $\hat{e}r_{min}^*$ (blue lines) and the MW test statistic (black lines). Figure 4-3 C (weight set 1) and D (weight set 2) depict the proportions of repetitions in which the p -values of $\hat{e}r_{min}^*$ (blue lines) and MW (black lines) were below the p -values of $\hat{e}r_{up}^*$. The dotted red line represents the 50% cut-off. The dashed blue lines represent points of reference as discussed in the text.

Leave-One-Out Error Rate Estimation

In the classification literature leave-one-out (LOO) cross-validation is often used to estimate error rates with less bias than the error rates used throughout this paper (Leeflang *et al.*, 2008). This leads to the question whether LOO error rates can also be used in a hypothesis testing role to find discriminating variables analogous to the approach presented above. However, lower bias of the LOO error rates does not automatically imply greater power in the testing context. We studied this issue and found that the LOO error rate based test statistics are also non-parametric but, on average, less powerful than their “full sample” (FS) counterparts which makes use of all available data. Sections 4.3.5 and 4.3.8 provide more details.

Application to Metabolomics Data

Finally, we illustrate ERp by applying it to data obtained from the metabolomics study reported in Mason *et al.* (2015). The data was generated through ^1H NMR spectroscopy from CSF (cerebral spinal fluid) samples obtained from subjects who suffered from meningitis, but not caused by *Mycobacterium tuberculosis* (Mtb) infection. They formed the control group. The experimental group consisted of patients who like-wise suffered from with meningitis, however confirmed to be caused by Mtb.

Mason *et al.* (2015) selected 12 variables in the CSF that were able to distinguish the control from the experimental subjects. These variables were selected based on PLS-DA VIP values as well as univariate statistics. Quantitatively, two metabolites that yielded the greatest measures of importance (i.e. those most responsible for the separation) were highly elevated lactate and decreased glucose in the TBM subjects relative to values observed for the controls. These two metabolites indicate the well-known disturbances in the energy metabolism of several neurological disorders (Leib *et al.*, 1999). Further selected variables also support the energy perturbation caused by the infection of the meninges by the tuberculosis bacterium.

We apply ERp to the same data excluding subjects identified as outliers in Mason *et al.* (2015) and including only identified metabolites, thus leaving us with 55 metabolites, 17 experimental and 30 control subjects. Next we split the dataset into two parts: a training set and a test set. The test set includes about 30% of subjects from each group, randomly selected, and are not used to find important metabolites. The test set contains 5 experimental and 9 control subjects and is used to show the classification ability of ERp given new subjects. The training set therefore contains $N_0 = 21$ control subjects while the experimental group consists of $N_1 = 12$ subjects.

Mason *et al.* (2015) made no weight assumptions so that using weight set 1 (our equal weights scenario) is likely to yield more comparable results. However, weight set 2 may be more appropriate in this application since untreated TBM is mostly fatal (Khan *et al.* 2003). Therefore, results are reported for both weight sets. Although in this context the shift directions are known for some variables we elect to make no directional assumptions so as to allow the data to speak in this regard.

List of Significantly Shifted Variables

Table 4-1 lists the variables selected by ERp for different choices of the FWER α together with their error rates and position in the NMR spectrum, thresholds (c), shift directions, p -values and the BH adjusted critical levels for the corresponding FWER. To illustrate how sensitive the list of selected variables is to the specified FWER, three choices are reported namely $\alpha=1\%$, 5% and 10%. As is to be expected, a more tolerant FWER results in a longer list of significant variables.

	Metabolite	ER	C	Direction	p-value (%)	1% BH (%)	5% BH (%)	10% BH (%)
Weight Set 1 (Equal Weights)	C(13)Lactate (1.27)	0.0000	0.0005	Up	0.0000	0.0182	0.0909	0.1818
	C(13)Lactate (1.283)	0.0000	0.0005	Up	0.0000	0.0182	0.0909	0.1818
	lactate (1.41)	0.0000	0.1875	Up	0.0000	0.0182	0.0909	0.1818
	glucose (3.21)	0.0000	0.0094	Down	0.0000	0.0182	0.0909	0.1818
	glucose (3.24)	0.0000	0.0325	Down	0.0000	0.0182	0.0909	0.1818
	glucose (3.38)	0.0000	0.0560	Down	0.0000	0.0182	0.0909	0.1818
	glucose (3.48)	0.0000	0.0632	Down	0.0000	0.0182	0.0909	0.1818
	lactate (4.36)	0.0000	0.0619	Up	0.0000	0.0182	0.0909	0.1818
	glucose (5.22)	0.0000	0.0216	Down	0.0000	0.0182	0.0909	0.1818
	glucose (4.63)	0.0238	0.0170	Down	0.0000	0.0217	0.1087	0.2174
	glucose (3.71)	0.0417	0.0973	Down	0.0000	0.0222	0.1111	0.2222
	glucose (3.86)	0.0417	0.1056	Down	0.0000	0.0222	0.1111	0.2222
	valine (1.04)	0.0655	0.0015	Up	0.0000	0.0233	0.1163	0.2326
	alanine/? (1.51) *	0.0655	0.0054	Up	0.0000	0.0233	0.1163	0.2326
	succinate (2.66)	0.0655	0.0002	Up	0.0000	0.0233	0.1163	0.2326
	glucose/glycerol (3.43) *	0.0655	0.0405	Down	0.0000	0.0233	0.1163	0.2326
	2-oxoglutarate (2.68)	0.0714	0.0001	Up	0.0001	0.0256	0.1282	0.2564
	creatinine (3.13)	0.0714	0.0027	Down	0.0001	0.0256	0.1282	0.2564
	C(13)Lactate (1.53)	0.0952	0.0013	Up	0.0013	0.0270	0.1351	0.2703
	val/iso (1.01) *	0.1369	0.0021	Up	0.0145	0.0278	0.1389	0.2778
	DMSO4 (3.14)	0.1369	0.0007	Down	0.0145	0.0278	0.1389	0.2778
	lysine (1.73)	0.1429	0.0033	Up	0.0234	0.0294	0.1471	0.2941
	phenylalanine/? (7.42) *	0.1548	0.0025	Up	0.0375	0.0303	0.1515	0.3030
	creatinine (4.29)	0.1726	0.0015	Down	0.0981	0.0313	0.1563	0.3125
	myo-inositol (3.285)	0.1786	0.0016	Down	0.1274	0.0323	0.1613	0.3226
	acetoacetate (2.30)	0.1905	0.0004	Up	0.2311	0.0333	0.1667	0.3333

	Metabolite	ER	C	Direction	p-value (%)	1% BH (%)	5% BH (%)	10% BH (%)
Weight Set 2 (Unequal Weights)	C(13)Lactate (1.27)	0.0000	0.0005	Up	0.0000	0.0182	0.0909	0.1818
	C(13)Lactate (1.283)	0.0000	0.0005	Up	0.0000	0.0182	0.0909	0.1818
	lactate (1.41)	0.0000	0.1875	Up	0.0000	0.0182	0.0909	0.1818
	glucose (3.21)	0.0000	0.0094	Down	0.0000	0.0182	0.0909	0.1818
	glucose (3.24)	0.0000	0.0325	Down	0.0000	0.0182	0.0909	0.1818
	glucose (3.38)	0.0000	0.0560	Down	0.0000	0.0182	0.0909	0.1818
	glucose (3.48)	0.0000	0.0632	Down	0.0000	0.0182	0.0909	0.1818
	lactate (4.36)	0.0000	0.0619	Up	0.0000	0.0182	0.0909	0.1818
	glucose (5.22)	0.0000	0.0216	Down	0.0000	0.0182	0.0909	0.1818
	glucose (4.63)	0.0159	0.0170	Down	0.0000	0.0182	0.0909	0.1818
	glucose (3.71)	0.0317	0.1173	Down	0.0000	0.0182	0.0909	0.1818
	glucose (3.86)	0.0317	0.1208	Down	0.0000	0.0182	0.0909	0.1818
	2-oxoglutarate (2.68)	0.0476	0.0001	Up	0.0000	0.0182	0.0909	0.1818
	creatinine (3.13)	0.0476	0.0027	Down	0.0000	0.0182	0.0909	0.1818
	C(13)Lactate (1.53)	0.0635	0.0013	Up	0.0001	0.0244	0.1220	0.2439
	succinate (2.66)	0.0635	0.000004	Up	0.0001	0.0244	0.1220	0.2439
	glucose/glycerol (3.43) *	0.0635	0.0463	Down	0.0001	0.0244	0.1220	0.2439
	valine (1.04)	0.0714	0.0015	Up	0.0012	0.0263	0.1316	0.2632
	alanine/? (1.51) *	0.0714	0.0054	Up	0.0012	0.0263	0.1316	0.2632
	lysine (1.73)	0.0952	0.0033	Up	0.0038	0.0278	0.1389	0.2778
	val/iso (1.01) *	0.1190	0.0021	Up	0.0297	0.0286	0.1429	0.2857
	DMSO4 (3.14)	0.1190	0.0007	Down	0.0297	0.0286	0.1429	0.2857
	acetoacetate (2.30)	0.1270	0.0004	Up	0.0390	0.0303	0.1515	0.3030
	myo-inositol (3.285)	0.1508	0.0018	Down	0.2014	0.0313	0.1563	0.3125
	phenylalanine/? (7.42) *	0.1587	0.0025	Up	0.2521	0.0323	0.1613	0.3226
* Overlap of known and/or unknown metabolites								

Table 4-1: Significant variables based on weight set 1 and 2.

Column 1 contains the variable names ordered from smallest to largest according to their error rates (column 2). Column 3 contains the threshold with the associated “up” or “down” classification rule indicated in column 4. Column 5 displays simulated p -values expressed as percentages. Columns 6 to 8 display the 1%, 5% and 10% FWERs respectively. The red, green and purple blocks group variables that were significantly shifted at a 1%, 5% and 10% FWERs, respectively. The alternating blocks of white and grey group variables with the same error rates and therefore the same p -values.

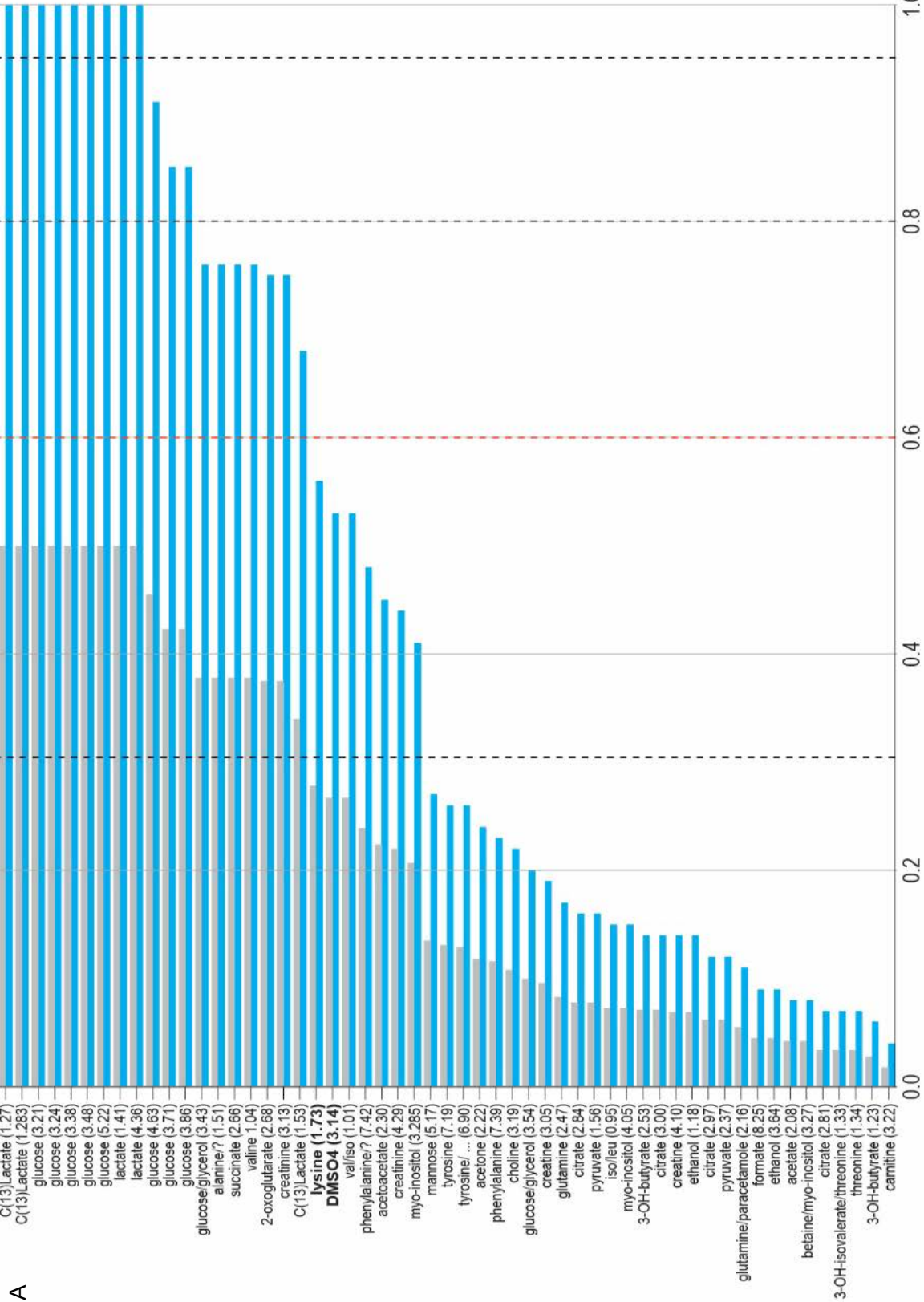
The first column provides the variable names ordered from smallest to largest according to their error rates which are shown in the second column. The third column provides the threshold estimates which can be used to classify new subjects by employing the “up” or “down” rule as indicated by the direction in the fourth column. The fifth column provides the p -values associated with the error rates, expressed as percentages. The significance of these values can be determined through comparison to the BH adjusted critical level. The last three columns provide

these levels for three different FWERs namely 1%, 5% and 10%. The red, green and purple blocks encapsulate the variables that were significantly shifted at a 1%, 5% and 10% FWERs, respectively. For groups of variables with the same error rates and therefore the same *p*-values the most conservative BH level is applied. These groups are indicated in alternating blocks of white and grey.

In addition to selecting the same metabolites as Mason *et al.* (2015), ERp also selected succinate as a significant metabolite. Since succinate is an intermediate in the tricarboxylic acid cycle, the finding is compatible with the increased aerobic energy metabolism required by the microglia to destroy the invading tuberculosis bacterium. Overall, we are able to draw the same conclusions presented in Mason *et al.* (2015) by using only one method as opposed to many different and more complex techniques.

For comparative purposes, we also modelled a standard CART classification tree, using SPSS (2015), with the following specifications: (i) all variables were selected as potential predictors; (ii) equal prior probabilities for the two groups; (iii) stopping when the final nodes contain only one subject or at a tree depth of 100; and (iv) the test set was used for validation purposes. These settings were selected to make ERp and CART results as comparable as possible. CART constructs a classification tree by recursively dividing the data into subsets until these subsets are as homogeneous as possible with regard to group labels (Breiman *et al.*, 1984). As a result, the tree stopped growing too soon and only included the variable 'C13 lactic (1.27)', overlooking other variables important for biological interpretation. CART does provide a measure of variable importance, i.e. the Normalized Importance (NI). Figure 4-4 displays these scores obtained when applying CART to the training data given the two weight sets.

The CART method provides a measure of importance for each variable. These values (grey bars), along with the normalized values (blue bars), are depicted here in the form of a bar chart for each variable (y-axis) in order of decreasing importance. Figure 4-4 A and B represent weight sets 1 and 2, respectively. The vertical black and red dashed lines represent values at which large drops in NI occur and therefore possible cut-off choices for variable selection. The dashed red lines represent the cut-offs chosen for comparison with ERp, while the dashed black lines represent alternative choices.



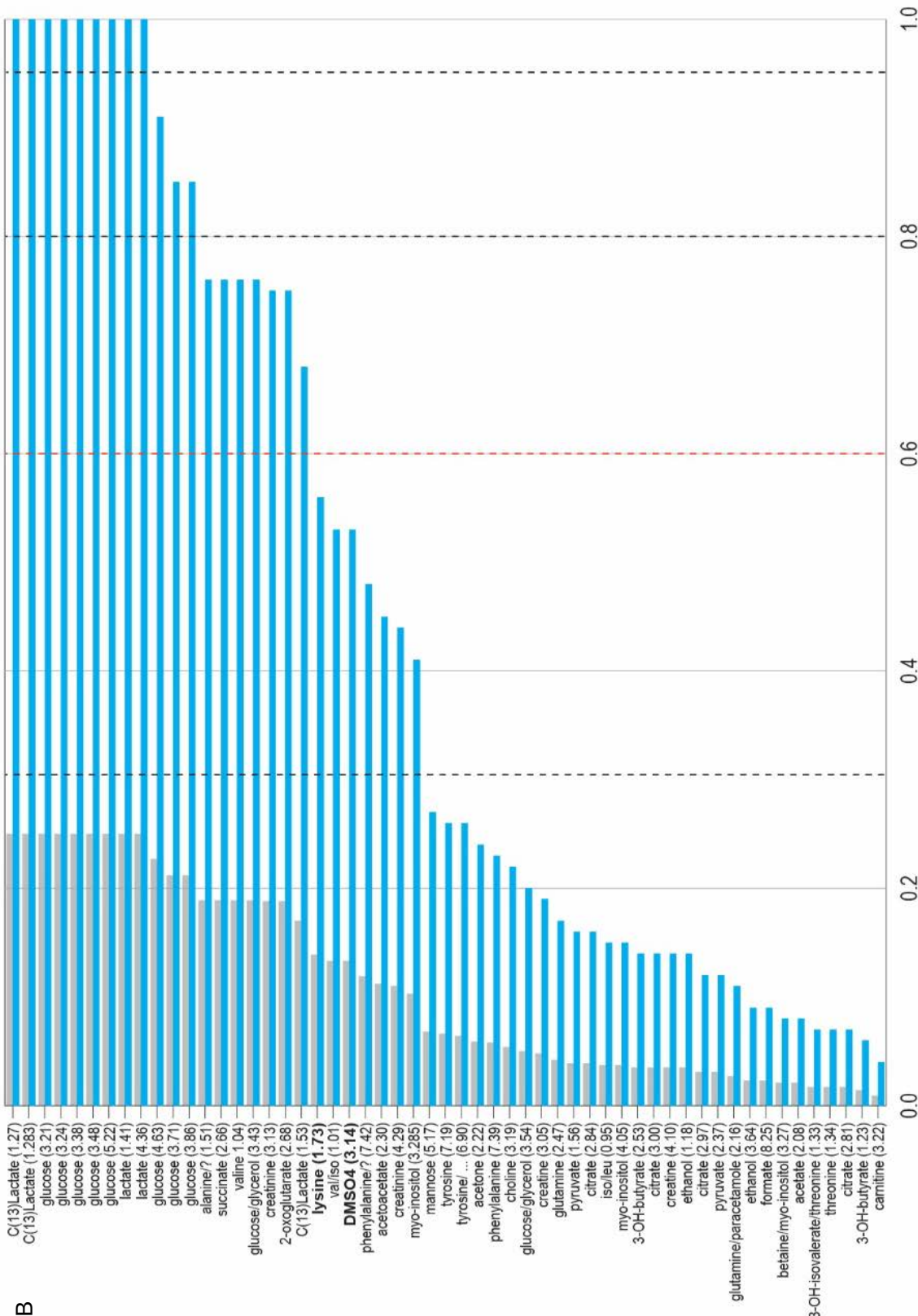


Figure 4-4: CART variable importance.

CART importance (grey bars) and normalized importance scores (blue bars) are depicted. Each bar represents a variable (y-axis). Figures A and B represent weight sets 1 and 2, respectively. The vertical black and red dashed lines represent large changes in importance scores. The dashed red lines represent the cut-offs chosen for comparison with ERp, while the dashed black lines represent alternative choices.

The suggested cut-off for variable selection corresponds to a point just before a large drop in NI. We chose a cut-off corresponding to an NI of 60%, that is, all variables with NI exceeding 60% were selected as important. At this cut-off CART did not flag two biologically important variables, compared to the original paper (Mason *et al.*, 2015), and missed DMSO (the depletion of DMSO is associated with oxidative stress) and lysine (increased levels of lysine are associated with mental retardation). CART does not select variables with similar information to those already in the tree structure and may overlook metabolites necessary for biological interpretation.

Classification of Unseen Subjects

We now make use of the test set of subjects to illustrate the classification feature of ERp. Table 4-2 shows the classification results based on the lists of variables in Table 4-1, taking $\alpha = 1\%$, using the corresponding thresholds, shift directions and classification rules.

The body of the table shows the classification result due to each significantly shifted variable for each subject, where 0 indicates the subject was classified into the control group and 1 indicates the subject was classified into the experimental group. Misclassifications are indicated in red. The last three rows (i) provide the final classification based on the majority vote; (ii) flag subjects that were misclassified; and (iii) flag potential outlying subjects based on the number of variables that misclassified it compared to the remaining subjects.

The variables in Table 4-2 are ordered based on their p -value, from smallest to largest. As is to be expected variables lower down the list are prone to making more misclassifications. However, in general there are very few misclassifications. The majority of variables made a single misclassification and mostly for the same subject (number 12), indicating that it may be an 'outlier', i.e. not representative of the group. Constructing classification tables for test as well as training subjects enables us to screen for potential outliers, another potentially useful application of ERp. However, since outlier detection is not the main aim of this paper, we do not explore the matter further.

Overall the second weight set, which puts more weight on correctly classifying subjects in the experimental group, was more successful and made no final misclassifications, based simply on majority vote, even in the presence of the potential outlier. No classification model was developed in the original paper (Mason *et al.*, 2015) and therefore no comparison is possible.

The classification results for the test set (assuming equal weights) are the same for the CART model as for ERp, with only one experimental subject misclassified. However, ERp outperforms the CART model for weight set 2 (assuming unequal weights), with CART again misclassifying one experimental subject while ERp made no misclassifications.

Subject		1	2	3	4	5	6	7	8	9	10	11	12	13	14
True Group Indicator		0	0	0	0	0	0	0	0	0	1	1	1	1	1
Variables		Control Group Classification									Experimental Group Classification				
Weight Set 1 (Equal Weights)	C(13)Lactate (1.27)	0	0	0	0	0	0	0	0	0	1	1	1	0	1
	C(13)Lactate (1.283)	0	0	0	0	0	0	1	0	0	1	1	1	1	1
	lactate (1.41)	0	0	0	0	0	0	0	0	0	1	1	1	1	1
	glucose (3.21)	0	0	0	0	0	0	0	0	0	1	1	0	1	1
	glucose (3.24)	0	0	0	0	0	0	0	0	0	1	1	0	1	1
	glucose (3.38)	0	0	0	0	0	0	0	0	0	1	1	0	1	1
	glucose (3.48)	0	0	0	0	0	0	0	0	0	1	1	0	1	1
	lactate (4.36)	0	0	0	0	0	0	0	0	0	1	1	1	0	1
	glucose (5.22)	0	0	0	0	0	0	0	0	0	1	1	0	1	1
	glucose (4.63)	0	0	0	0	0	0	0	0	0	1	1	1	1	1
	glucose (3.71)	0	0	0	0	0	0	0	0	0	1	1	0	1	1
	glucose (3.86)	0	0	0	0	0	0	0	0	0	1	1	0	1	1
	valine (1.04)	0	0	0	0	0	0	0	0	0	1	0	0	1	0
	alanine/? (1.51)	0	0	0	0	0	0	0	0	0	1	0	0	1	0
	succinate (2.66)	0	0	0	0	0	0	0	0	0	1	0	1	0	0
	glucose/glycerol (3.43)	0	0	0	0	0	0	0	0	0	1	1	1	1	1
	2-oxoglutarate (2.68)	1	0	0	0	0	0	1	1	1	1	1	1	0	1
	creatinine (3.13)	0	0	0	0	1	0	0	0	0	1	0	0	1	1
	C(13)Lactate (1.53)	0	0	0	0	0	0	0	0	0	1	1	1	1	1
	val/iso (1.01)	0	0	0	0	0	1	0	0	0	1	0	0	1	0
	DMSO4 (3.14)	0	0	0	1	1	0	0	0	0	1	1	0	0	1
	lysine (1.73)	1	0	0	0	0	0	0	0	0	1	1	0	1	0
Final Classification		0	0	0	0	0	0	0	0	0	1	1	0	1	1
Misclassified Subjects															
Potential Outliers															

	Subject	1	2	3	4	5	6	7	8	9	10	11	12	13	14
	True Group Indicator	0	0	0	0	0	0	0	0	0	1	1	1	1	1
	Variables	Control Group Classification									Experimental Group Classification				
Weight Set 2 (Unequal Weights)	C(13)Lactate (1.27)	0	0	0	0	0	0	0	0	0	1	1	1	0	1
	C(13)Lactate (1.283)	0	0	0	0	0	0	1	0	0	1	1	1	1	1
	lactate (1.41)	0	0	0	0	0	0	0	0	0	1	1	1	1	1
	glucose (3.21)	0	0	0	0	0	0	0	0	0	1	1	0	1	1
	glucose (3.24)	0	0	0	0	0	0	0	0	0	1	1	0	1	1
	glucose (3.38)	0	0	0	0	0	0	0	0	0	1	1	0	1	1
	glucose (3.48)	0	0	0	0	0	0	0	0	0	1	1	0	1	1
	lactate (4.36)	0	0	0	0	0	0	0	0	0	1	1	1	0	1
	glucose (5.22)	0	0	0	0	0	0	0	0	0	1	1	0	1	1
	glucose (4.63)	0	0	0	0	0	0	0	0	0	1	1	1	1	1
	glucose (3.71)	0	1	0	0	0	0	0	0	0	1	1	1	1	1
	glucose (3.86)	0	1	0	0	0	0	0	0	0	1	1	1	1	1
	2-oxoglutarate (2.68)	1	0	0	0	0	0	1	1	1	1	1	1	0	1
	creatinine (3.13)	0	0	0	0	1	0	0	0	0	1	0	0	1	1
	C(13)Lactate (1.53)	0	0	0	0	0	0	0	0	0	1	1	1	1	1
	succinate (2.66)	0	0	0	0	0	0	1	1	0	1	1	1	1	1
	glucose/glycerol (3.43)	0	1	0	0	0	1	0	1	0	1	1	1	1	1
	valine (1.04)	0	0	0	0	0	0	0	0	0	1	0	0	1	0
	alanine/? (1.51)	0	0	0	0	0	0	0	0	0	1	0	0	1	0
	lysine (1.73)	1	0	0	0	0	0	0	0	0	1	1	0	1	0
Final Classification		0	0	0	0	0	0	0	0	0	1	1	1	1	1
Misclassified Subjects															
Potential Outliers															

Table 4-2: Group classification and outlier detection using significant variables based on weight set 1 and 2.

The table shows the classification result based on significantly shifted variables for each subject, with 0 indicating classification into the control group, 1 indicating classification into the experimental group and misclassifications in red. The table consists of two sections representing two weigh sets. The last three rows in each section provide the final classification based on the majority vote, flags misclassified subjects, and flags potential outliers, respectively.

Though this is only a single application and not a comprehensive comparison, the choice of cut-off for CART remains difficult to interpret. In contrast, ERp makes use of a controlled FWER (α) which has a direct interpretation as the probability of having included one or more variables which do not discriminate between the groups.

4.2.6 Conclusion

Our main contribution has been to show that non-parametric hypothesis testing, based on minimum error rates, can find statistically significantly shifted variables. We found that the discriminatory ability of variables becomes more apparent when error rates were evaluated based on their corresponding p -values as relatively high error rates can still be significant. The power simulations performed concluded that the MW test is more powerful for small shifts in distribution, while ERp is competitive for larger shifts. An exploratory application of ERp indicated that markers of the disease state of patients suffering from TBM were successfully selected and used for the classification of patients with meningitis due to Mtb infection relative to other causes.

In summary, ERp can accommodate unequal and small group sizes while accounting for the cost of misclassification into either group. ERp retains (if known) or reveals (if unknown) the shift direction, aiding biological interpretation. The thresholds resulting in the minimum error rates can be used to classify new subjects or to identify potential outliers.

ERp is a useful addition to the range of methods used for binary discrimination and classification. Future research aims to explain how ERp can evaluate interactions and extend ERp to accommodate data with a large proportion of observations below the detection limit.

4.2.7 Additional Information

Competing Interests

The authors declare that they have no competing interests

Authors' Contributions

MvR and JHV developed the method and wrote the paper. JAW provided direction with regard to which aspects of the method be developed in this initial paper. JAW also reviewed and edited the paper. CJR provided the data as well as the biological interpretation of metabolites selected by ERp. All authors read and approved the final manuscript.

Acknowledgements

We would like to acknowledge Shayne Mason for making the data available to us. Shayne Mason is a recipient of a Desmond Tutu-NRF-VU doctoral fellowship. Research funding for his project

was provided by the Technological Innovation Agency (TIA) of the Department of Science and Technology of South Africa.

4.3 Supplementary Information: Variable selection for binary classification using error rate p -values applied to metabolomics data

This section provides a summary of additional results that may be useful to the reader and should be read in conjunction with the main manuscript (Section 4.2).

4.3.1 Estimating error rates from data

Denote the number of subjects observed in the control and experimental groups by N_0 and N_1 respectively, so that $N = N_0 + N_1$ denotes the total number of subjects in both groups combined. Denote the observed value of X for the n -th subject by x_n for $n = 1, 2, \dots, N$. The grouping variable is denoted by the binary indicator Y , taking the value 0 for the control group and 1 for the experimental group and denoting the value of Y for the n -th subject by y_n . The (y_n, x_n) pairs then represent the observed data given the variable X .

The empirical estimates of $F_0(c)$ and $F_1(c)$ are then given by $\frac{1}{N_0} \sum_{n=1}^N (1 - y_n) I(x_n \leq c)$ and $\frac{1}{N_1} \sum_{n=1}^N y_n I(x_n \leq c)$ respectively, where $I(A)$ is the indicator function of the event A . Replacing $F_0(c)$ and $F_1(c)$ in equation 4.2-2 by their estimates, the estimated combined error rate is

$$\widehat{er}_{down}(c) = \frac{w_0}{N_0} \sum_{n=1}^N (1 - y_n) I(x_n \leq c) + \frac{w_1}{N_1} \sum_{n=1}^N y_n I(x_n > c)$$

Denote by $\hat{c}_{down}$, the corresponding estimate of c_{down} , a choice of c that minimises the equation above with minimized value $\widehat{er}_{down}^*$. This minimization can be performed using the same approach as described in the main paper. These two items provide estimates of c_{down} and er_{down}^* based on the available data when using the downward rule. If an downward shift in the values of the variable X is of interest and $\widehat{er}_{down}^*$ turns out to be small, the variable X can be used to classify subjects following the downward rule with threshold $\hat{c}_{down}$.

However, in the estimation context, we may not have a preference for either one of the two directions and allow the data to ‘speak’ in this regard, i.e. to consider the smaller of the two directional error rates.

Parallel to the notation in equations 4.2-3 and 4.2-4:

$$\widehat{er}_{min}^* = \widehat{er}_{up}^*, \hat{c}_{min} = \hat{c}_{up} \text{ and } \hat{d}_{min} = "up" \text{ if } \widehat{er}_{up}^* < \widehat{er}_{down}^* \text{ and}$$

$$\widehat{er}_{min}^* = \widehat{er}_{down}^*, \hat{c}_{min} = \hat{c}_{down} \text{ and } \hat{d}_{min} = "down" \text{ if } \widehat{er}_{up}^* \geq \widehat{er}_{down}^*$$

represent the corresponding estimators of er_{min}^* , c_{min} and d_{min} .

4.3.2 Using classification error rates as test statistics

Let $F(x) = F_0(x) = F_1(x)$ denote the common CDF of X under this null hypothesis. We only consider the case where $F(x)$ is a continuous and strictly increasing function in x starting at 0 for some sufficiently small value of its argument x . For the purpose of deriving the null distributions we assume here that the x_n 's are independent and identically distributed (*iid*) according to F . Putting $u_n = F(x_n)$ the probability integral transform states that the u_n 's are independent and identically uniformly distributed on $[0, 1]$, henceforth abbreviated as *IIUD* $[0, 1]$. Therefore, from equation 4.2-5, putting $b = F(c)$ we have

$$\begin{aligned} \widehat{er}_{up}(c) &= \frac{w_0}{N_0} \sum_{n=1}^N (1 - y_n) I(F(x_n) > F(c)) + \frac{w_1}{N_1} \sum_{n=1}^N y_n I(F(x_n) \leq F(c)) \\ &= \frac{w_0}{N_0} \sum_{n=1}^N (1 - y_n) I(u_n > b) + \frac{w_1}{N_1} \sum_{n=1}^N y_n I(u_n \leq b) \\ &= \widetilde{er}_{up}(b) \end{aligned}$$

Since this expression is only a function of b , minimising over c is equivalent to minimising over b , giving

$$\widehat{er}_{up}^* = \min_c \{\widehat{er}_{up}(c)\} = \min_b \{\widetilde{er}_{up}(b)\}.$$

This expresses $\widehat{er}_{up}^*$ as a function of the u_n 's whose distribution does not depend on F so that the distribution of $\widehat{er}_{up}^*$ also does not depend on F , i.e. $\widehat{er}_{up}^*$ is a non-parametric test statistic. Similarly

$$\begin{aligned} \widehat{er}_{down}(c) &= \frac{w_0}{N_0} \sum_{n=1}^N (1 - y_n) I(u_n \leq b) + \frac{w_1}{N_1} \sum_{n=1}^N y_n I(u_n > b) \\ &= \widetilde{er}_{down}(b) \end{aligned}$$

and again $\widehat{er}_{down}^* = \min_c \{\widehat{er}_{down}(c)\} = \min_b \{\widehat{er}_{down}(b)\}$ which shows that $\widehat{er}_{down}^*$ is also a non-parametric test statistic.

Putting $u'_n = 1 - u_n$ and $b' = 1 - b$, $\widehat{er}_{down}(c)$ can be expressed as

$$\widehat{er}_{down}(c) = \frac{w_0}{N_0} \sum_{n=1}^N (1 - y_n) I(u'_n \geq b') + \frac{w_1}{N_1} \sum_{n=1}^N y_n I(u'_n < b') = \widehat{er}_{down}(b')$$

so that $\widehat{er}_{down}^* = \min_c \{\widehat{er}_{down}(c)\} = \min_{b'} \{\widehat{er}_{down}(b')\}$. The expressions for $\widehat{er}_{down}^*$ and $\widehat{er}_{up}^*$ are therefore the same except for possible equalities which carry zero probability and since the u'_n 's have the same distribution as the u_n 's, it follows that $\widehat{er}_{down}^*$ has the same distribution as $\widehat{er}_{up}^*$.

Finally, under the null hypothesis $\widehat{er}_{min}^* = \min[\min_b \{\widehat{er}_{up}(b)\}, \min_b \{\widehat{er}_{down}(b)\}]$ which expresses $\widehat{er}_{min}^*$ as a function of the u_n 's as well so that it too is a non-parametric test statistic.

4.3.3 Calculation of the null distributions by simulation

The supplementary material provides five easy steps following steps can be used to calculate the null distribution of $\widehat{er}_{up}^*$:

1. Generate N *IIUD*[0, 1] u_n 's.
2. Assign the first N_0 y_n 's as 0 and the remainder as 1.
3. Minimize $\frac{w_0}{N_0} \sum_{n=1}^N (1 - y_n) I(u_n > b) + \frac{w_1}{N_1} \sum_{n=1}^N y_n I(u_n \leq b)$ over an adequate grid of b values (as previously discussed) to obtain $\widehat{er}_{up}^*$.
4. Repeat these steps M times to build up a file of *iid* copies of $\widehat{er}_{up}^*$, say $\widehat{er}_{up}^*(m)$, $m = 1, \dots, M$. The empirical CDF of the $\widehat{er}_{up}^*(m)$'s then provides a simulation approximation of the CDF of the null distribution.
5. If T of the $\widehat{er}_{up}^*(m)$'s fall below an actually observed $\widehat{er}_{up}^*$ its associated p -value is approximately T/M . These approximations can be made accurate by taking M large enough.

Similarly for $\widehat{er}_{down}^*$ and $\widehat{er}_{min}^*$.

4.3.4 Asymptotic estimation of the null distribution of the directional full sample error rate test statistic

A large sample approximation of the null distribution can be obtained by letting:

$$F_{N_0}(b) = \frac{1}{N_0} \sum_{n=1}^N (1 - y_n) I(u_n \leq b) \text{ and } F_{N_1}(b) = \frac{1}{N_1} \sum_{n=1}^N y_n I(u_n \leq b)$$

denote the respective empirical CDFs in the argument b of the u_n 's corresponding to the control and experimental groups, then $\widehat{e\tau}_{up}(b)$ may be written as:

$$\widehat{e\tau}_{up}(b) = w_0 [1 - F_{N_0}(b)] + w_1 F_{N_1}(b).$$

It is well known that when sample sizes increase $\sqrt{N_0}\{F_{N_0}(b) - b\}$, taken as a stochastic process in b , tends in distribution to a standard Brownian bridge (BB) process and the same holds for $\sqrt{N_1}\{F_{N_1}(b) - b\}$ (Breiman, 1968; Dudley, 1999). Now:

$$\widehat{e\tau}_{up}(b) = w_0 + b(w_1 - w_0) + \frac{w_1}{\sqrt{N_1}} \sqrt{N_1}\{F_{N_1}(b) - b\} - \frac{w_0}{\sqrt{N_0}} \sqrt{N_0}\{F_{N_0}(b) - b\}$$

and therefore

$$\widehat{e\tau}_{up}(b) \approx_D w_0 + b(w_1 - w_0) + \frac{w_1}{\sqrt{N_1}} B_1 - \frac{w_0}{\sqrt{N_0}} B_0$$

where B_0 and B_1 are two independent standard BB's and $\approx_D$ denotes approximation in distribution.

Since a linear combination of independent BBs is again a BB, the difference of the last two terms in the above equation can also be approximated by another BB, denoted by $aB(b)$ with $B(b)$ a standard BB and $a = \sqrt{\frac{w_0^2}{N_0} + \frac{w_1^2}{N_1}}$.

The above equation can then be restated as: $\widehat{e\tau}_{up}(b) \approx_D w_0 + b(w_1 - w_0) + aB(b)$.

Consequently, under the null hypothesis, the distribution of $\widehat{e\tau}_{up}^* = \min_d \{\widehat{e\tau}_{up}(d)\}$ can be approximated by:

$$\widehat{e\tau}_{up}^* \approx_D w_0 + \min_b \{b(w_1 - w_0) + aB(b)\}.$$

The stochastic process $\widehat{e}r_{up}(b)$ is a Brownian bridge **with drift** $b(w_1 - w_0)$ and to the best of our knowledge an explicit expression for the CDF of its minimum, i.e. the CDF of $\widehat{e}r_{up}^*$, is not known.

However, it is easy to calculate it by simulation using the following steps:

1. Generate $B(b)$ over a fine grid of d values.
2. Find the minimum value of $\{b(w_1 - w_0) + aB(b)\}$ over the grid.
3. Repeat these two steps M times to build up a file of *iid* copies of the minimum values.
4. Calculate the corresponding CDF, providing an asymptotic approximation of the null distribution.
5. Observed values of $\widehat{e}r_{up}^*$ or $\widehat{e}r_{down}^*$ can again be compared to the file entries to get asymptotic approximations of their p -values.

Assuming the costs of misclassifying subjects are equal for each group, i.e. $w_0 = w_1 = \frac{1}{2}$, there is no longer a drift present. Now the associated asymptotic p -values can be calculated without simulation since the relevant distribution is known. In this case:

$$\widehat{e}r_{up}^* \approx_D \frac{1}{2} + \sqrt{N} \min_b \{B(b)\} / 2\sqrt{N_0 N_1}$$

$$\text{and therefore: } P(\widehat{e}r_{up}^* \leq x) \approx P(\min_b \{B(b)\} \leq 2\sqrt{N_0 N_1}(x - \frac{1}{2})/\sqrt{N}).$$

Now, the CDF of $\min_d \{B(b)\}$ is given by the simple formula

$P(\min_d \{B(b)\} \leq t) = \exp(-2t^2)$ for $t \leq 0$, refer to Breiman (1968) and Dudley (1999), so that we end up with the simple approximation:

$$P(\widehat{e}r_{up}^* \leq x) \approx \exp\{-8 \frac{N_0 N_1}{N} \left(x - \frac{1}{2}\right)^2\}$$

It is evident from the expression for $\widehat{e}r_{up}^*$ that the numbers N_0 and N_1 and the weights w_0 and w_1 act on the null distribution in large samples through the two quantities: $(w_1 - w_0)$ and $a = \sqrt{w_0^2/N_0 + w_1^2/N_1}$. In particular, if N_0 and N_1 are both large while w_0 and w_1 differ substantially, then the stochastic part $aB(b)$ has little influence since the factor a is small.

This simulation approximation has the advantage that its computation time does not increase for increasing sample sizes. However, this comes at the cost of two approximations being involved, namely the simulation approximation as well as the large sample approximation.

Assuming equal weights simplifies the approach even further. To see this, substitute $w_0 = w_1 = \frac{1}{2}$ into 4.2-5 and 4.2-6 and write the sums in terms of empirical CDFs $F_{N_0}(c)$ and $F_{N_1}(c)$ to get:

$$\widehat{er}_{up}(c) = \frac{1}{2} + \frac{1}{2}[F_{N_1}(c) - F_{N_0}(c)] \text{ and}$$

$$\widehat{er}_{down}(c) = \frac{1}{2} + \frac{1}{2}[F_{N_0}(c) - F_{N_1}(c)].$$

Consequently:

$$\widehat{er}_{up}^* = \frac{1}{2} + \frac{1}{2} \min_c [F_{N_1}(c) - F_{N_0}(c)] = \frac{1}{2} - \frac{1}{2} \max_c [F_{N_0}(c) - F_{N_1}(c)] = \frac{1}{2}(1 - D^+)$$

$$\widehat{er}_{down}^* = \frac{1}{2} + \frac{1}{2} \min_c [F_{N_0}(c) - F_{N_1}(c)] = \frac{1}{2} - \frac{1}{2} \max_c [F_{N_1}(c) - F_{N_0}(c)] = \frac{1}{2}(1 - D^-)$$

where D^+ and D^- are the right and left one-sided two-sample Kolmogorov-Smirnov (K-S) test statistics for comparing the experimental with the control group (Breiman, 1968).

The exact null distributions of the K-S statistics are available in standard statistical packages such as SAS (SAS Institute Inc., 2011) from which the relevant p -values may be found. Note that when D^+ and D^- are multiplied by the factor $\sqrt{N_0 N_1 / (N_0 + N_1)}$ they have the asymptotic CDF $H(x) = 1 - \exp(-2x^2), x \geq 0$, from which the asymptotic formula for the p -values again follows.

4.3.5 Tests based on leave-one-out error rates

Consider first the amendments required for the error rate of the upward rule. Let $\widehat{er}_{up}(c, i)$ denote the error rate in 4.2-5, but with the i^{th} subject excluded from the summations on the right hand side of the equation. Also let $\hat{c}_{up}(i)$ denote the threshold level c that minimizes $\widehat{er}_{up}(c, i)$ over c and let $\widehat{er}_{up}^*(i)$ denote the minimized value of $\widehat{er}_{up}(c, i)$. Using the upward rule the i^{th} subject is classified as an experimental subject if $x_i > \hat{c}_{up}(i)$ and as a control subject if $x_i \leq \hat{c}_{up}(i)$.

Weighting and summing over all the excluded cases, the LOO error rate for the upward rule can be expressed as

$$\hat{el}_{up} = \frac{w_0}{N_0} \sum_{i=1}^N (1 - y_i) I(x_i > \hat{c}_{up}(i)) + \frac{w_1}{N_1} \sum_{i=1}^N y_i I(x_i \leq \hat{c}_{up}(i))$$

One generally finds that $\hat{el}_{up}$ is somewhat larger than $\hat{er}_{up}^*$ which is to be anticipated since x_i is not able to influence the choice of $\hat{c}_{up}(i)$ and this leads to reduction in bias of $\hat{el}_{up}$ as an estimator of er_{up}^* as compared to $\hat{er}_{up}^*$. For the downward shift rule $\hat{el}_{down}$ is defined analogously. There is more than one way to define the minimum LOO error rate. However, we restrict our discussion to the simplest choice, namely to take $\hat{el}_{min} = \min\{\hat{el}_{up}, \hat{el}_{down}\}$ and the associated direction as $\hat{dl}_{min} = "up"$ if $\hat{el}_{up} < \hat{el}_{down}$ and $\hat{dl}_{min} = "down"$ otherwise.

The test statistics derived above are all non-parametric, with $\hat{el}_{up}$ and $\hat{el}_{down}$ having the same distribution. To see this, consider first $\hat{el}_{up}$ and put $u_n = F(x_n)$ and $b = F(c)$ as before, so that $\hat{el}_{up}$ can be expressed as

$$\hat{el}_{up} = \frac{w_0}{N_0} \sum_{i=1}^N (1 - y_i) I(u_i > F(\hat{c}_{up}(i))) + \frac{w_1}{N_1} \sum_{i=1}^N y_i I(u_i \leq F(\hat{c}_{up}(i)))$$

Suppose firstly that $y_i = 0$, i.e. subject i is in the control group. Then

$$\begin{aligned} \hat{er}_{up}(c, i) &= \frac{w_0}{N_0 - 1} \sum_{n=1, n \neq i}^N (1 - y_n) I(x_n > c) + \frac{w_1}{N_1} \sum_{n=1}^N y_n I(x_n \leq c) \\ &= \frac{w_0}{N_0 - 1} \sum_{n=1, n \neq i}^N (1 - y_n) I(u_n > b) + \frac{w_1}{N_1} \sum_{n=1}^N y_n I(u_n \leq b) \end{aligned}$$

Minimising over c is equivalent to minimising over $b = F(c)$ so that $F(\hat{c}_{up}(i)) = \hat{b}_{up}(i)$ which minimises $\hat{er}_{up}(c, i)$ and is therefore only a function of the u_n 's (other than the u_i 's). Similarly for the case $y_i = 1$. Hence $\hat{el}_{up}$ can be restated as

$$\hat{el}_{up} = \frac{w_0}{N_0} \sum_{i=1}^N (1 - y_i) I(u_i > \hat{b}_{up}(i)) + \frac{w_1}{N_1} \sum_{i=1}^N y_i I(u_i \leq \hat{b}_{up}(i))$$

which shows that $\widehat{el}_{up}$ is also only a function of the u_n 's and its null distribution therefore does not depend on the true common CDF F under the null hypothesis. A similar argument holds for $\widehat{el}_{down}$. It can also be seen that the expressions involved in the upward and downward cases differ only on sets carrying zero probability under the null hypothesis assumption. Finally, $\widehat{el}_{min}$ can also be expressed as a function of the u_n 's and is therefore also a non-parametric test statistic.

4.3.6 Summary of the ERp Approach

Here we describe the ERp approach as a variable selection and binary classification method. The method was programmed separately in SAS (2011) and MATLAB (2012b) and their outputs were compared to ensure the accuracy and reliability of the reported results. The MATLAB scripts are available in the ZIP folder accompanying the electronic version of this thesis or on the compact disc accompanying the printed version of this thesis.

As depicted in Figure 4-5, the following inputs are required: (i) A training data matrix $\mathbf{X}$ of which the rows provide the measurements for each subject and the columns contain the observed values of any number of non-categorical variables as potential discriminators. (ii) The variable names in a separate vector V . (iii) A categorical vector Y whose entries indicate the group membership of the subjects in each row of $\mathbf{X}$, with 0 representing the control group and 1 the experimental group. (iv) The cost of misclassification into one group relative to the other, i.e. the weight pair w_0 and w_1 specified, as separate variables, such that the pair sums to 1. (v) A vector D containing the shift direction of interest for each variable named in V . Therefore, D specifies the test statistic to be used and assumes the following values: (i) The value 1 indicates that an upward shift is of interest (i.e. represents the alternative hypothesis) requiring the use of $\widehat{er}_{up}^*$ as test statistic; (ii) The value -1 indicates that a downward shift is of interest (i.e. represents the alternative hypothesis) requiring the use of $\widehat{er}_{down}^*$ as test statistic; (iii) The value -1 indicates that a downward shift is of interest (i.e. represents the alternative hypothesis) requiring the use of $\widehat{er}_{down}^*$ as test statistic; (vi) Lastly, the value of the preferred FWER α as a separate variable.

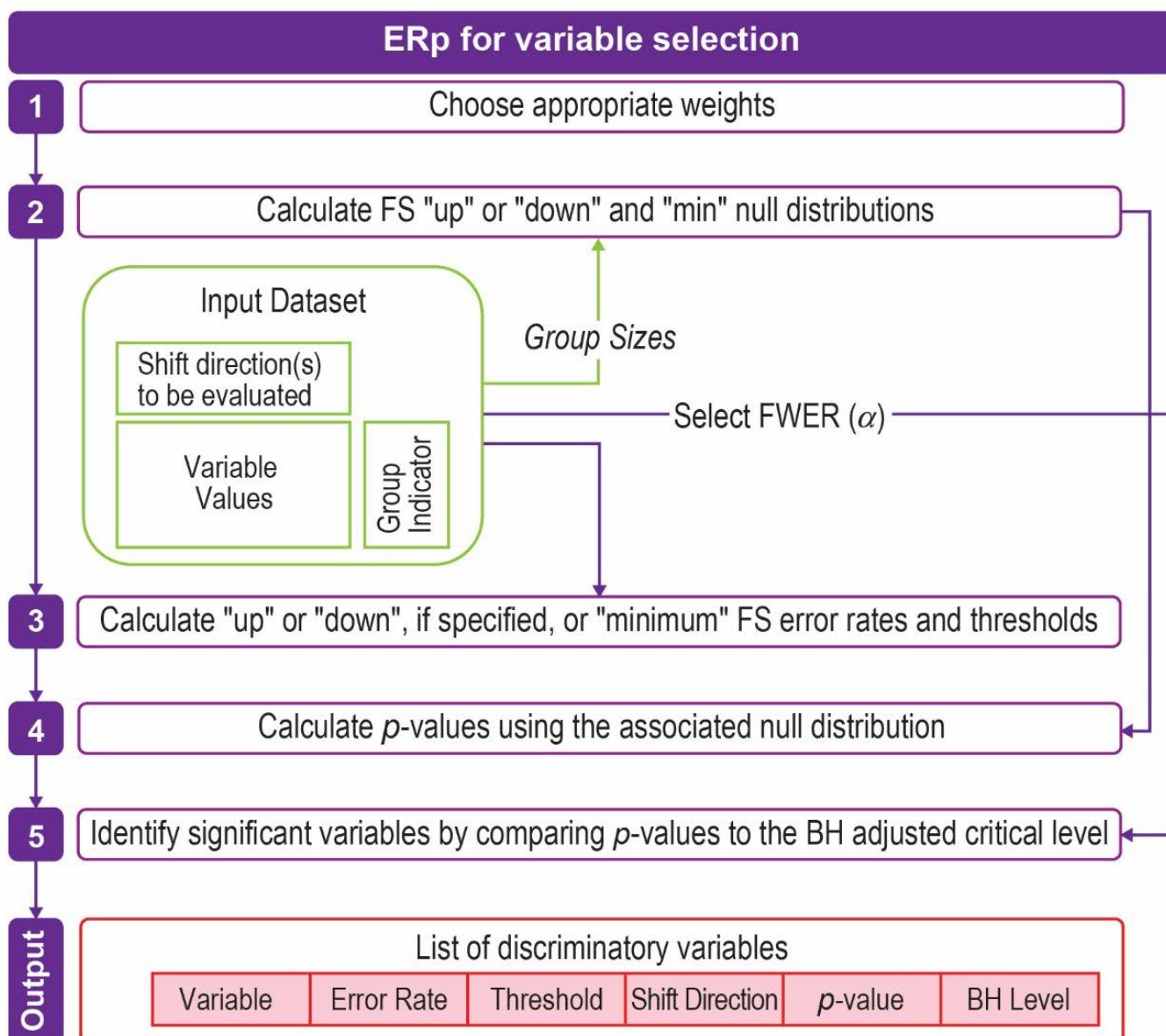


Figure 4-5: The ERp method for variable selection.

Once these inputs are presented to the program, the number of observations within each group is determined (N_0 and N_1) and used along with the weight pair (w_0 and w_1) to generate the null distributions required, depending on the shift direction indicated. As mentioned, ERp based on the FS error rates is more powerful so that only the FS null distributions are generated. Note that this is done only once regardless of the number of variables since the null distributions depend

only N_0 , N_1 , w_0 and w_1 . Next the desired FS minimised error rate as well as the corresponding threshold value is calculated for each variable. Minimised error rates are converted to their corresponding p -values by referencing the appropriate null distribution. Finally, the list of significantly shifted variables is produced with cut-off provided by the BH method to control the FWER at the level α . This list may then be used for biological interpretation.

In summary, the algorithm produces a table containing: (a) the list of names of selected variables; (b) the error rate used to test the specified directional hypothesis, i.e. $\widehat{er}_{up}^*$, if an upward shift was specified, $\widehat{er}_{down}^*$, if a downward shift was specified, or $\widehat{er}_{min}^*$, if no shift direction was specified; (c) the threshold associated with each error rate; (d) the shift direction as specified or, if no direction was specified, the direction assigned by ERp ("up" if $\widehat{er}_{up}^* > \widehat{er}_{down}^*$ or "down" if $\widehat{er}_{up}^* < \widehat{er}_{down}^*$); (e) the p -value associated with the error rate; and (f) the BH critical levels for multiple testing to control the FWER.

If variables expected to be significantly shifted, based on subject knowledge, are not found in the list, one may look at the sensitivity of the list when the FWER α is varied. This may be especially relevant when we have small sample sizes and thus insufficient power to detect small shift effects. If the sample size is larger, one should investigate alternative reasons before compromising the preferred FWER. After the researcher is comfortable that the results are biologically relevant, clinical consideration can be taken into account to select the final list used to classify new subjects.

Each listed variable can provide a classification of a new subject using the corresponding threshold value and shift direction (provided or calculated). If the variables are not unanimous in their classification we use a majority vote. However, this implies giving equal credence to all variables in the list, while variables with smaller p -values should be considered better discriminators. It may be more reasonable to use a summary classifier that takes this into account. We will not discuss such a 'weighted vote' in this paper, but will address it in future research.

4.3.7 Comparing null distributions

In this section we graphically compare the three approaches to calculating the null CDF, i.e. FS, LOO and Asymptotic approach. To gain some insight into how each approach differs given various combinations of group sizes, weight selections and shift assumptions, we plot $\widehat{er}_{up}^*$, $\widehat{er}_{min}^*$, $\widehat{el}_{up}$ and $\widehat{el}_{min}$ as well as the asymptotic approximation of the null distribution for the nine scenarios outlined in Table 4-3. Note that these scenarios are used throughout this supplementary to illustrate findings described in the paper.

Scenario		Control Group		Experimental Group	
		Group Size	Weight	Group Size	Weight
1) $w_0 = w_1$	(a) $N_0 = N_1$	20	$\frac{1}{2}$	20	$\frac{1}{2}$
	(b) $N_0 < N_1$	10	$\frac{1}{2}$	20	$\frac{1}{2}$
	(c) $N_0 > N_1$	20	$\frac{1}{2}$	10	$\frac{1}{2}$
2) $w_0 > w_1$	(a) $N_0 = N_1$	20	$\frac{3}{4}$	20	$\frac{1}{4}$
	(b) $N_0 < N_1$	10	$\frac{3}{4}$	20	$\frac{1}{4}$
	(c) $N_0 > N_1$	20	$\frac{3}{4}$	10	$\frac{1}{4}$
3) $w_0 < w_1$	(a) $N_0 = N_1$	20	$\frac{1}{4}$	20	$\frac{3}{4}$
	(b) $N_0 < N_1$	10	$\frac{1}{4}$	20	$\frac{3}{4}$
	(c) $N_0 > N_1$	20	$\frac{1}{4}$	10	$\frac{3}{4}$

Table 4-3: List of Scenarios.

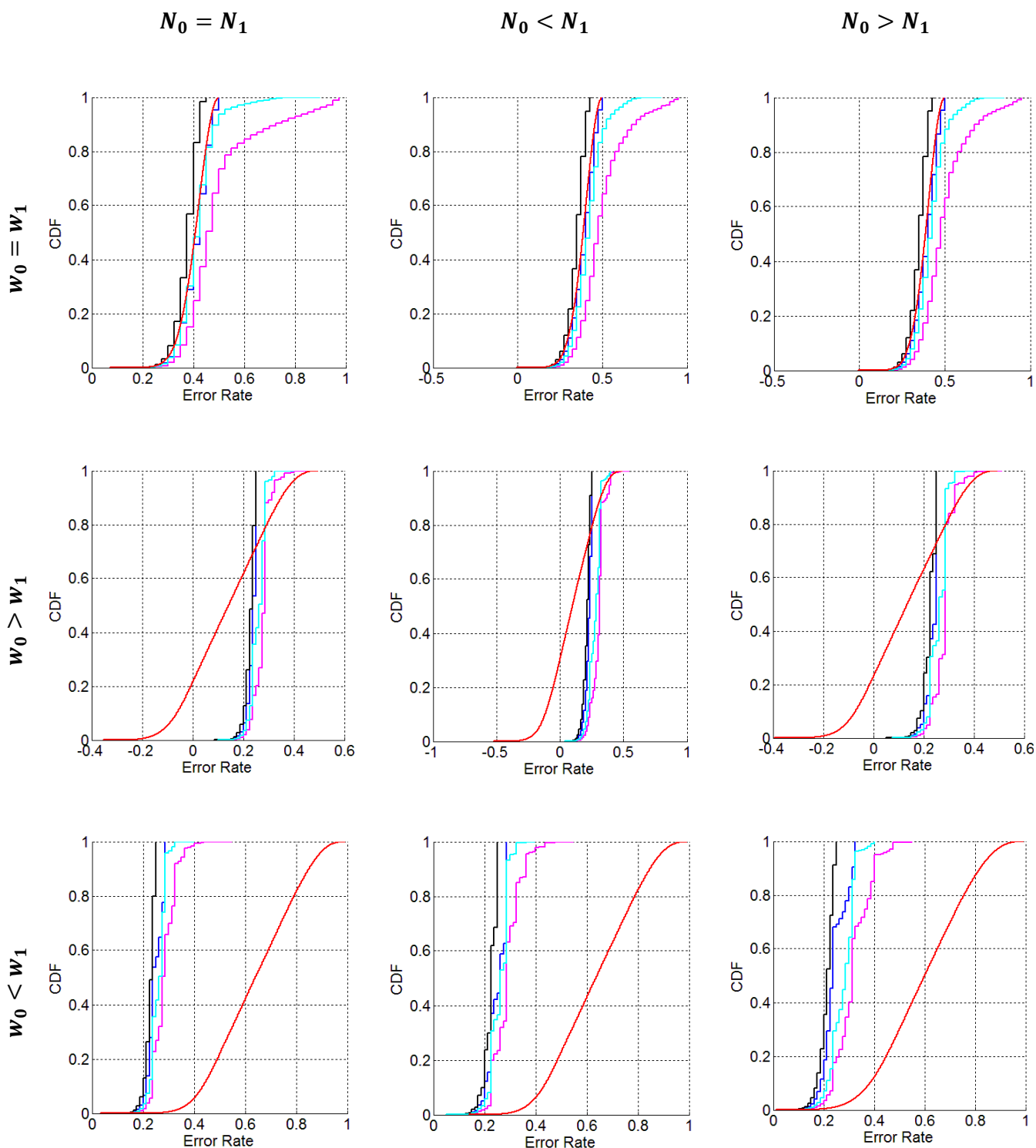


Figure 4-6: A graphically comparison of approaches to calculating the null CDF.

Each image represents a scenario from Table 4-3. The blue and black lines represent the FS null CDF's for $\hat{e}r_{up}^*$ and $\hat{e}r_{min}^*$, respectively. The magenta and cyan lines represent the LOO null CDF's for $\hat{e}l_{up}$ and $\hat{e}l_{min}$, respectively. Finally, the red lines represent the asymptotic approximation of the null CDF.

Figure 4-6 shows how the null CDFs of the FS and LOO error rates starts to turn sharply as the error rate increases to the smaller of the two weights. For equal weights this turn is less sharp and the null CDF is close to 1 when the error rate is close to 0.5. In general, LOO CDF reaches the values of the corresponding FS CDF at larger error rates. It is also apparent that the asymptotic approximation is only appropriate when weights are equal. It is therefore evident from Figure 4-6 that the shape of the null distribution, regardless of the ERp approach (LOO or FS) selected to derive it, depends on the selected weights as well as the group sizes. This proves that the ERp approach accounts for group sizes and selected weight sets. As discussed in the main paper, the discriminatory significance of a variable becomes more apparent when its error rate is evaluated based on its corresponding p -value.

However, as is apparent from Figure 4-6, the p -values for different error rates will differ even if their values are equal. To decide which error rate (FS or LOO) performs better (i.e. has the smallest p -value) we need to investigate the behaviour of the tails of their distributions in greater detail. Section 4.3.8 does this through a power comparison.

4.3.8 Power comparisons of test statistics

This section compares the power of the following four error rate estimators or test statistics: $\hat{er}_{up}^*$; $\hat{er}_{min}^*$; $\hat{el}_{up}$ and $\hat{el}_{min}$. That is, when an upward shift is assumed and explicitly tested ($\hat{er}_{up}^*$ and $\hat{el}_{up}$), as well as when no assumption is made and both shift directions are evaluated ($\hat{er}_{min}^*$ and $\hat{el}_{min}$). Furthermore, we also compare our four error rate estimators or test statistics to the Mann-Whitney (MW) test statistic as it represent a standard non-parametric two-sample test which is known to be powerful.

The scenarios listed in Table 4-3 were evaluated under three different distributional assumptions to incorporate some of the known characteristics of metabolomics data.

Scenario 1: Values for control subjects followed a $N(0,1)$ distribution while those of experimental subject where drawn from a $N(\mu, 1)$ distribution where μ assumed different values.

Scenario 2: Values for control subjects followed a $LN(0,1)$ distribution while those of experimental subject where drawn from a $LN(\mu, 1)$ distribution where μ assumed different values.

Scenario 3: Values for control subjects followed a $\Gamma(0,1,1)$ distribution while those of experimental subject were drawn from a $\Gamma(0, \beta, 1)$ distribution where β (the scale parameter) assumed different values.

As a first comparison, the resulting p -values were averaged over ten thousand repetitions to measure the expected power of the test statistics. As a second measure, we calculated the proportion of the ten thousand p -values resulting from each simulation for the test statistics $\widehat{er}_{min}^*$, $\widehat{el}_{up}$, $\widehat{el}_{min}$ and MW which turned out to be lower than those of $\widehat{er}_{up}^*$, at each shift magnitude. These proportions were evaluated against a 50% threshold which if exceeded indicates that this particular error rate estimator or test statistic outperforms $\widehat{er}_{up}^*$.

Legend for Figures 4-7 to 4-15:

The left panel shows the average p -values associated with $\widehat{er}_{up}^*$ (red lines), $\widehat{er}_{min}^*$ (blue lines), $\widehat{el}_{up}$ (magenta lines), $\widehat{el}_{min}$ (cyan lines) as well as the MW test statistic (black lines). The right panel shows the proportion of $\widehat{er}_{min}^*$ (blue lines), $\widehat{el}_{up}$ (magenta lines), $\widehat{el}_{min}$ (cyan lines) as well as the MW (black lines) p -value below those associated with $\widehat{er}_{up}^*$. The dotted red line represents the 50% cut-off. Each figure has three rows representing the distributional assumptions outlined above.

From Figure 4-7 it is evident $\widehat{er}_{up}^*$ is on average the most powerful or at least as powerful as the MW test statistic. Furthermore, $\widehat{er}_{up}^*$ also has a high probability of being more powerful than $\widehat{el}_{up}$, $\widehat{er}_{min}^*$, $\widehat{el}_{min}$ and MW for larger shifts. The MW is the only test statistic that outperforms $\widehat{er}_{up}^*$ and only does so for shifts below 1 given a gamma distribution, which are known to indicate insignificant shifts in the distribution of the experimental group.

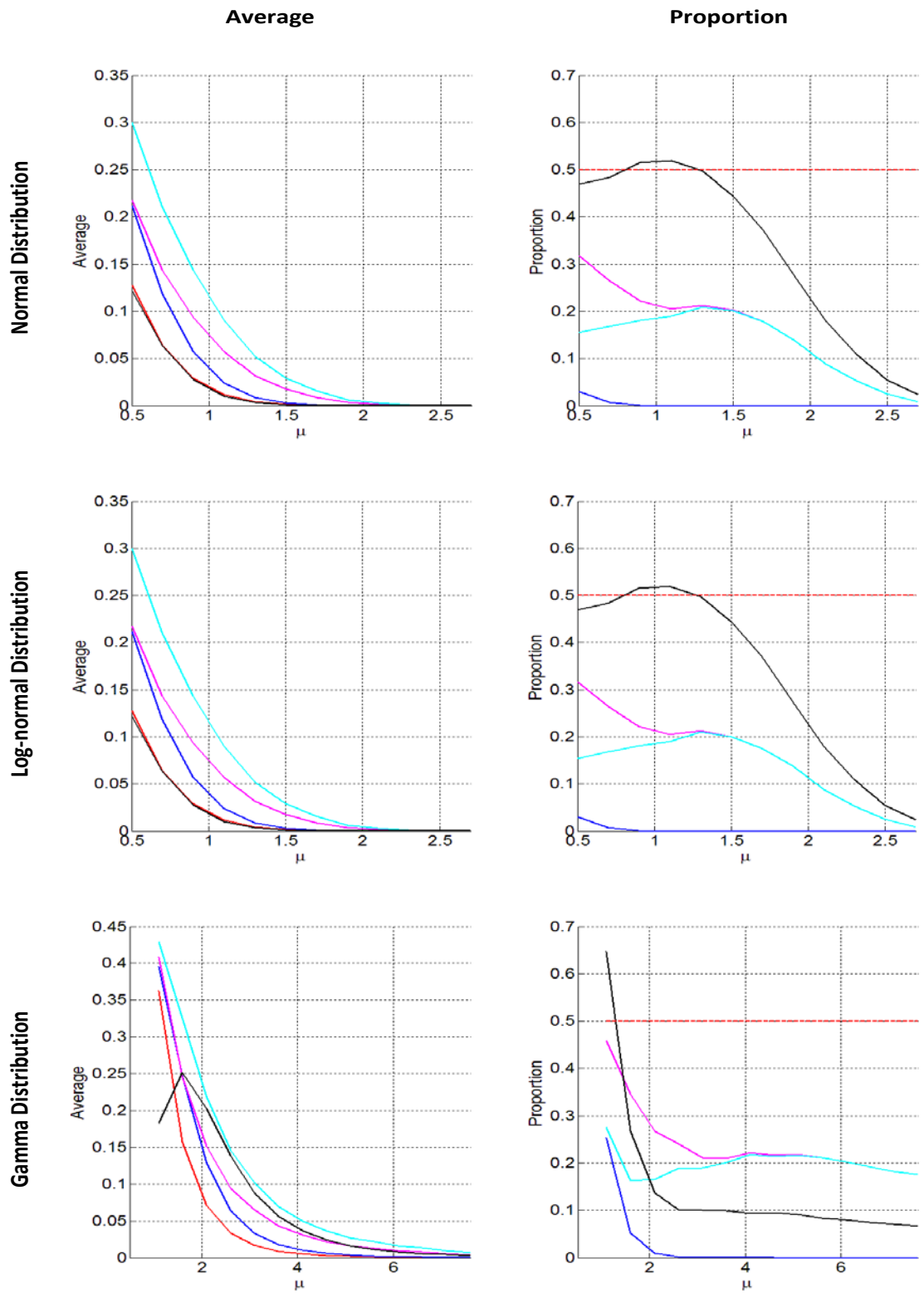


Figure 4-7: Power comparison between test statistics for scenario 1a: $N_0 = 20$ and $N_1 = 20$ with $w_0 = \frac{1}{2}$ and $w_1 = \frac{1}{2}$.

Figure 4-8 shows the results of the power comparison between test statistics for $N_0 = 10$ and $N_1 = 20$ with $w_0 = \frac{1}{2}$ and $w_1 = \frac{1}{2}$. From Figure 4-8 it is evident $\widehat{er}_{up}^*$ is on average the most powerful or at least as powerful as the MW test statistic. Furthermore, $\widehat{er}_{up}^*$ also has a high probability of being more powerful than $\widehat{el}_{up}$, $\widehat{er}_{min}^*$, $\widehat{el}_{min}$ and MW for larger shifts. The MW is the only test statistic that outperforms $\widehat{er}_{up}^*$ and only does so for shifts below 1 given a gamma distribution, which are known to indicate insignificant shifts in the distribution of the experimental group.

Figure 4-9 shows the results of the power comparison between test statistics for $N_0 = 20$ and $N_1 = 10$ with $w_0 = \frac{1}{2}$ and $w_1 = \frac{1}{2}$. From Figure 4-9 it is evident $\widehat{er}_{up}^*$ is on average the most powerful or at least as powerful as the MW test statistic. Furthermore, $\widehat{er}_{up}^*$ also has a high probability of being more powerful than $\widehat{el}_{up}$, $\widehat{er}_{min}^*$, $\widehat{el}_{min}$ and MW for larger shifts. The MW is the only test statistic that outperforms $\widehat{er}_{up}^*$ and only does so for shifts below 1 given a gamma distribution, which are known to indicate insignificant shifts in the distribution of the experimental group.

Figure 4-10 shows results of the power comparison between test statistics for $N_0 = 20$ and $N_1 = 20$ with $w_0 = \frac{3}{4}$ and $w_1 = \frac{1}{4}$. From Figure 4-10 it is evident the MW test statistic is on average the most powerful given a normal or log-normal distribution, while $\widehat{er}_{up}^*$ is on average the most powerful when assuming a gamma distribution. MW has a high probability of being the most powerful for small to medium shifts. The $\widehat{er}_{up}^*$ test statistic is more probable to outperform MW for larger shifts in distribution. However, it is important to note that the MW test cannot account for the weight selection which is no longer equal.

Figure 4-11 shows results of the power comparison between test statistics for $N_0 = 10$ and $N_1 = 20$ with $w_0 = \frac{3}{4}$ and $w_1 = \frac{1}{4}$. From Figure 4-11 it is evident the MW test statistic is on average the most powerful given a normal or log-normal distribution, while $\widehat{er}_{up}^*$ is on average the most powerful when assuming a gamma distribution. MW has a high probability of being the most powerful for small to medium shifts. The $\widehat{er}_{up}^*$ test statistic is more probable to outperform MW for larger shifts in distribution. However, it is important to note that the MW test cannot account for the weight selection which is no longer equal.

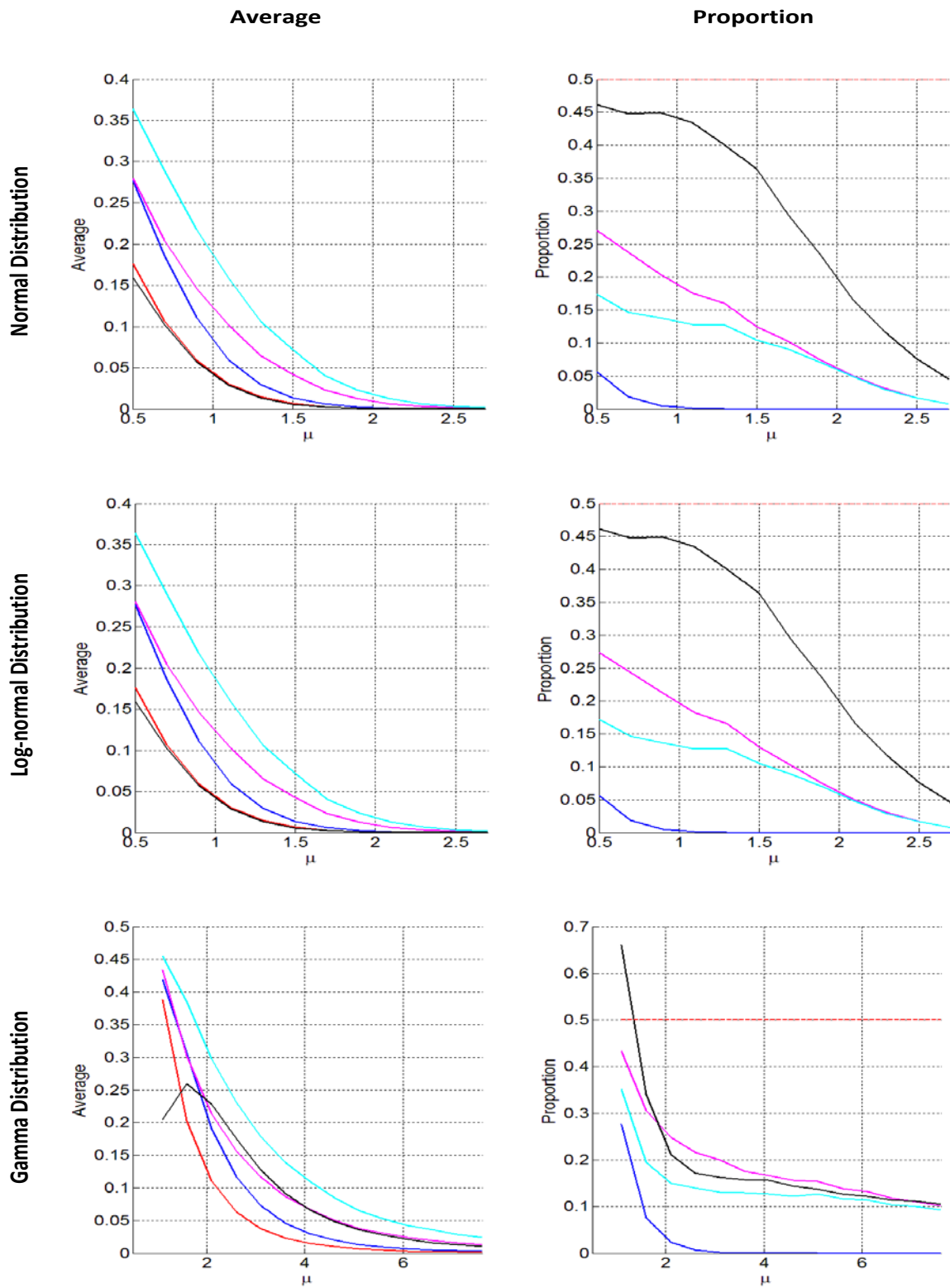


Figure 4-8: Power comparison between test statistics for scenario 1b: $N_0 = 10$ and $N_1 = 20$ with $w_0 = \frac{1}{2}$ and $w_1 = \frac{1}{2}$.

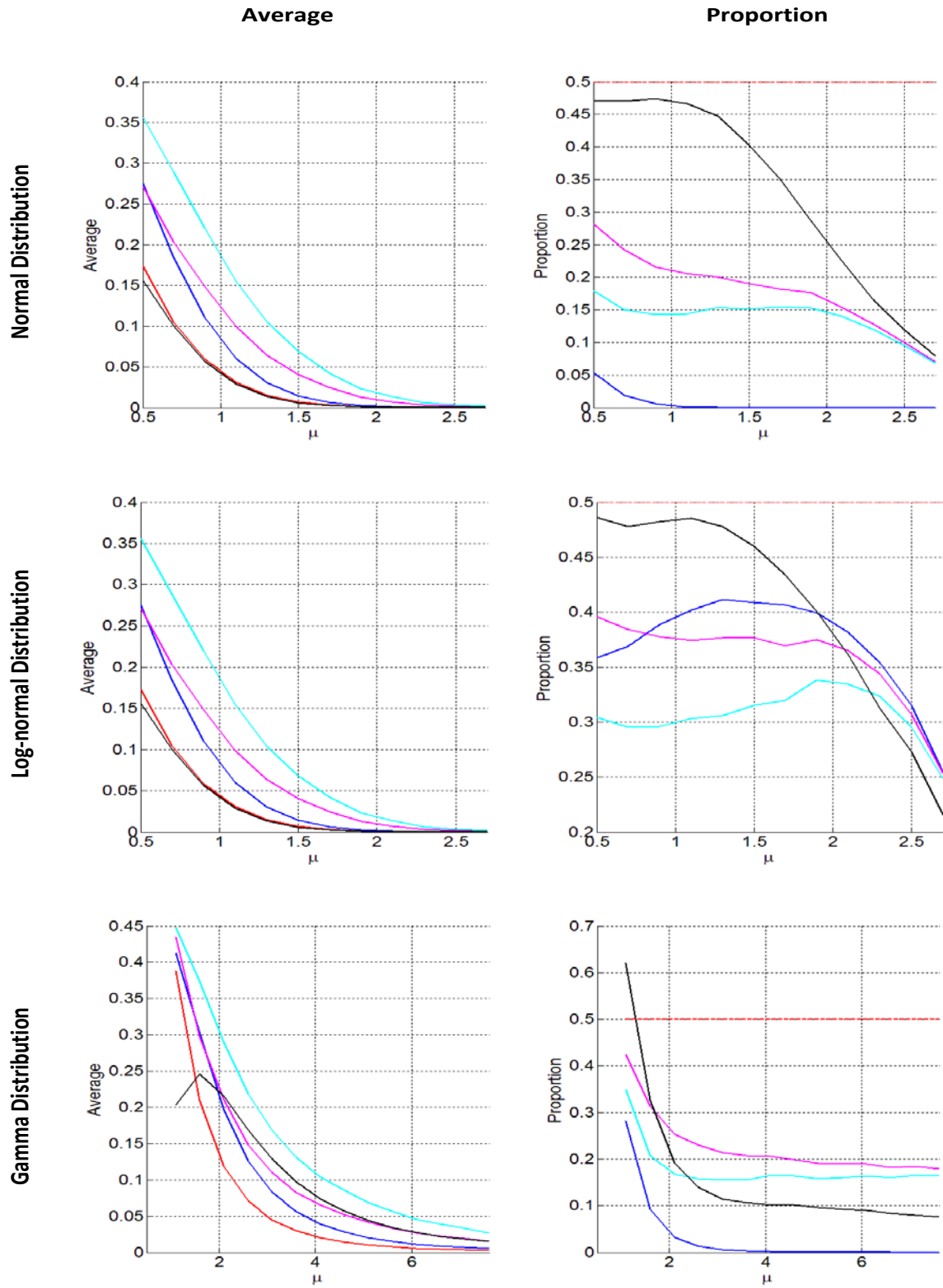


Figure 4-9: Power comparison between test statistics for scenario 1c: $N_0 = 20$ and $N_1 = 10$ with $w_0 = \frac{1}{2}$ and $w_1 = \frac{1}{2}$.

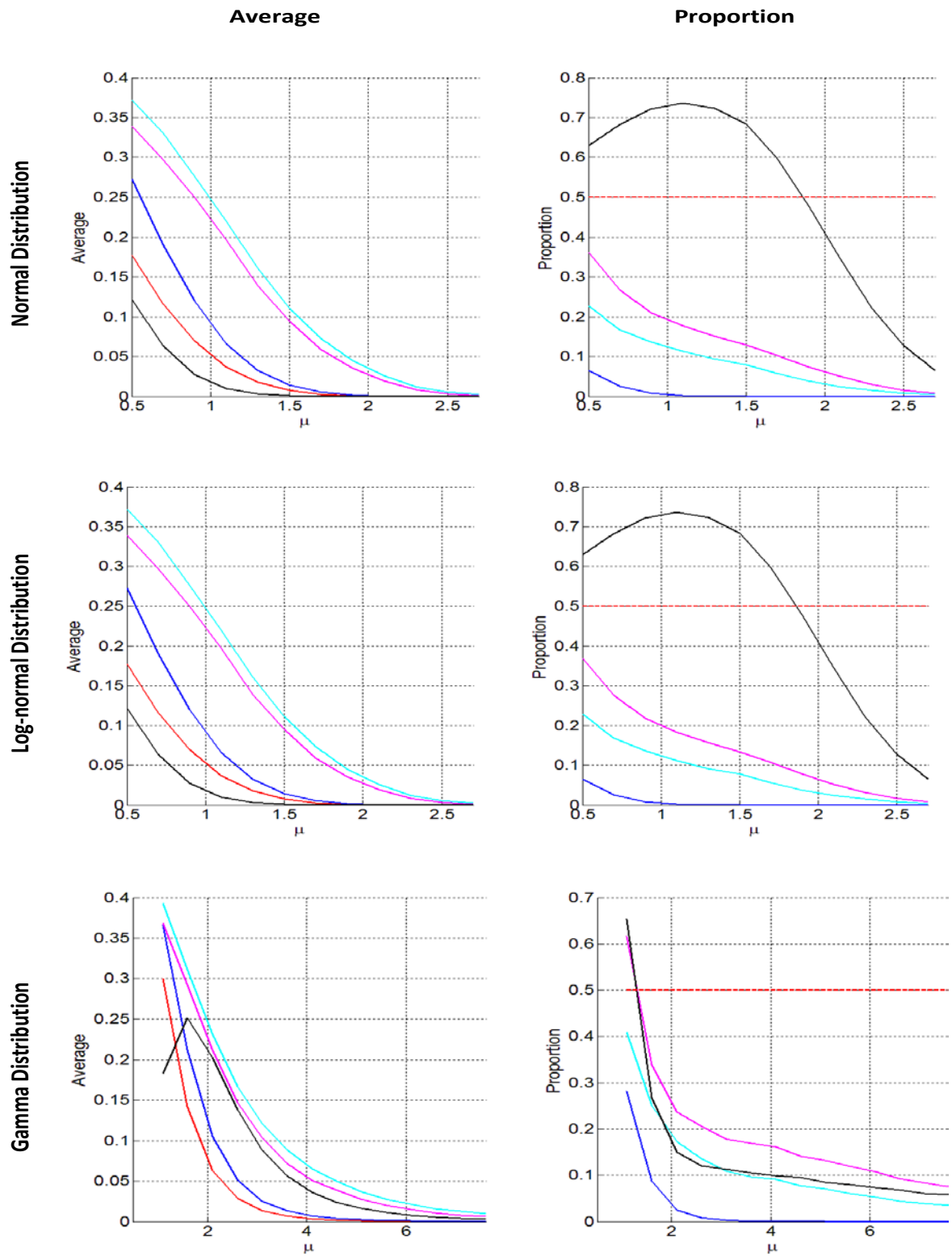


Figure 4-10: Power comparison between test statistics for scenario 2a: $N_0 = 20$ and $N_1 = 20$ with $w_0 = \frac{3}{4}$ and $w_1 = \frac{1}{4}$.

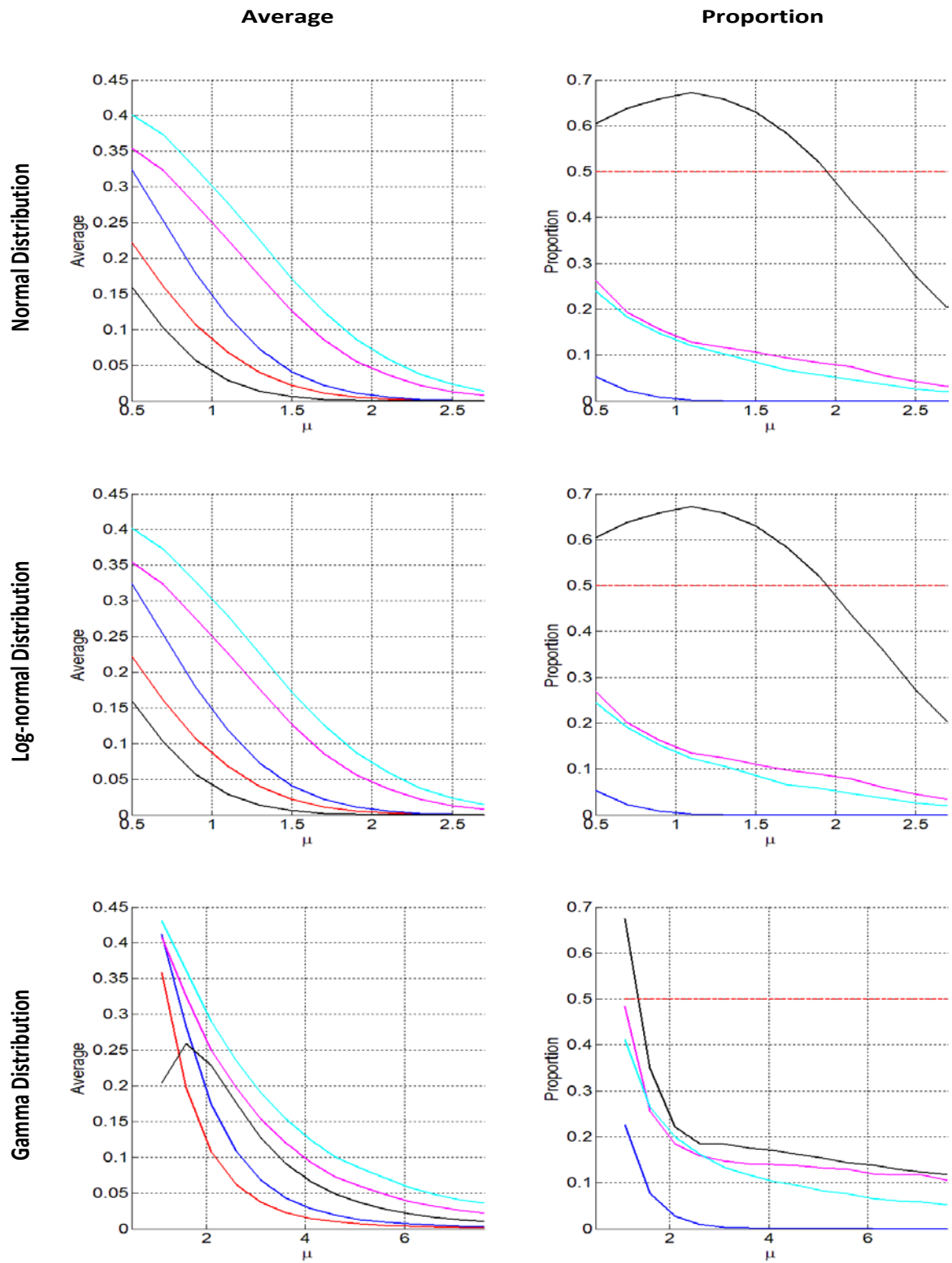


Figure 4-11: Power comparison between test statistics for scenario 2b: $N_0 = 10$ and $N_1 = 20$ with $w_0 = \frac{3}{4}$ and $w_1 = \frac{1}{4}$.

Figure 4-12 shows results of the power comparison between test statistics for $N_0 = 20$ and $N_1 = 10$ with $w_0 = \frac{3}{4}$ and $w_1 = \frac{1}{4}$. From Figure 4-12 it is evident the MW test statistic is on average the most powerful given a normal or log-normal distribution, while $\widehat{er}_{up}^*$ is on average the most powerful when assuming a gamma distribution. MW has a high probability of being the most powerful for small to medium shifts. The $\widehat{er}_{up}^*$ test statistic is more probable to outperform MW for larger shifts in distribution. However, it is important to note that the MW test cannot account for the weight selection which is no longer equal.

Figure 4-13 shows results of the power comparison between test statistics for $N_0 = 20$ and $N_1 = 20$ with $w_0 = \frac{1}{4}$ and $w_1 = \frac{3}{4}$. From Figure 4-13 it is evident the MW test statistic is on average the most powerful. MW has a high probability of being the most powerful for small to medium shifts. The $\widehat{er}_{up}^*$ test statistic is more probable to outperform MW for larger shifts in distribution given a normal or log-normal distribution. However, it is important to note that the MW test cannot account for the weight selection which is no longer equal.

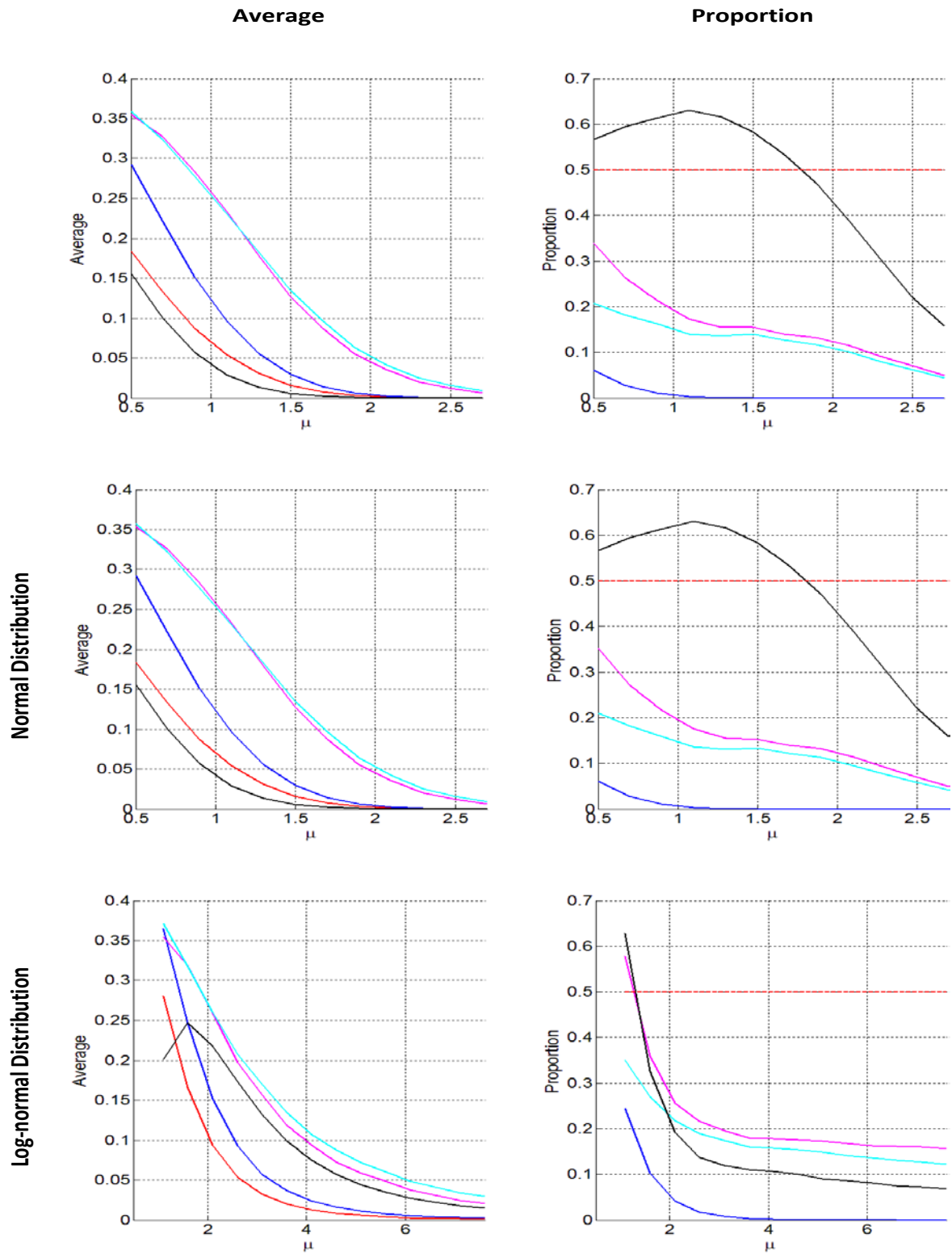


Figure 4-12: Power comparison between test statistics for scenario 2c: $N_0 = 20$ and $N_1 = 10$ with $w_0 = \frac{3}{4}$ and $w_1 = \frac{1}{4}$.

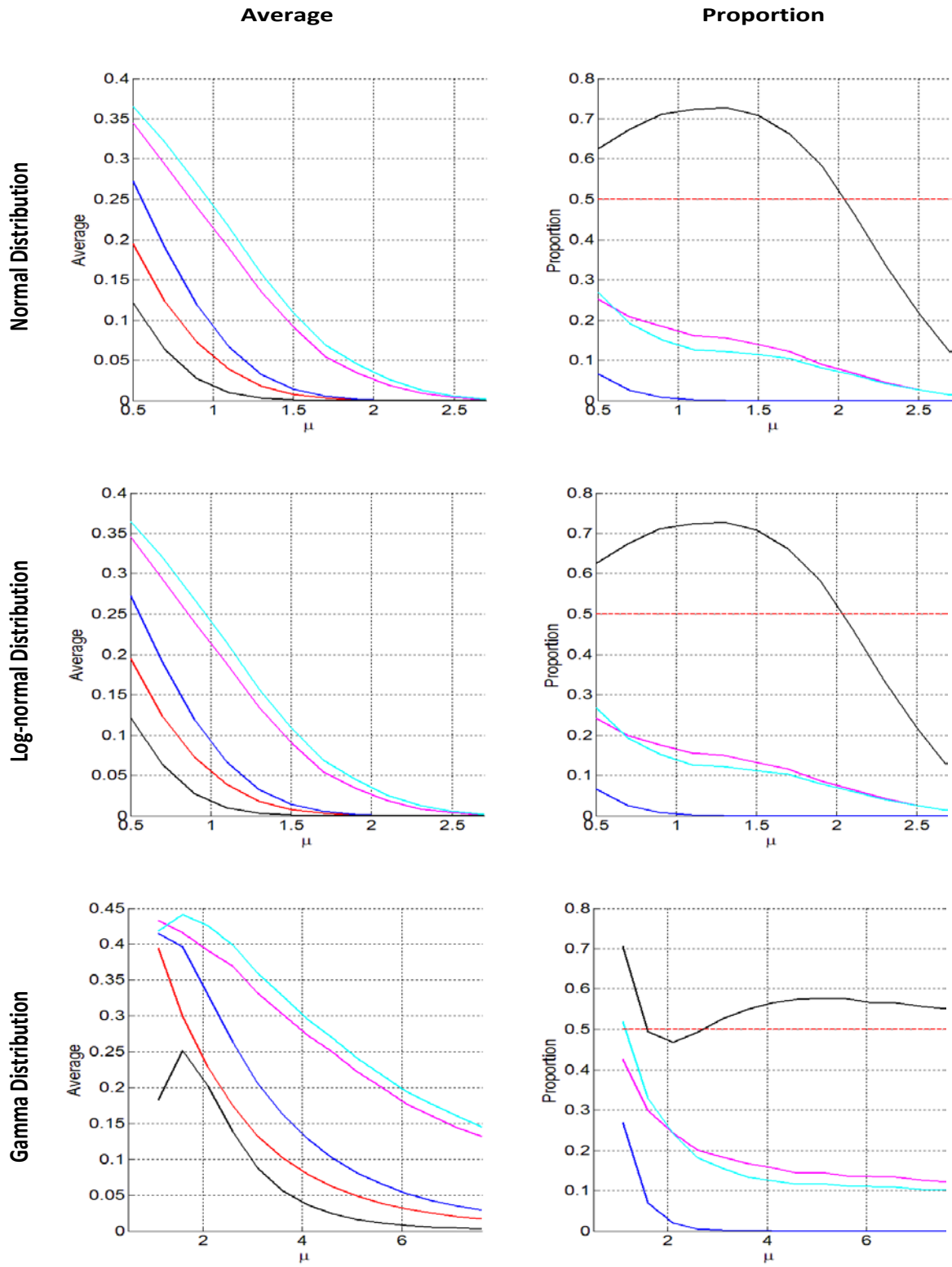


Figure 4-13: Power comparison between test statistics for scenario 3a: $N_0 = 20$ and $N_1 = 20$ with $w_0 = \frac{1}{4}$ and $w_1 = \frac{3}{4}$.

Figure 4-14 shows results of the power comparison between test statistics for $N_0 = 10$ and $N_1 = 20$ with $w_0 = \frac{1}{4}$ and $w_1 = \frac{3}{4}$. From Figure 4-14 it is evident the MW test statistic is on average the most powerful given a normal or log-normal distribution, while $\widehat{er}_{up}^*$ is on average the most powerful when assuming a gamma distribution. MW has a high probability of being the most powerful for small to medium shifts. The $\widehat{er}_{up}^*$ test statistic is more probable to outperform MW for larger shifts in distribution. However, it is important to note that the MW test cannot account for the weight selection which is no longer equal.

Figure 4-15 shows results of the power comparison between test statistics for $N_0 = 20$ and $N_1 = 10$ with $w_0 = \frac{1}{4}$ and $w_1 = \frac{3}{4}$. From Figure 4-15 it is evident the MW test statistic is on average the most powerful. MW has a high probability of being the most powerful for small to medium shifts. The $\widehat{er}_{up}^*$ test statistic is more probable to outperform MW for larger shifts in distribution given a normal or log-normal distribution. However, it is important to note that the MW test cannot account for the weight selection which is no longer equal.

Summary

The figures presented in this section all indicate that $\widehat{er}_{up}^*$ is the most powerful of the four error rate estimates. Though outperformed by the MW test statistic in some instances, the MW test cannot classify new subjects or account for unequal cost of misclassification into the two groups.

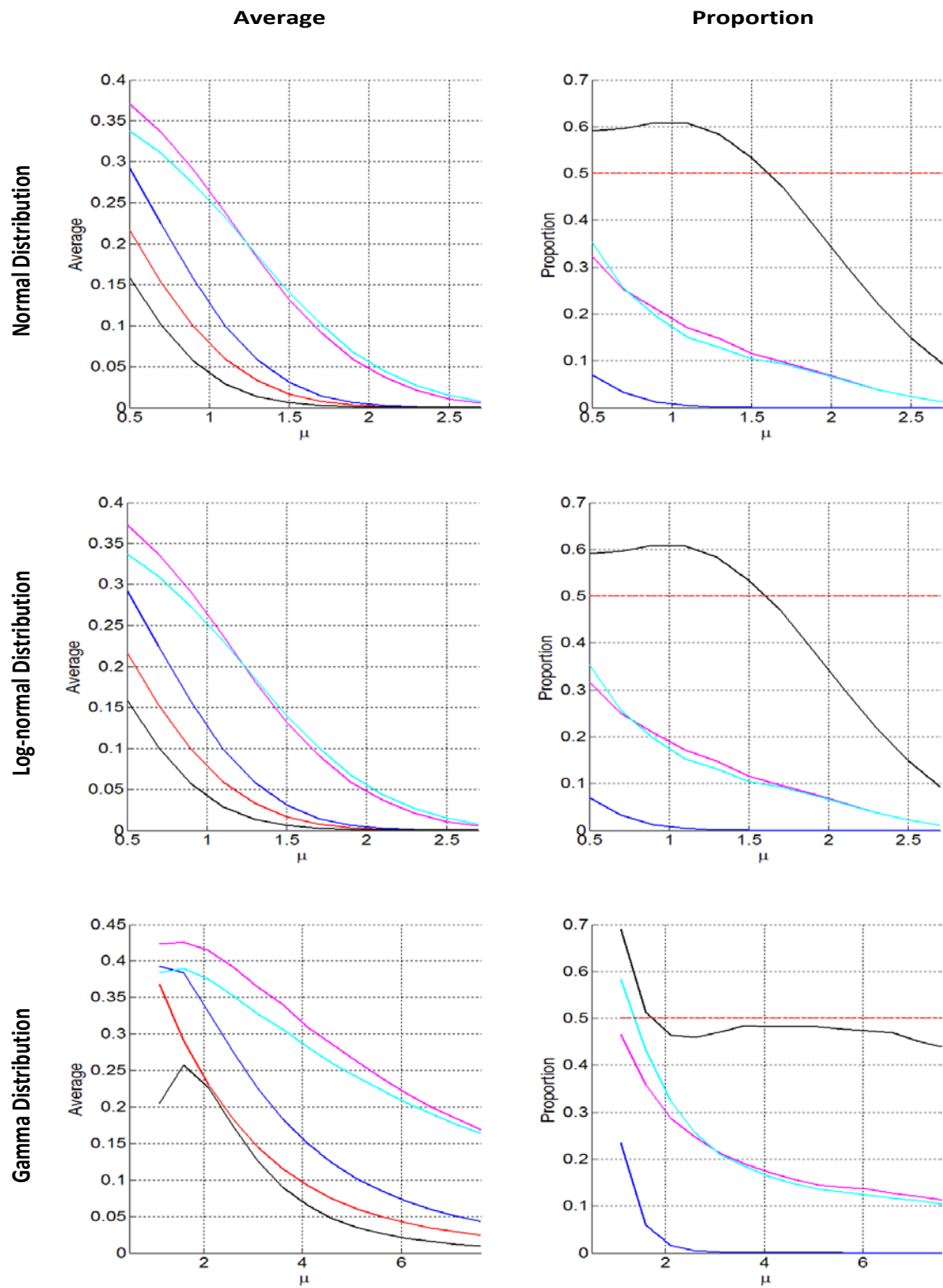


Figure 4-14: Power comparison between test statistics for scenario 3b: $N_0 = 10$ and $N_1 = 20$ with $w_0 = \frac{1}{4}$ and $w_1 = \frac{3}{4}$.

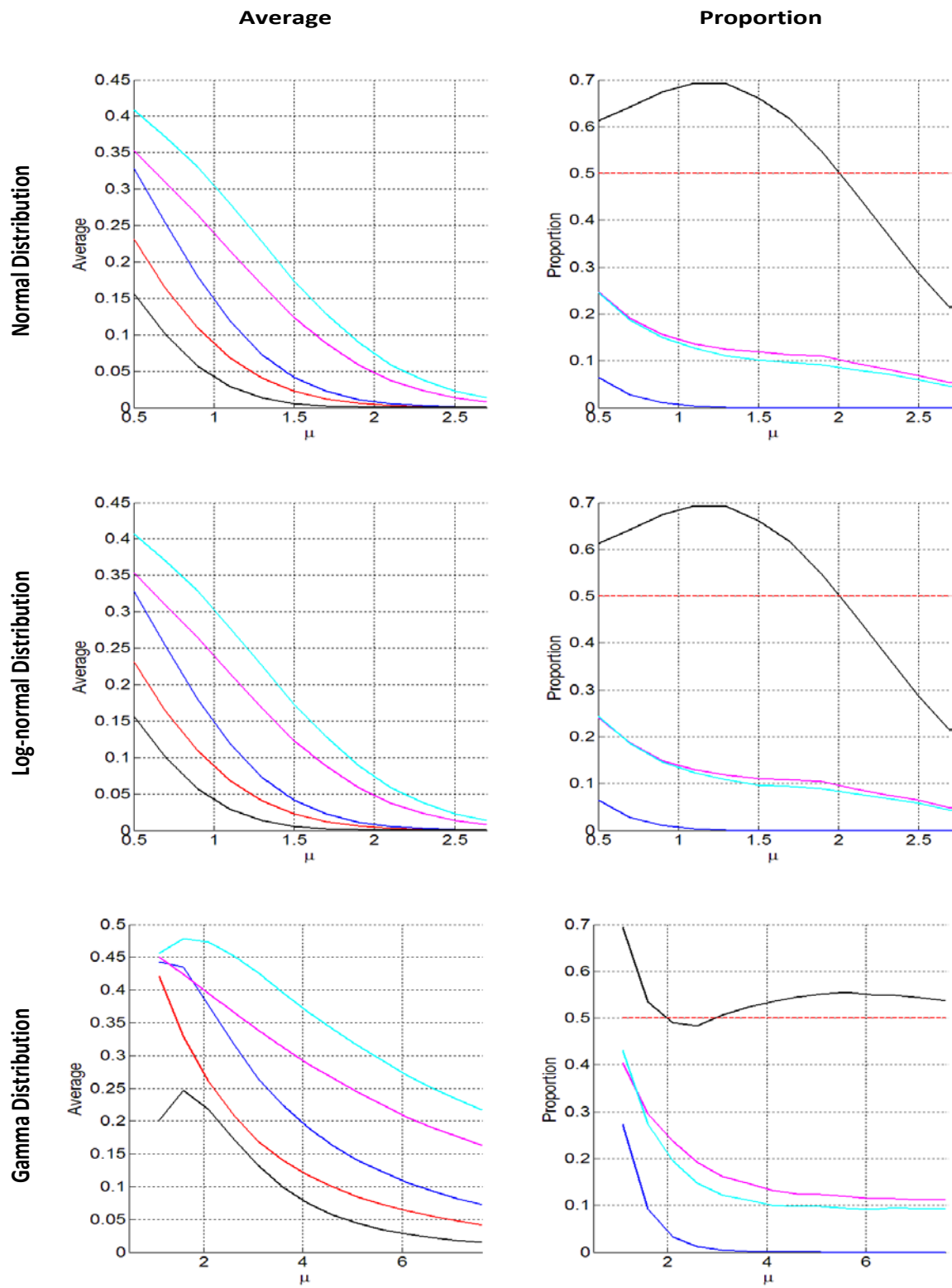


Figure 4-15: Power comparison between test statistics for scenario 3c: $N_0 = 20$ and $N_1 = 10$ with $w_0 = \frac{1}{4}$ and $w_1 = \frac{3}{4}$.

4.3.9 Comparing threshold estimators

When considering an upwards shift, the ERp software produces a full sample error rate $\widehat{er}_{up}^*$ along with its classification threshold $\hat{c}_{up}$. ERp can also produce a LOO error rate $\widehat{el}_{up}$ which in turn produces a threshold for each iteration. To obtain a single LOO threshold two alternatives were considered, the mean ($\widehat{cl}_{up,1}$) and the median ($\widehat{cl}_{up,2}$). To decide on the relative merits of the three threshold estimates - $\hat{c}_{up}$, $\widehat{cl}_{up,1}$ and $\widehat{cl}_{up,2}$ - we compared them to the population optimal threshold (c_{up}) using a simulation study similar to that reported in the previous section. As described earlier, c_{up} is the choice of c that minimizes $er_{up}(c)$ given by 4.2-1. Differentiating 4.2-1 and setting the result equal to 0, c_{up} is the solution of the equation

$$w_1 f_1(c) = w_0 f_0(c)$$

where $f_0(c)$ and $f_1(c)$ are the densities (derivatives) of the population CDFs $F_0(c)$ and $F_1(c)$ respectively. Let $\tau = \frac{w_1}{w_0} = \frac{f_0(c)}{f_1(c)}$ represent the population threshold can be calculated for each distribution scenario:

Scenario 1:

Values for control subjects followed a $N(0,1)$ distribution while those of experimental subject were drawn from a $N(\mu, 1)$ distribution where μ assumed different values

$$f_0(c) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-c^2}{2}\right)$$

$$f_1(c) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(c-\mu)^2}{2}\right)$$

$$\tau = \frac{\frac{1}{\sqrt{2\pi}} \exp\left(\frac{-c^2}{2}\right)}{\frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(c-\mu)^2}{2}\right)} = \exp\left(c\mu - \frac{1}{2}\mu^2\right)$$

$$\ln \tau = c\mu - \frac{1}{2}\mu^2$$

$$c = \left(-\frac{1}{\mu} \ln \tau + \frac{1}{2}\mu\right)$$

Scenario 2:

Values for control subjects followed a LN(0,1) distribution while those of experimental subject were drawn from a LN(μ , 1) distribution where μ assumed different values

$$f_0(c) = \frac{1}{c\sqrt{2\pi}} \exp\left(\frac{-(\ln c)^2}{2}\right)$$

$$f_1(c) = \frac{1}{c\sqrt{2\pi}} \exp\left(\frac{-(\ln c - \mu)^2}{2}\right)$$

$$\tau = \frac{\frac{1}{c\sqrt{2\pi}} \exp\left(\frac{-(\ln c)^2}{2}\right)}{\frac{1}{c\sqrt{2\pi}} \exp\left(\frac{-(\ln c - \mu)^2}{2}\right)} = \exp\left(\mu \ln c - \frac{1}{2}\mu^2\right)$$

$$\ln \tau = \mu \ln c - \frac{1}{2}\mu^2$$

$$c = \exp\left(-\frac{1}{\mu} \ln \tau + \frac{1}{2}\mu\right)$$

Scenario 3:

Values for control subjects followed a $\Gamma(0,1,1)$ distribution while those of experimental subject were drawn from a $\Gamma(0, \beta, 1)$ distribution where β (the scale parameter) assumed different values.

$$f_0(c) = \exp(-c)$$

$$f_1(c) = \exp\left(\frac{-c}{b}\right) \frac{1}{b}$$

$$\tau = \exp\left(\frac{c}{b} - c\right) b$$

$$\ln \tau = c\left(\frac{1}{b} - 1\right) + \ln b$$

$$c = \frac{\ln \tau - \ln b}{\left(\frac{1}{b} - 1\right)}$$

The bias and mean squared error (MSE) were estimated by averaging the differences and squared differences from c_{up} over the ten thousand estimated values of $\hat{c}_{up}$, $\hat{c}_{l_{up,1}}$ and $\hat{c}_{l_{up,2}}$ at each shift magnitude (μ).

Legend for Figures 4-16 to 4-24:

The three rows within each figure represent the distributional assumptions as outlined above. The left panel shows the bias of $\hat{c}_{up}$, (black lines), $\hat{c}_{l_{up,1}}$ (magenta lines) and $\hat{c}_{l_{up,2}}$ (blue lines) when compared to the optimal threshold c_{up} . The right panel shows the MSE for $\hat{c}_{up}$, (black lines), $\hat{c}_{l_{up,1}}$ (magenta lines) and $\hat{c}_{l_{up,2}}$ (blue lines).

Figure 4-16 compares the threshold estimates for $N_0 = 20$ and $N_1 = 20$ with $w_0 = \frac{1}{2}$ and $w_1 = \frac{1}{2}$. The left panel of Figure 4-16 indicate that $\hat{c}_{l_{up,2}}$ is the least bias, however the bias of $\hat{c}_{l_{up,1}}$ is very comparable. From the right panel it is evident that $\hat{c}_{l_{up,2}}$ is preferable as it has a smaller MSE, however the MSE of $\hat{c}_{l_{up,1}}$ is again very comparable.

Figure 4-17 compares the threshold estimates for $N_0 = 10$ and $N_1 = 20$ with $w_0 = \frac{1}{2}$ and $w_1 = \frac{1}{2}$. The left panel of Figure 4-17 indicate that $\hat{c}_{l_{up,2}}$ is the least bias. However from the right panel it is evident that $\hat{c}_{l_{up,1}}$ is preferable as it has a smaller MSE

Figure 4-19 compares the threshold estimates for $N_0 = 20$ and $N_1 = 20$ with $w_0 = \frac{3}{4}$ and $w_1 = \frac{1}{4}$. Figure 4-19 indicates that $\hat{c}_{l_{up,1}}$ is the least bias (left panel) and is preferable as it has a smaller MSE (right panel).

Figure 4-20 compares the threshold estimates for $N_0 = 10$ and $N_1 = 20$ with $w_0 = \frac{3}{4}$ and $w_1 = \frac{1}{4}$. The left panel of Figure 4-20 indicate that $\hat{c}_{l_{up,2}}$ and $\hat{c}_{l_{up,1}}$ are less bias than $\hat{c}_{up}$. Furthermore, from the right panel, it is evident that $\hat{c}_{l_{up,1}}$ is preferable as it has to lowest MSE.

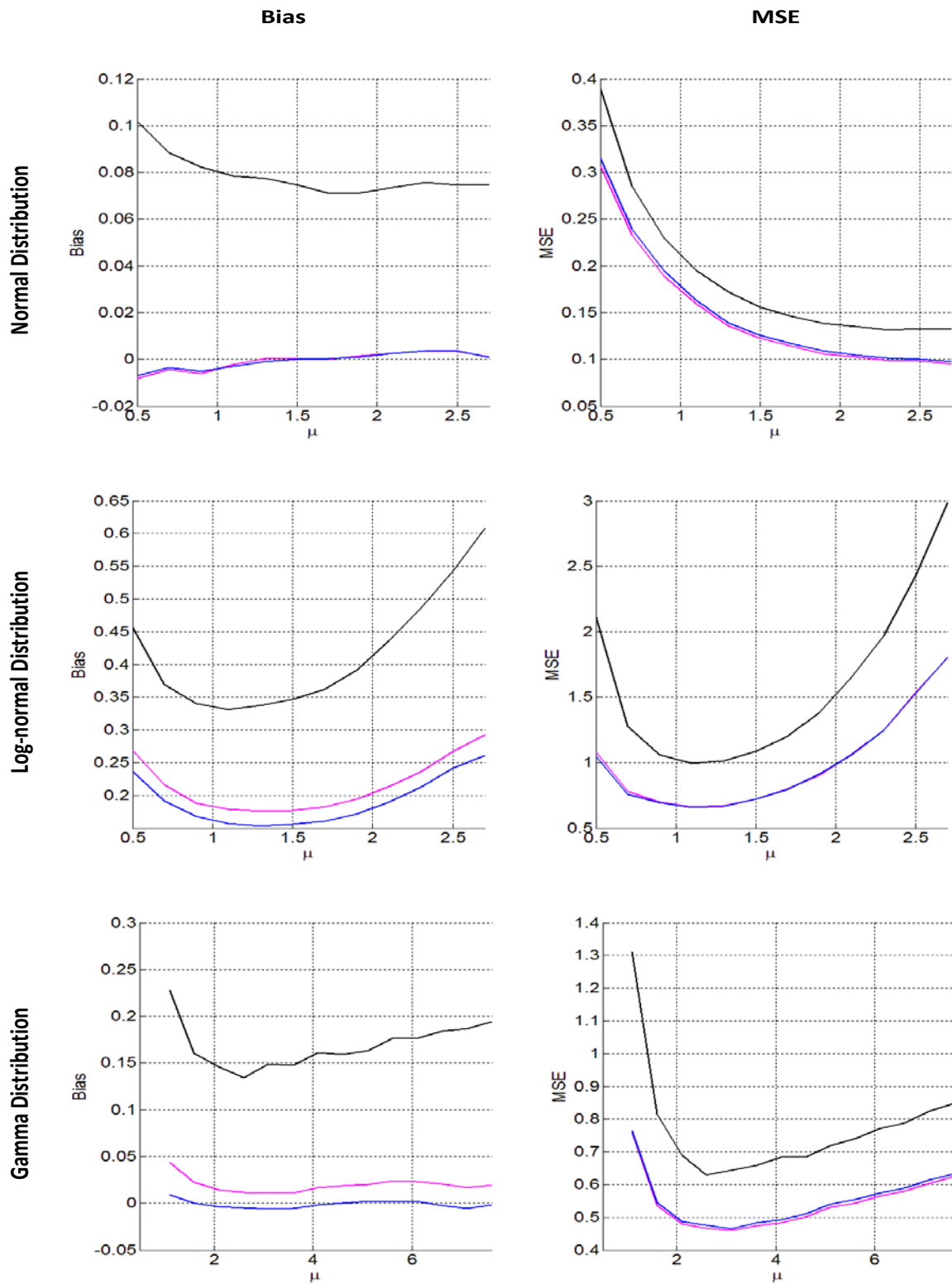


Figure 4-16: Comparing threshold estimates for scenario 1a: $N_0 = 20$ and $N_1 = 20$ with $w_0 = \frac{1}{2}$ and $w_1 = \frac{1}{2}$.

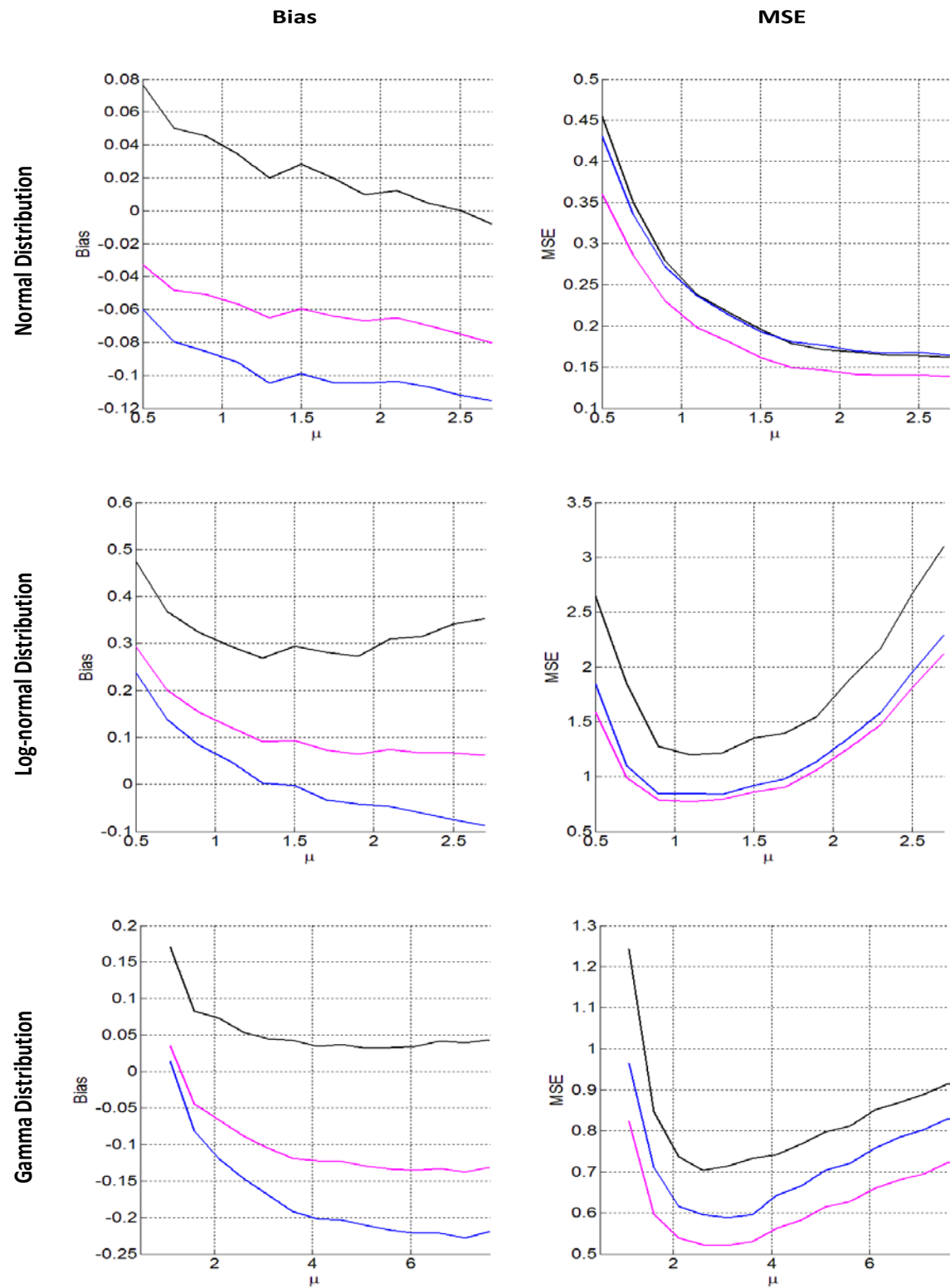


Figure 4-17: Comparing threshold estimates for scenario 1b: $N_0 = 10$ and $N_1 = 20$ with $w_0 = \frac{1}{2}$ and $w_1 = \frac{1}{2}$.

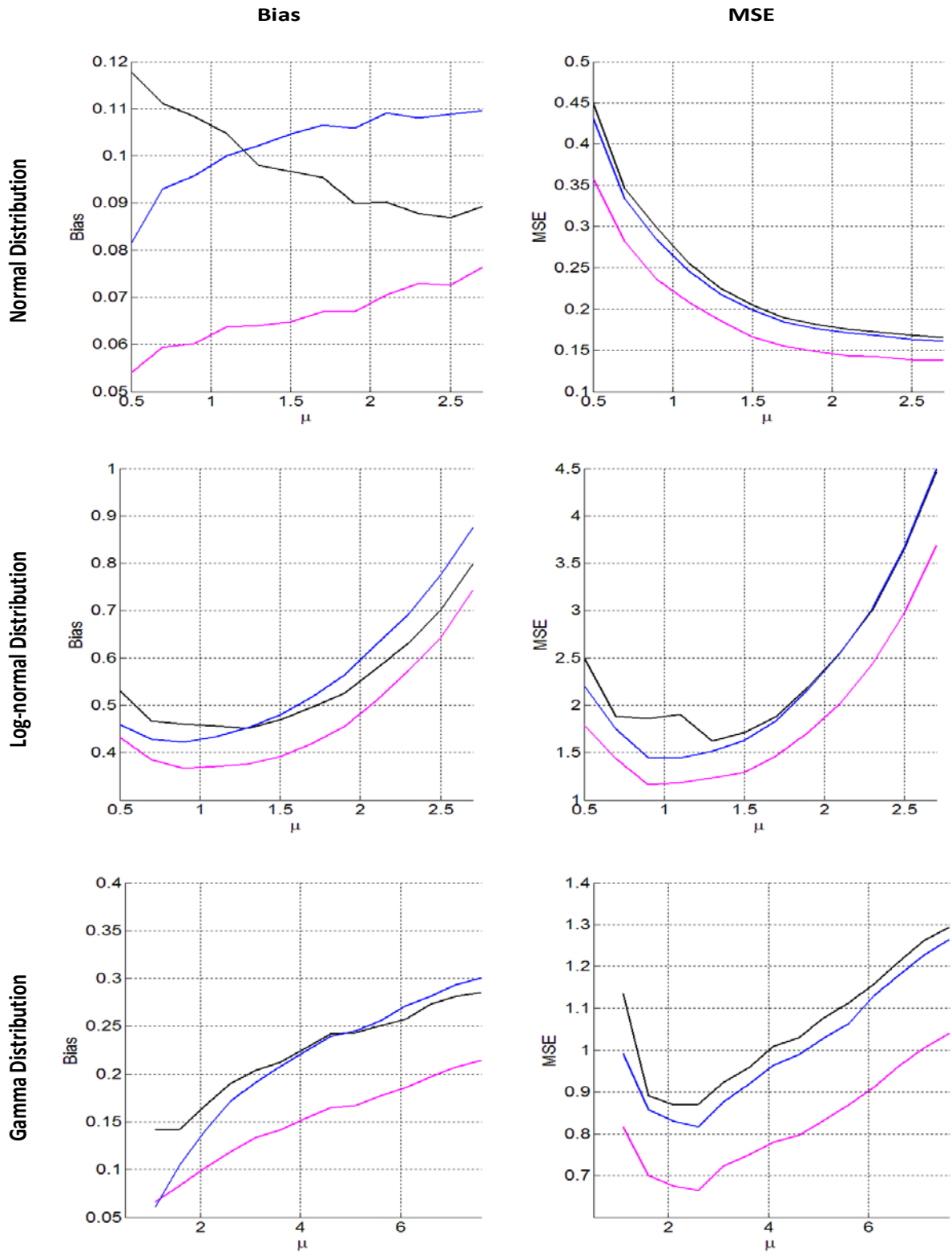


Figure 4-18: Comparing threshold estimates for scenario 1c: $N_0 = 20$ and $N_1 = 10$ with $w_0 = \frac{1}{2}$ and $w_1 = \frac{1}{2}$.

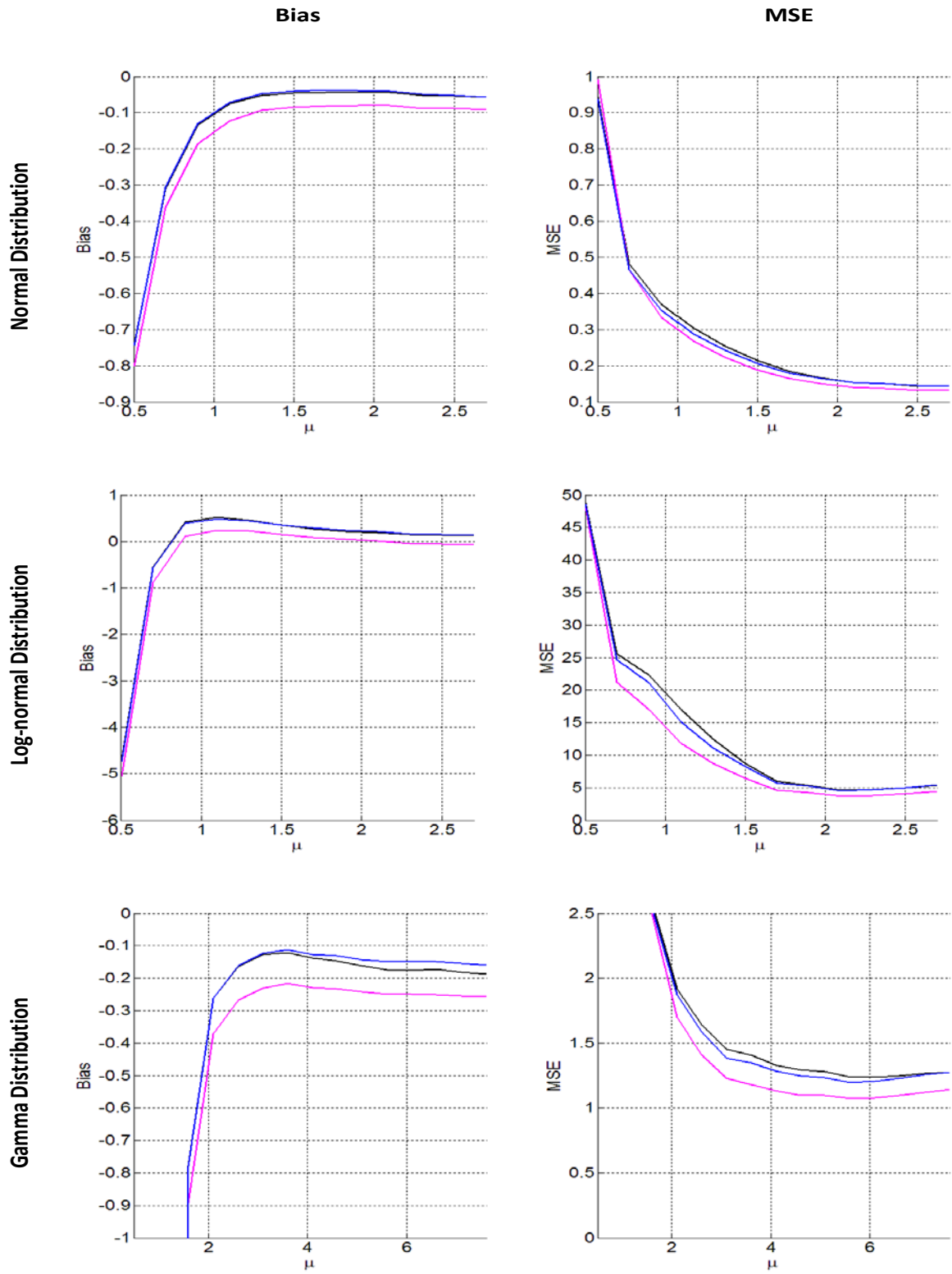


Figure 4-19: Comparing threshold estimates for scenario 2a: $N_0 = 20$ and $N_1 = 20$ with $w_0 = \frac{3}{4}$ and $w_1 = \frac{1}{4}$.

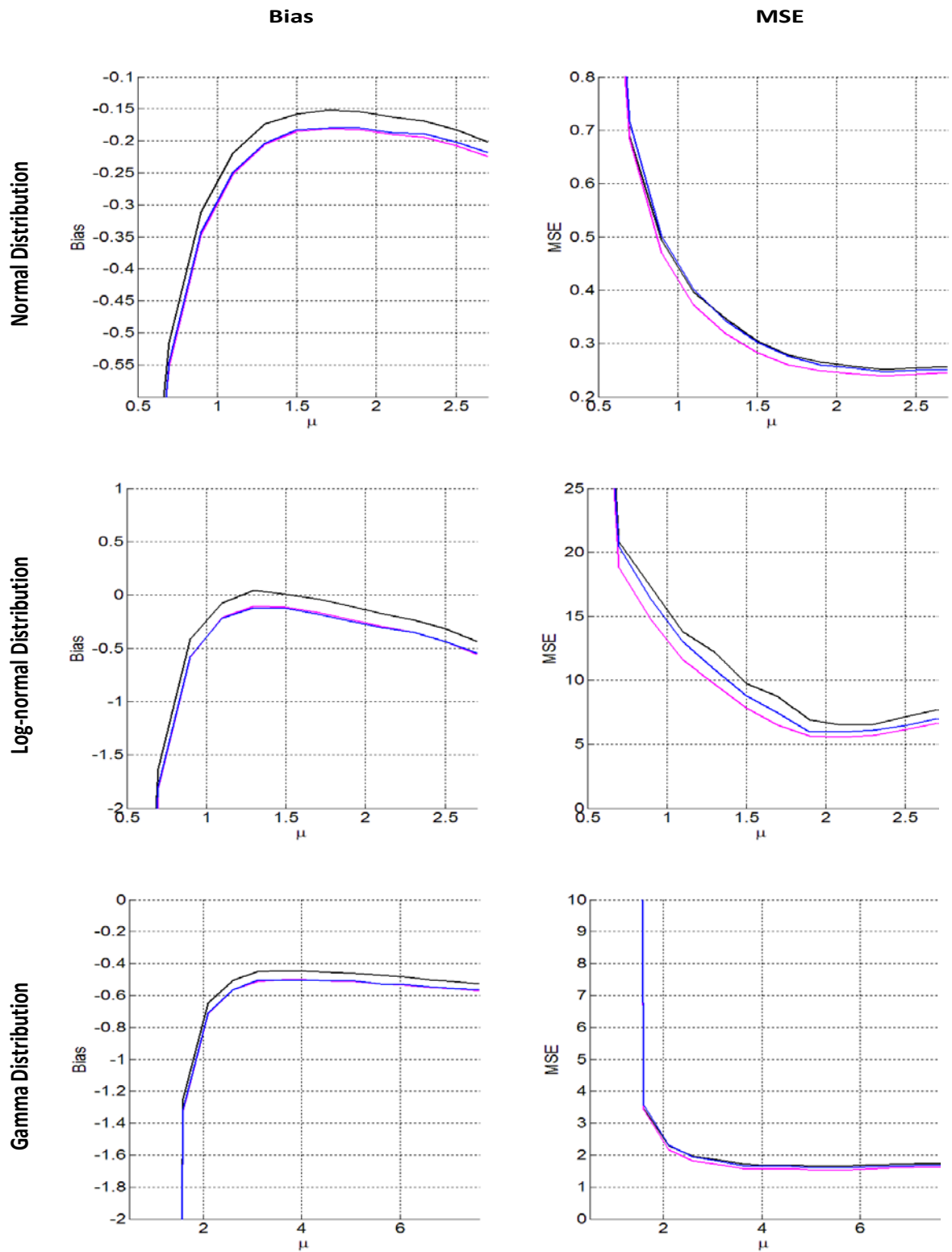


Figure 4-20: Comparing threshold estimates for scenario 2b: $N_0 = 10$ and $N_1 = 20$ with $w_0 = \frac{3}{4}$ and $w_1 = \frac{1}{4}$.

Figure 4-21 compares the threshold estimates for $N_0 = 20$ and $N_1 = 10$ with $w_0 = \frac{3}{4}$ and $w_1 = \frac{1}{4}$. Figure 4-21 indicates that $\hat{c}l_{up,1}$ is the least bias (left panel) and is preferable as it has a lower MSE (right panel).

Figure 4-22 compares the threshold estimates for $N_0 = 20$ and $N_1 = 20$ with $w_0 = \frac{1}{4}$ and $w_1 = \frac{3}{4}$. Figure 4-22 indicates that $\hat{c}l_{up,2}$ is the least bias (left panel) and is preferable as it has a lower or very comparable MSE (right panel).

Figure 4-23 compares the threshold estimates for $N_0 = 10$ and $N_1 = 20$ with $w_0 = \frac{1}{4}$ and $w_1 = \frac{3}{4}$. Figure 4-23 indicates that $\hat{c}l_{up,2}$ is the least bias (left panel), however $\hat{c}l_{up,1}$ may be preferable as it has a slightly lower MSE (right panel).

Figure 4-24 compares the threshold estimates for $N_0 = 20$ and $N_1 = 10$ with $w_0 = \frac{1}{4}$ and $w_1 = \frac{3}{4}$. Figure 4-24 indicates that $\hat{c}l_{up,2}$ is the least bias (left panel), however $\hat{c}l_{up,1}$ may be preferable as it has a slightly lower MSE (right panel).

Summary

The left panels of the figures presented in this section show a degree bias for all threshold estimators regardless of the sample size, weight or distribution scenario. The two LOO estimators consistently outperform the FS estimator in term of bias as well as MSE. From the right panel, it is evident that $\hat{c}l_{up,1}$ is a more consistent estimator based on MSE. The MSE for both LOO estimators are very comparable for small shifts in distribution, however, as shifts become larger, $\hat{c}l_{up,1}$ outperforms $\hat{c}l_{up,2}$ in the majority of scenarios. We therefore advocate the use of $\hat{c}l_{up,1}$ as an estimator of the population threshold.

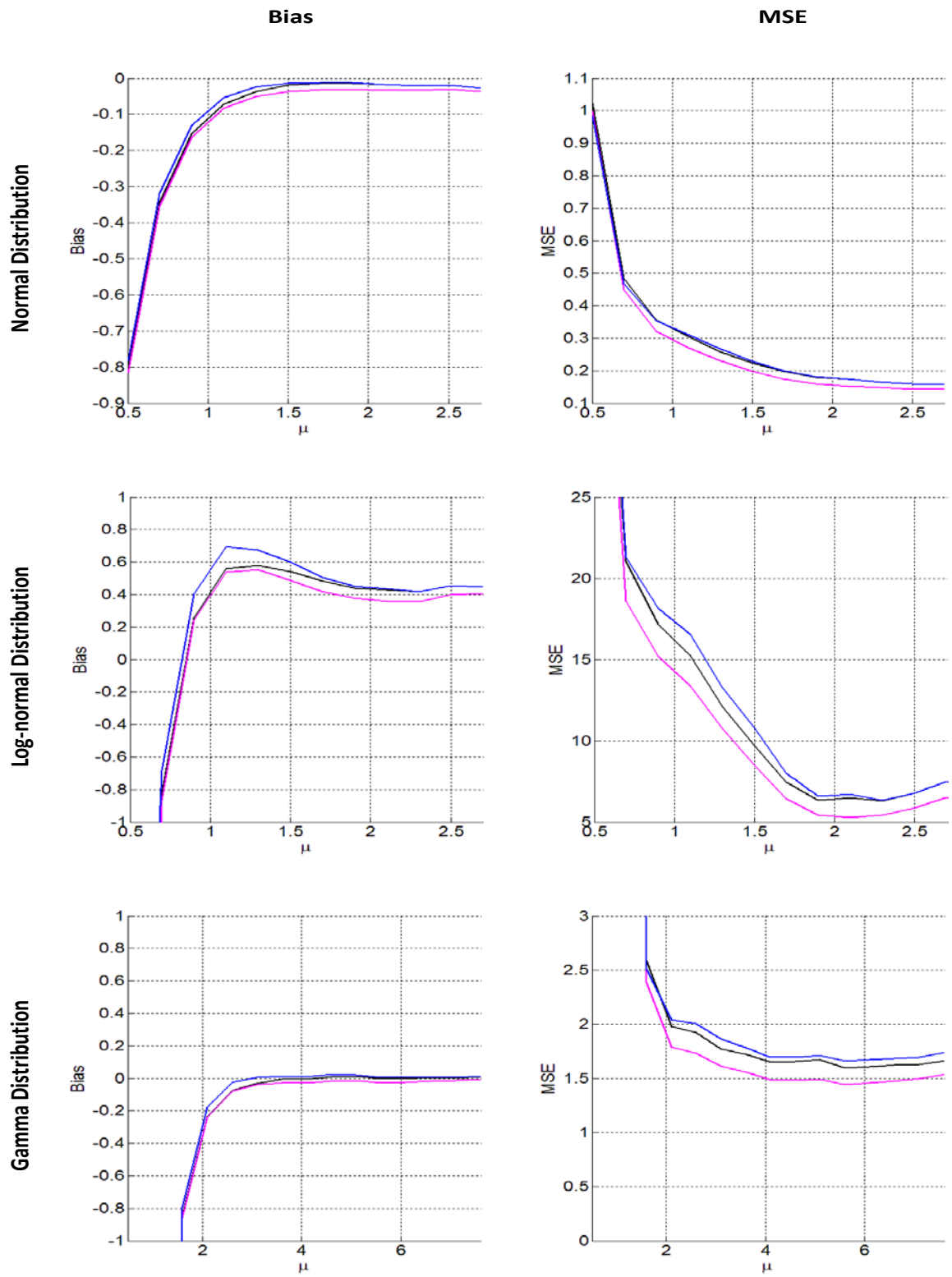


Figure 4-21: Comparing threshold estimates for scenario 2c: $N_0 = 20$ and $N_1 = 10$ with $w_0 = \frac{3}{4}$ and $w_1 = \frac{1}{4}$.

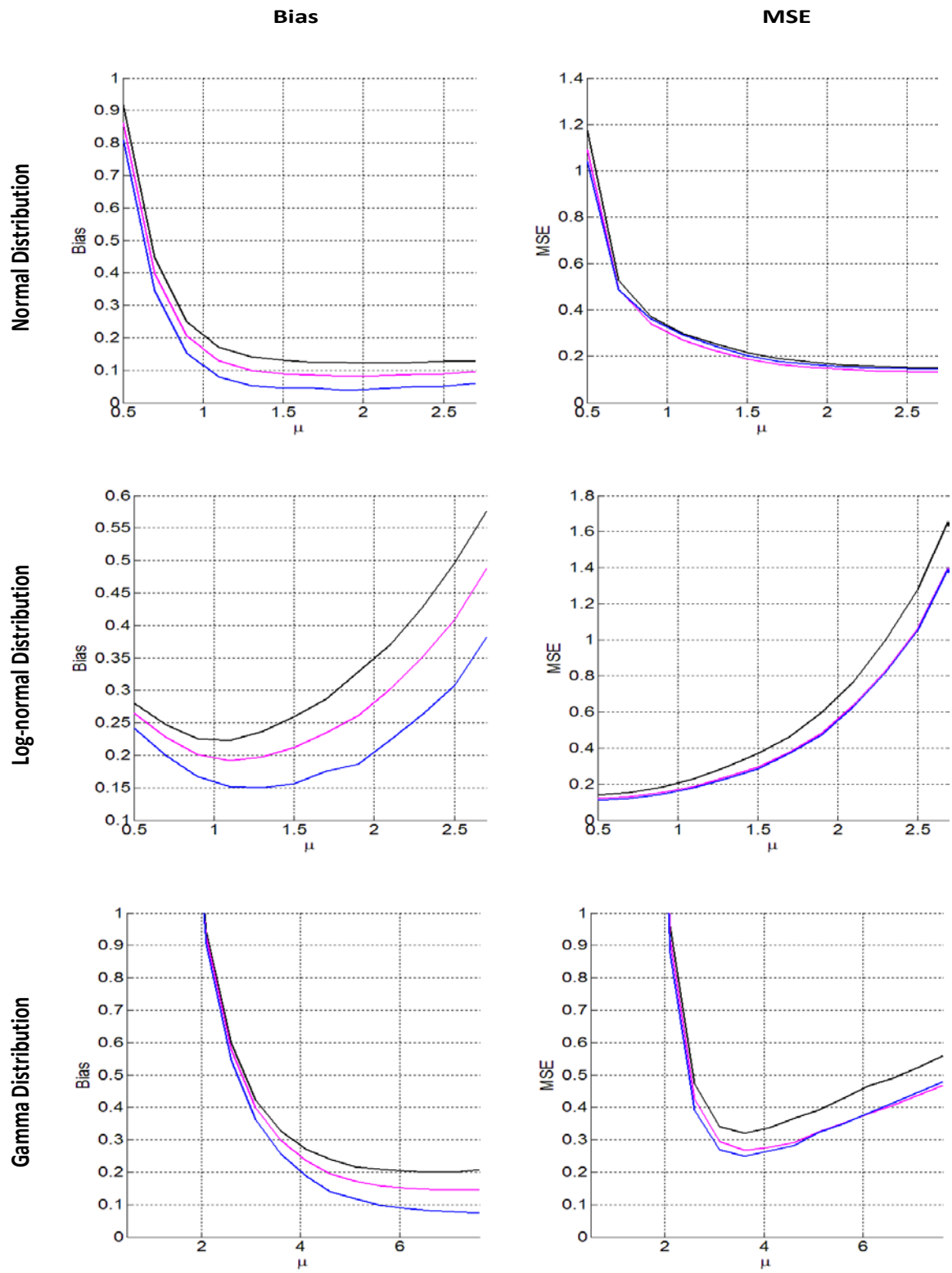


Figure 4-22: Comparing threshold estimates for scenario 3a: $N_0 = 20$ and $N_1 = 20$ with $w_0 = \frac{1}{4}$ and $w_1 = \frac{3}{4}$.

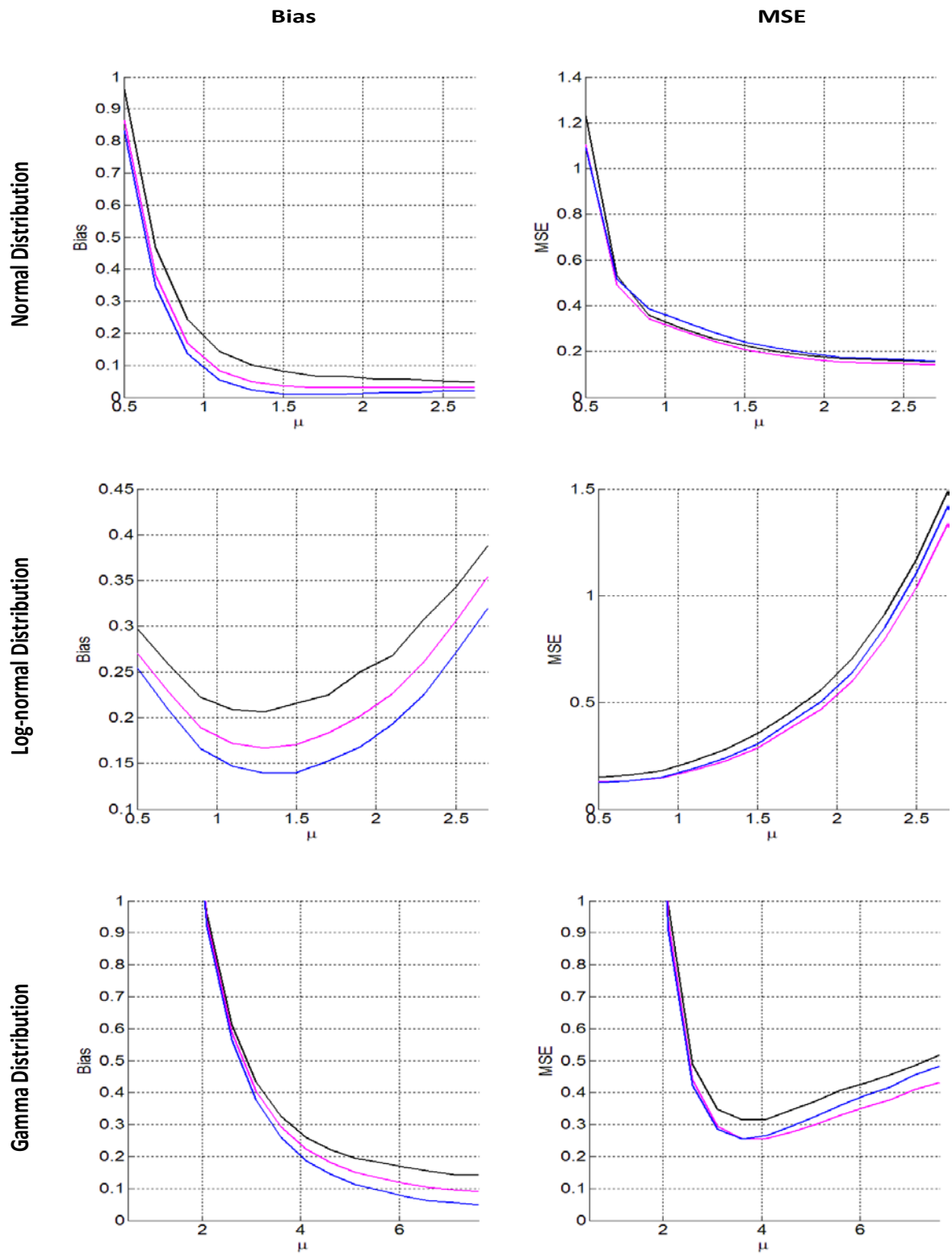


Figure 4-23: Comparing threshold estimates for scenario 3b: $N_0 = 10$ and $N_1 = 20$ with $w_0 = \frac{1}{4}$ and $w_1 = \frac{3}{4}$.

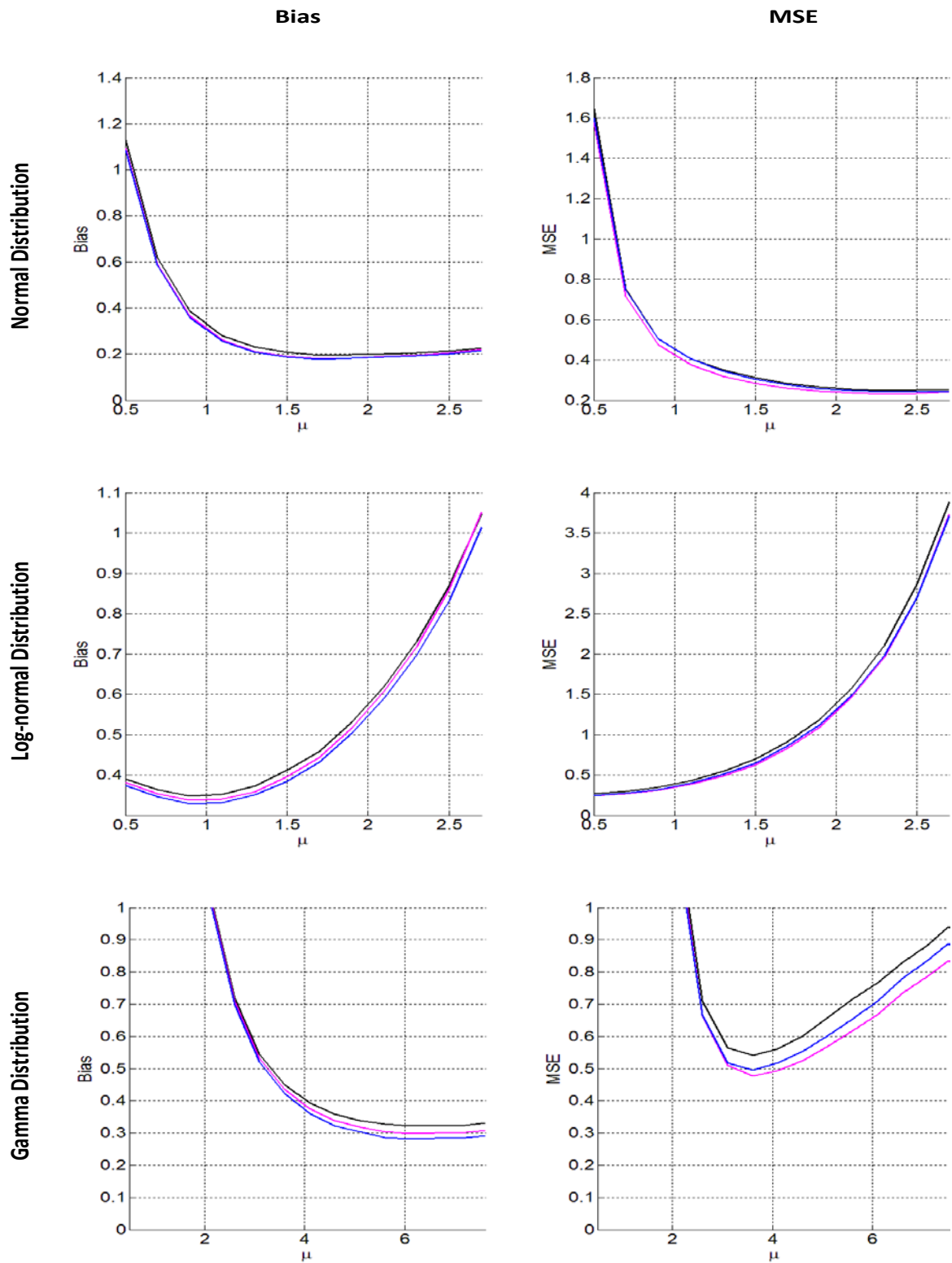


Figure 4-24: Comparing threshold estimates for scenario 3c: $N_0 = 20$ and $N_1 = 10$ with $w_0 = \frac{1}{4}$ and $w_1 = \frac{3}{4}$.

4.3.10 Comparisons of LOO and FS error rate estimates per se

We have shown that FS error rate estimates are better able to detect that a variable has discriminatory content than the LOO versions following the testing route. However, one may also be interested in the error rates as entities in their own right and for this reason we compared the FS and LOO versions in terms of their biases and MSEs as estimators of the corresponding population error rates. For this purpose we used of the same simulation as outlined in the paper for threshold comparison. Since we simulate an upward shift in distribution, the population error rate is given by $er_{up}^* = er_{up}(c_{up})$. We compare the bias and mean squared error (MSE) of $\widehat{er}_{up}^*$, $\widehat{er}_{min}^*$, $\widehat{el}_{up}$ and $\widehat{el}_{min}$ over ten thousand repetitions assuming various distributional shifts.

A simulation-based comparison between error rate estimates and population error rate was performed and results are presented in Figure 4-25. Observations for the control group were drawn from a $LN(0,1)$ distribution with $N_0 = 21$, while observations for the experimental group were drawn from an upward shifted $LN(\mu, 1)$ distribution with $N_1 = 12$. The left panels show the bias and the right panels show the MSE of $\widehat{er}_{up}^*$ (black lines), $\widehat{el}_{up}$ (magenta lines), $\widehat{er}_{min}^*$ (blue lines) and $\widehat{el}_{min}$ (cyan line) as estimates of the optimal error rate er_{up}^* . The top and bottom panel show the results for weight set 1 and 2 respectively.

We find that the FS estimates are overly optimistic while the LOO estimates are overly pessimistic from a bias point of view (Figure 4-25). At shifts exceeding 1.2 (as required to get the p -values low enough to suggest discriminatory content in the variable) the LOO estimators are less biased than the FS estimators, while the FS estimators have lower MSE.

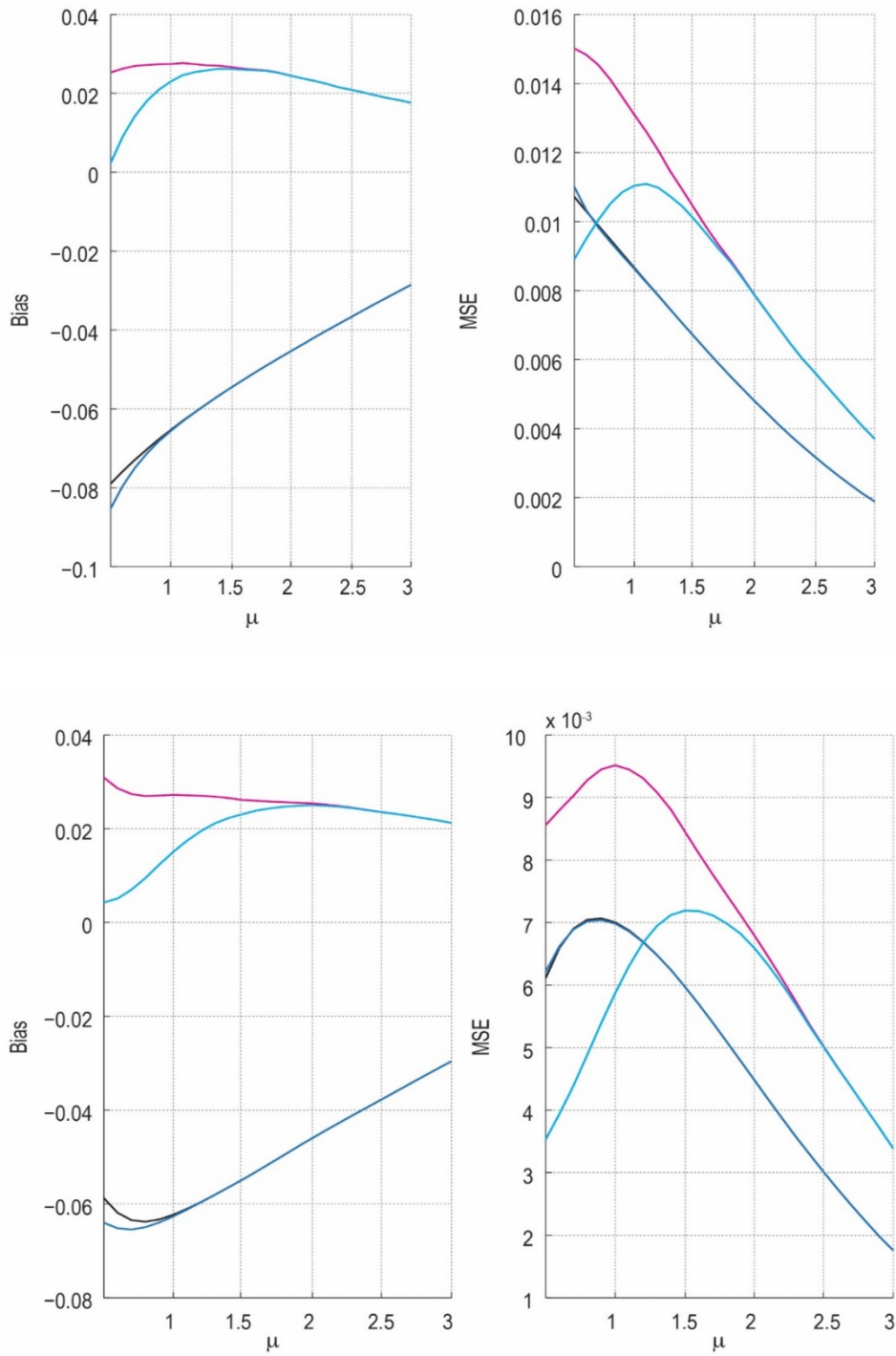


Figure 4-25: Simulation comparison of error rate estimates with the population error rate.

The left panels show the bias and the right panels show the MSE of $\hat{er}_{up}^*$ (black lines), $\hat{el}_{up}$ (magenta lines), $\hat{er}_{min}^*$ (blue lines) and $\hat{el}_{min}$ (cyan line) as estimates of the optimal error rate er_{up}^* . The top and bottom panel show the results for weight set 1 and 2 respectively.

We also compared these error rate estimates for the simulations outlined in Table 4-3.

Legend for Figures 4-26 to 4-33:

The left panels of the figures presented in this section show the bias of $\widehat{er}_{up}^*$, (black lines), $\widehat{el}_{up}$ (magenta lines) and $\widehat{er}_{min}^*$ (blue lines) and $\widehat{el}_{min}$ (cyan line) when compared to the optimal error rate er_{up}^* . The right panels show the MSE for $\widehat{er}_{up}^*$, (black lines), $\widehat{el}_{up}$ (magenta lines), $\widehat{er}_{min}^*$ (blue lines), and $\widehat{el}_{min}$ (cyan line).

From these figures it is evident that the FS error rate estimates, $\widehat{er}_{up}^*$ and $\widehat{er}_{min}^*$, consistently outperform the LOO error rate estimates, $\widehat{el}_{up}$ and $\widehat{el}_{min}$, in terms of bias (left panels). We find that the MSE for $\widehat{er}_{min}^*$ is consistently the lowest at larger shifts in distribution. However, in some instance where shifts are smaller, $\widehat{el}_{min}$ outperforms $\widehat{er}_{min}^*$ in terms of MSE.

Figure 4-26 compares error rate estimates for $N_0 = 20$ and $N_1 = 20$ with $w_0 = \frac{1}{2}$ and $w_1 = \frac{1}{2}$. It is evident from Figure 4-26 that the FS error rate estimates are very comparable and outperform the LOO error rate estimates both in term of bias (left panel) and MSE (right panel). However, for smaller shifts in distribution, $\widehat{er}_{min}^*$ is preferable.

Figure 4-27 compares error rate estimates for $N_0 = 10$ and $N_1 = 20$ with $w_0 = \frac{1}{2}$ and $w_1 = \frac{1}{2}$. It is evident from Figure 4-27 that the FS error rate estimates are very comparable and outperform the LOO error rate estimates both in term of bias (left panel) and MSE (right panel). However, for smaller shifts in distribution, $\widehat{er}_{min}^*$ is preferable.

Figure 4-28 compares error rate estimates for $N_0 = 20$ and $N_1 = 10$ with $w_0 = \frac{1}{2}$ and $w_1 = \frac{1}{2}$. It is evident from Figure 4-28 that the FS error rate estimates are very comparable and outperform the LOO error rate estimates both in term of bias (left panel) and MSE (right panel). However, for smaller shifts in distribution, $\widehat{er}_{min}^*$ is preferable.

Figure 4-29 compares error rate estimates for $N_0 = 20$ and $N_1 = 20$ with $w_0 = \frac{3}{4}$ and $w_1 = \frac{1}{4}$. It is evident from Figure 4-29 that the FS error rate estimates are very comparable and outperform the LOO error rate estimates both in term of bias (left panel) and MSE (right panel), specifically, for larger shifts in distribution..

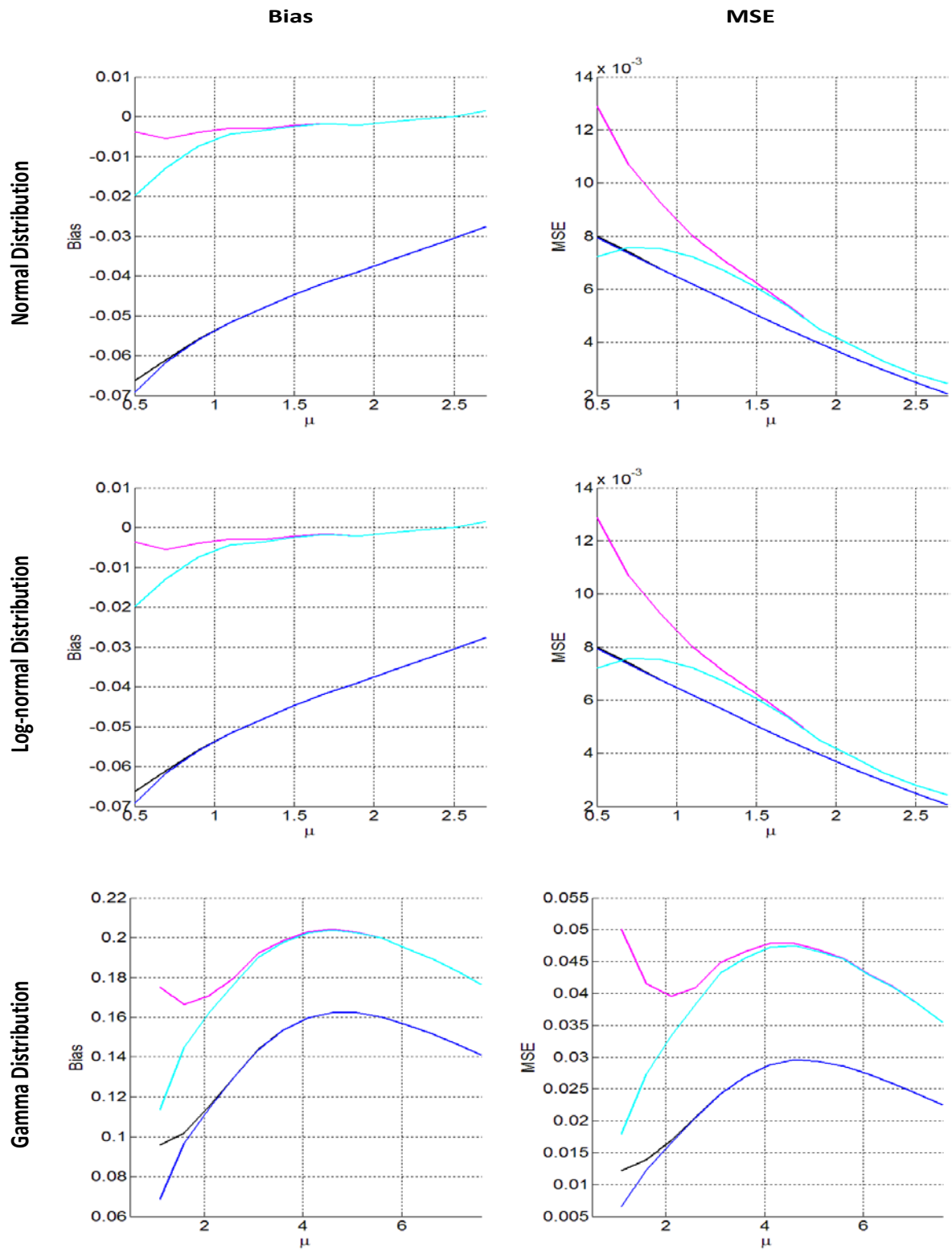


Figure 4-26: Comparing error rate estimates for $N_0 = 20$ and $N_1 = 20$ with $w_0 = \frac{1}{2}$ and $w_1 = \frac{1}{2}$.

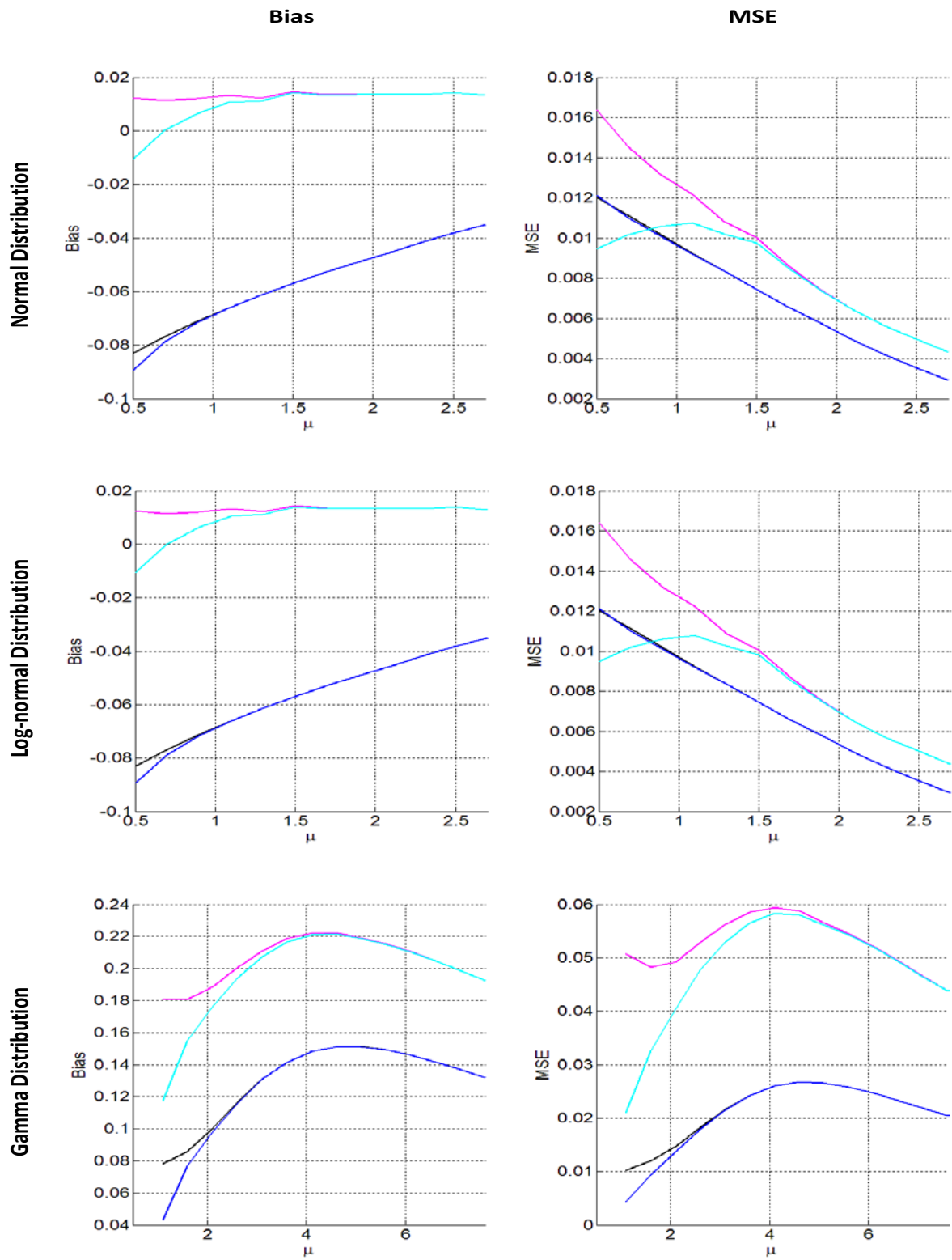


Figure 4-27: Comparing error rate estimates for $N_0 = 10$ and $N_1 = 20$ with $w_0 = \frac{1}{2}$ and $w_1 = \frac{1}{2}$.

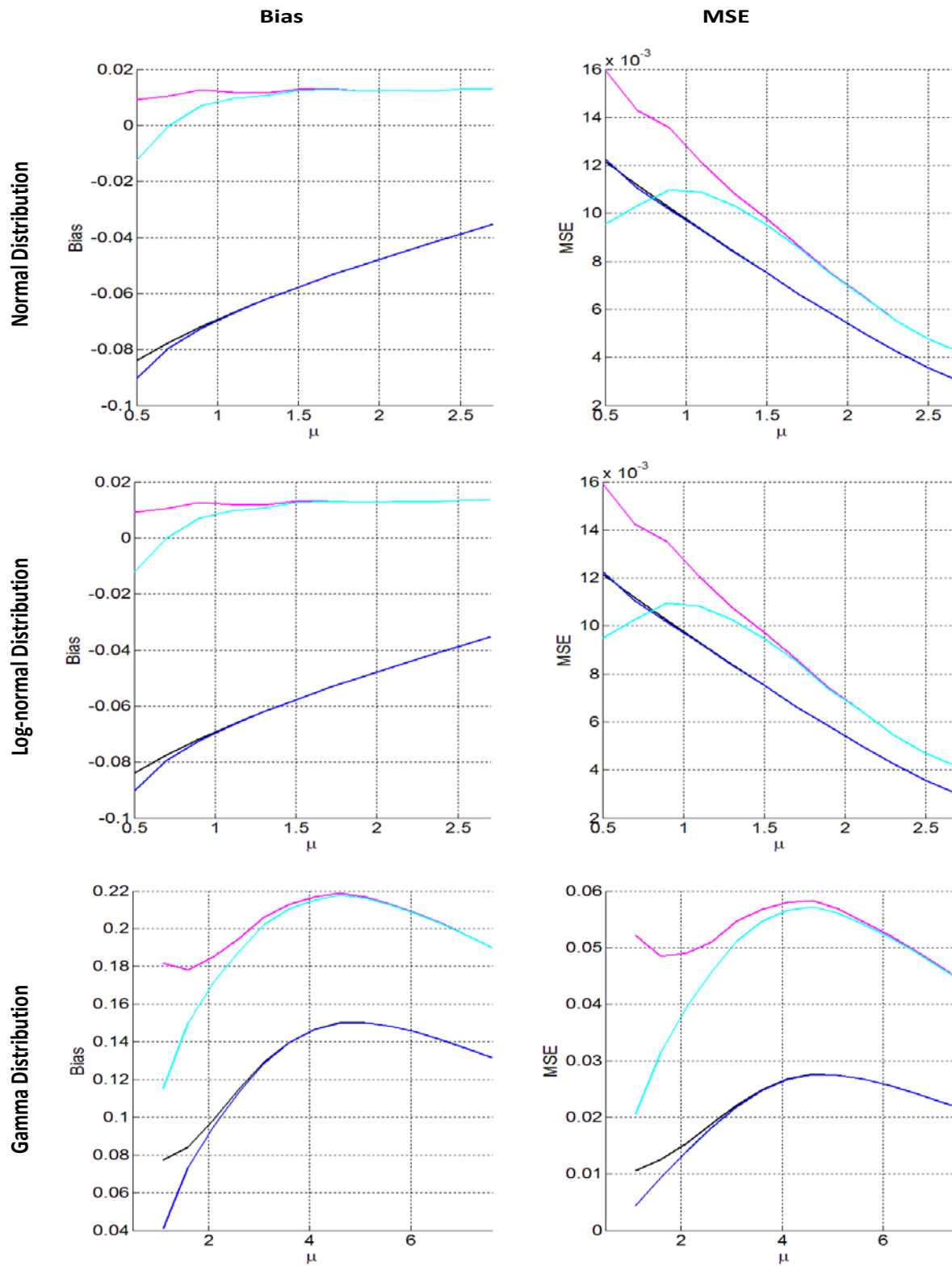


Figure 4-28: Comparing error rate estimates for $N_0 = 20$ and $N_1 = 10$ with $w_0 = \frac{1}{2}$ and $w_1 = \frac{1}{2}$.

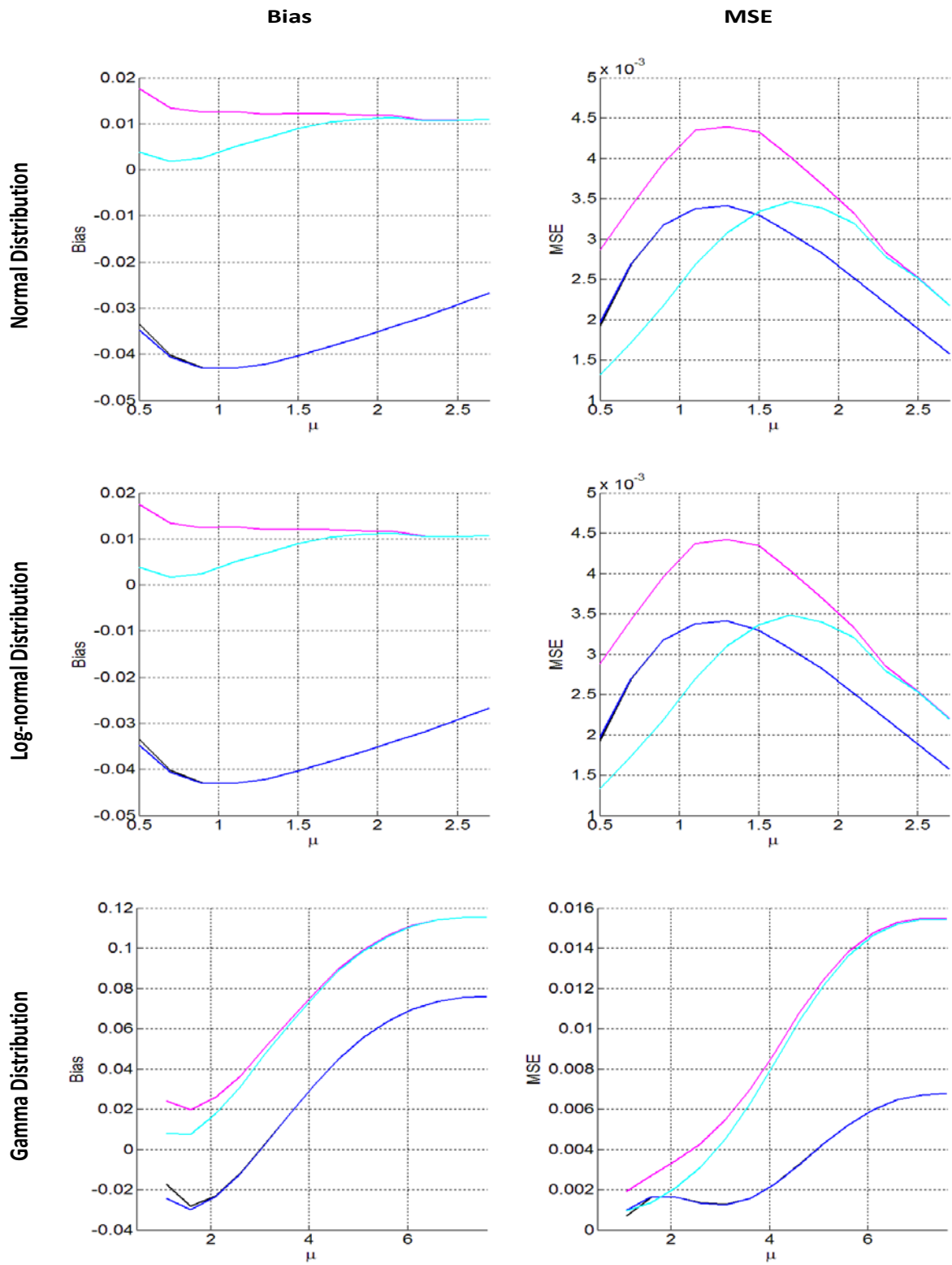


Figure 4-29: Comparing error rate estimates for $N_0 = 20$ and $N_1 = 20$ with $w_0 = \frac{3}{4}$ and $w_1 = \frac{1}{4}$.

Figure 4-30 compares error rate for $N_0 = 10$ and $N_1 = 20$ with $w_0 = \frac{3}{4}$ and $w_1 = \frac{1}{4}$. It is evident from Figure 4-30 that the FS error rate estimates are very comparable and outperform the LOO error rate estimates both in term of bias (left panel) and MSE (right panel), specifically, for larger shifts in distribution.

Figure 4-31 compares error rate estimates for $N_0 = 20$ and $N_1 = 10$ with $w_0 = \frac{3}{4}$ and $w_1 = \frac{1}{4}$. It is evident from Figure 4-31 that the FS error rate estimates are very comparable and outperform the LOO error rate estimates both in term of bias (left panel) and MSE (right panel), specifically, for larger shifts in distribution..

Figure 4-32 compares error rate estimates for $N_0 = 20$ and $N_1 = 20$ with $w_0 = \frac{1}{4}$ and $w_1 = \frac{3}{4}$. It is evident from Figure 4-32 that the FS error rate estimates are very comparable and outperform the LOO error rate estimates both in term of bias (left panel) and MSE (right panel), specifically, for larger shifts in distribution..

Figure 4-33 compares error rate estimates for $N_0 = 20$ and $N_1 = 10$ with $w_0 = \frac{1}{4}$ and $w_1 = \frac{3}{4}$. It is evident from Figure 4-33 that the FS error rate estimates are very comparable and outperform the LOO error rate estimates both in terms of bias (left panel) and MSE (right panel), specifically, for larger shifts in distribution.

Summary

Overall, we find that the FS error rate estimates ($\widehat{er}_{up}^*$ and $\widehat{er}_{min}^*$) consistently outperform the LOO error rate estimates ($\widehat{el}_{up}$ and $\widehat{el}_{min}$) in terms of bias. The MSE for $\widehat{er}_{min}^*$ is consistently the lowest at larger shifts in distribution, however, for smaller shifts, $\widehat{el}_{min}$ outperforms $\widehat{er}_{min}^*$ in terms of MSE. Therefore, when the overall rate of misclassification is of interest these properties should be kept in mind when choosing between the FS and LOO error rate estimators.

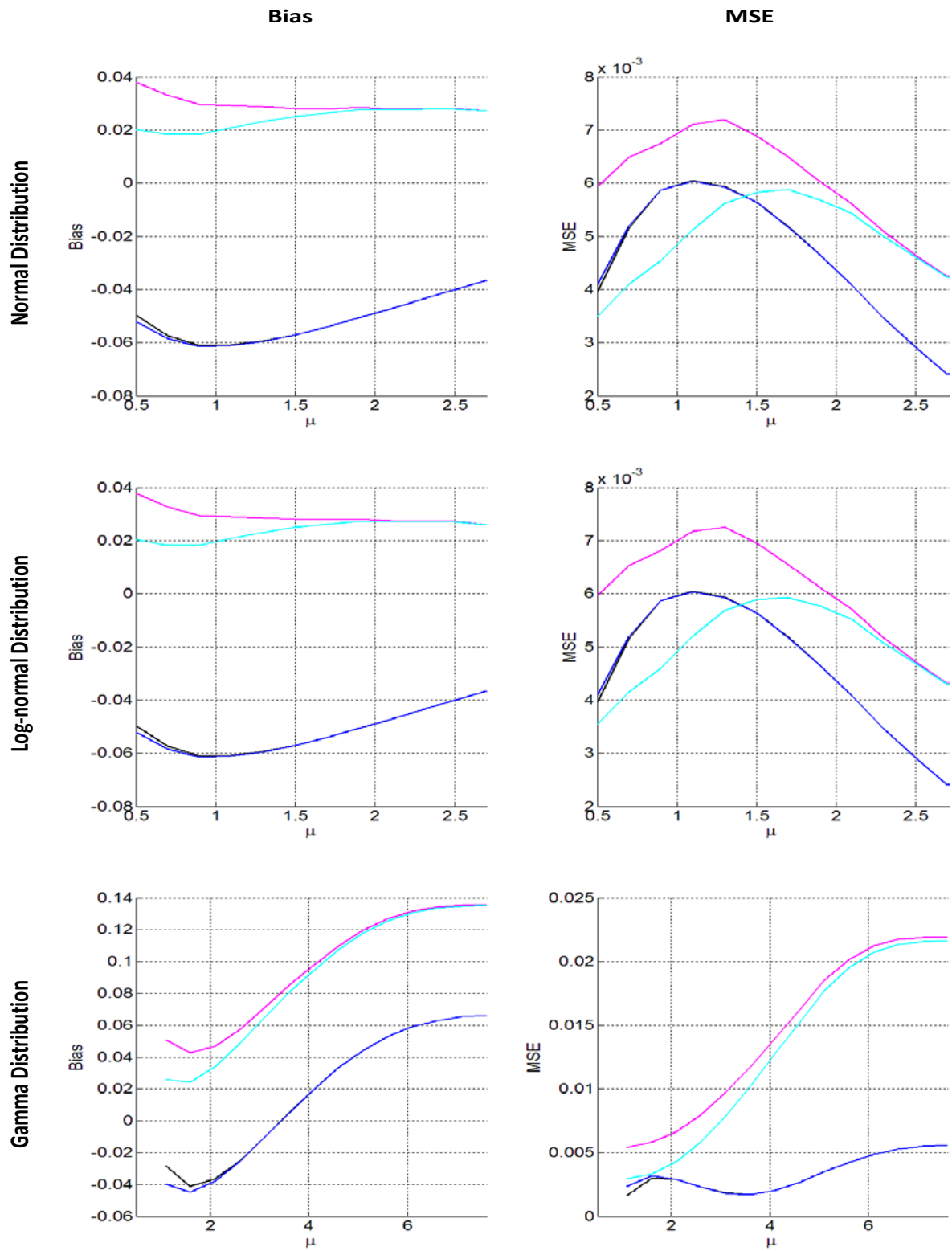


Figure 4-30: Comparing error rate for $N_0 = 10$ and $N_1 = 20$ with $w_0 = \frac{3}{4}$ and $w_1 = \frac{1}{4}$.

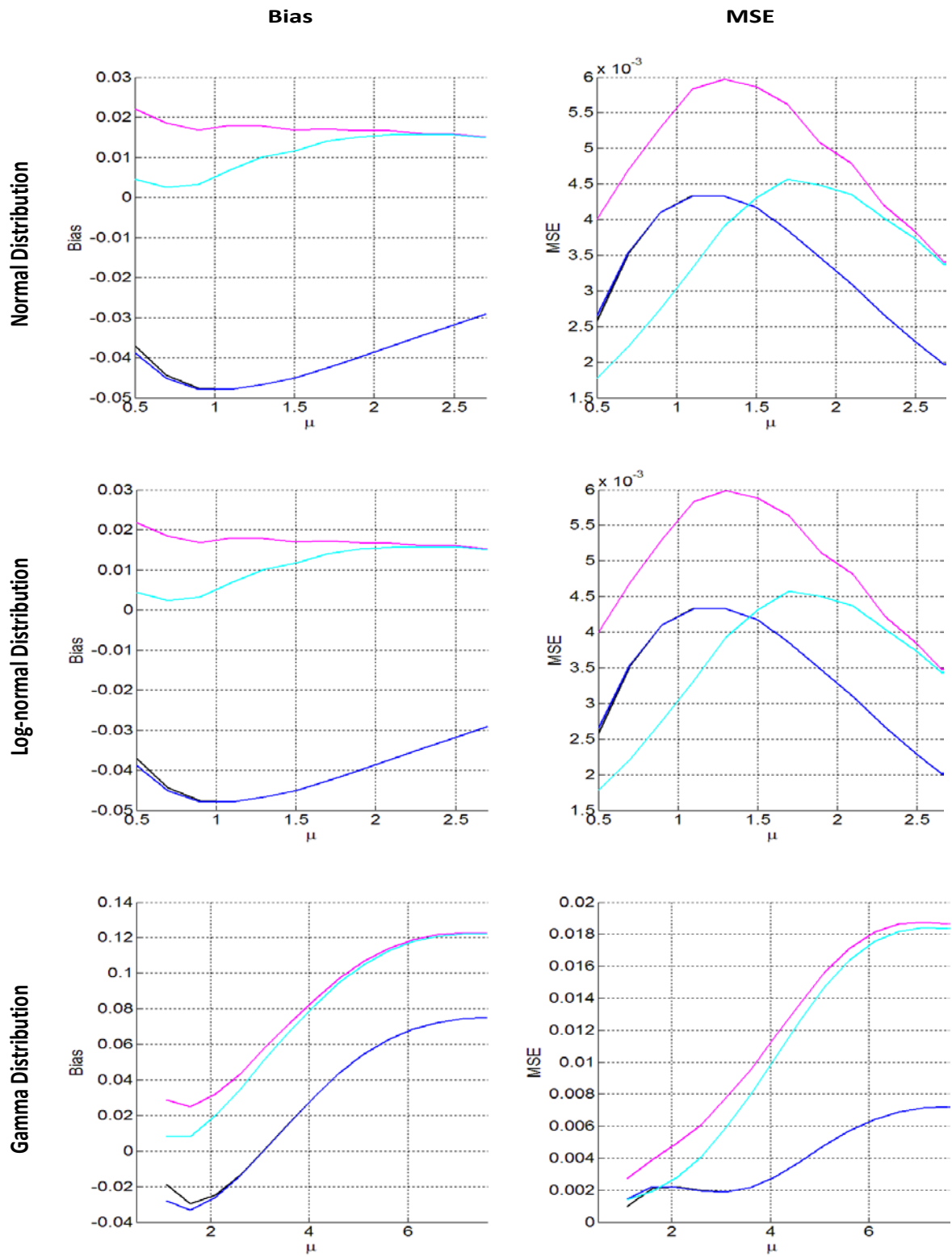


Figure 4-31: Comparing error rate estimates for $N_0 = 20$ and $N_1 = 10$ with $w_0 = \frac{3}{4}$ and $w_1 = \frac{1}{4}$.

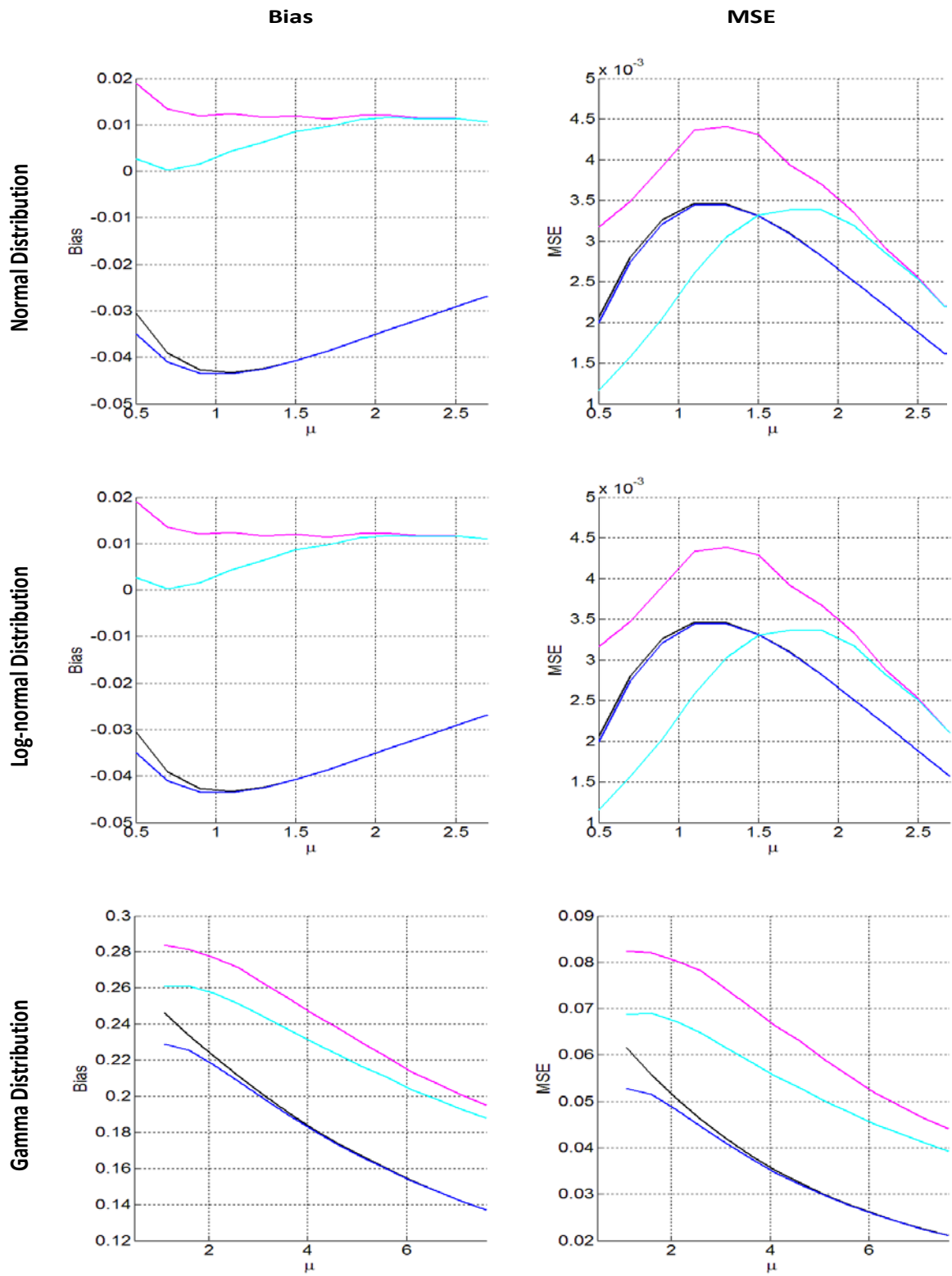


Figure 4-32: Comparing error rate estimates for $N_0 = 20$ and $N_1 = 20$ with $w_0 = \frac{1}{4}$ and $w_1 = \frac{3}{4}$.

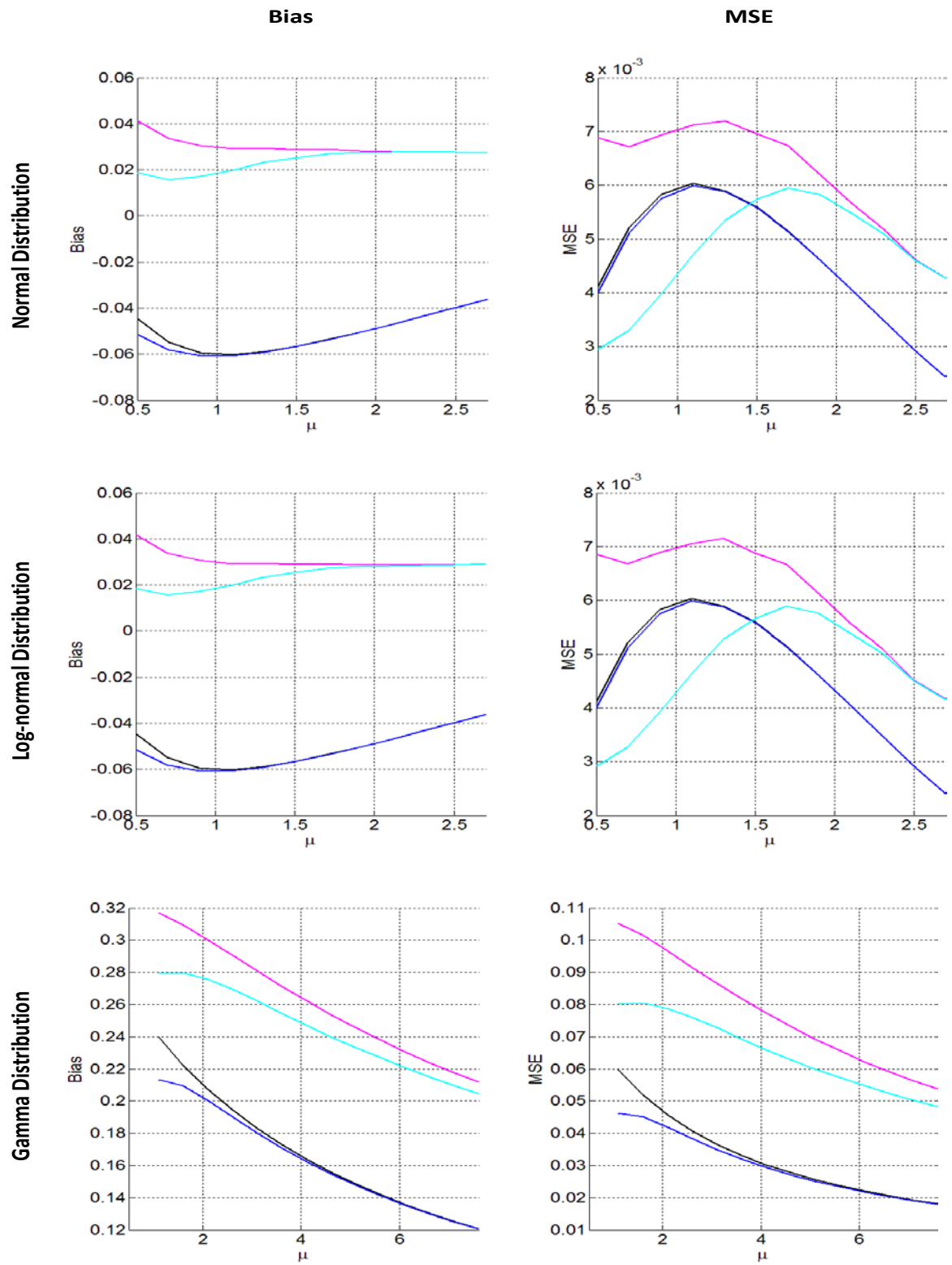


Figure 4-33: Comparing error rate estimates for $N_0 = 20$ and $N_1 = 10$ with $w_0 = \frac{1}{4}$ and $w_1 = \frac{3}{4}$.

4.3.11 Complete results for metabolomics study

Table 4-7 and Table 4-8 provide the results for both the LOO ERp approach when applied to a real world metabolomics dataset, as outlined in the paper. The first column provides the variable names ordered from smallest to largest error rate, as listed in the second column (ER). The third column provides the mean LOO estimate for the threshold (i.e. C) which can be used to classify new subjects by employing the “up” or “down” rule as indicated by the fourth column (i.e. Direction). The fifth column provides the p -value associated with the error rate expressed as a percentage. The significance of these values can be determined when compared to the selected FWER (after adjustment for multiple testing). That is, any p -value lower than the corresponding Bonferroni-Holm adjusted alpha (also expressed as a percentage) is significant. The three last columns provide these values for three different FWER of 1%, 5% and 10%. The red, green and purple squares encapsulate the variables that were significantly shifted at a 1%, 5% and 10% level, respectively. Lines are coloured white or grey to group variables with the same error rates and therefore p -value, since the most conservative FWER correction is applied to each such grouping. No directional assumptions were made for any of the variables and two weighting scenarios are demonstrated (i) $w_0 = w_1 = \frac{1}{2}$; and (ii) $w_0 = \frac{1}{3}, w_1 = \frac{2}{3}$.

Table 4-4 contains an excerpt of results when applying $\hat{el}_{min}$ with $w_0 = w_1 = \frac{1}{2}$. The first column provides the variable names. The second column list the associated error rate (ER). The third column provides the mean LOO estimate for the threshold (i.e. C). The direction producing the lowest error rate is indicated in the fourth column (i.e. Direction). The fifth column provides the p -value associated with the error rate expressed as a percentage. The three last columns provide these values for three different FWER of 1%, 5% and 10%. The red, green and purple squares encapsulate the variables that were significantly shifted at a 1%, 5% and 10% level, respectively.

Metabolite	ER	C	Direction	p-value (%)	1% BH (%)	5% BH (%)	10% BH (%)
glucose (3.21)	0.0000	0.0094	Down	0.0000	0.0182	0.0909	0.1818
glucose (3.24)	0.0000	0.0325	Down	0.0000	0.0182	0.0909	0.1818
glucose (3.38)	0.0000	0.0559	Down	0.0000	0.0182	0.0909	0.1818
glucose (3.48)	0.0000	0.0632	Down	0.0000	0.0182	0.0909	0.1818
glucose (5.22)	0.0000	0.0216	Down	0.0000	0.0182	0.0909	0.1818
glucose (4.63)	0.0238	0.0170	Down	0.0000	0.0200	0.1000	0.2000
C(13)Lactate (1.27)	0.0417	0.0005	Up	0.0000	0.0204	0.1020	0.2041
lactate (1.41)	0.0417	0.1882	Up	0.0000	0.0204	0.1020	0.2041
lactate (4.36)	0.0417	0.0621	Up	0.0000	0.0204	0.1020	0.2041
C(13)Lactate (1.283)	0.0655	0.0005	Up	0.0000	0.0217	0.1087	0.2174
alanine/? (1.51)	0.0655	0.0054	Up	0.0000	0.0217	0.1087	0.2174
succinate (2.66)	0.0655	0.0002	Up	0.0000	0.0217	0.1087	0.2174
creatinine (3.13)	0.0714	0.0027	Down	0.0000	0.0233	0.1163	0.2326
glucose (3.71)	0.0893	0.0984	Down	0.0000	0.0238	0.1190	0.2381
glucose (3.86)	0.0893	0.1064	Down	0.0000	0.0238	0.1190	0.2381
valine 1.04)	0.1071	0.0015	Up	0.0002	0.0250	0.1250	0.2500
glucose/glycerol (3.43)	0.1310	0.0404	Down	0.0029	0.0256	0.1282	0.2564
C(13)Lactate (1.53)	0.1369	0.0013	Up	0.0044	0.0263	0.1316	0.2632
2-oxoglutarate (2.68)	0.1369	0.0001	Up	0.0044	0.0263	0.1316	0.2632
lysine (1.73)	0.1845	0.0033	Up	0.0496	0.0278	0.1389	0.2778
acetoacetate (2.30)	0.1905	0.0004	Up	0.0633	0.0286	0.1429	0.2857
val/iso (1.01)	0.2202	0.0022	Up	0.2201	0.0294	0.1471	0.2941
creatinine (4.29)	0.2202	0.0015	Down	0.2201	0.0294	0.1471	0.2941
phenylalanine/? (7.42)	0.2202	0.0025	Up	0.2201	0.0294	0.1471	0.2941

Table 4-4: Results using $\hat{e}l_{min}$ with $w_0 = w_1 = \frac{1}{2}$.

Column 1 contains the variable names ordered from smallest to largest according to their error rates (column 2). Column 3 contains the threshold with the associated “up” or “down” classification rule indicated in column 4. Column 5 displays simulated p -values expressed as percentages. Columns 6 to 8 display the 1%, 5% and 10% FWERs respectively. The red, green and purple blocks group variables that were significantly shifted at a 1%, 5% and 10% FWERs, respectively. The alternating blocks of white and grey group variables with the same error rates and therefore the same p -values.

Similarly, Table 4-5 contains an excerpt of results when applying $\hat{e}l_{min}$ with $w_0 = \frac{1}{3}$ and $w_1 = \frac{2}{3}$. The results of LOO ERp and of FS ERp (reported in the main paper) are very comparable, with only minor changes to estimated error rate, p and threshold values.

Metabolite	ER	C	Direction	p-value (%)	1% BH (%)	5% BH (%)	10% BH (%)
glucose (3.21)	0.0000	0.0094	Down	0.0000	0.0182	0.0909	0.1818
glucose (3.24)	0.0000	0.0325	Down	0.0000	0.0182	0.0909	0.1818
glucose (3.38)	0.0000	0.0559	Down	0.0000	0.0182	0.0909	0.1818
glucose (3.48)	0.0000	0.0632	Down	0.0000	0.0182	0.0909	0.1818
glucose (5.22)	0.0000	0.0216	Down	0.0000	0.0182	0.0909	0.1818
glucose (4.63)	0.0159	0.0170	Down	0.0000	0.0200	0.1000	0.2000
creatinine (3.13)	0.0476	0.0027	Down	0.0000	0.0204	0.1020	0.2041
C(13)Lactate (1.27)	0.0556	0.0005	Up	0.0000	0.0208	0.1042	0.2083
lactate (1.41)	0.0556	0.1882	Up	0.0000	0.0208	0.1042	0.2083
lactate (4.36)	0.0556	0.0621	Up	0.0000	0.0208	0.1042	0.2083
C(13)Lactate (1.283)	0.0714	0.0005	Up	0.0003	0.0222	0.1111	0.2222
alanine/? (1.51)	0.0714	0.0054	Up	0.0003	0.0222	0.1111	0.2222
glucose (3.71)	0.0873	0.1167	Down	0.0006	0.0233	0.1163	0.2326
glucose (3.86)	0.1032	0.1203	Down	0.0031	0.0238	0.1190	0.2381
val/iso (1.01)	0.1190	0.0021	Up	0.0072	0.0244	0.1220	0.2439
C(13)Lactate (1.53)	0.1190	0.0013	Up	0.0072	0.0244	0.1220	0.2439
succinate (2.66)	0.1190	0.0000	Up	0.0072	0.0244	0.1220	0.2439
2-oxoglutarate (2.68)	0.1190	0.0001	Up	0.0072	0.0244	0.1220	0.2439
valine (1.04)	0.1270	0.0015	Up	0.0098	0.0270	0.1351	0.2703
acetoacetate (2.30)	0.1270	0.0004	Up	0.0098	0.0270	0.1351	0.2703
lysine (1.73)	0.1508	0.0033	Up	0.0471	0.0286	0.1429	0.2857

Table 4-5: Results using $\hat{e}l_{min}$ with $w_0 = \frac{1}{3}$ and $w_1 = \frac{2}{3}$.

Column 1 contains the variable names ordered from smallest to largest according to their error rates (column 2). Column 3 contains the threshold with the associated “up” or “down” classification rule indicated in column 4. Column 5 displays simulated p -values expressed as percentages. Columns 6 to 8 display the 1%, 5% and 10% FWERs respectively. The red, green and purple blocks group variables that were significantly shifted at a 1%, 5% and 10% FWERs, respectively. The alternating blocks of white and grey group variables with the same error rates and therefore the same p -values.

Table 4-6 contains an excerpt of results when applying the Mann-Whitney Test.

Metabolite	p-value (%)	1% BH (%)	5% BH (%)	10% BH (%)
C(13)Lactate (1.27)	0.0001	0.0182	0.0909	0.1818
C(13)Lactate (1.283)	0.0001	0.0182	0.0909	0.1818
lactate (1.41)	0.0001	0.0182	0.0909	0.1818
glucose (3.21)	0.0001	0.0182	0.0909	0.1818
glucose (3.24)	0.0001	0.0182	0.0909	0.1818
glucose (3.38)	0.0001	0.0182	0.0909	0.1818
glucose (3.48)	0.0001	0.0182	0.0909	0.1818
lactate (4.36)	0.0001	0.0182	0.0909	0.1818
glucose (5.22)	0.0001	0.0182	0.0909	0.1818
glucose (3.71)	0.0002	0.0217	0.1087	0.2174
glucose (3.86)	0.0002	0.0217	0.1087	0.2174
glucose (4.63)	0.0002	0.0217	0.1087	0.2174
glucose/glycerol (3.43)	0.0003	0.0233	0.1163	0.2326
alanine/? (1.51)	0.0009	0.0238	0.1190	0.2381
creatinine (3.13)	0.0009	0.0238	0.1190	0.2381
succinate (2.66)	0.0013	0.0250	0.1250	0.2500
2-oxoglutarate (2.68)	0.0018	0.0256	0.1282	0.2564
valine 1.04)	0.0029	0.0263	0.1316	0.2632
C(13)Lactate (1.53)	0.0114	0.0270	0.1351	0.2703
val/iso (1.01)	0.0132	0.0278	0.1389	0.2778
myo-inositol (3.285)	0.0176	0.0286	0.1429	0.2857
DMSO4 (3.14)	0.0308	0.0294	0.1471	0.2941
lysine (1.73)	0.0463	0.0303	0.1515	0.3030
creatinine (4.29)	0.0889	0.0313	0.1563	0.3125
phenylalanine/? (7.42)	0.2360	0.0323	0.1613	0.3226

Table 4-6: Results using the Mann-Whitney Test.

The first column provides the variable names ordered from smallest to largest M-W p -value (second column). The last three columns provide corrected FWER of 1%, 5% and 10%, respectively. The red, green and purple squares encapsulate the variables that were significantly shifted at a 1%, 5% and 10% level, respectively. The alternating blocks of white and grey group variables with the same error rates and therefore the same p -values.

From Table 4-6 it is evident that the significantly shifted variables when making use of the Mann-Whitney test, FS ERp and LOO ERp, respectively, are very comparable. However, the Mann-Whitney test cannot classify new subjects or account for unequal misclassification costs, while the LOO approach is much more time-consuming. In the main paper we illustrated how the FS threshold estimates can be used to classify new subjects.

Table 4-7 represents the classification results based on the median LOO threshold with $w_0 = w_1 = \frac{1}{2}$. The body of the table shows the classification result of each significantly shifted variable for each subject, where 0 indicates the subject was classified into the control group and 1 indicates the subject was classified into the experimental group. Misclassifications are indicated in red. The number of misclassification s per variable is reported in the last column. The last three rows (i) provide the final classification based on majority vote; (ii) flag subjects that were misclassified; and (ii) flag potential outliers based on the number of variable that misclassified a give subject compared to the remaining subjects.

Similarly, Table 4-8 represents the classification results based on the median LOO threshold with $w_0 = \frac{1}{3}$ and $w_1 = \frac{2}{3}$. The body of the table shows the classification result of each significantly shifted variable for each subject, where 0 indicates the subject was classified into the control group and 1 indicates the subject was classified into the experimental group. Misclassifications are indicated in red. The number of misclassification s per variable is reported in the last column. The last three rows (i) provide the final classification based on majority vote; (ii) flag subjects that were misclassified; and (ii) flag potential outliers based on the number of variable that misclassified a give subject compared to the remaining subjects.

The results of LOO ERp and of FS ERp (reported in the main paper) are again almost identical while the calculation of the FS threshold is much less onerous.

Subject	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Number of Misclassified Subjects
True Group Indicator	0	0	0	0	0	0	0	0	0	1	1	1	1	1	
Variables	Control Group Classification									Experimental Group Classification					
glucose (3.21)	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1
glucose (3.24)	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1
glucose (3.38)	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1
glucose (3.48)	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1
glucose (5.22)	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1
glucose (4.63)	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1
C(13)Lactate (1.27)	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0
lactate (1.41)	0	0	0	0	0	0	0	0	0	1	1	1	0	1	1
lactate (4.36)	0	0	0	0	0	0	0	0	0	1	1	1	0	1	1
C(13)Lactate (1.283)	0	0	0	0	0	0	0	1	0	1	1	1	1	1	1
alanine/? (1.51)	0	0	0	0	0	0	0	0	0	1	0	0	1	0	3
succinate (2.66)	0	0	0	0	0	0	0	0	0	1	0	1	0	0	3
creatinine (3.13)	0	0	0	0	1	0	0	0	0	1	0	0	1	1	3
glucose (3.71)	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1
glucose (3.86)	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1
valine 1.04)	0	0	0	0	0	0	0	0	0	1	0	0	1	0	3
glucose/glycerol (3.43)	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0
C(13)Lactate (1.53)	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0
2-oxoglutarate (2.68)	1	0	0	0	0	0	0	1	1	1	1	1	0	1	5
Final Classification	0	0	0	0	0	0	0	0	0	1	1	0	1	1	
Misclassified Subjects															
Potential Outliers															

Table 4-7: Classification results using the median LOO threshold with $w_0 = w_1 = \frac{1}{2}$.

The table shows the classification result based on significantly shifted variables for each subject, with 0 indicating classification into the control group, 1 indicating classification into the experimental group and misclassifications in red. The table consists of two sections representing two weigh sets. The last three rows in each section provide the final classification based on the majority vote, flags misclassified subjects, and flags potential outliers, respectively.

Subject	1	2	3	4	5	6	7	8	9	Experimental Group Classification				Number of Misclassified Subjects			
True Group Indicator	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
Variables	Control Group Classification									Experimental Group Classification							
glucose (3.21)	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1	1	1
glucose (3.24)	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1	1	1
glucose (3.38)	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1	1	1
glucose (3.48)	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1	1	1
glucose (5.22)	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1	1	1
glucose (4.63)	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	0
creatinine (3.13)	0	0	0	0	1	0	0	0	0	1	0	0	1	1	1	3	3
C(13)Lactate (1.27)	0	0	0	0	0	0	0	0	0	1	1	1	1	0	1	1	1
lactate (1.41)	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0
lactate (4.36)	0	0	0	0	0	0	0	0	0	1	1	1	1	0	1	1	1
C(13)Lactate (1.283)	0	0	0	0	0	0	1	0	0	1	1	1	1	1	1	1	1
alanine/? (1.51)	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	3	3
glucose (3.71)	0	1	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
glucose (3.86)	0	1	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
val/iso (1.01)	0	0	0	0	0	1	0	0	0	1	0	0	1	0	4	4	4
C(13)Lactate (1.53)	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0
succinate (2.66)	0	0	0	0	0	0	1	1	0	1	1	1	1	1	1	2	2
2-oxoglutarate (2.68)	1	0	0	0	0	0	0	1	1	1	1	1	1	0	1	5	5
valine 1.04)	0	0	0	0	0	0	0	0	0	1	0	0	1	0	3	3	3
acetoacetate (2.30)	1	1	0	0	0	0	1	0	1	1	1	1	1	0	6	6	6
Final Classification	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
Misclassified Subjects																	
Potential Outliers																	

Table 4-8: Classification results using the median LOO threshold with $w_0 = \frac{1}{3}$ and $w_1 = \frac{2}{3}$.

The table shows the classification result based on significantly shifted variables for each subject, with 0 indicating classification into the control group, 1 indicating classification into the experimental group and misclassifications in red. The table consists of two sections representing two weigh sets. The last three rows in each section provide the final classification based on the majority vote, flags misclassified subjects, and flags potential outliers, respectively.

4.4 Accepted Publication: Metabolomics variable selection and classification in the presence of observations below the detection limit using an extension of ERp

4.4.1 Title Page

Metabolomics variable selection and classification in the presence of observations below the detection limit using an extension of ERp

Authors:

Mari van Reenen^{3,§}, Johan A Westerhuis^{1,3}, Carolus J Reinecke³, J Hendrik Venter²

§Corresponding author:

Mari van Reenen email: 12791733@nwu.ac.za

Addresses:

¹Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands

²Centre for Business Mathematics and Informatics, Faculty of Natural Sciences, North-West University (Potchefstroom Campus), Private Bag X6001, Potchefstroom, South Africa.

³Centre for Human Metabolomics, Faculty of Natural Sciences, North-West University (Potchefstroom Campus), Private Bag X6001, Potchefstroom, South Africa

Email addresses:

Mari van Reenen: 12791733@nwu.ac.za

Johan A Westerhuis: j.a.westerhuis@uva.nl

Carolus J Reinecke: carools.reinecke@nwu.ac.za

J Hendrik Venter: hennie.venter@nwu.ac.za

4.4.2 Abstract

Background

ERp is a variable selection and classification method for metabolomics data. ERp uses minimized classification error rates, based on data from a control and experimental group, to test the null hypothesis of no difference between the distributions of variables over the two groups. If the associated p-values are significant they indicate discriminatory variables (i.e. informative metabolites). The p-values are calculated assuming a common continuous strictly increasing cumulative distribution under the null hypothesis. This assumption is violated when zero-valued observations can occur with positive probability, a characteristic of GC–MS metabolomics data, disqualifying ERp in this context. This paper extends ERp to address two sources of zero-valued observations: (i) zeros reflecting the complete absence of a metabolite from a sample (true zeros); and (ii) zeros reflecting a measurement below the detection limit. This is achieved by allowing the null cumulative distribution function to take the form of a mixture between a jump at zero and a continuous strictly increasing function. The extended ERp approach is referred to as XERp.

Results

XERp is no longer non-parametric, but its null distributions depend only on one parameter, the true proportion of zeros. Under the null hypothesis this parameter can be estimated by the proportion of zeros in the available data. XERp is shown to perform well with regard to bias and power. To demonstrate the utility of XERp, it is applied to GC–MS data from a metabolomics study on tuberculosis meningitis in infants and children. We find that XERp is able to provide an informative shortlist of discriminatory variables, while attaining satisfactory classification accuracy for new subjects in a leave-one-out cross-validation context.

Conclusion

XERp takes into account the distributional structure of data with a probability mass at zero without requiring any knowledge of the detection limit of the metabolomics platform. XERp is able to identify variables that discriminate between two groups by simultaneously extracting information from the difference in the proportion of zeros and shifts in the distributions of the non-zero observations. XERp uses simple rules to classify new subjects and a weight pair to adjust for unequal sample sizes or sensitivity and specificity requirements.

Keywords

Detection limit; probability mass at zero; variable selection; classification; metabolomics.

4.4.3 Background

Feature selection and classification in metabolomics can be problematic due to the large number of missing values often present in the data. Specifically, metabolomics data generated through gas chromatography–mass spectrometry (GC–MS), are known to contain many missing values (Gromski *et al.*, 2014). To complicate matters further, values can be missing at random or not, depending on the source of the missing values. Missing values can result from technical limitations, for example peak misalignment, deconvolution errors resulting from the indistinct shape of a peak, the detection limit of the platform or any combination of these sources. A missing value can also have a biological origin, i.e. a metabolite which is a marker for some disease can be truly absent from a healthy sample. Datasets containing missing values can become cumbersome when performing statistical analysis. Kang (2013) lists lack of power and estimation bias as some of the concerns. As a result, many different imputation strategies have been tried and tested, from very basic strategies like replacing missing values by some fraction of the minimum observed value, to more advanced techniques that aim to impute values based on the remaining data. However, imputation is not always ideal or straightforward. The more successful techniques, such as k-nearest-neighbour and random forest (Gromski *et al.*, 2014b; Armitage *et al.*, 2015), require larger sample sizes, a known limitation of metabolomics studies. Other approaches, even the most elementary ones where missing values are replaced by a fixed number, require the tuning or estimation of parameters. It is also ill-advised to make use of a one-size-fits-all approach to dealing with missing values, especially given the different sources of missing values. Armitage *et al.* (2015) propose a combination of zero value imputation when the metabolite is assumed absent from a sample for biological reasons and k-nearest-neighbour when missing values are believed to be the result of the technical limitations of the platform (Armitage *et al.*, 2015). Specifically, applying ERp to data with missing values imputed by random numbers can have some unwanted effects without any real gains, as we show in a comparative study reported in Section 4.5.7 of the supplementary information (SI).

The research into missing value replacement is vast and we do not go into further detail here, instead we propose a somewhat different approach. Ensuring that the data contains as few as possible missing values due to technical errors must be the first line of defence. Again, we do not go into detail here since software packages are continually being improved and developed to

reduce the number of false positives and negatives during peak identification and quantification (Behrends *et al.*, 2011; Lu *et al.*, 2008). The remaining missing values, predominantly resulting from the detection limit of the platform and biological sources, can then reasonably be replaced by zeros. Since a large proportion of zero values still poses a challenge for hypothesis testing, we devise a new test statistic which can accommodate zeros. We first discuss why a new test statistic is needed.

Traditional statistical tests make distributional assumptions or are sensitive to skewed distributions such as those resulting from data with a pronounced frequency of zero values. To control the proportion of zero values a 'zero filter' can be applied and entails the removal of variables containing too many zeros from the data (Smilde *et al.*, 2005). The minimum proportion of zero values required to remove a variable is rather arbitrary, but the common consensus is that the proportion should be high (e.g. at least 50 %). This proportion is now a tuning parameter which forms part of any resulting model and this is not ideal. Also, we cannot guarantee that important metabolites will not be discarded even if the group structure is taken into account. Alternatively, an equal number of zero observations can be removed from each group, e.g. the Chop-Lump approach (Follmann *et al.*, 2009) which proved powerful when combined with the Wilcoxon or t-test. However, this approach will further reduce already small group sizes for which metabolomics research is known. More complex approaches have also been proposed and we group them into three categories: (i) one-part tests that account for the mixture distributions of data with a positive probability of zero values (Hallstrom, 2010; Zhan *et al.*, 2015); (ii) two-part tests that compare the proportions of zeros and the non-zero values separately (Taylor & Pollard, 2009); and (iii) inverted survival analysis methods (Zhang *et al.*, 2009). However, these approaches have their limitations. One- and two-part tests are known to explain the presence of zero values either due to technical (e.g. below detection limit) or biological (e.g. metabolite not present) reasons, but not necessarily both (Gleiss *et al.*, 2015). One- and two-part tests have proved valuable in the variable selection context, but do not have the ability to classify new subjects and constructing classification models as a second phase has been criticized (Berk *et al.*, 2013; van Reenen *et al.*, 2016). Furthermore, two-part tests still rely on independently derived and equal weighted test statistics for the zero and non-zero data and, as a result, may lack power (Dakna *et al.*, 2010). Methods derived from survival analysis have also proven valuable, but require knowledge of the actual detection limit (Zhang *et al.*, 2009).

Our proposed new test statistic is derived from ERp, a recently introduced approach for variable selection and classification with application to metabolomics data (van Reenen *et al.*, 2016). In its

current form, ERp makes use of p-values associated with minimized classification error rates to identify variables that can discriminate between a control and experimental group. These p-values are calculated based on the assumption that the cumulative distribution function (CDF) for the two groups, common under the null hypothesis, starts at zero at the lower limit of its range of values and continuously increases to one, at the upper limit of its range. This assumption does not allow for jumps in the CDF and in particular, does not cater for a positive probability at zero. In this paper we introduce XERp, an extension of ERp, which takes the presence of two sources of zeros into account: (i) zeros representing instances where the metabolite is truly absent from the biological sample (e.g. metabolites depleted or expressed by a disease); and (ii) zeros representing observations below the detection limit of the metabolomics platform used. Low level metabolites of some diseases, which do not have a monogenetic origin, are mostly not reflected as major metabolic perturbations characterized by dominant metabolic biomarkers. These diseases, of which tuberculosis meningitis is an example and the data used in this study, are of particular interest due to their importance in community health. ERp in its current form can already accommodate many of the characteristics of metabolomics data such as small sample sizes, unequal group sizes and data without dominant biomarkers, making XERp an important and logical extension through its accommodation of zero values.

In the Methods Section, we show how the null distributions used in ERp can be extended to take account of a probability mass at zero. We find that though ERp is robust to small proportions of zeros, XERp is more appropriate when these proportions become larger. We explore the impact of having to estimate p-values as the true proportion of zeros is unknown. We also outline the XERp software accompanying this paper. The Results & Discussion demonstrates the sensitivity of the null distributions to the proportion of zeros via simulation and reports the bias and power associated with various p-value estimates. Next, we demonstrate XERp by applying it to a GC–MS metabolomics dataset. The experimental group represents patients suffering from tuberculosis meningitis (TBM) – a disease which is not expressed through one or more dominant diagnostic biomarkers. We find that XERp is able to select biologically relevant metabolites by extracting information from the frequency of zeros, as well as from the distributional shift. In addition, XERp retains the classification ability of ERp and performs well for new subjects, as well as in a leave-one-out (LOO) cross-validation context. We also discuss a comparison to imputation with non-zero values reported in the SI. Finally, we discuss the utility and future prospects of XERp.

4.4.4 Methods

Notation, Terminology & Null Distribution Assumptions

ERp, introduced in van Reenen *et al.* (2016), aims to identify variables with significantly higher (upward shift) or lower (downward shift) values in the experimental group relative to the control group. Upward and downward classification rules are constructed based on a threshold value. The rates of misclassification for both shift directions are minimised over the thresholds, resulting in two minimised error rates $\widehat{er}_{up}^*$ and $\widehat{er}_{down}^*$ for each variable. These minimised error rates are then used as test statistics to test the hypothesis that the distribution of the variable is the same for the control and experimental groups, while the associated minimising thresholds are used to classify new subjects. ERp assumes a common continuous strictly increasing CDF under the null hypothesis. This does not cater for the possibility that the underlying variable assumes the value zero with positive probability. Here we extend the notation and terminology used in van Reenen *et al.* (2016), to account for such a jump in the CDF at zero.

More specifically, consider a variable $W \geq 0$. It may be that the relevant metabolite is not present in the biological sample in which case $W = 0$. It may also be that there is a detection limit $\delta > 0$ and if $W \leq \delta$ then the exact value of W is unknown and the value 0 is recorded instead, while if $W > \delta$ then the exact value of W is recorded. Calling the actually recorded variable X , it is related to the underlying variable W by the equations $X = 0$ if $W \leq \delta$ and $X = W$ if $W > \delta$.

Denote the population CDF of W by H and let H take the value π' in the point 0 (i.e. $H(0) = \pi'$) and assume that $H(x)$ is continuous and increasing in x for $x \geq 0$, where x denotes the argument at which the CDF is evaluated. The jump π' at 0 caters for the possibility that the underlying variable may take the value 0 with positive probability, representing instances where the metabolite is not present at all. Using the relation between the underlying variable W and the observed variable X , it follows that the CDF of X is given by $F(x) = \pi = H(\delta)$ for $0 \leq x \leq \delta$ and $F(x) = H(x)$ for $x > \delta$. Notice that $\pi \geq \pi'$ since π accounts for both the possibilities that the metabolite is not present and that it may be positive but below the detection limit. Next, let G denote the conditional CDF of X given that $X > 0$, i.e. the positive part of the CDF of X . Formally,

$$G(x) = P(X \leq x | X > 0) = \frac{P(0 < X \leq x)}{P(X > 0)} = \frac{F(x) - F(0)}{1 - F(0)} = \frac{F(x) - \pi}{1 - \pi}$$

Taking into account that $F(x) = \pi$ for $0 \leq x \leq \delta$:

$$G(x) = 0 \text{ for } 0 \leq x \leq \delta \text{ and } G(x) = \frac{F(x) - \pi}{1 - \pi} \text{ for } x > \delta \quad (4.4-1)$$

Then F can be expressed in terms of G by:

$$F(x) = \pi \text{ for } 0 \leq x \leq \delta \text{ and } F(x) = \pi + (1 - \pi)G(x) \text{ for } x > \delta \quad (4.4-2)$$

Figure 4-34 below illustrates these assumptions and notation.

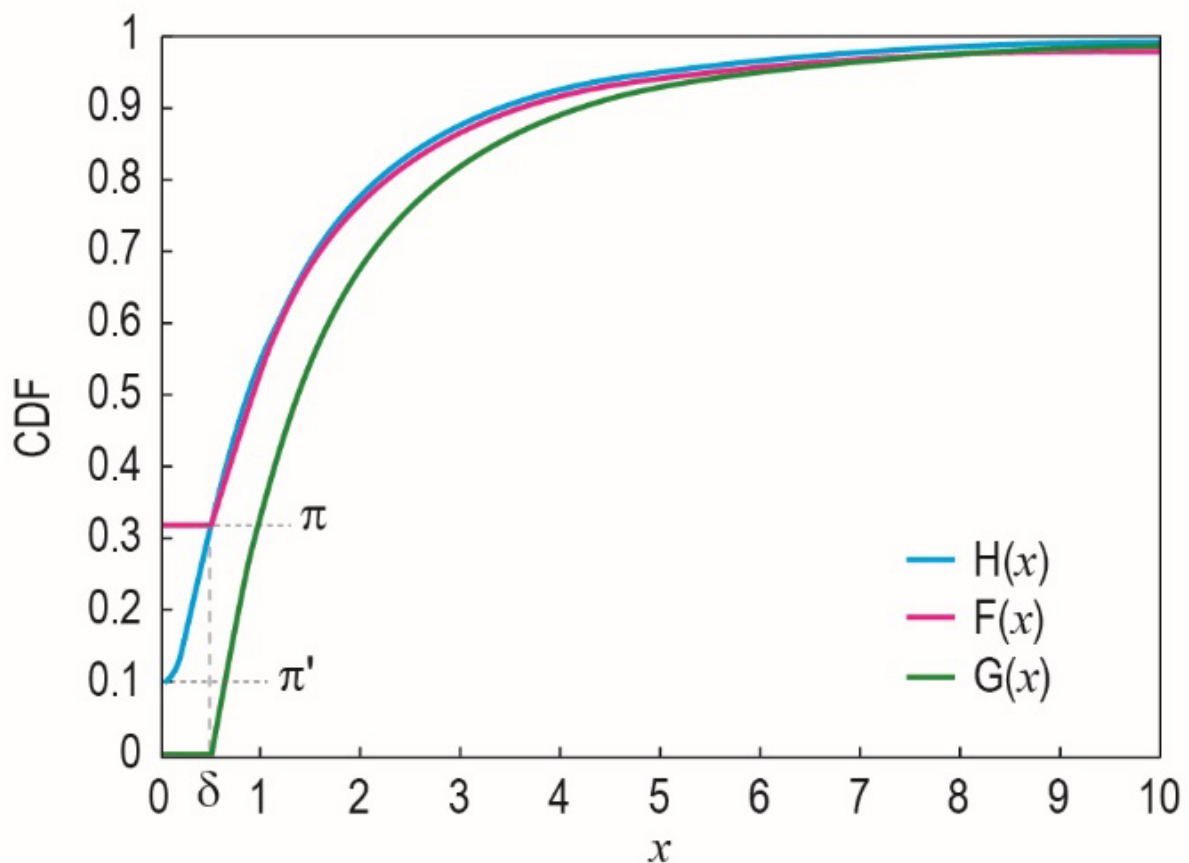


Figure 4-34: An illustration of the CDFs discussed.

The graph illustrates the distributions of the variables W and X using a standard log-normal CDF to depict the positive part of $H(x)$, represented by the blue line, with π' set to 0.1. The assumptions on $H(x)$ imply that $G(x)$, represented by the green line, is continuous and increasing over $x > \delta$ with δ set to 0.5. For $x \leq \delta$, $F(x) = \pi$ which is equal to 0.32 in this illustration, represented by the pink line. For $x > \delta$, $F(x) = H(x)$, but their graphs were slightly shifted for clarity.

The CDF of X is a mixture of a jump of size π at $x = 0$ and a continuous CDF $G(x)$ over $x > 0$. Similar mixed distributions were used by Schisterman *et al.* (2008) in the context of estimating the Youden Index. In their estimation context, two distributions of this type are required, one for the control group and another for the experimental group. The XERp context is simpler in that only one such distribution is required since the distributions of the two groups are the same under the null hypothesis. XERp is also related to the developments of Ruopp *et al.* (2008) for estimating the Youden Index in the presence of observations below the detection limit. However, XERp bases variable selection on the well-known concept of a p-value, providing the added benefit of a clear interpretation of the variable selection threshold (i.e. the significance level α). The Youden Index, on the other hand, has no practical interpretation - as discussed in van Reenen *et al.* (2016). In the next section calculation of the null distributions of the error rate test statistics, required for conversion to p-values, are discussed in the XERp context.

The null distribution of $\widehat{er}_{up}^*$

Denote the actual observed data by (y_n, x_n) , $n = 1, 2, \dots, N$. Here N is the total number of subjects present and y_n represents the group label of the n -th subject, taking the value 0 if a subject is in the control group or 1 if it is in the experimental group. Also x_n represents the observed value of X for the n -th subject. For the time being we restrict attention to the error rate associated with the upward rule. With $c \geq 0$ denoting a generic threshold, this rule classifies the n -th subject into the control group if $x_n \leq c$ and into the experimental group otherwise. Let w_0 and w_1 (with $w_0 + w_1 = 1$) represent the respective relative costs of misclassification of control and experimental subjects. Then the error rate for the upward rule with threshold c is

$$\widehat{er}_{up}(c) = \frac{w_0}{N_0} \sum_{n=1}^N I(y_n = 0, x_n > c) + \frac{w_1}{N_1} \sum_{n=1}^N I(y_n = 1, x_n \leq c) \quad (4.4-3)$$

Here N_0 and N_1 are the numbers of subjects in the control and experimental groups respectively. The minimised error rate is $\widehat{er}_{up}^* = \min_{c \geq 0} \{\widehat{er}_{up}(c)\}$ and this is still used as the test statistic to test the hypothesis that $F_0(x) = F_1(x) = F(x)$, where $F(x)$ is the common CDF of X under the null hypothesis for the control and experimental groups as in van Reenen *et al.* (2016). However, $F(x)$ must now take the form 4.4-2, requiring a revised calculation of the null distribution of $\widehat{er}_{up}^*$.

In the XERp context some x_n 's may be zero, while the non-zero x_n 's are all greater than the detection limit δ . The event $x_n > c$ implies that x_n is non-zero and therefore also that $x_n > \delta$.

Hence 4.4-3 can be written as:

$$\begin{aligned}\widehat{er}_{up}(c) = & \frac{w_0}{N_0} \sum_{n=1}^N I(y_n = 0, x_n > c, x_n > \delta) + \frac{w_1}{N_1} \sum_{n=1}^N I(y_n = 1, x_n \leq c, x_n > \delta) \\ & + \frac{w_1}{N_1} \sum_{n=1}^N I(y_n = 1, x_n \leq c, x_n = 0) \end{aligned} \quad (4.4-4a)$$

where we have split the second term according to the two events that $x_n > \delta$ and $x_n = 0$. Next, evaluate each term in equation 4.4-4a for $c > \delta$ and $c \leq \delta$. Considering the first term in 4.4-4a, if $c > \delta$ the intersection of the events $x_n > c$ and $x_n > \delta$ is equivalent to $x_n > c$, which in turn is equivalent to $G(x_n) > G(c)$. If $c \leq \delta$ the intersection of the events $x_n > c$ and $x_n > \delta$ is equivalent to $x_n > \delta$, which in turn is equivalent to $G(x_n) > G(\delta) = G(0) = G(c)$. Over all c , the intersection of the events $x_n > c$ and $x_n > \delta$ is thus equivalent to $G(x_n) > G(c)$.

Considering the second term in 4.4-4a, if $c > \delta$ the intersection of the events $x_n \leq c$ and $x_n > \delta$ is equivalent to $\delta < x_n \leq c$, which in turn is equivalent to $0 < G(x_n) \leq G(c)$. If $c \leq \delta$ the intersection of $x_n \leq c$ and $x_n > \delta$ is vacuous. Over all c , the intersection of the events $x_n \leq c$ and $x_n > \delta$ is thus equivalent to $0 < G(x_n) \leq G(c)$. Finally, consider the third term in 4.4-4a, regardless of whether $c > \delta$ or $c \leq \delta$, the intersection of the events $x_n \leq c$ and $x_n = 0$ is equivalent to $x_n = 0$, which in turn is equivalent to $G(x_n) = G(0) = 0$. Equation 4.4-4a therefore reduces to:

$$\begin{aligned}\widehat{er}_{up}(c) = & \frac{w_0}{N_0} \sum_{n=1}^N I(y_n = 0, G(x_n) > G(c)) + \frac{w_1}{N_1} \sum_{n=1}^N I(y_n = 1, 0 < G(x_n) \leq G(c)) \\ & + \frac{w_1}{N_1} \sum_{n=1}^N I(y_n = 1, G(x_n) = 0) \end{aligned} \quad (4.4-4b)$$

Putting $u_n = G(x_n)$ and $b = G(c)$, it follows that $\widehat{er}_{up}(c)$ in 4.4-4b can be restated as:

$$\begin{aligned}\widehat{er}_{up}(b) = & \frac{w_0}{N_0} \sum_{n=1}^N I(y_n = 0, u_n > b) + \frac{w_1}{N_1} \sum_{n=1}^N I(y_n = 1, 0 < u_n \leq b) \\ & + \frac{w_1}{N_1} \sum_{n=1}^N I(y_n = 1, u_n = 0) \end{aligned} \quad (4.4-5)$$

Note that the terms in equation 4.4-5 neatly address the group differences we want to investigate. The first two terms evaluate the presence of an upward shift in the distribution of the experimental group, while the third term evaluates the presence of a difference in the proportion of zero observations.

The range $c \geq 0$ is equivalent to $0 \leq b \leq 1$ so that $\widehat{er}_{up}^* = \min_{0 \leq b \leq 1} \{\widehat{er}_{up}(b)\}$. To derive the null distribution of $\widehat{er}_{up}^*$ from this expression requires the common CDF of the u_n 's. Note that the u_n 's are independent and identically distributed (*iid*), since the x_n 's were assumed to be *iid* for the purpose of calculating the null distribution. By definition the common CDF of the u_n 's is $P(u_n \leq u)$, where u denotes the argument of the CDF. Consider separately the cases $u = 0$ and $u > 0$:

$$P(u_n = 0) = P(G(x_n) = 0) = P(x_n = 0) = F(0) = \pi \quad (4.4-6a)$$

$$\begin{aligned} P(u_n \leq u) &= P(G(x_n) \leq u) = P(x_n \leq G^{-1}(u)) = F(G^{-1}(u)) \\ &= \pi + (1 - \pi)G(G^{-1}(u)) = \pi + (1 - \pi)u \end{aligned} \quad (4.4-6b)$$

The common CDF of the u_n 's is also a mixture between a jump at zero of size π and a uniform distribution on (0,1). The distribution of the u_n 's only depends on π and, since $\widehat{er}_{up}(b)$ is only a function of the u_n 's, the same holds for $\widehat{er}_{up}^* = \min_{0 \leq b \leq 1} \{\widehat{er}_{up}(b)\}$.

To conclude, the null distribution of $\widehat{er}_{up}^*$ 4.4-3 depends only on the parameter π and not on any of the other parameters (π' , δ or the positive part CDF $G(x)$). Though $\widehat{er}_{up}^*$ is no longer fully non-parametric, as was the case in van Reenen *et al.* (2016), the only remaining unknown parameter is π . Not needing to take into account π' , δ or $G(x)$ when calculating p-values is a major advantage since these quantities are typically all unknown. Furthermore, the dependence on π only becomes pronounced for larger π , as shown in Figure 4-35 below. To calculate the null-distribution of $\widehat{er}_{up}^*$ for any given value of the parameter π via simulation, the algorithm in Table 1 of van Reenen *et al.* (2016) can still be followed, with the exception that the u_n 's must be sampled from the CDF given by 4.4-6a and 4.4-6b. This is easily achieved by drawing v_n from a uniform (0,1) distribution and setting $u_n = 0$ if $v_n \leq \pi$ and $u_n = (v_n - \pi)/(1 - \pi)$ otherwise (refer to Section 4.5.2 of the SI for more detail). In the next section, several solutions are proposed to finding p-values in the presence of the remaining unknown parameter π .

A simulation study was performed to assess the sensitivity of the null distribution of $\widehat{er}_{up}^*$ to changing values of π . Figure 4-35 illustrates these null distributions for π varying from 0 to 0.8 in steps of 0.2. One million simulation repetitions, group sizes $N_0 = N_1 = 20$ and the weight set $w_0 = w_1 = \frac{1}{2}$ were used to calculate the null-distributions.

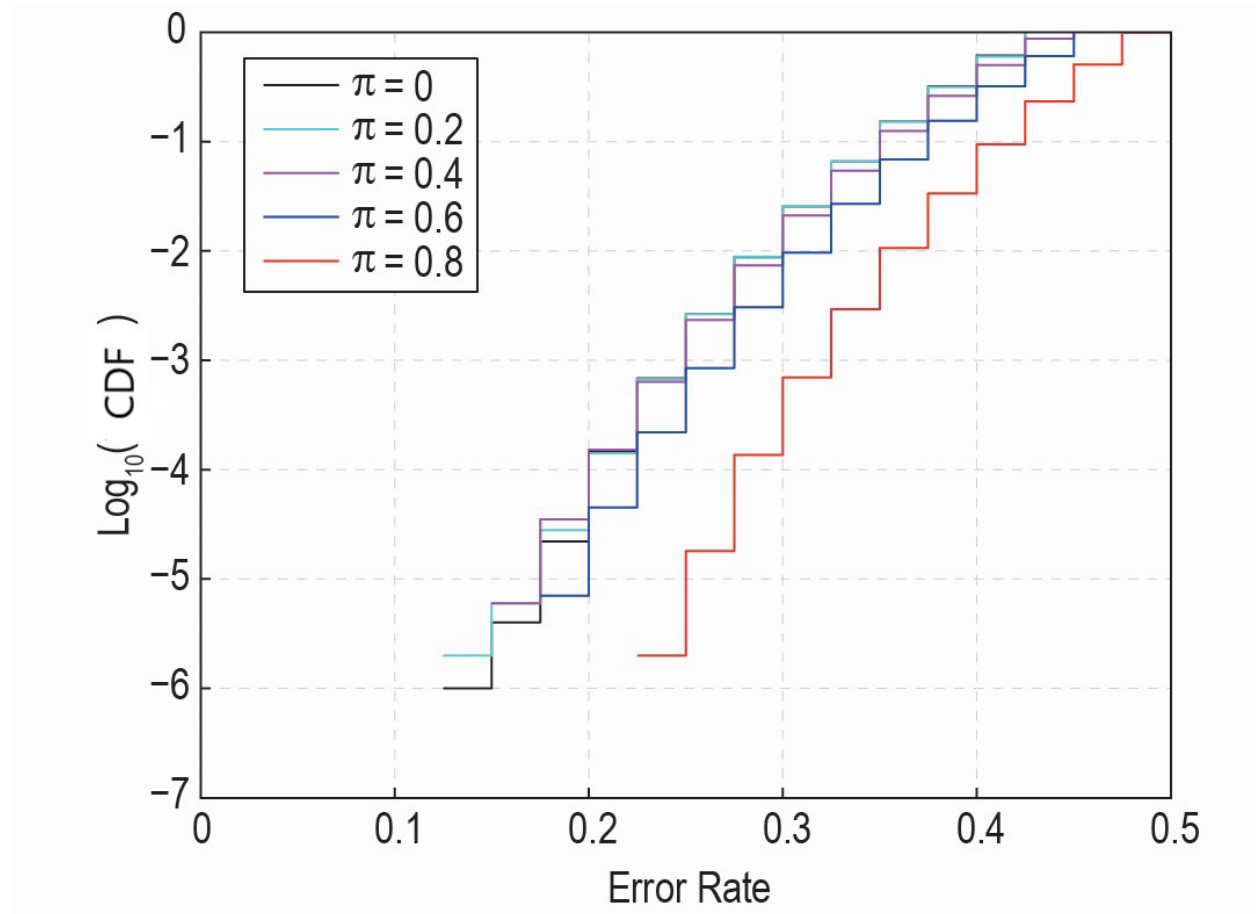


Figure 4-35: Null CDF for $\widehat{er}_{up}^*$ with π taking on different values.

Each line represents the CDF for a different value of π , $\pi = 0$ (black), $\pi = 0.2$ (light blue), $\pi = 0.4$ (purple), $\pi = 0.6$ (dark blue) and $\pi = 0.8$ (red). The CDFs are plotted on a log10 scale for clarity purposes, since interest centres on the extreme left tail.

Note that the distributions are discrete, with jumps at the possible values of $\widehat{er}_{up}^*$, being the different numbers in the list $\{(w_0 n_0 + w_1 n_1)/N : n_0 = 1, 2, \dots, N_0; n_1 = 1, 2, \dots, N_1\}$. The null distribution of $\widehat{er}_{up}^*$ changes slowly when π is small - in fact the graphs for $\pi = 0$ and 0.2 are almost indistinguishable in Figure 4-35. Even at $\pi = 0.4$ the differences are quite small, but become increasingly pronounced for larger π . Note that for $\pi = 0.8$ the million simulation repetitions yielded no error rate values below 0.225, implying that the probability below 0.225 rapidly becomes very small for the sample sizes used in this illustration. In fact, under the null hypothesis

the probability of the event $\widehat{er}_{up}^* = 0$ can be calculated analytically. This is shown in Section 4.5.3 of the SI, which also gives this probability as a function of N_0 , N_1 and π in a reference table.

Computing p-values

Next p-values need to be computed for observed values of $\widehat{er}_{up}^*$. This is no longer straightforward, since the null distribution depends on π and we do not know its true value. In this section three possible choices of p-values are discussed.

Let p_π denote the p-value of an observed $\widehat{er}_{up}^*$ when referred to the null CDF with the true parameter π . Since we do not actually know the value of π , we cannot use p_π in practice and need specific choices. The first choice is p_0 , which implies referring $\widehat{er}_{up}^*$ to the null CDF with $\pi = 0$. This amounts to reverting back to the original ERp p-value and ignoring the possible effects of the zeros.

The second choice is the maximum p-value, defined as $p_{max} = \max_{0 \leq \pi \leq 1} \{p_\pi\}$, and is aimed at ensuring the Type I error rate is controlled since rejection of the null hypothesis at significance level α using p_{max} implies $p_{max} \leq \alpha$. Therefore $p_\pi \leq p_{max} \leq \alpha$ and the null hypothesis will also be rejected if the true p_π were used. The calculation of this estimate requires some additional simulation: (i) calculate the null distributions over a grid of all possible π values given the sample sizes in each group (i.e. N_0 and N_1); (ii) list the individual p-values, pertaining to the observed error rate, as π varies over this grid; (iii) then p_{max} is the maximum in this list. In practice it is not necessary to use a fine grid since the maximum usually occurs at small choices of π where the null-distribution does not change dramatically (Figure 4-35).

The third choice estimates π by the proportion of zeros observed in the dataset (denoted by $\hat{\pi}$) and uses the corresponding p-value, i.e. $p_{obs} = p_{\hat{\pi}}$.

In datasets with a low frequency of zeros (reflected in $\hat{\pi}$ being small), p_0 , p_{max} and p_{obs} should yield similar results since the null distribution changes only slowly for small values of π (Figure 4-35). If one is determined to control the Type I error rate, p_{max} would be a reasonable choice, but this may also imply loss of power. To investigate this further a comparative study was performed between the three proposed p-values, the details of which are reported in the Results & Discussion (Section 4.4.5). The outcomes of this investigation lead to our recommendation of p_{obs} as the best choice.

Other error rates

The developments and results discussed above are based on the upward rule error rate test statistic $\widehat{er}_{up}^*$, but can easily be extended to the downward and minimum error rates, $\widehat{er}_{down}^*$ and $\widehat{er}_{min}^*$, as described in more detail in Section 4.5.1. Importantly, the null distributions of $\widehat{er}_{up}^*$ and $\widehat{er}_{down}^*$ are no longer the same if $\pi > 0$, as was the case in van Reenen *et al.* (2016).

XERp Software

XERp was programmed in MATLAB (2012b) and all scripts and functions are provided as supplementary material along with an example application. The software allows the user to (i) generate the null distributions; (ii) rank variables based on XERp p-values; (iii) select variables for any given significance level after correcting for multiple testing by controlling either the family-wise error rate or the false discovery rate; and (iv) predict group membership of new samples or perform leave-one-out cross-validation. A discussion of the software, as well as a description of the results produced, are included in Section 4.5.4, with a graphical overview provided as Figure 4-38.

4.4.5 Results & Discussion

Comparison of the p-values

Here we report the results of a comparative simulation study to assess the bias and power of the three p-value alternatives. Five performance metrics were used. Firstly, to assess estimation accuracy, the bias and mean squared error (MSE) were used. The bias and MSE were calculated by comparing the different p-value alternatives to p_π , conditioning on a 10 % significance level. Next, the test size (referred to as the size) was used to assess the Type I error probability. The size represents the fraction of times the estimate falsely rejected the null hypothesis given a 10 % significance level (i.e. $\alpha = 0.1$). These three metrics were used to evaluate the performance of p_0 , p_{max} and p_{obs} under the null hypothesis. Finally, the performance of the three p-values were also assessed under the alternative hypothesis. To do so the average p-value and the proportion of null hypothesis rejections (presented in Section 4.5.6) were used to assess the discriminatory power of the different p-values.

The alternative hypothesis was simulated using a log-normal (0,1) distribution for the control group and a shifted log-normal (μ ,1) distribution, with shift equal to μ , for the experimental group. A jump component was added to these distributions by assuming different proportions of zeros in

each group. The simulations were repeated a hundred thousand times for each shift. Null distributions were based on one million repetitions. A more detailed description of the steps to calculate all five performance metrics is provided in sections 4.5.5 and 4.5.6 of the SI. Two group size and weight scenarios were used. The first (scenario 1) corresponds to equal group sizes ($N_0 = N_1 = 20$) and weights ($w_0 = w_1 = \frac{1}{2}$), while the second (scenario 2) is motivated by the metabolomics dataset used to illustrate XERp, namely, $N_0 = 31$; $N_1 = 17$; $w_0 = 0.35$; $w_1 = 0.65$.

Figure 4-36 shows graphs of the bias, MSE and size as functions of π for scenarios 1 and 2. The number of simulation repetitions was a hundred thousand with π varying from 0 to 0.7 and the group size and weigh choices as indicated (left panel: $N_0 = N_1 = 20$ and $w_0 = w_1 = 0.5$ and right panel: $N_0 = 31$; $N_1 = 17$ and $w_0 = 0.35$; $w_1 = 0.65$). The lines represent the results for the three p-value (i) p_0 in dark blue; (ii) p_{max} in purple; and (iii) p_{obs} in light blue. The top row depicts the bias, while the middle row depicts the MSE. The red line in the size graphs (bottom row) corresponds to p_π . The size was calculated for a significance level of 10 %.

The results for p_0 and p_{max} differ so little that they are almost indistinguishable. Moreover, they have rapidly increasing bias and MSE for increasing values of π so that they are not recommendable unless one is quite certain that π is small. The p-value alternative p_{obs} does much better, having small bias and MSE, while remaining so for increasing values of π . Note that the bias and MSE graphs for p_0 and p_{max} were truncated to make the bias and MSE of p_{obs} more visible. The size results, given a 10 % significance level, are presented in the final row of Figure 4-36. Additional size results, given significance levels of 1 % and 5 %, are reported in Figure 4-39 of the SI. It is evident that p_{obs} is better able to retain the significance level specified, compared to the other estimates regardless of whether π is small or large.

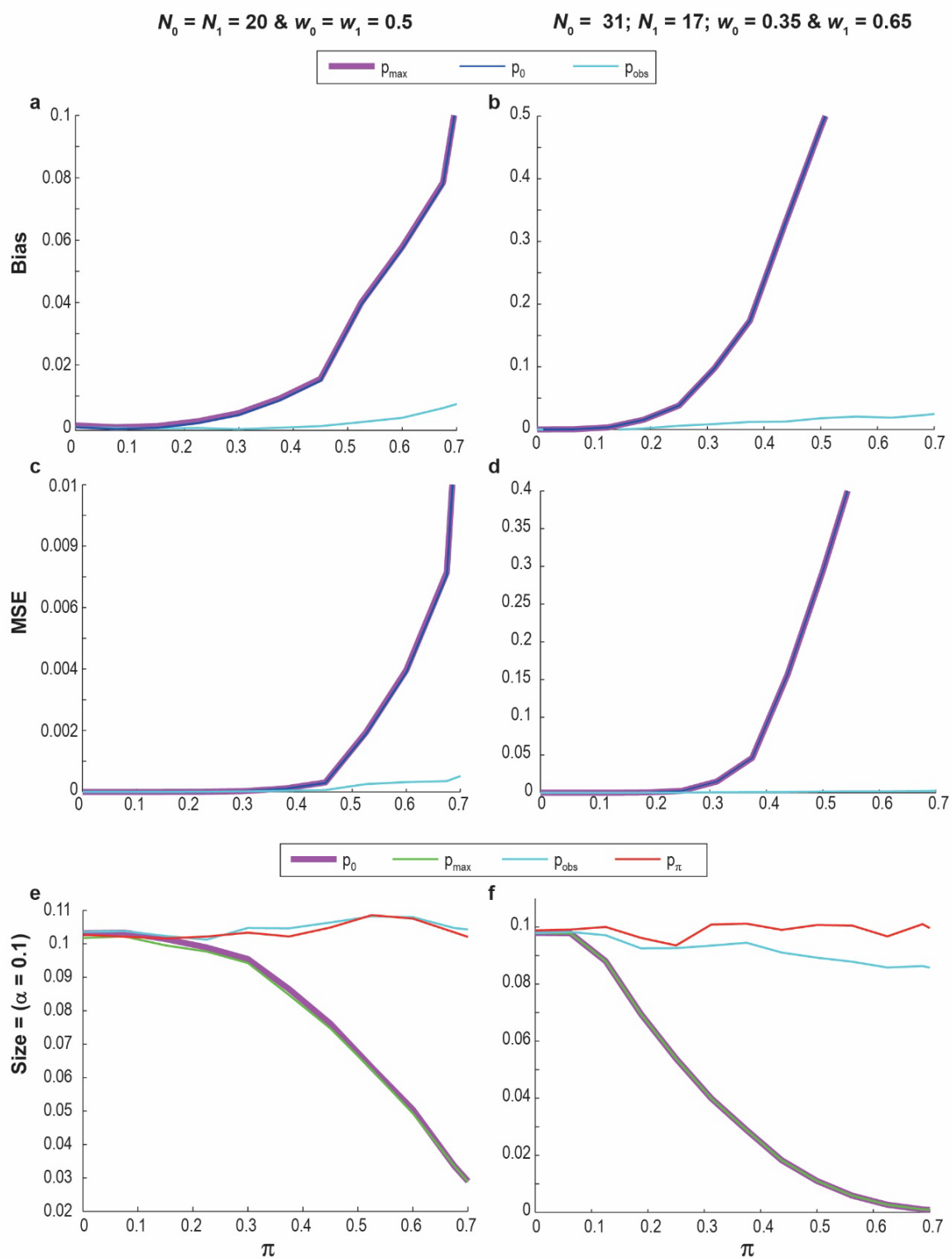


Figure 4-36: Bias, MSE and size of the three p-value alternatives for $\hat{e}r_{up}^*$.

The lines represent the results for the three p-value (i) p_0 in dark blue; (ii) $p_{\max}$ in purple; and (iii) p_{obs} in light blue. The top row depicts the bias, the middle row depicts the MSE and the bottom row the size ($\alpha=10\%$), with the red line corresponds to p_π .

Next, the performance of the three p-value alternatives is compared under the alternative hypothesis, i.e. there is either a difference in the proportion of zeros (the jump part) or in the continuous part of the distributions or in both. Define π_0 and π_1 as the population proportions of zeros among control experimental subjects respectively. The power was assessed given three π_0 and π_1 combinations. The first two represent instances where the proportion of zeros: (i) contains discriminatory information coinciding with the distributional shift - i.e. consonant variables where the group with the larger proportion of zeros has a lower mean (Taylor & Pollard, 2009; Gleiss *et al.*, 2015; Dakna *et al.*, 2010) - represented by the choice $\pi_0 = 0.25$; $\pi_1 = 0$; and (ii) does not contain discriminatory information in that they are equal, represented by the choice $\pi_0 = \pi_1 = 0.25$. A third combination is included in Figure 4-40 to evaluate the power given dissonant variables - i.e. where the proportion of zeros contains discriminatory information different from the distributional shift (Taylor & Pollard, 2009; Gleiss *et al.*, 2015; Dakna *et al.*, 2010) - represented by the choice $\pi_0 = 0$; $\pi_1 = 0.25$. Though this is an unusual scenario (i.e. where the group with the larger proportion of zeros also has a higher mean) it is not unheard of as subjects react differently in the presence of disease and some diseases are known to cause different metabolic changes at different stages.

Figure 4-37 shows the average p-value as a measure of testing power for different zero proportions in the two groups and increasing distributional shifts in the experimental group (μ). Observations for control subjects followed a log-normal (0,1) distribution while those of experimental subjects were drawn from a log-normal (μ ,1) distribution where μ assumed values ranging between 0 and 2. The number of simulation repetitions was a hundred thousand with π_0 and π_1 pairs selected as indicated (i.e. the top row $\pi_0 = 0.25$; $\pi_1 = 0$ and the bottom row $\pi_0 = \pi_1 = 0.25$). The graphs represent the two group size and weigh scenarios as indicated (left panel: $N_0 = N_1 = 20$; $w_0 = w_1 = 0.5$ and right panel: $N_0 = 31$; $N_1 = 17$; $w_0 = 0.35$; $w_1 = 0.65$). The lines represent the averages of the three p-value estimates (i) p_0 in dark blue; (ii) p_{max} in purple; and (iii) p_{obs} in light blue, conditioned on a 10 % significance level. The red lines correspond to p_π .

To make the differences between the p-value alternatives more visible the average p-value was calculated conditioning on $p \leq \alpha$ with α set to 10 %. For the same reason the graphs are not displayed for the entire range of shift values (μ). Once no differences between p-value alternatives are apparent and the average p-values have achieved sufficiently low levels the graphs are no longer displayed. The graphs comparing the proportions of null hypothesis rejections can be found in Figure 4-41 and Figure 4-42.

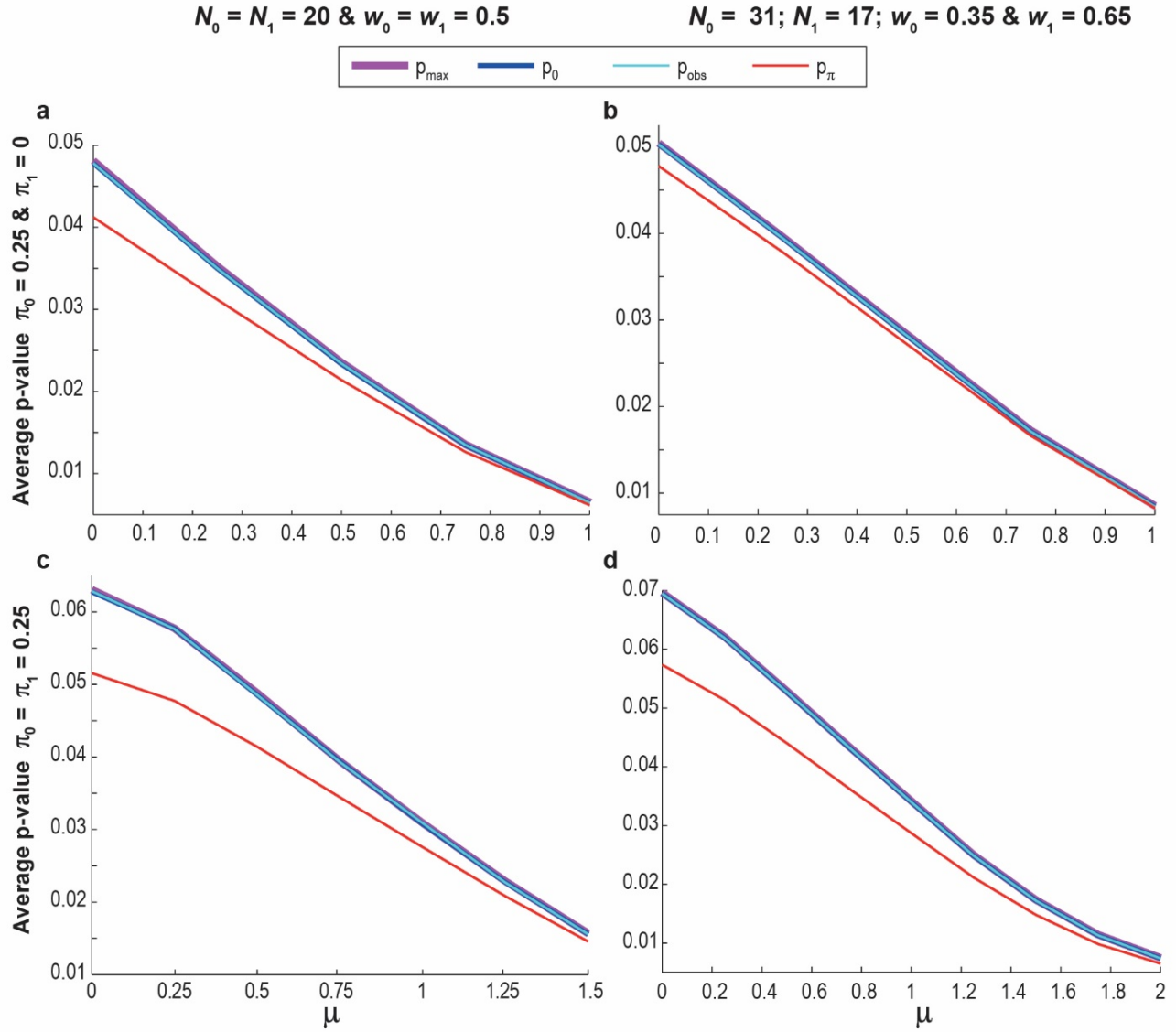


Figure 4-37: Measures of power of the three p-value alternatives for $\hat{e}r_{up}^*$.

The lines represent the averages of the three p-value estimates (i) p_0 in dark blue; (ii) p_{max} in purple; and (iii) p_{obs} in light blue, conditioned on a 10 % significance level. The red lines correspond to p_π .

To interpret Figure 4-37, it is important to note that when comparing p-value alternatives, the p-value with the lower average is able to detect difference faster given the conditions specified, i.e. it has greater power or smaller Type II errors. We find that the three p-value alternatives are almost indistinguishable in most instances and when differences are noted, p_{obs} always outperforms p_{max} and p_0 . The differences between the three p-value alternatives and p_π show the price paid in terms of power for not knowing the true value of π . On the basis of the evidence

presented in Figure 4-36 and Figure 4-37, p_{obs} is a better choice compared to p_0 and p_{max} . The p-value recommended for use in XERp is therefore p_{obs} .

Comparison to imputation

As a further assessment of XERp, a limited comparison to imputation was done. We only briefly discuss the setup and outcome of the comparison here, but report it in greater detail in Section 4.5.7 of the SI. The results from XERp, applied to the data with zeros, was compared to the results from ERp applied to the same data, but with zeros values replaced by positive random numbers. We find that replacing zeros with numbers imputed between some lower bound and the smallest non-zero value can have unwanted effects without real gains. The grid of classification thresholds evaluated, the minimum error rate statistic and corresponding minimising classification threshold and ultimately the associated p-value all become random to some extent when imputing zero values. If the p-value happens to be in the critical area an important variable may be missed or an unimportant variable may be selected by mistake, all depending on numbers chosen at random. Moreover, if a new subject has a zero value for a selected variable, the imputation rule used must first be applied. Classification is then based on whether this imputed number is larger or smaller than the threshold which may also have been estimated from imputed numbers. All this adds to unnecessary uncertainty about conclusions drawn without any obvious gains compared to simply using XERp.

Application to metabolomics data

Here we report the results of XERp as applied to data generated from the GC–MS organic acid analysis of urine samples. Refer to the paper and SI of Mason *et al.* (2016) for a full description of the processing of the data and clinical profiles of patients and controls. The dataset contains concentration levels for 185 variables observed in one experimental and two control groups. Variables refer to 180 substances that could be unequivocally annotated as metabolites and five with insufficient analytical-chemical information to be identified as metabolites, and thus designated as non-annotated variables. The experimental group consisted of 17 children diagnosed with TBM, referred to as the TBM group. The first control group consisted of 31 healthy infants, referred to as the Healthy Controls group. The second control group consisted of 21 seriously ill children whose initial clinical presentation was similar to the TBM cases, but subsequently proved to be negative for TBM, referred to as the Sick Controls group.

The results reported here firstly show the list of metabolites identified by XERp as important based on their ability to discriminate between TBM and Healthy Controls. The weight pair was used to adjust for the differences in sample size by setting them proportional to the inverses of the group sizes which leads to $w_0 = \frac{N_1}{N_0+N_1}$ and $w_1 = \frac{N_0}{N_0+N_1}$ when taking into account that they must add up to 1. One million simulation repetitions were performed to build the null distributions for the different comparisons, while $\widehat{er}_{min}^*$ was used as test statistic.

Table 4-9 shows selected output from the XERp software (as describe in the SI) for the metabolites selected based on p_{obs} .

The table lists the discriminatory variables based on their significant p_{obs} -values when compared to B-H adjusted significance level ($\alpha = 0.05$). The first column lists the names of these variables (labelled 'Variable'), the second the minimized error rate (labelled 'ER') and the third the associated threshold (labelled 'C'). The directions of the minimized error rates are given in the fourth column labelled 'Direction'. Column five contains the p_{obs} -values. The last three columns also report the percentages of zeros in each group and in the combined data set.

The variable selection cut-off was based on the Bonferroni-Holm (B-H) multiple significance testing method, as explained in van Reenen *et al.* (2016), at a family-wise error rate of 5 %. Note that the software accompanying this paper can alternatively correct for multiple testing by controlling the false discovery rate, please refer to Section 4.5.4 of the SI for more detail.

Variable	ER	C	Direction	Observed p-value	Percentage zeros		
					Control	Experimental	Overall
2-hydroxybutyric acid	0.08	0.3	Up	0	68%	0%	44%
3-hydroxyisovaleric acid	0.09	20.05	Up	0	0%	0%	0%
4-hydroxyphenylpyruvic acid	0.11	0	Up	0	100%	18%	71%
methylcitric acid	0.1	0.45	Up	0	55%	6%	38%
quinolinic acid	0.05	2.38	Up	0	45%	0%	29%
2-hydroxyvaleric acid	0.12	0	Up	0	77%	6%	52%
non-annotated-1	0.15	0	Up	0	100%	24%	73%
uracil	0.13	0.67	Up	0.00001	48%	0%	31%
1,4-dihydroxycyclohexane	0.14	0.23	Up	0.00003	55%	6%	38%
non-annotated-3	0.19	0	Up	0.00005	100%	29%	75%
2-ketoglutaric acid	0.16	0.56	Up	0.00006	55%	6%	38%
phenylacetylglutamine	0.17	1.02	Up	0.00008	90%	24%	67%
hexanoic acid	0.16	0.34	Down	0.00013	35%	35%	35%
pyruvic acid	0.17	0.25	Up	0.00014	29%	0%	19%
isocitric acid	0.17	1.6	Up	0.00021	32%	0%	21%
glycolic acid	0.17	12.98	Up	0.00021	0%	0%	0%
pyroglutamic acid	0.17	4.69	Up	0.00022	52%	18%	40%
vanillylmandelic acid	0.18	7.5	Up	0.00027	13%	0%	8%

Table 4-9: XERp Results for TBM vs Healthy Controls.

The first column lists the significant variables (labelled 'Variable'), the second the minimized error rate (labelled 'ER') and the third the associated threshold (labelled 'C'). The directions of the minimized error rates are given in the fourth column labelled 'Direction'. Column five contains the p_{obs} -values. The last three columns also report the percentages of zeros in each group and in the combined data set.

It is beyond the scope of this paper to provide an extensive interpretation of the biological significance of the eighteen variables listed in Table 4-9, but note that fifteen of the sixteen annotated metabolites bear some relationship to the clinical consequences in patients suffering from of TBM. The variables 2-hydroxybutyric, 3-hydroxyisovaleric, 4-hydroxyphenylpyruvic and 2-hydroxyvaleric acids are indicative of a perturbed energy metabolism in the patients due to the disease itself or as a consequence of antibiotic treatment. These variables are of low diagnostic value towards TBM. However, methylcitric acid most likely originated from the well-characterized methylcitrate cycle of *Mycobacterium tuberculosis* (*Mtb*), the bacterium known to induce TBM in

the human host (Muñoz-Elías *et al.*, 2006; Savvi *et al.*, 2008). The presence of quinolinic acid in the urine samples of the TBM patients is likely due to perturbations in the serotonin-tryptophane-pyridoxal phosphate pathways caused by TBM (Blumenthal *et al.*, 2002). In accordance with these biological observations, methylcitric acid and quinolinic acid were recently proposed as two of four metabolites with high diagnostic potential for TBM (Mason *et al.*, 2016). Noteworthy: (1) A variable of unknown chemical structure (non-annotated-1) clearly highlights the importance of chemical characterization of unknown substances associated with infectious diseases in man, given the potential important diagnostic and translational value of these substances; (2) A gut metabolite, 4-hydroxyhipuric acid, included as an important indicator by Mason *et al.* (2016), is not included in this list, but is selected when controlling the false discovery rate rather than the family-wise error rate.

Aside from the biological relevance of the variables listed in Table 4-9, we also note some valuable aspects of XERp. According to Table 4-9, 4-hydroxyphenylpyruvic acid contained 71 % zero values overall of which 100 % occurred in the control and 18 % in the experimental group, while hexanoic acid contained 35 % zero values in both groups. Both are listed as important by XERp, while both could just as easily have been excluded had zero-filtering been applied beforehand, even if the filter accounted for the group structure. Most importantly XERp is able to identify discriminatory variables regardless of whether π is small (e.g. 3-hydroxyisovaleric acid with no zero observations) or large (e.g. non-annotated-3 with 75 % zero observations) and without requiring any knowledge of the detection limit.

Secondly, the classification ability of the resulting list was assessed in two ways: (i) based on LOO cross-validation; and (ii) using the second control group as a hold-out set. This approach was followed for three reasons: (i) given the small group sizes we felt it unwise to select a hold-out or test set from the TBM and Healthy Control groups; (ii) small group sizes are common to metabolomics studies and as such it is important to make available and illustrate the LOO approach to assessing classification ability; and (iii) using a group of difficult to classify subjects (i.e. Sick Controls) allows us to assess the clinical practicality of the list, specifically, whether the list can distinguish between patients with TBM and patients with similar symptoms but not having TBM, indicated by the absence of *Mtb* infection.

The LOO cross-validation results are reported in Table 4-10. The table displays an excerpt of the LOO XERp results. The table is sorted in descending order of the second column ('% Selected'), i.e. the percentage of times a variable (as listed in the first column under 'Variable') was selected

out of the $N = N_0 + N_1$ LOO iterations. In addition, the table reports the specificity (accurate classification of control subjects) and sensitivity (accurate classification of experimental subjects) percentages in columns 3 and 4, as well as the overall classification accuracy in column 5. Columns 6 and 7 contain the threshold values averaged over all instances the variable was selected ('Average Threshold') and the direction of the shift found to be significant ('Direction').

Variable	% Selected	Classification Accuracy %			Average threshold	Direction
		Control Group	Experimental Group	Overall		
4-hydroxyphenylpyruvic acid	100	100	82	94	0	Up
non-annotated-1	100	100	76	92	0	Up
methylcitric acid	100	94	88	92	0.45	Up
quinolinic acid	100	87	94	90	2.38	Up
2-hydroxybutyric acid	100	77	94	83	0.31	Up
2-hydroxyvaleric acid	100	77	94	83	0	Up
phenylacetylglutamine	100	90	71	83	1.02	Up
3-hydroxyisovaleric acid	100	71	100	81	20.02	Up
1,4-dihydroxycyclohexane	100	71	88	77	0.23	Up
2-ketoglutaric acid	100	65	82	71	0.53	Up
uracil	100	61	88	71	0.69	Up
non-annotated-3	75	100	71	90	0	Up
hexanoic acid	67	65	94	75	0.35	Down
isocitric acid	33	52	94	67	1.67	Up
pyruvic acid	17	81	76	79	0.25	Up

Table 4-10: LOO XERp Results for TBM vs Healthy Controls.

The first column lists the significant variables, sorted in descending order of the percentage of times a variable was selected (column 2). The specificity and sensitivity percentages are reported in columns 3 and 4, as well as the overall classification accuracy in column 5. Columns 6 and 7 contain the threshold values averaged over all instances the variable was selected and the direction of the shift.

In each iteration of the $N = N_0 + N_1$ LOO iterations, variables were selected based on p-values derived from the null CDF specific to the reduced group sizes and revised proportions of zeros. The CDFs were constructed using a hundred thousand simulation repetitions and the weight pairs did not change with the changing group sizes. Classification was performed in the exact same

manner as described in van Reenen *et al.* (2016). We made use of the threshold resulting from the corresponding iteration to classify the subject left out. More details on how LOO cross-validation was performed are provided in the SI.

Table 4-10 illustrates the variable selection stability of XERp, as is evident from the '% Selected' column. The top 11 variables (i.e. variables achieving a '% Selected' of 100) were consistently selected regardless of which subject was excluded. Notably, 10 of the top 11 correspond to the top 10 variables selected when no cross-validation is performed, with non-annotated-1, methylcitric and quinolinic acid in the top 4. The selection frequency of lower ranking variables dwindles quickly (75 to 17 %), while the majority of variables were never selected (not shown). The classification accuracy of XERp is high with the top 9 variables achieving sensitivity and specificity levels exceeding 70 % when classifying 'left out' and therefore unseen subjects. The top 4 variables all had overall classification error rates (1-classification accuracy) of 10 % or less, which may be a good argument to reduce the complexity of the list as a classification model to only 4. In addition, the average thresholds do not differ dramatically from those obtained using all available subjects (column 3 of Table 4-9), illustrating the robustness of XERp.

Table 4-11 contains the results when classifying the Sick Controls based on the variable list in Table 4-9. The table lists the classification results for Sick Controls using variable selected when XERp is applied (without LOO cross-validation) to the TBM and Healthy Controls data. The first column lists the variable, while the second and third list the prediction error in terms of absolute count and percentage respectively.

A prediction error is made when a Sick Control is classified into the TBM group. Interestingly non-annotated-3 made no classification errors, while non-annotated-1 only classified 3 Sick Controls as TBM and was also the second highest ranking variable in the LOO cross-validation. Though 4-hydroxyphenylpyruvic acid performed the best in the LOO cross-validation context, it is a marker of disease in general rather than a TBM-specific marker, as it only classified 62 % of Sick Controls correctly.

The potential diagnostic value of methylcitric acid and non-annotated variables was already discussed above. Phenylacetylglutamine and pyroglutamic acid also occurred in the list of important variables summarized in Table 4-9. Interesting, phenylacetylglutamine has been implicated in autism (Al-Ansary *et al.*, 2013), while a metabolomics study highlighted pyroglutamic acid as one of 13 metabolites that differentiate between post-stroke patients group and healthy control subjects (Ding *et al.*, 2016). Both these neuropathological conditions are not related to

TBM and indicate potential importance in clinical chemical studies of diseases that resemble meningitis, but are not caused by *Mtb*.

Variables	Prediction Error	
	Count	Rate
non-annotated-3	0	0%
non-annotated-1	3	14%
phenylacetylglutamine	3	14%
pyroglutamic acid	3	14%
methylcitric acid	4	19%
2-hydroxyvaleric acid	7	33%
quinolinic acid	7	33%
3-hydroxyisovaleric acid	8	38%
4-hydroxyphenylpyruvic acid	8	38%
vanillylmandelic acid	9	43%
pyruvic acid	11	52%
1,4-dihydroxycyclohexane	12	57%
2-ketoglutaric acid	13	62%
glycolic acid	13	62%
uracil	13	62%
2-hydroxybutyric acid	16	76%
isocitric acid	16	76%
hexanoic acid	18	86%

Table 4-11: Classification Results for Sick Controls.

4.4.6 Conclusion

We extended the ERp testing approach to take account of zeros occurring with positive probability by introducing a jump component into the CDF of the underlying variable and named it XERp. Though XERp is no longer nonparametric, it only requires the estimation of one parameter, the proportion of zeros, which can easily be estimated from the available data. XERp is able to simultaneously extract information from differences in the proportion of zeros between two groups as well as the distributional shifts. XERp does not require any knowledge of the detection limit.

The most distinctive feature of XERp is that it is not only a variable selection tool, but also has the ability to directly classify new subjects.

XERp is favourable compared to combining ERp with random imputation of zero values. The latter may lead to threshold values that are only based on the random imputed values and therefore also would cause randomness in the classification of new subjects.

Though XERp does not account for zero-valued observations resulting from other sources, e.g. misalignment during spectral processing, the two sources of the most prominent are indeed addressed, namely, true zeros and observation below the detection limit.

Future research will develop XERp in two ways. The two-part testing approaches briefly mentioned in the Background may hold some benefit as they are able to exploit both consonant as well as dissonant variables. Our first endeavour is therefore to develop XERp along the lines of a two-part test to ensure dissonant variables are correctly evaluated. Secondly, XERp is a univariate approach and only addresses the multivariate nature of metabolomics data by correcting for multiple testing. Future research will aim to generalize XERp to the multivariate setting.

4.4.7 Additional Information

Abbreviations

B-H	Bonferroni-Holm multiple significance testing method
CDF	Cumulative distribution function
ERp	Variable selection for binary classification using error rate p-values.
GC–MS	Gas chromatography–mass spectrometry
iid	Independently and identically distributed
LOO	Leave-one-out
MSE	Mean squared error
<i>Mtb</i>	<i>Mycobacterium tuberculosis</i>
SI	Supplementary information
XERp	Variable selection for binary classification based error rate p-values using an extension of ERp that accounts for zero.

Ethics and Consent to Participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The datasets analysed during the current study are available from the corresponding author on reasonable request and with permission from the authors of the original paper (Mason *et al.*, 2016).

Competing Interests

The authors declare that they have no competing interests.

Funding

Research funding for this project was provided by the Technological Innovation Agency (TIA) of the Department of Science and Technology of South Africa.

Authors' Contributions

MvR and JHV developed the method. JAW provided direction with regard to which aspects of the method to develop. JAW also reviewed and edited the paper. CJR provided the data as well as the biological interpretation of metabolites selected by ERp.

All authors read and approved the final manuscript.

Acknowledgements

We would like to acknowledge Shayne Mason for making the data available to us.

4.5 Supplementary Information: Metabolomics variable selection and classification in the presence of observations below the detection limit using an extension of ERp

4.5.1 The null distribution of $\widehat{er}_{down}^*$ and $\widehat{er}_{min}^*$

Similar to the notation used in the main text, denote the actual observed data by (y_n, x_n) , $n = 1, 2, \dots, N$. Here y_n represents the group label of the n -th subject, taking the value 0 if it is in the control group and 1 if it is in the experimental group. Also x_n represents the observed value of X for the n -th subject. The definitions of minimized error rates for the down rule remains unchanged in the presence of zeros, i.e. $\widehat{er}_{down}^* = \min_{c \geq 0} \{\widehat{er}_{down}(c)\}$ where:

$$\widehat{er}_{down}(c) = \frac{w_0}{N_0} \sum_{n=1}^N I(y_n = 0, x_n \leq c) + \frac{w_1}{N_1} \sum_{n=1}^N I(y_n = 1, x_n > c)$$

Again $F(x)$ is the common CDF of X under the null hypothesis for both the control and experimental groups, but $F(x)$ must now take the form 4.4-2 in the main paper. That is, some x_n 's may be zero, while the non-zero x_n 's are all greater than the detection limit $\delta > 0$:

$$\begin{aligned} \widehat{er}_{down}(c) = & \frac{w_0}{N_0} \sum_{n=1}^N I(y_n = 0, x_n \leq c, x_n > \delta) + \frac{w_0}{N_0} \sum_{n=1}^N I(y_n = 0, x_n \leq c, x_n = 0) \\ & + \frac{w_1}{N_1} \sum_{n=1}^N I(y_n = 1, x_n > c, x_n > \delta) + \frac{w_1}{N_1} \sum_{n=1}^N I(y_n = 1, x_n > c, x_n = 0) \end{aligned}$$

Now evaluate each term in the above equation for $c > \delta$ and $c \leq \delta$. Considering the first term, if $c > \delta$ the intersection of the events $x_n \leq c$ and $x_n > \delta$ is equivalent to $\delta < x_n \leq c$, which in turn is equivalent to $0 < G(x_n) \leq G(c)$. If $c \leq \delta$ the intersection of $x_n \leq c$ and $x_n > \delta$ is vacuous. Considering the second term, regardless of whether $c > \delta$ or $c \leq \delta$, the intersection of the events $x_n = 0$ and $x_n \leq c$ is equivalent to $x_n = 0$, which in turn is equivalent to $G(x_n) = G(0) = 0$.

Consider the third term, if $c > \delta$ the intersection of the events $x_n > c$ and $x_n > \delta$ is equivalent to $x_n > c$, which in turn is equivalent to $G(x_n) > G(c)$. If $c \leq \delta$ the intersection of the events $x_n > c$ and $x_n > \delta$ is equivalent to $x_n > \delta$, which in turn is equivalent to $G(x_n) > G(\delta)$ or $G(x_n) > 0$. Consider the fourth term, regardless of whether $c > \delta$ or not, the intersection of the events $x_n > c$ and $x_n = 0$ is vacuous and this term is equal to zero.

Putting $u_n = G(x_n)$ and $b = G(c)$, it follows that $\widehat{er}_{down}(c)$ can be restated as:

$$\begin{aligned}\widetilde{er}_{down}(b) &= \frac{w_0}{N_0} \sum_{n=1}^N I(y_n = 0, 0 < u_n \leq b) + \frac{w_0}{N_0} \sum_{n=1}^N I(y_n = 0, u_n = 0) \\ &\quad + \frac{w_1}{N_1} \sum_{n=1}^N I(y_n = 1, u_n > b)\end{aligned}$$

The range $c \geq 0$ is equivalent $0 \leq b \leq 1$ so that $\widehat{er}_{down}^* = \min_{0 \leq b \leq 1} \{\widetilde{er}_{down}(b)\}$. To derive the null distribution of $\widehat{er}_{down}^*$ from this expression requires the common CDF of the u_n 's. Note that the u_n 's are independent and identically distributed (*iid*), since the x_n 's were assumed to be *iid* for the purpose of calculating the null distribution. The common CDF of the u_n 's, which by definition is equal to $P(u_n \leq u)$, was derived in the main text as $P(u_n \leq u) = \pi$ for $u = 0$ while $P(u_n \leq u) = \pi + (1 - \pi)u$ for $u > 0$. Again, the common distribution of the u_n 's is also a mixture between a jump at zero of the size π and a uniform distribution on $(0,1)$. The distribution of the u_n 's only depends on π and since $\widetilde{er}_{down}(b)$ is only a function of the u_n 's, the same holds for $\widehat{er}_{down}^* = \min_{0 \leq b \leq 1} \{\widetilde{er}_{down}(b)\}$.

To conclude, the null distribution of $\widehat{er}_{down}^*$ depends only on the parameter π and can be simulated in the same way as $\widehat{er}_{up}^*$. However, since the differences between the equation for $\widetilde{er}_{up}(b)$ and $\widetilde{er}_{down}(b)$ are no longer just in terms of the inequalities, i.e. $>$ vs $\geq$ and $<$ vs $\leq$ as described in van Reenen *et al.* (2016), the null distributions are no longer the same.

Finally, the distribution of the minimum rule can be obtained by simulating

$$\widehat{er}_{min}^* = \min[\min_b \{\widetilde{er}_{up}(b)\}, \min_b \{\widetilde{er}_{down}(b)\}]$$

4.5.2 Simulating the null distribution

The original ERp approach required the following steps to simulate the null distributions needed to convert error rates (in this instance $\widehat{er}_{up}^*$) into p-values:

1. Generate N independent observations from a uniform $(0,1)$ distribution, i.e. the u_n 's.
2. Assign the first N_0 y_n 's as 0 and the remainder as 1.

3. Minimize $\frac{w_0}{N_0} \sum_{n=1}^N (1 - y_n) I(u_n > b) + \frac{w_1}{N_1} \sum_{n=1}^N y_n I(u_n \leq b)$ by varying b over the midpoints of the increasingly ordered u_n 's to obtain $\widehat{er}_{up}^*$.
4. Repeat these steps M times to build up a file of *iid* copies of $\widehat{er}_{up}^*$, say $\widehat{er}_{up}^*(m)$, $m = 1, \dots, M$, providing a simulation approximation of the null CDF.
5. If T of the $\widehat{er}_{up}^*(m)$'s fall below an actually observed $\widehat{er}_{up}^*$ its associated p -value is approximately T/M . Approximations are more accurate for large M .

In the XERp context the u_n 's must be sampled from the CDF given by 4.4-6a and 4.4-6b, as explained in the main manuscript. We do so by first defining v_n 's as *iid* observations from a uniform (0,1) distribution and express the u_n 's relative to the v_n 's by setting $u_n = 0$ if $v_n \leq \pi$ and $u_n = (v_n - \pi)/(1 - \pi)$. To prove that this substitution has the desired effect, we again look at the common CDF of the u_n 's (i.e. $P(u_n \leq u)$, where u denotes the argument of the CDF). Considering separately the cases $u = 0$ and $u > 0$:

$$P(u_n = 0) = P(v_n \leq \pi) = \pi$$

and for $u > 0$

$$\begin{aligned}
 P(u_n \leq u) &= P(u_n \leq u, v_n \leq \pi) + P(u_n \leq u, v_n > \pi) \\
 &= P(v_n \leq \pi) + P\left(\frac{v_n - \pi}{1 - \pi} \leq u, v_n > \pi\right) \\
 &= \pi + P(v_n \leq \pi + (1 - \pi)u, v_n > \pi) \\
 &= \pi + (1 - \pi)u
 \end{aligned}$$

4.5.3 The null hypothesis probability of getting a zero error rate

Here we provide an explicit formula for $P_0(\widehat{er}_{up}^* = 0)$. The event $\widehat{er}_{up}^* = 0$ happens if and only if all the x_n -values of the experimental group are all larger than those of the control group. Under the null hypothesis this amounts to all the u_n -values in 4.4-5 corresponding to the experimental group being larger than those corresponding to the control group. Under the null hypothesis, the u_n 's are *iid* according to 4.4-6a and 4.4-6b and then

$$\begin{aligned}
P_0(\widehat{er}_{up}^* = 0) &= \int_0^1 [\pi + (1 - \pi)u]^{N_0} [1 - \pi - (1 - \pi)u]^{N_1-1} N_1(1 - \pi) du \\
&= N_1(1 - \pi)^{N_1} \sum_{i=0}^{N_0} \frac{N_0!}{i! (N_0 - i)!} \pi^i (1 - \pi)^{N_0-i} \int_0^1 u^{N_0-i} (1 - u)^{N_1-1} du \\
&= N_1(1 - \pi)^{N_1} \sum_{i=0}^{N_0} \frac{N_0!}{i! (N_0 - i)!} \pi^i (1 - \pi)^{N-i} \frac{(N_0 - i)! (N_1 - 1)!}{(N_0 - i + N_1)!} \\
&= N_0! N_1! \sum_{i=0}^{N_0} \frac{1}{i! (N - i)!} \pi^i (1 - \pi)^{N-i}
\end{aligned}$$

This formula gives the null hypothesis probability for a zero upward rule error rate value. In case of a downward rule error rate the same formula applies but with the roles of N_0 and N_1 interchanged. For the minimum rule the two probabilities must be added. Below, Table 4-12 shows the upward rule zero probabilities for N_0 and N_1 ranging from 5 to and 10 and π varying from 0 to 0.9 in steps of 0.1.

N_0 & N_1		π									
		0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
5	5	0.003968	0.003968	0.003943	0.003780	0.003309	0.002472	0.001456	0.000596	0.000130	0.000006
5	6	0.002165	0.002164	0.002139	0.001995	0.001631	0.001082	0.000534	0.000169	0.000025	0.000001
5	7	0.001263	0.001262	0.001238	0.001114	0.000840	0.000489	0.000200	0.000049	0.000005	0.000000
5	8	0.000777	0.000776	0.000754	0.000648	0.000446	0.000226	0.000076	0.000014	0.000001	0.000000
5	9	0.000500	0.000499	0.000478	0.000390	0.000243	0.000106	0.000029	0.000004	0.000000	0.000000
5	10	0.000333	0.000332	0.000313	0.000240	0.000134	0.000050	0.000011	0.000001	0.000000	0.000000
6	5	0.002165	0.002164	0.002160	0.002118	0.001949	0.001571	0.001011	0.000455	0.000109	0.000006
6	6	0.001082	0.001082	0.001078	0.001040	0.000911	0.000663	0.000362	0.000128	0.000021	0.000001
6	7	0.000583	0.000583	0.000579	0.000546	0.000449	0.000291	0.000133	0.000036	0.000004	0.000000
6	8	0.000333	0.000333	0.000329	0.000302	0.000231	0.000132	0.000050	0.000010	0.000001	0.000000
6	9	0.000200	0.000200	0.000196	0.000174	0.000122	0.000061	0.000019	0.000003	0.000000	0.000000
6	10	0.000125	0.000125	0.000122	0.000103	0.000066	0.000028	0.000007	0.000001	0.000000	0.000000

N_0 & N_1		π									
		0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
7	5	0.001263	0.001263	0.001262	0.001251	0.001190	0.001018	0.000709	0.000349	0.000092	0.000005
7	6	0.000583	0.000583	0.000582	0.000572	0.000526	0.000413	0.000248	0.000096	0.000018	0.000001
7	7	0.000291	0.000291	0.000291	0.000282	0.000248	0.000176	0.000090	0.000027	0.000003	0.000000
7	8	0.000155	0.000155	0.000155	0.000148	0.000122	0.000078	0.000033	0.000008	0.000001	0.000000
7	9	0.000087	0.000087	0.000087	0.000081	0.000063	0.000035	0.000012	0.000002	0.000000	0.000000
7	10	0.000051	0.000051	0.000051	0.000046	0.000033	0.000016	0.000005	0.000001	0.000000	0.000000
8	5	0.000777	0.000777	0.000777	0.000774	0.000752	0.000673	0.000503	0.000269	0.000077	0.000005
8	6	0.000333	0.000333	0.000333	0.000330	0.000314	0.000262	0.000171	0.000073	0.000015	0.000000
8	7	0.000155	0.000155	0.000155	0.000153	0.000141	0.000108	0.000061	0.000020	0.000003	0.000000
8	8	0.000078	0.000078	0.000078	0.000076	0.000067	0.000046	0.000022	0.000006	0.000001	0.000000
8	9	0.000041	0.000041	0.000041	0.000039	0.000033	0.000021	0.000008	0.000002	0.000000	0.000000
8	10	0.000023	0.000023	0.000023	0.000021	0.000017	0.000009	0.000003	0.000000	0.000000	0.000000
9	5	0.000500	0.000500	0.000499	0.000499	0.000491	0.000455	0.000360	0.000208	0.000065	0.000005
9	6	0.000200	0.000200	0.000200	0.000199	0.000193	0.000170	0.000119	0.000056	0.000012	0.000000
9	7	0.000087	0.000087	0.000087	0.000087	0.000082	0.000068	0.000041	0.000015	0.000002	0.000000
9	8	0.000041	0.000041	0.000041	0.000041	0.000037	0.000028	0.000015	0.000004	0.000000	0.000000
9	9	0.000021	0.000021	0.000021	0.000020	0.000018	0.000012	0.000005	0.000001	0.000000	0.000000
9	10	0.000011	0.000011	0.000011	0.000010	0.000009	0.000005	0.000002	0.000000	0.000000	0.000000
10	5	0.000333	0.000333	0.000333	0.000333	0.000330	0.000313	0.000261	0.000161	0.000055	0.000004
10	6	0.000125	0.000125	0.000125	0.000125	0.000122	0.000112	0.000084	0.000042	0.000010	0.000000
10	7	0.000051	0.000051	0.000051	0.000051	0.000050	0.000043	0.000028	0.000012	0.000002	0.000000

N_0 & N_1		π									
		0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
10	8	0.000023	0.000023	0.000023	0.000023	0.000022	0.000017	0.000010	0.000003	0.000000	0.000000
10	9	0.000011	0.000011	0.000011	0.000011	0.000010	0.000007	0.000004	0.000001	0.000000	0.000000
10	10	0.000005	0.000005	0.000005	0.000005	0.000005	0.000003	0.000001	0.000000	0.000000	0.000000

Table 4-12: Null hypothesis probabilities of the event that $\widehat{er}_{up}^* = 0$.

The table shows the upward rule zero probabilities for N_0 and N_1 ranging from 5 to and 10 and π varying from 0 to 0.9 in steps of 0.1.

The zero probabilities decrease as π increases, slowly at first but then more rapidly. It also decreases with increasing sample sizes. For sample sizes both 10 or larger, the zero probability is below 0.000005 regardless of π . In particular, if all the control measurements are zero and all the experimental measurements are positive in such a case, then $\widehat{er}_{up}^* = 0$ and the corresponding exact p-value is below 0.000005.

4.5.4 XERp Software

Here we describe the XERp software, its variations, the inputs required and results produced. XERp was programmed in MATLAB (2012b) and the scripts are provided in addition to this document of which Figure 4-38 provides a graphical overview. The following inputs are required:

- A data matrix containing the observed values for all subjects (represented by the rows of the matrix) and variables (represented by the columns of the matrix).
- The variable names in a separate column vector of type character and the same length as the second dimension of the data matrix (i.e. each variable must have a name).
- A vector containing the shift direction of interest for each variable and determining the test statistic to be used. The value 1 indicates that an upward shift should be evaluated, using $\widehat{er}_{up}^*$ as a test statistic; the value -1 indicates that a downward shift should be evaluated, using $\widehat{er}_{down}^*$ as a test statistic; and the value 0 indicates that both shift directions should be evaluated, using $\widehat{er}_{min}^*$ as a test statistic.

- A categorical row vector indicating the group membership of each subjects. The vector must have the same length as the first dimension of the data matrix (i.e. each subject must have been assigned to a group).
- The two string variables used as group labels in the group membership vector.
- The cost of misclassification into the control group. The cost of misclassification into the experimental group is derived from the constraint on the weights, i.e. $w_0 + w_1 = 1$. Alternatively, the user can set $w_0 = 0$ which prompts the software to calculate the weights from the group sizes, as explained in the main text, i.e. $w_0 = \frac{N_1}{N_0 + N_1}$ and $w_1 = \frac{N_0}{N_0 + N_1}$.
- The value of the preferred significance level α as a separate numeric input.
- The method to use when correcting for multiple testing as a separate string variable. There are two options (i) 'FDR' controls the false discovery rate and makes use of the `fdr` script of Arnaud Delorme (2008) sourced through MathWorks (<https://uk.mathworks.com/matlabcentral/fileexchange/27960-resampling-statistical-toolkit/content/statistics/fdr.m>). (ii) 'BH' controls the familywise error rate using the Bonferroni-Holm approach, as described in van Reenen *et al.* (2016).
- The number of simulations to use when calculating the null distributions.
- The file names to use when saving and exporting results.

These user inputs must then be presented to the different scripts provided to (i) generate the null distributions; (ii) perform variable selection based on XERp; (iii) predict group membership based on the results from the variable selection function; or (iv) perform leave-one-out (LOO) cross-validation. The scripts must be called in sequence as results files from one script are required as input to other scripts. The processing sequence is illustrated in Figure 4-38 as well as in the example provided (the `Example.m` file). In addition, the 'help' command can also be used to ascertain the required inputs and resulting output of each script.

Note that a different data may be provided when performing prediction to allow for the classification of subjects other than those used to perform variable selection. Doing so requires a data matrix of observations for the same variables originally presented, but for the new subjects. If the group membership of the new subjects is known, this can also be provided to the prediction script (`XERp_Predict.m`) which will add this information to the results for easier interpretation.

Further, all predictions are performed on a per variable basis and the final classification is left to the user. One way to achieve this is by simply using majority vote as explained in van Reenen *et al.* (2016).

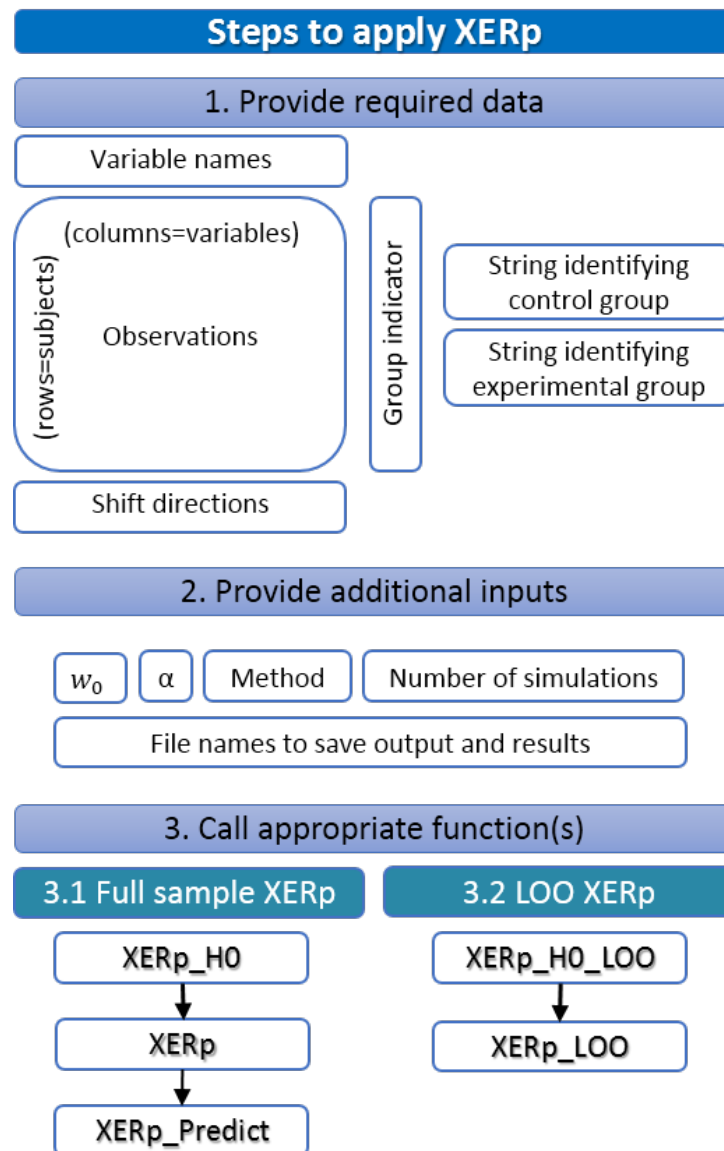


Figure 4-38: Overview of the XERp software

The results from the software are automatically exported to an Excel workbook with the name provided. Two sets of results can be generated, namely, the results when using all subjects (i.e. the full sample or FS results) and the results when applying leave-one-out cross-validation (i.e. the LOO results). The FS results workbook contains the variable selection and classification (i.e. group membership prediction) results. The LOO results workbook contains the results for each

iteration, namely: (i) a binary indicator taking the value 1 if a variable was significant and selected as predictor; (ii) the minimizing threshold for each variable; (iii) the shift direction producing the smallest error rate, if the directional vector was set to 0, or the direction specified; and (iv) the group membership prediction for each variable for the subject 'left out'. In addition, a summary of the LOO results is provided.

4.5.5 Comparison of the p-values under the null hypothesis

A comparative simulation study to assess the bias and power of the three p-value alternatives was done as discussed in the main text. The detailed simulation steps used to evaluate the bias, MSE and size are as follows:

1. Choose values for N_0, N_1, w_0 and w_1 .
2. Let $\{\pi_1, \pi_2, \dots, \pi_J\}$ denote a grid of π values. It is convenient to choose $J = N = N_0 + N_1$ and $\pi_j = (j - 1)/J$ (guaranteeing a sufficient grid). For each π in the grid compute the null CDF by simulation and table the results, referred to as the CDF-Table below.
3. For $j = 1, 2, \dots, J$ and for $m = 1, 2, \dots, M$
 - a. Generate $\{v_n(j, m), n = 1, 2, \dots, N\}$ iid from the uniform (0,1) distribution and transform them to $u_n(j, m) = 0$ if $v_n(j, m) \leq \pi_j$ and $u_n(j, m) = (v_n(j, m) - \pi_j)/(1 - \pi_j)$ otherwise.
 - b. Put $y_n(j, m) = 0$ for $n = 1, 2, \dots, N_0$ and $y_n(j, m) = 1$ for $n = N_0 + 1, N_0 + 2, \dots, N$.
 - c. Compute $\widehat{er}_{up}^*(j, m)$ by minimizing based on the data $\{y_n(j, m), u_n(j, m)\}$, $n = 1, 2, \dots, N$.
 - d. Convert $\widehat{er}_{up}^*(j, m)$ to the three p-value alternatives as defined in the main text. Also convert $\widehat{er}_{up}^*(j, m)$ to the true p-value by referring the $\widehat{er}_{up}^*(j, m)$ to the column corresponding to π_j in the CDF-table, calling the result $p_{\pi_j}(j, m)$.
4. Finally, $B(j) = \frac{1}{M} \sum_{m=1}^M \{p'(j, m) - p_{\pi_j}(j, m)\}$; $MS(j) = \frac{1}{M} \sum_{m=1}^M \{p'(j, m) - p_{\pi_j}(j, m)\}^2$ and $S(j) = \frac{1}{M} \sum_{m=1}^M I(p'(j, m) \leq p_{\pi_j}(j, m))$ estimates the bias, MSE and size of the three p-values (represented here in general by $p'(j, m)$), respectively, when π_j is the true value of π .

Figure 4-39 shows the test size (referred to as the size) used to assess the Type I error probability, as was done in the main text. Again, the size represents the fraction of times the estimate falsely rejected the null hypothesis, but for significance levels of 5 % (Figure 4-39 a and b) and 1 % (Figure 4-39 c and d).

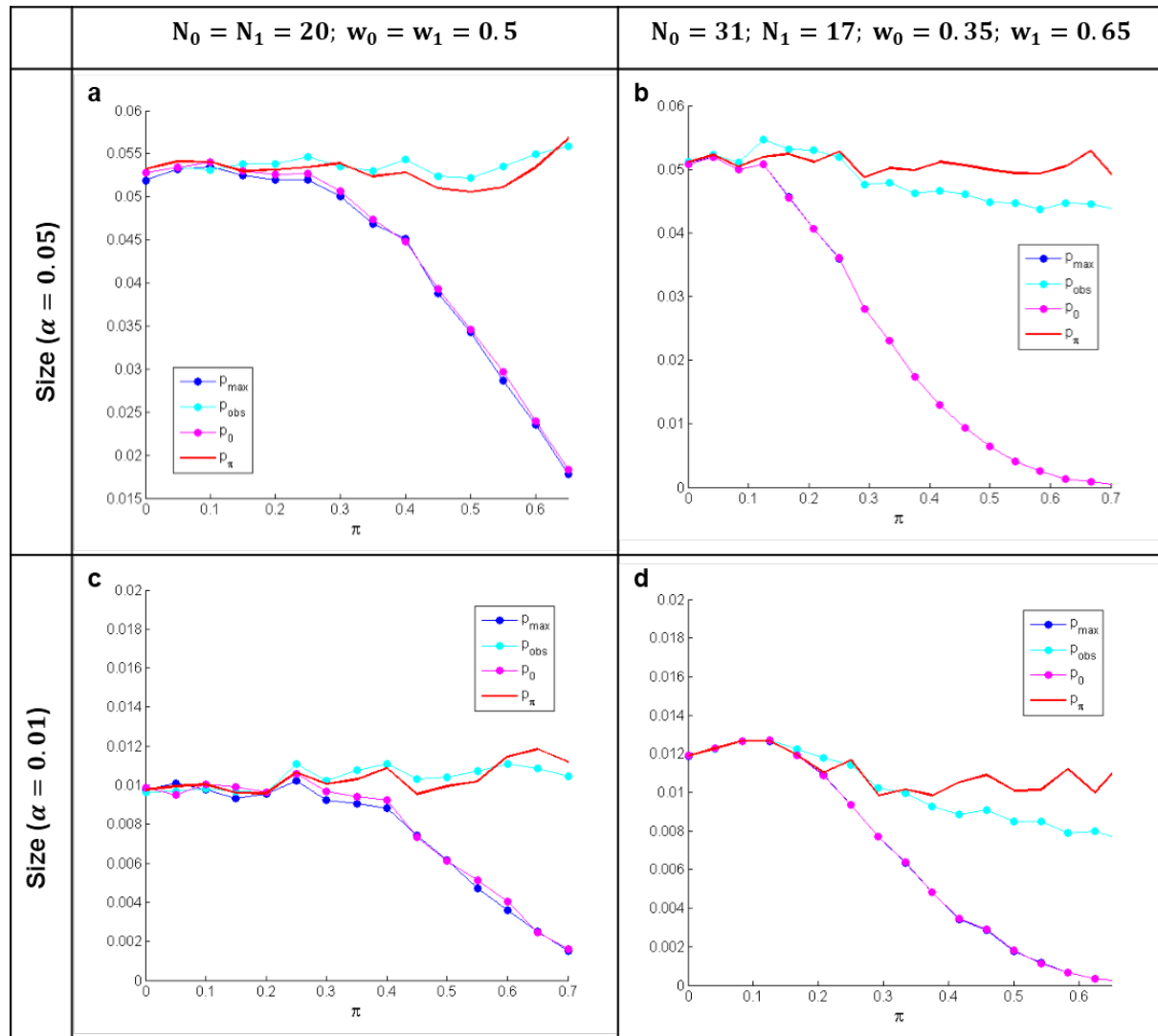


Figure 4-39: Size results of the three p-value alternatives.

The size represents the fraction of times the estimate falsely rejected the null hypothesis for significance levels of 5 % (a and b) and 1 % (c and d).

It is again evident that p_{obs} is better able to retain the significance level specified, compared to the other estimates regardless of whether π is small or large. The reader may note the seemingly unexpected instability of p_{π} , but this is simply due to the discrete nature of the distributions which become especially apparent in the tails.

4.5.6 Comparison of the p-values under the alternative hypothesis

The comparative simulation study done to assess the power of the three p-value alternatives followed the following simulation steps:

1. Choose values for N_0, N_1, w_0 and w_1 .
2. Choose a grid of distributional shift values $\{\mu_k, k = 1, 2, \dots, K\}$
3. For $k = 1, 2, \dots, K$ and for $m = 1, 2, \dots, M$
 - a. Generate $\{X_n^0(k, m), n = 1, 2, \dots, N_0\}$ from the LN (0,1) distribution. Now, generate $\{v_n(k, m), n = 1, 2, \dots, N_0\}$ from the uniform (0,1) distribution and set $X_n^0(k, m) = 0$ if $v_n(k, m) \leq \pi_0$ and $X_n^0(k, m) = (X_n^0(k, m) - \pi_0)/(1 - \pi_0)$ otherwise. The $X_n^0(k, m)$ represent the observations for the control group.
 - b. Generate $\{X_n^1(k, m), n = 1, 2, \dots, N_1\}$ from the LN $(0, \mu_k)$ distribution. Now, generate $\{v_n(k, m), n = 1, 2, \dots, N_1\}$ from the uniform (0,1) distribution and set $X_n^1(k, m) = 0$ if $v_n(k, m) \leq \pi_1$ and $X_n^1(k, m) = (X_n^1(k, m) - \pi_1)/(1 - \pi_1)$ otherwise. The $X_n^1(k, m)$ represent the observations for the experimental group.
 - c. Combine the $\{X_n^0(k, m), n = 1, 2, \dots, N_0\}$ and $\{X_n^1(k, m), n = 1, 2, \dots, N_1\}$ observations as $\{X_n(k, m), n = 1, 2, \dots, N\}$ and put $y_n(k, m) = 0$ for $n = 1, 2, \dots, N_0$ and $y_n(j, m) = 1$ for $n = N_0 + 1, N_0 + 2, \dots, N$.
 - d. Compute $\widehat{er}_{up}^*(k, m)$ by minimizing based on the data $\{y_n(k, m), X_n(k, m)\}$, $n = 1, 2, \dots, N$.
 - e. Convert $\widehat{er}_{up}^*(k, m)$ to the three p-values as defined in the main text. Also convert $\widehat{er}_{up}^*(k, m)$ to the true p-value by referring the $\widehat{er}_{up}^*(k, m)$ to the column corresponding to $\pi = \frac{\pi_0 N_0 + \pi_1 N_1}{N_0 + N_1}$ in the CDF-table, calling the result $p_\pi(k, m)$.
4. Finally, $A(k) = \frac{1}{M} \sum_{m=1}^M \{p'(k, m)\}$ and $PR(j) = \frac{1}{M} \sum_{m=1}^M I(p'(k, m) \leq \alpha)$ estimates the average p-value and proportion of rejection for the three p-values, respectively, for a given significance level α .

Next we report further results on the performance of the p-value alternatives under the alternative hypothesis based on the average p-value and the proportion of null hypothesis rejections. The alternative hypothesis was simulated using a shifted log-normal (LN) distribution for the experimental groups compared to a LN(0,1) distribution for the control group. A jump component was added to these distributions by assuming different proportions of zero values in each group.

Figure 4-40 illustrates the average p-value, given the value is below 10 %, for the three p-value estimates as well as p_π . The slight spike in the graphs illustrates how XERp first identifies the difference between π_0 and π_1 , the difference between the shift direction and these proportions result in a slight loss of power which is quickly regained once the shift is more pronounced. The three p-value alternatives perform equally well with p_{obs} displaying slightly more power. The slight advantage of p_{obs} over the other alternatives will become more pronounced for more accurate estimation of π , as is evident from the red lines indicating the somewhat greater power of p_π .

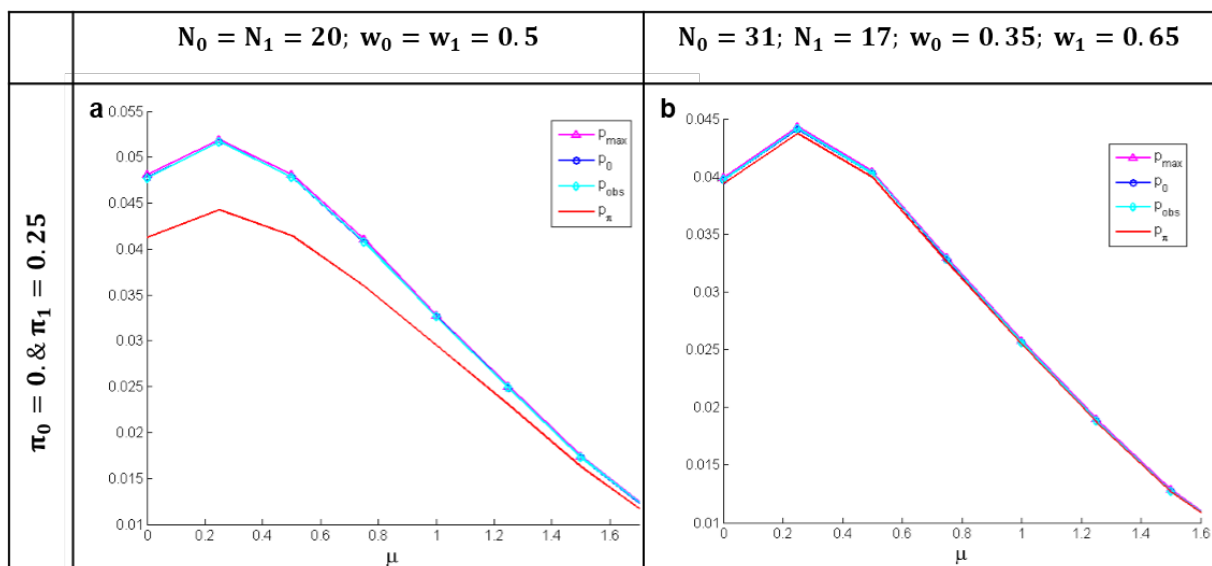


Figure 4-40: Average p-value ($p \leq 0.1$) for the dissonant case.

The figure plots the average p-value (y-axis), given the value is below 10 %, for the three p-value estimates as well as p_π for increasing shift (x-axis)

Figure 4-41 and Figure 4-42 illustrate the proportion of times the null hypothesis is rejected given significance levels of 5 and 1 %, respectively. To make the graphs more legible they are not displayed for the entire range of shift values (μ). Once no differences between p-value alternatives are apparent and sufficiently high rejection rates are achieved (i.e. the average p-values have achieved sufficiently low levels) the graphs are no longer displayed.

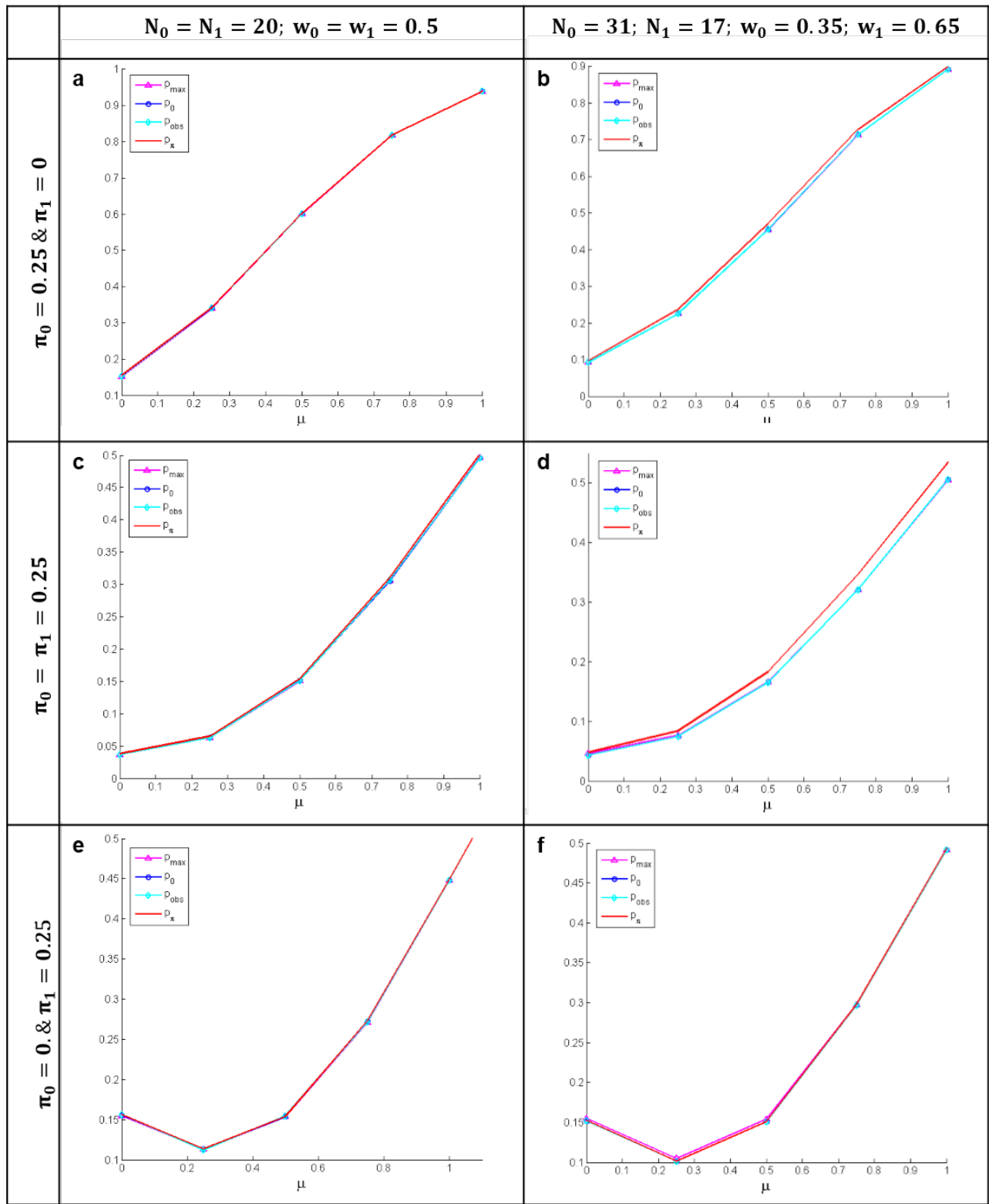


Figure 4-41: The proportion of null hypotheses rejected given $\alpha = 5\%$.

The figures display the proportion of times the null hypothesis is rejected (y-axis), given significance levels of 5%, for the three p-value estimates as well as p_π for increasing shift (x-axis)

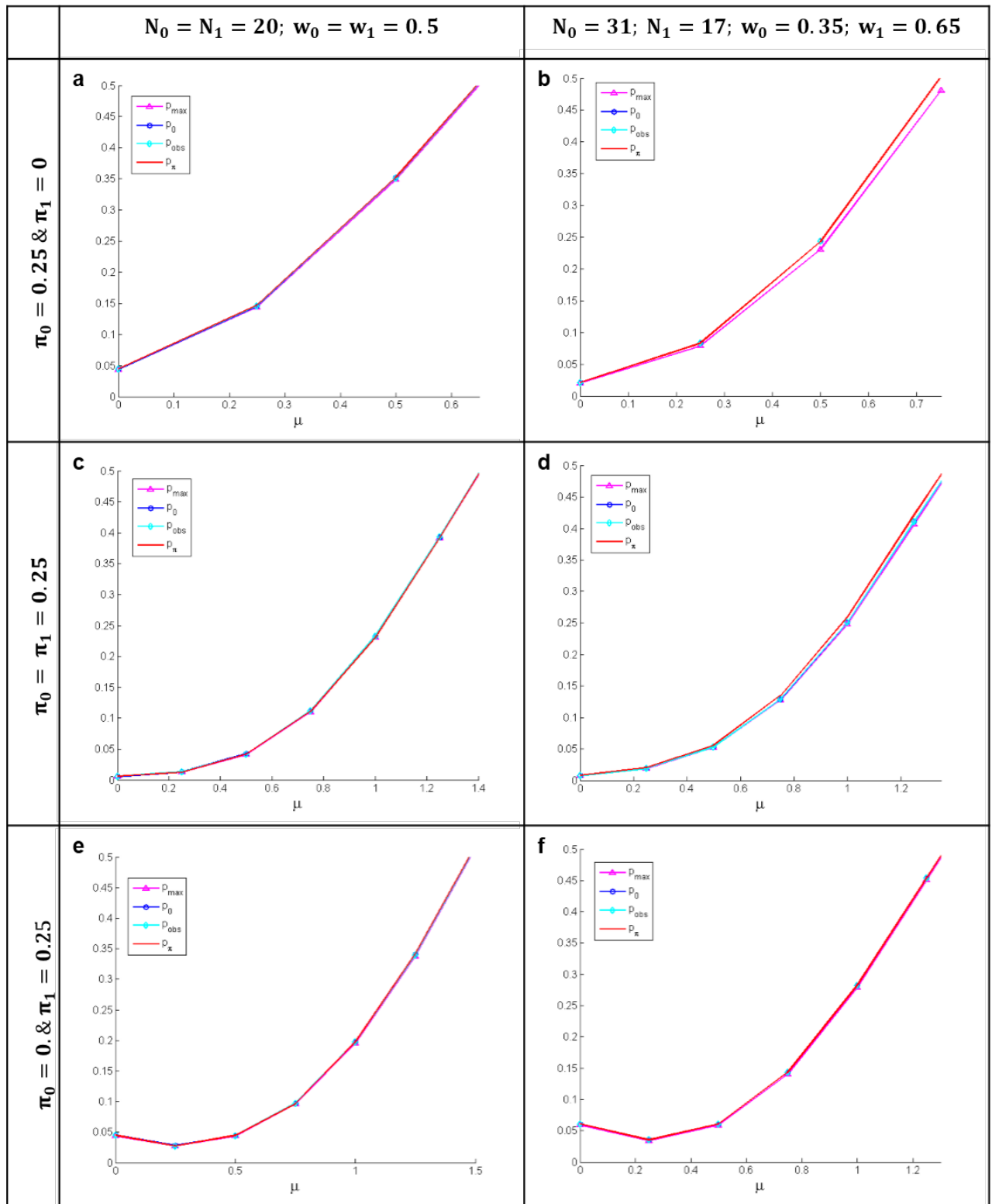


Figure 4-42: The proportion of null hypotheses rejected given $\alpha = 1\%$.

The figures display the proportion of times the null hypothesis is rejected (y-axis), given significance levels of 1%, for the three p-value estimates as well as p_π for increasing shift (x-axis)

Again the three p-value alternatives perform equally well. Notably, the significance level is retained reasonably well when there is no difference between the groups with $\pi_0 = \pi_1$ and $\mu = 0$ (Figure 4-41 c and d, as well as Figure 4-42 c and d). XERp again loses some power in the dissonant case, but only until the distributional shift becomes more pronounced.

4.5.7 Comparison to random imputation

This section takes a quick look at a random imputation strategy for replacing zero values and the use of XERp. The simulation made use of three variables selected based on their XERp results as provided in Table 4-13.

Variable	ER	C	Direction	Observed p-value	Observed π		
					Control	Experimental	All
3,4-dihydroxyphenylacetic acid	0.24	0.99	Up	0.01	29%	18%	25%
2,6-dihydroxybenzoic acid	0.27	0.00	Up	0.05	55%	18%	42%
4-hydroxycyclohexanecarboxylic acid	0.24	0.00	Down	0.01	35%	82%	52%

Table 4-13: XERp significant variables, performance & proportion of zeroes.

From Table 4-13 we see that the variables are well balanced with regard to the placement and proportion of zero values. The p-values calculated by XERp round up to significance levels of 1 and 5 % meaning that, for illustrative purposes, the selection of the variable is sensitive to small fluctuations in the p-value calculated. Finally, both up- and downward shift scenarios are represented, as well as both threshold scenarios $c = 0$ and $c > 0$.

To perform the simulation, the three variables were passed to the XERp software in two forms: (i) data with zeros (as was done in the main paper and corresponding to the results in Table 4-13); and (ii) data with zeros replaced by random values from a beta distribution, after which ERp (van Reenen *et al.*, 2016) was applied. To average out the impact of random imputation a hundred thousand iterations were carried out. During each iteration random numbers were generated from a beta distribution with location and scale parameters set to half the minimum and the minimum of the non-zero data, respectively, while the shape parameters were both set equal to 2.

The XERp estimated threshold for 3,4-dihydroxyphenylacetic acid is greater than zero, therefore the threshold and error rate are not affected by random imputation of zero values. The same XERp error rate as for the data with zeros of 0.2434 is obtained.

The associated p-values did however, differ as different null distributions were referenced, i.e. where there are 25 % zeros $p_{obs} = 0.0123$ (data with zeros) and no zeros $p_{obs} = 0.0126$ (random imputation). It is important to note that in instance where $c > 0$ for the data with zeros, the error rates will only be equal if the imputed numbers are not allowed to exceed the minimum of the non-zero values. More advanced imputations strategies such as k-nearest-neighbour and random forest may not conform to this rule and will produce different results. Also, in the event that the error rates are the same, the p-value for the data with zeros may not always be smaller as it depends on the proportion of zeros (refer to Figure 4-35 in the main manuscript).

The results for the second and third variables, 2,6-dihydroxybenzoic acid and 4-hydroxycyclohexanecarboxylic acid, are very similar since both have thresholds $c = 0$ for the data containing zeros (Table 4-13). In instances where the XERp estimated threshold is equal to zero, ERp combined with imputation by random numbers smaller than the minimum non-zero value, will result in a threshold estimate somewhere between two imputed numbers. In other words, the classification rule depends on some random value between the ranges set by the imputation rule, e.g. zero and the minimum observed value. A threshold that falls in random territory introduces unwanted randomness into the associated error rate and p-value. Most importantly, a threshold in random territory implies that a new subject with a measured value of zero, therefore requiring imputation, would be classified by comparing the random value from imputation to the threshold value, which is also based on random imputations. Classification is thus based on whether one random number is larger or smaller than another random number.

Concluding, imputation with random numbers has no effect on the error rate when the XERp threshold is already above the detection limit, while it can have a dramatic effect when the XERp threshold value is lower than the detection limit. Moreover, random imputation introduces unwanted randomness into the thresholds, error rates and p-values without any obvious gain.

4.6 References

- Armitage, E.G., Godzien, J., Alonso-Herranz, V., Lopez-Gonzalvez, A. & Barbas, C. 2015. Missing value imputation strategies for metabolomics data. *Electrophoresis*, 36:3050-3060.
- Behrends, V., Tredwell, G.D. & Bundy, J.G.. 2011. A software complement to AMDIS for processing GC-MS metabolomic data. *Analytical Biochemistry*, 415:206-208.
- Berk, R., Brown, L., Buja, A., Zhang, K. & Zhao, L. 2013. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837.
- Berk, R., Brown, L., Buja, A., Zhang, K. & Zhao, L. 2013. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–37.
- Bewick, V., Cheek, L. & Ball, J. 2005. Statistics review 14: logistic regression. *Critical Care*, 9(1):112-118.
- Blumenthal, A., Nagalingam, G., Huch, J.H., Walker, L., Guillemin, G.J. & Smythe, G.A. 2002. M. tuberculosis induces potent activation of IDO-1, but this is not essential for the immunological control of infection. *PloS one*, 7:e37314.
- Breiman, L. 1968. Probability. Reading, Massachusetts: Addison-Wesley.
- Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. 1984. Classification and regression Trees. California: Wadsworth Inc.
- Brereton, R.G. & Lloyd, G.R. 2014. Partial least squares discriminant analysis: taking the magic away. *Journal of Chemometrics*, 28:213–225.
- Bundy, J.G., Davey, M.P. & Viant, M.R. 2009. Environmental metabolomics: a critical review and future perspectives. *Metabolomics*, 5:3–21.
- Dakna, M., Harris, K., Kalousis, A., Carpentier, S., Kolch, W., Schanstra, J.P., Haubitz, M., Vlahou, A., Mischak, H. & Girolami, M. 2010. Addressing the challenge of defining valid proteomic biomarkers and classifiers. *BMC Bioinformatics*, 11:594.
- Ding, X., Liu, R., Li, W., Ni, H., Liu, Y., Wu, D., Yang, S., Liu, J., Xiao, B. & Liu, S. 2016. A metabonomic investigation on the biochemical perturbation in post-stroke patients with depressive disorder (PSD). *Metabolic Brain Disease*, 31(2):279-287.

- Dudley, R.M. 1999. Uniform central limit theorems. (*In Cambridge Studies in Advanced Mathematics*, 63. Cambridge, UK: Cambridge University Press. p. 129 & 334-335.
- El-Ansary, A., Shake, G.H. & Rizk, M.Z. 2013. Role of gut-brain axis in the aetiology of neurodevelopmental disorders with reference to autism. *Journal of Clinical Toxicology*, S6(005).
- Fluss, R., Faraggi, D. & Reisei, B. 2005. Estimation of the Youden index and its associated cutoff point. *Biometrical Journal*, 47(4):458–47.
- Follmann, D., Fay, M.P. & Proschan, M. Chop-lump tests for vaccine trials. *Biometrics*, 65:885–93.
- Gleiss, A., Dakna, M., Mischak, H. & Heinze, G. 2015. Two-group comparisons of zero-inflated intensity values: the choice of test statistic matters. *Bioinformatics*, 1–8.
- Gromski, P.S., Xu, Y., Correa, E., Ellis, D.I., Turner, M.L. & Goodacre, R. 2014a. A comparative investigation of modern feature selection and classification approaches for the analysis of mass spectroscopy data. *Analytica Chimica Acta*, 829:1-8.
- Gromski, P.S., Xu, Y., Kotzé, H.L., Correa, E., Ellis, D.I., Armitage, E.G., Turner, M.L. & Goodacre, R. 2014b. Influence of missing values substitutes on multivariate analysis of metabolomics data. *Metabolites*, 4:433-52.
- Hallstrom, A.P. 2010. A modified Wilcoxon test for non-negative distributions with a clump of zeros. *Statistics in Medicine*, 29:39-40.
- Holm, S. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65-70.
- Kaddurah-Daouk, R., Kristal, B.S. & Weinshilboum, R.M. 2008. Metabolomics: a global Biochemical approach to drug response and disease. *Annual Review of Pharmacology and Toxicology*, 48:653–83.
- Kang, H. 2013. The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, 64(5):402-6.
- Khan, I.M., Khan, S. & Laaser, U. 2003. Tuberculous meningitis: a disease of fatal outcome in children. *European Journal Pediatrics*, 162:281–282.

- Kvalheim, O.M., Arneberg, R., Bleie, O., Rajalahti, T., Smilde, A.K. & Westerhuis, J.A. 2014. Variable importance in latent variable regression models. *Journal of Chemometrics*, 28:615–622.
- Lai, C-Y., Tian, L. & Schisterman, E.F. 2012. Exact confidence interval estimation for the Youden index and its corresponding optimal cut-point. *Computational Statistics and Data Analysis*, 56:1103–1114.
- Leeflang, M.M.G., Moons, K.G.M., Reitsma, J.B. & Zwinderman, A.H. 2008. Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: mechanisms, magnitude, and solutions. *Clinical Chemistry*, 54(4):729–737.
- Leib, S.L., Bosacci, R., Gratzl, O. & Zimmerli, W. 1999. Predictive value of cerebrospinal fluid (CSF) lactate level versus CSF/blood glucose ratio for the diagnosis of bacterial meningitis following neurosurgery. *Clinical Infectious Diseases*, 29:69–74.
- Lu, H., Dunn, W.B., Shen, H., Kell, D.B. & Liang, Y. 2008. Comparative evaluation of software for deconvolution of metabolomics data based on GC-TOF-MS. *Trends in Analytical Chemistry*, 27(3):215-227.
- Madsen, R., Lundstedt, T. & Trygg, J. 2010. Chemometrics in metabolomics—A review in human disease diagnosis. *Analytica Chimica Acta*, 659:23–33..
- Manso, A. & Machado, C. 2015. Dependence aliasing and the control of family-wise error rate in multiple hypothesis testing. *Statistics and Computing*, 25(3):669–681.
- Mason, S., Tutu van Furth, A.M., Solomons, R., Wevers, R.A., van Reenen, M. & Reinecke, C.J. 2016. A putative urinary biosignature for diagnosis and follow-up of tuberculous meningitis in children: outcome of a metabolomics study disclosing host–pathogen responses. *Metabolomics*, 12(110).
- Mason, S., Van Furth, A.M., Mienie, L.J., Engelke, U.F.H., Wevers, R.A., Solomons, R. & Reinecke, C.J. 2015. A hypothetical astrocyte-microglia lactate shuttle derived from a ^1H NMR metabolomics analysis of cerebrospinal fluid from a cohort of South African children with tuberculous meningitis. *Metabolomics*, 11(4):822-837.

- Muñoz-Elías, E.J., Upton, A.M., Cherian, J. & McKinney, J.D. 2006. Role of the methylcitrate cycle in *Mycobacterium tuberculosis* metabolism, intracellular growth, and virulence. *Molecular Microbiology*, 60(5):1109–1122.
- Romanoa, J.P. & Wolf, M. 2005. Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469):94-108.
- Ruopp, M.D., Perkins, N.J., Whitcomb, B.W. & Schisterman, E.F. 2008. Youden index and optimal cut-point estimated from observations affected by a lower limit of detection. *Biometrical Journal*, 50(3):419–30.
- SAS Institute Inc. 2011 The SAS System for Windows Release 9.3 TS Level 1M0 Copyright© by SAS Institute Inc., Cary, NC, USA.
- Savvi, S., Warner, D.F., Kana, B.D., McKinney, J.D., Mizrahi, V. & Dawes, S.S. 2008. Functional characterization of a vitamin B12-dependent methylmalonyl pathway in *mycobacterium tuberculosis*: implications for propionate metabolism during growth on fatty acids. *Journal of Bacteriology*, 190(1):3886–3895.
- Schisterman, E.F., Faraggi, D., Reiser, B. & Hu, J. 2008. Youden index and the optimal threshold for markers with mass at zero. *Statistics in Medicine*, 27(2):297–315.
- Smilde, A.K., Van der Werf, M.J., Bijlsma, S., Van der Werff-Van der Vat, B.J.C. & Jellema, R.H. 2005. Fusion of mass spectrometry-based metabolomics data. *Analytical Chemistry*, 77(20):6729-36.
- SPSS Inc. (2015). IBM SPSS Statistics Version 22, Release 22.0.0, Copyright© IBM Corporation and its licensors. <http://www-01.ibm.com/software/analytics/spss/>
- Steel, S.J., Oosthuizen, S. & Uys, D.W. A measure of post variable selection error in multiple linear regression, and its estimation. *Journal of Statistical Computation and Simulation*, 74(2):123-133.
- Taylor, S. & Pollard, K. 2009. Hypothesis tests for point-mass mixture data with application to omics data with many zero values. *Statistical Applications in Genetics and Molecular Biology*, 8(1).

- van Reenen, M., Reinecke, C.J., Westerhuis, J.A. & Venter, J.H. 2016. Variable selection for binary classification using error rate p-values applied to metabolomics data. *BMC bioinformatics*, 17(1):33.
- van Reenen, M., Westerhuis, J.A., Reinecke, C.J. & Venter, J.H. 2017. Metabolomics variable selection and classification in the presence of observations below the detection limit using an extension of ERp. *BMC bioinformatics*, 18(1):83.
- Xia, J., Broadhurst, D.I., Wilson, M. & Wishart, D.S. 2013. Translational biomarker discovery in clinical metabolomics: an introductory tutorial. *Metabolomics*, 9:280–299.
- Yin, J. & Tian, L. 2014. Joint confidence region estimation for area under ROC curve and Youden index. *Statistics in Medicine*, 33:985–1000.
- Zhan, X., Patterson, A.D. & Ghosh, D. 2015. Kernel approaches for differential expression analysis of mass spectrometry-based metabolomics data. *BMC Bioinformatics*, 16:77.
- Zhang, D., Fan, C., Zhang, J. & Zhang, C. 2009. Nonparametric methods for measurements below detection limit. *Statistics in Medicine*, 28:700–15.

CHAPTER 5 – DISCUSSION & FUTURE PROSPECTS

5.1 Concluding Remarks

This chapter is my reflection on the contributions, shortcomings, and most importantly, the future potential of the work presented in this thesis. To tie in with my opinionated discourse at the beginning of this thesis, I again raise my own opinions, observations and comments. I question the status quo with reflection on the work done. These questions are sometimes rhetorical, at other times accompanied by my own retort, and even on a few occasions descriptive of my inner debate, all to convey the complex nature of the topics under review. This deliberation then provides a backdrop against which an amalgamated thesis such as the present may be evaluated.

5.1.1 Valid, valuable and generalizable results – a view on evaluating Chapter 3 and similar contributions

In my transfer from consultant, somewhere in the middle of ‘the spectrum’, to a member of a multi-disciplinary human metabolomics team, right at ‘The Gap’ (Chapter 2), the importance of making clear distinctions between valid, valuable and generalizable results became apparent. Such distinctions are essential if the publications presented in Chapter 3 are to be fairly appraised.

If generalizable results do not always translate to results conveying value then the converse is also true, the fact that a finding cannot be generalized does not immediately diminish its value. Validity on the other hand, is a one-way street and underpins both the generalizability and the value of results. To understand this riddle, requires a context-specific understanding of the terms used and perhaps also an open mind.

The validity of results is best described through synonyms for the term ‘valid’, which include: sound, reasonable, logical, justifiable, defensible, arguable, and sustainable (English Oxford Living Dictionaries, 2018). From these terms, three themes became apparent to me: (i) the probability of errors should be calculated to judge the sustainability of results; (ii) supporting documentation is needed to defend and argue the soundness of results; and (iii) all is lost without transparency, for logic and sound judgment require complete or, at the very least reasonable, disclosure of how the results were generated.

The first theme centres on drawing the wrong conclusions, focusing on the level of statistical certainty, that is, the quantification of the probability of false discoveries, but also the probability

of drawing conclusions based only on partial information due to a lack of power. Unfortunately, Type I and II error rates tend to make these uncertainties appear more simplistic than they actually are. We should not forget that Type I and II error rates are also estimated and these estimates can make it appear as though two halves make a whole to quantify surety precisely. But, it is rarely the case that if the results tell half the story, i.e. an estimated proportion of the story, it is a coherent half. This is then precisely why the more traditional, one-dimensional, purely statistical probability-based approach to declaring results valid and valuable by declaring them generalizable, is not sufficient. If we want results to lead us to an understanding of the context which generated those results, we need to allow our results to make more than a simple statement on whether or not to reject some hypothesis. We need a wider lens to get a panoramic view, even though we may be less sure of what we are seeing along the far edges.

The validity of results is routed in each aspect of the planning and execution of a human metabolomics research study. One aspect and the one receiving attention here is the statistical analysis. The sheer volume of data generated makes graphical exploration of the data extremely time-consuming and even impossible for more than three dimensions at a time. So, the intensive use of different statistical methods, in isolation and in combination, to create shortlists of important variables, as done in Chapter 3, is not a new concept and is considered essential to explore all aspects of the data (Saccenti *et al.*, 2013). That said, the more comprehensive methods, specifically methods that can accommodate the $n \ll p$ scenario and/or complex experimental designs, are less frequently combined. Indeed, I could find no publication that combined methods such as reported in Chapter 3.

Combining more comprehensive methods may increase overall power, but may also increase the risks on interpretation. Model overfit becomes more likely and so do false discoveries. This is one of the primary reasons the following set of safeguards were implemented in Chapter 3: (i) corrections for multiple testing were retained for univariate methods; (ii) some method overlap, with regard to variable importance, was required before variables were shortlisted; (iii) key findings were biologically interpreted and in so doing validated against existing biological knowledge; (iv) the limitation of findings, for example, uncertainties with regard to the level of peak identification, were not hidden; (v) we always reverted back to the original unscaled data to judge discovered effects; and (vi) results constituting new contributions to existing knowledge on the human metabolome were flagged for further validation.

The paragraphs that follow provide some rationale for the safeguards selected in the context of the remaining two themes of validity. It is important to note, however, that these safeguards are always provisional, given the contextual and experimental constraints of a given study design. One lesson I take forward is also to guard against becoming uncritical. Perhaps wanting to keep all variables measured should not be the approach of choice when considering: (i) some variables cannot yet be accurately measured, at least to any reasonable extent; (ii) the level of accuracy varies dramatically across variables; and (iii) any effect these inaccuracies may introduce is compounded by scaling variables to compete on equal footing in multivariate methods (van den Berg *et al*, 2006). The same can be said for wanting to study immediately the interaction effect of NAD and alcohol, while in retrospect we should have first obtained more experimental information on the efficacy and timing of NAD supplementation. Perhaps using common sense to reduce the dimensionality of the data and complexity of the design, rather than wanting to achieve a comprehensive view which was likely not yet possible, would have been a shrewder tactic.

The second theme is supporting documentation and addresses the problematic idea that statistical probability overrides logic. In human metabolomics there must be a balance between controlling for false discoveries and ranking variables according to their ability to answer the question posed, which is mostly not tied to a defined hypothesis in explorative investigations. This balance is only achieved if the statistical Type I error rate is not the only thing considered, specifically when the aim is to inform on the understanding of a biological system. In human metabolomics, statistical analysis should aim to rank metabolites based on the size of the effect. Variables can then be shortlisted based on some threshold for the minimum effect considered, which is not necessarily adversely affected by the number of variables measured. Finally, biological interpretation should verify the list based on prior knowledge of the role shortlisted metabolites play in the human body, that is, how unique they are and how well they fit with the expected results of the study. Selecting important variables based solely on either statistical significance or biological fit is risky.

Selecting metabolites based on biological prominence alone can lead to subjective explanations and in truth, merely opinions on the effect measured. The vast number of relatively quantified metabolites; the level of provisional annotation of some variables; and insufficient interdisciplinary research teams, leaves the door wide open for chance patterns to be interpreted as indicative of biological function. This has been well illustrated prior to the advent of metabolomics (Erratum to Pauling *et al.*, 1971). This notorious example of a blunder in science, made by working in

separation, underscores the repeated emphasis on the requirement of a multidisciplinary team with unique and overlapping expertise.

Selecting metabolites based on statistical significance alone may lead to theoretical leaps on the side of biology to try and explain what may lead to misleading interpretations, even if it is with low probability. Statistical selection alone may also keep the puzzle less clear than it needs to be, less significantly perturbed metabolites may help to elucidate metabolic pathways: *“The omission of nonsignificant results from publications is undesirable for both scientific and ethical reasons...”* (Garamszegi, 2006). The curse of dimensionality may be given undue notoriety in exploratory human metabolomics research due to the small sample sizes associated with hypothesis generating research and the insistence on correcting for multiple testing. Bender and Lange go as far as to state that *“...in exploratory studies, in which data are collected with an objective but not with a prespecified key hypothesis, multiple test adjustments are not strictly required”* (Bender & Lange, 2001). One should not forget that small samples combined with corrections for multiple testing may very well discard truly significant findings.

To explain further the rationale behind the safeguards set in Chapter 3, requires a distinction between valid and generalizable result. Valid results may or may not be generalizable. Human metabolomics research is still for the most part in the business of generating hypotheses, which in turn is commonly associated with small sample sizes for ethical and financial reasons. This was discussed in Chapter 2, so I focus on a second, more controversial argument. Do we still need to generalize findings? Treatment is moving towards personalized/precision medicine, focusing on an individual subject and not as a member of a group with a predefined treatment set at group level (Collins & Varmus, 2015). One human body often reacts greatly differently from the next, due to a combination of genetic and lifestyle factors, that even when a seemingly homogeneous or well-selected group of individuals are compared, large inter-individual differences are still present (Manach *et al.*, 2017), as was also observed in Chapter 3. Should we then not rather include a novel statistical focus on understanding biological mechanisms - the small, medium and large perturbations - and the variability across individuals? In his comments on Breiman's publication on the philosophy of the science of statistics, Cox said it so well: *“One of our failings has, I believe, been, in a wish to stress generality, not to set out more clearly the distinctions between different kinds of application and the consequences for the strategy of statistical analysis”* (Breiman, 2001, comment by D.R. Cox)

To understand a mechanism as complex as the human body, which cannot be studied all at once, some level of surety remains essential. Even if the surety relates only to the ability to differentiate randomness or noise from a signal, as opposed to moving beyond the current sample. Larger sample sizes remain the best option to avoid statistical errors and reduce uncertainty and so one cannot ignore the small sample sizes (24 individuals represented in the NMR data and only 12 in the GC–MS data) compared to the vast number of variables measured and the questions this may raise. Yet, as discussed at some length through-out this thesis, large sample sizes are often not realistic. An alternative to these analytical realities is participation in a consortium of transdisciplinary groups (Collins, 2011; Perkins *et al.*, 2018), which, however, is beyond the scope of our present metabolomics centre.

Therefore, returning to my initial argument, I believe the best way forward is to strike a balance between chemical certainty, statistical accuracy, and existing knowledge of human biology. This may require some assumptions of validity until proved otherwise, but is this not how knowledge production works – independent researches building on each other to become knowledge? Is this not also the underlying idea behind generalization or null hypothesis testing, to assume the status quo until sufficient evidence accumulates to change it? I am not suggesting that sample size and statistical power are no longer of great importance and can be set aside simply by arguing incremental learning. No, on the contrary, I plead simply against the outright dismissal of results based purely on the size of the sample. Of course sample size must be kept within reason, but even single patient case studies have informed greatly in the past: *“Discovery may sometimes arise from small investigations or even unanticipated case observations”* (Ioannidis, 2008).

How would this work in practice? By implementing safeguards: *“...it is all about the use of sound, rational thinking and responsibility instead of push-the-button automated data analysis”* (Kjeldahl & Bro, 2010). As demonstrated in Chapter 3, even a small sample size, as small as 12, could shed sufficient light on how alcohol is metabolised that the larger research community welcomed the contribution (reviewer’s comment: *“The paper deals with an interesting and timely topic”*). This was achieved by keeping an open mind and enforcing rational thinking. For example, results should cohere to current knowledge, but complete conformation must not be a prerequisite for validity or a high appraisal of value, as this disowns the power of new discovery from statistical analysis. If results piece together well with existing knowledge, this must be allowed to add to their validity. The message is then that logic and balance must prevail against any set rule or one-directional relationship, and above all the need of further validation must remain a recurring objective.

The third theme is the most essential, most trivial, but can also be the most difficult to achieve: transparency. Transparency underpins all. Excluding key information hinders all aspects of the intersubjective evaluation of results, specifically validity, value and generalizability. The reason for variable selection or exclusion must be clearly stated, along with the limitations of design and levels of uncertainty - uncertainty pertaining to all assumptions made, but specifically to chemical quantification and identification, statistical evaluation, and alternative biological interpretations. *"Flexible analyses will not cause a problem if they are accompanied by complete and transparent reporting of all results"* (Ioannidis, 2008). The validity, value and generalizability of findings can then be fairly judged by a jury of peers. The GC–MS publication presented in Section 3.6 is an unfortunate example; wanting to be too concise led to an oversight that could have discredited any biological deduction, but fortunately it did not in this instance. Scrupulously stating the shortcomings of a study or the lack of significant results (however defined) is essential, yet also seems the brave thing to do. Without venturing into an investigation here, I strongly advocate further debate by the broader research community on the profitability of research. How should researchers be rewarded, does the 'publish or perish' approach ensure quantity and quality? Can a profit-driven academic publication industry remain objective? These seem to be evidence to the contrary, for example: (i) the increasing number of pirate journals pursuing open access publications in return for payment from authors; (ii) journals favouring success stories, while well-founded studies with negative outcomes remain unpublished, presumably to boost readership; (iii) journals wanting to broaden their readership base beyond academia by reducing publications in length and in so doing, in detail, as most supplementary material are exempt from peer review; and (iv) journals pushing volume by reducing response times. A lesson I will most certainly carry forward is that asking for an extension, even if it results in a resubmission, is infinitely better than submitting a correction.

To conclude, if letting reason trump rule will turn research into opinion, then a balance must be struck between reason and rule. Research in human metabolomics requires that reason and rule be given equal weight. Human metabolomics is rooted in compromise, from the design of a study to judging the validity and value of its results. A purely statistical or biological or chemical evaluation of significance could slow the knowledge production process. Instead, a transdisciplinary approach is again essential, not only to judge the validity and value of research, but also to whom it applies and most importantly how it can be validated: *"The only way to go is to take responsibility: decide what is relevant by support of biological/chemical knowledge and sound reasoning and always keep the purpose of the modeling in focus!"* (Kjeldahl and Bro, 2010).

5.1.2 Statistics for purpose or prestige - a view on evaluating Chapter 4 and similar contributions

The review process for the ERp publication (Chapter 4) was an inspiring learning experience. The wide variety of views on our contribution was at once interesting and perplexing. Some reviewers felt the contribution that ERp made was too trivial, while others could immediately share in our vision for the current and, even more so, the future possibilities of the approach. From a theoretical point of view, some judged ERp as too simplistic to justify application to human metabolomics data and perhaps rightly so when considering only the technical advancement made: the primary statistical contribution of ERp was the conversion of classification error rates to p-values. As ERp is a univariate approach the application to human metabolomics was perhaps not yet justified. But this would have been a cynical review of ERp, disregarding improvements made to the interpretability, the level of user interaction, and ease of implementation, seeing as no data transformation is required due to ERp's non-parametric nature. It took almost a year's worth of correspondence, through the formal review process, to satisfy all parties on the contributions of the ERp groundwork and its potential for further development.

I found this hesitance to publish rather revealing and it gave me some understanding as to why statistical research has become so far removed from application (Chapter 2). The decision on which statistical methods to research can be a difficult one and should take purpose or utility into consideration – although it appears that sometimes prestige carries more weight, as this seems to increase the likelihood of publication. The development of increasingly advanced methods is promoted regardless of whether an application is clear or not. This is not a misplaced policy and should continue, but reciprocity is definitely required. How contributions to research in statistics are evaluated must go beyond technical advancement. Improvements such as the ease of interpretation or the level of data preparation required should likewise be considered as contributions in their own right. We cannot continue to further separate statistical development from societal need, else we will remove the need for statistics from society. Breiman summarizes this conundrum with his comment on clinicians' preference: *"There is no way they can interpret a black box containing fifty trees hooked together. In a choice between accuracy and interpretability, they'll go for interpretability"* (Breiman, 2001).

Most importantly, the future potential of new approaches must be seen as a key contribution. Further development was part and parcel of ERp from the start, but all could have been lost had we been unable to publish the foundational theory. ERp was an attempt to bridge 'The Gap',

(Chapter 2) and paved the way for its extension, XERp, to succeed. XERp, and its predecessor ERp, are then proof-of-concept for the advancement of statistical theory through exposure to real-world problems. We may not have closed ‘The Gap’ but we have certainly made headway. We have taken the idea that the best way to evaluate a model is through prediction error (Xia *et al.*, 2013; Kjeldahl & Bro, 2010), made it a method, and customized it for a context - the context being GC–MS data sets known to contain many missing values. The approach followed in Chapter 3 is not uncommon: missing values are replaced by zeros; reduced by excluding variables with too many zeros; and treated through various imputation strategies. However, the choice of imputation strategy “...may have a dramatic impact on GC–MS based metabolomics data, both for unsupervised and supervised learning” (Gromski *et al.*, 2014).

XERp approaches this data characteristic from a unique angle, accepting zeros rather than mediating them away. Specifically, the null distributions used in ERp were modified to account for the proportion of zeros resulting from values below detection or true zeros (Chapter 4). This modification made XERp robust to small and large proportions of zeros. This customisation for human metabolomics presents XERp as a novel method with a distinct contribution and immediate applicability as recognized in a recent ‘food-omics’ paper, referring to ERp and XERp: “One of the advantages of this approach, together with its ease of use and interpretation, is the ability to deal with unequal and small group sizes, and the possibility to explicitly specify the cost of misclassification which is desirable when choosing the threshold for each variable” (Bevilacqua *et al.*, 2017).

5.2 Future Prospects

5.2.1 Continuing the ALC2013 investigation

As noted in Section 3.9, there remains numerous possibilities for the analysis of the GC–MS and even more so, of the NMR data. Different statistical methods combined with different subsets of the data or even the entire NMR or GC–MS dataset, might inevitably reveal novel insights into the workings of the human metabolome when confronted with alcohol and NAD.

The strengths and weaknesses of different PCA-based approaches – for example, multilevel PCA, unfolded PCA and ASCA - are not yet fully understood given diverse designs and applications. Neither are how these approaches respond to different data transformations and

scaling, more specifically how and when they are applied in the data analysis pipeline, i.e. prior to any analysis on the entire dataset, or considering only factors or factor levels presented for a given analysis, or even specific for each analysis. I believe there is value in understanding these methods better as this will make future application more robust and may even result in a new approach or method.

5.2.2 Experimental design at ‘The Gap’

The transdisciplinary approach towards experimental design was briefly discussed in Chapter 2. Future work will focus more intently on facilitating negotiations (i.e. ‘the practice of compromise’) described in Chapter 2.

To recap, designing a human metabolomics research study demands that recursive attention be given to the three pillars of study design - the research aim, experimental design, and measurement design – by the design team, which forms the fourth pillar. The design of research studies is a concept that has received much attention (Blaise *et al.*, 2016; Saccenti and Timmerman, 2015; Nyamundanda *et al.*, 2013), focusing mostly on the estimation of a sufficient sample size to ensure a predefined statistical model is not under-powered. The design literature predominantly focusses on experimental design and often assumes: (i) some hypothesis is being tested; (ii) the researcher has extensive knowledge of the expected size of the effect being measured; (iii) confounding factors are known; and (iv) measurement error can be minimized or randomized away. Making these assumptions is required when research questions are absolute and indeed try to prove some hypothesis, making design literature ill-suited for human metabolomics research.

Research into the human metabolome will remain exploratory for some time, given the limited understanding of the response of the human metabolome to different endo- and exogenous conditions. Research aims flowing from such human metabolomics research are therefore less absolute and the primary objective is not to test but to generate hypotheses. Confounding factors are sundry and their individual impact difficult to estimate. Designing a human metabolomics study is not only a question of the expected size of the effect, but also the size of the effect relative to measurement error. Therefore the right answers to design questions, as concluded from the statistical literature, may not be realistically achievable in the context of human metabolomics, leaving researchers to wonder: Can the design be compromised without selling out the validity of the subsequent findings?

Future research will explore this question with special emphasis on how compromise can be guided by simulation studies to minimise the potential impact. More specifically, we will investigate how preliminary experiments on the size of the treatment effect and screening experiments (to quantify inter- and intra-individual variation and set future selection criteria) can be used to estimate the reproducibility of the treatment effect. The research will likely conclude that: (i) compromise is only possible for less absolute research aims; (ii) the entire design team must form part of negotiations pertaining to the research aim and design; and (iii) simulation studies can make difficult choices easier but cannot trump subject matter knowledge.

5.2.3 Extending and expanding ERp beyond XERp

The characteristics of ERp made it a good univariate method which could be applied to numerous data sets. To make it more attractive to the metabolomics community, ERp was developed further to take zero-valued observations into account. Future endeavours will aim to take more of the characteristics of human metabolomics data (as described in Chapter 2) into account.

Current work beyond this thesis focuses on a refinement of XERp to increase its accuracy in terms of minimum error rate determination and better estimation of the associated best-performing threshold of classification. TERp (Two-part ERp) restructures XERp into a test of two parts, similar to that of Taylor & Pollard (2009), but with the added benefit of retaining key XERp features, that is, the error rate weights and its almost non-parametric nature.

Future prospects include:

- Evaluating permutation as an alternative to the simulated distributions currently used for p-value estimation.
- Pairwise error rate estimation (PERp), which moves ERp towards a multivariate approach. PERp will aim to estimate joint error rates for all pairs of variables.

5.3 In conclusion

The aim of this thesis was two-fold: (i) to apply existing statistical and data analysis techniques in a novel way; and (ii) to develop new statistical approaches able to take the unique characteristics of metabolomics data and design into account. Chapter 3 convincingly achieved the first aim, which resulted in three publications accepted in reputable journals. Chapter 2 partly addressed the second aim as this chapter describes new approaches to the design of metabolomics data.

Chapter 4 addresses the remainder of the second aim and describes a new statistical method which already accommodates some of the characteristics of metabolomics data and can be further developed to accommodate even more

The accompanying objectives were also achieved:

- Chapter 2 documents the insights gained into the drivers of development in the field of statistics and the role of statistics in human metabolomics. A knowledge production process appropriate for research in human metabolomics was developed and described in Chapter 2.
- Chapter 3 documents the publications resulting from the application of advanced statistical methods to an intervention-based, longitudinal, crossover human metabolomics case study. Chapter 3 also demonstrates the use of and need for multiple statistical methods to maximize discovery and interpretation.
- All chapters emphasize how essential a transdisciplinary approach is in human metabolomics research.
- A new statistical approach was developed in Chapter 4, which can take some of - and can be developed to take more of - the characteristics of human metabolomics data into account. The ERp approach and its extension, XERp, were also custom-built to identify metabolites that can discriminate between two groups.
- Finally, the current chapter proposes a way forward for the application of statistical methods, as well as the development of core statistical theory, in the context of human metabolomics research.

5.4 References

- Bender, R. & Lange, S. 2001. Adjusting for multiple testing—when and how? *Journal of clinical epidemiology*, 54(4):343-349.
- Bevilacqua, M., Bro, R., Marini, F., Rinnan, Å., Rasmussen, M.A. and Skov, T. 2017. Recent chemometrics advances for foodomics. *Trends in analytical chemistry*, 96:42e51.
- Blaise, B.J., Correia, G., Tin, A., Young, J.H., Vergnaud, A.C., Lewis, M., Pearce, J.T., Elliott, P., Nicholson, J.K., Holmes, E. & Ebbels, T.M. 2016. Power analysis and sample size determination in metabolic phenotyping. *Analytical chemistry*, 88(10):5179-5188.
- Collins, F.S. 2011. Re-engineering Translational Science: The Time Is Right. *Commentary in www.ScienceTranslationalMedicine.org*, 3(90):1.
- Collins, F. S. & Varmus, H. 2015. A New Initiative on Precision Medicine. *New England Journal of Medicine*, 372:793-795.
- Emwas, A.H., Saccenti, E., Gao, X., McKay, R.T., dos Santos, V.A.M., Roy, R. and Wishart, D.S., 2018. Recommended strategies for spectral processing and post-processing of 1D ¹H-NMR data of biofluids with a particular focus on urine. *Metabolomics*, 14(3):31.
- English Oxford Living Dictionaries. *Oxford University Press Online thesaurus*, 2018. Available at: <https://en.oxforddictionaries.com/thesaurus/valid> [Accessed: 17 April 2018]
- Garamszegi, L.Z. 2006. Comparing effect sizes across variables: generalization without the need for Bonferroni correction. *Behavioral Ecology*, 17(4):682-687.
- Gromski, P.S., Xu, Y., Kotze, H.L., Correa, E., Ellis, D.I., Armitage, E.G., Turner, M.L. & Goodacre, R. 2014. Influence of missing values substitutes on multivariate analysis of metabolomics data. *Metabolites*, 4(2):433-452.
- Ioannidis, J.P. 2008. Why most discovered true associations are inflated. *Epidemiology*, 19(5):640-648.
- Kjeldahl, K. & Bro, R. 2010. Some common misunderstandings in chemometrics. *Journal of chemometrics*, 24(7-8):558-564.

Manach, C., Milenkovic, D., Wiele, T., Rodriguez-Mateos, A., Roos, B., Garcia-Conesa, M.T., Landberg, R., Gibney, E.R., Heinonen, M., Tomás-Barberán, F. & Morand, C. 2017. Addressing the inter-individual variation in response to consumption of plant food bioactives: Towards a better understanding of their role in healthy aging and cardiometabolic risk reduction. *Molecular nutrition & food research*, 61(6).

Nyamundanda, G., Gormley, I.C., Fan, Y., Gallagher, W.M. & Brennan, L. 2013. MetSizeR: selecting the optimal sample size for metabolomic studies using an analysis based approach. *BMC bioinformatics*, 14(1):338.

Pauling, L., Robinson, A.B., Teranishi, R. & Cary, P. 1971. Quantitative analysis of urine vapor and breath by gas-liquid partition chromatography. *Proceedings of the National Academy of Sciences USA*, 68(10):2374–2376.

Perkins, B.A., Caskey, C.T., Brar, P., Dec, E., Karow, D.S., Kahn, A.M., Hou, Y.C.C., Shah, N., Boeldt, D., Coughlin, E. & Hands, G. 2018. Precision medicine screening using whole-genome sequencing and advanced imaging to identify disease risk in adults. *Proceedings of the National Academy of Sciences*, 115(14):3686-3691.

Saccenti, E., Hoefsloot, H.C.J., Smilde, A.K., Westerhuis, J.A. & Hendriks, M.M.W.B. 2013. Reflections on univariate and multivariate analysis of metabolomics data. *Metabolomics*, 10(3):361-374.

Saccenti, E. & Timmerman, M.E. 2016. Approaches to sample size determination for multivariate data: Applications to PCA and PLS-DA of omics data. *Journal of proteome research*, 15(8):2379-2393.

Taylor, S. & Pollard, K. Hypothesis Tests for Point-Mass Mixture Data with Application to `Omics Data with Many Zero Values. *Statistical Applications in Genetics and Molecular Biology*. 2009. 8(1):1–43.

van den Berg, R.A., Hoefsloot, H.C., Westerhuis, J.A., Smilde, A.K. & Van der Werf, M.J. 2006. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC genomics*, 7(1):142.

Xia, J., Broadhurst, D.I., Wilson, M. & Wishart, D.S. 2013. Translational biomarker discovery in clinical metabolomics: an introductory tutorial. *Metabolomics*, 9(2):280-299.

ADDENDUMS

6.1 Permission from Co-authors

None of the co-authors contributing to the publications included in this thesis had any objections regarding the reproduction of said publication in this thesis.

6.2 Copyright Licences of Journals

6.2.1 PloS one

Irwin, C., van Reenen, M., Mason, S., Mienie, L.J., Westerhuis, J.A. & Reinecke, C.J. 2016. Contribution towards a metabolite profile of the detoxification of benzoic acid through glycine conjugation: an intervention study. *PloS one*, 11(12):e0167309. doi:10.1371/journal.pone.0167309.

Copyright: © Irwin *et al.* 2016. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Irwin, C., van Reenen, M., Mason, S., Mienie, L.J., Wevers, R.A., Westerhuis, J.A. & Reinecke, C.J. 2018. The ¹H-NMR-based metabolite profile of acute alcohol consumption: a metabolomics intervention study. *PloS one*. (Still in press)

Copyright: © Irwin *et al.* 2018. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Terms of Use | PLOS

For Authors

These Terms of Use are a contract between you and the Public Library of Science ("PLOS"). PLOS operates www.PLOS.org and other websites ("PLOS Sites") through which it strives to make the world's scientific and medical literature a freely available public resource. By using the PLOS Sites and any services accessible from the PLOS Sites (including the PLOS API, as defined below) you agree to be bound by the following terms and conditions ("Terms of Use"). If you do not agree to these Terms of Use, do not visit or use the PLOS Sites.

Reproduction of Articles

Articles and accompanying materials published by PLOS on the PLOS Sites, unless otherwise indicated, are licensed by the respective authors of such articles for use and distribution by you subject to citation of the original source in accordance with the [Creative Commons Attribution \(CC BY\) license](#).

Ownership of Content

The copyright in the material contained on the PLOS Sites belongs to PLOS or its licensors. The trademarks and other elements appearing on the PLOS Sites are protected by California, US, and international copyright, trade dress, patent, and trademark laws, international conventions and other relevant and applicable laws.

USE OF PLOS' ONLINE TOOLS AND SERVICES

PLOS provides access on its websites to certain software tools and functionality ("Services") for the purposes of allowing and facilitating the submission or posting of research outputs and results in various forms including articles in whole or part, manuscripts, and comments on and reviews of such content, and the like. By using our website(s) and/or Services, you agree to not interfere with or otherwise disrupt any of the Services, the operation or proper functioning of the Services, or any person's use of or access to the Services. This includes but is not limited to employing trojan horses, viruses, and other malware, breaches of or tampering with the Services, or attempts to probe, scan, or test for vulnerabilities in the Services or PLOS' websites or our computer systems, networks, usage rules, or any of our security components, authentication measures or any other protection measures applicable to our websites or the Services, or the content or any part thereof.

PLOS API

PLOS provides certain application programming interfaces for the PLOS Sites (the "PLOS API"). You agree not to make any unauthorized use of the PLOS API or interfere with the functionality or proper working of the PLOS API. Through your use of the PLOS API, you agree not to, at PLOS' sole discretion, use an

unreasonable amount of bandwidth or unreasonably burden PLOS' servers. The PLOS API is provided "as is" without any support or warranty of any kind. PLOS makes no guarantees that the PLOS API will be available. PLOS reserves the right to modify or discontinue, temporarily or permanently, the PLOS API with or without notice at any time. You agree that PLOS shall not be liable to you or to any third party for any modification, suspension or discontinuance of the PLOS API. You can find out more about the PLOS API [here](#).

Ownership of Sites

Unless otherwise indicated, the PLOS Sites are the property of PLOS. You acknowledge and agree that the PLOS Sites, and any necessary software used in connection with the PLOS Sites, may contain proprietary and confidential information that is protected by applicable laws governing intellectual property, proprietary rights and the like. All PLOS Sites code and specifications are subject to their stated or implied licenses.

Privacy

You confirm that you have read and accept our [Privacy Policy](#).

General Practices Regarding Use

PLOS reserves the right to limit, suspend or terminate your access and use of the PLOS Sites at any time without notice.

Article and Content Submission

By submitting articles for publication or other content for posting on the PLOS sites, you agree to (i) comply with these Terms of Use, (ii) have the Creative Commons Attribution (CC BY) license applied to material you submit, and (iii) provide any information required by PLOS during or as part of the submission process. You are responsible for all content that you submit for posting or publication on the PLOS Sites.

Impermissible Acts

As a condition to your use of the PLOS Sites, you agree not to:

- Upload, post, email, transmit or otherwise make available any information, materials or other content that is illegal, harmful, threatening, abusive, harassing, defamatory, obscene, pornographic or offensive; or that infringes another's rights, including any intellectual property rights
- Impersonate any person or entity or falsely state or otherwise misrepresent your affiliation with a person or entity; or obtain, collect, store or modify personal information about other users
- Upload, post, email, transmit or otherwise make available any unsolicited or unauthorized advertising, promotional materials, "junk mail," "spam," "chain letters," "pyramid schemes" or any other form of solicitation
- Modify, adapt or hack the PLOS Sites or falsely imply that some other site is

associated with the PLOS Sites or PLOS

- Use the PLOS Sites or PLOS API for any illegal or unauthorized purpose. You must not, in the use of the PLOS Sites or PLOS API, violate any laws in your jurisdiction (including but not limited to copyright laws)

Violation of these Terms of Use

You acknowledge that PLOS has no obligation to pre-screen or monitor your access to or use of the PLOS Sites or PLOS API or any information, materials or other content provided or made available through the PLOS Sites or PLOS API, but has the right to do so. You hereby agree that PLOS may, in the exercise of PLOS' sole discretion, remove or delete any entries, information, materials or other content that violates these Terms of Use or that is otherwise objectionable.

Modifications to the PLOS Sites and Terms of Use

PLOS reserves the right to modify or discontinue, temporarily or permanently, the PLOS Sites (or any part thereof) with or without notice at any time. You agree that PLOS shall not be liable to you or to any third party for any modification, suspension or discontinuance of the PLOS Sites or PLOS API. PLOS reserves the right to change these Terms of Use by posting changes on this page of the PLOS Sites. By using the PLOS Sites or PLOS API, you agree to be bound by any such revisions and should therefore periodically visit this page to determine the then-current Terms of Use to which you are bound.

Links

The PLOS Sites may provide links to other web sites or resources. You acknowledge and agree that PLOS is not responsible for the availability of such external sites or resources, and that PLOS does not endorse and is not responsible or liable for any content, advertising, products or other materials on or available from such sites or resources. Your business dealings with any third-parties through the PLOS Sites, including payment and delivery of related goods or services, and any other terms, conditions, warranties or representations associated with such dealings, are solely between you and such third party. You further acknowledge and agree that PLOS shall not be responsible or liable, directly or indirectly, for any damage or loss caused or alleged to be caused by or in connection with use of or reliance on any such content, goods or services available on or through any such site or resource.

Notification of Claims of Infringement

If you believe that a creative work that you or an entity you represent owns has been used on a PLOS website in violation of copyright laws, and you want to request that the material be taken down, please go to <http://journals.plos.org/plosone/s/licenses-and-copyright>.

DISCLAIMER OF WARRANTIES

YOUR USE OF THE PLOS SITES, PLOS API OR ANY WEB SITE TO WHICH THEY LINK IS AT YOUR SOLE RISK. THE PLOS SITES, PLOS API AND THEIR CONTENT ARE PROVIDED FOR USE ON AN "AS IS" AND "AS AVAILABLE"

BASIS. PLOS MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE PLOS SITES, PLOS API OR THEIR CONTENTS AND HEREBY DISCLAIMS ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING WITHOUT LIMITATION THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE, OR NON-INFRINGEMENT RELATING TO THE PLOS SITES, THEIR CONTENT, PLOS API OR ANY WEB SITE TO WHICH THEY ARE LINKED.

NO ADVICE OR INFORMATION, WHETHER ORAL OR WRITTEN, OBTAINED BY YOU FROM PLOS OR THROUGH OR FROM THE PLOS SITES OR PLOS API SHALL CREATE ANY WARRANTY NOT EXPRESSLY STATED IN THE TERMS OF USE.

Indemnification, Limitation of Liability

You agree to indemnify and hold PLOS, and its subsidiaries, affiliates, officers, agents, co-branders or other partners and employees, harmless from any claim or demand, including reasonable attorney fees, made by any third party due to or arising out of your use of the PLOS Sites, PLOS API, your violation of these Terms of Use, your violation of any rights of another, or any breach of the foregoing representations, warranties and covenants. You are solely responsible for your actions when using the PLOS Sites, including, but not limited to, costs incurred for Internet access. PLOS reserves the right, at our own expense, to assume the exclusive defense and control of any matter for which you are required to indemnify us, and you agree to cooperate with our defense of these claims.

IN NO EVENT SHALL PLOS OR ITS EMPLOYEES, AGENTS, SUPPLIERS OR CONTRACTORS BE LIABLE FOR ANY DAMAGES OF ANY NATURE, INCLUDING WITHOUT LIMITATION ANY CONSEQUENTIAL LOSS, DAMAGES FOR LOSS OF INCOME OR PROFIT, LOSS OF OR DAMAGE TO PROPERTY, LOSS OF GOODWILL, USE, DATA OR OTHER INTANGIBLE LOSSES, CLAIMS OF THIRD PARTIES, OR ANY OTHER LOSS, COST, CLAIM OR EXPENSE OF ANY KIND OR CHARACTER ARISING OUT OF OR IN CONNECTION WITH THE USE OF THE PLOS SITES, THE PLOS API, THEIR CONTENT OR ANY WEB SITE WITH WHICH THEY ARE LINKED.

Special Considerations for International use

Recognizing the global nature of the Internet, you agree to comply with all local rules regarding online conduct. Specifically, you agree to comply with all applicable laws regarding the transmission of technical data exported from the United States or the country in which you reside.

General Terms

These Terms of Use constitute the entire agreement between you and PLOS and govern your use of the PLOS Sites, superseding any prior agreements between you and PLOS (including, but not limited to, any prior versions of the Terms of Use) and any prior representations by PLOS. These Terms of Use shall be governed by and construed in accordance with the laws of the State of California, without regard to conflicts of law principles. The failure of PLOS to exercise or enforce any right or provision of the Terms of Use shall not constitute a waiver of such right or provision. If any provision of the Terms of Use is found by a court of competent

jurisdiction to be invalid, the parties nevertheless agree that the court should endeavor to give effect to the parties' intentions as reflected in the provision, and the other provisions of the Terms of Use remain in full force and effect. You agree that regardless of any statute or law to the contrary, any claim or cause of action brought against PLOS and its subsidiaries, affiliates, officers, agents, co-branders or other partners and employees, arising out of or related to use of the PLOS Sites or the Terms of Use must be filed within one (1) year after such claim or cause of action arose or be forever barred. The section titles in the Terms of Use are for convenience only and have no legal or contractual effect.

6.2.2 Scientific Reports

Irwin, C., Mienie, L.J., Wevers, R.A., Mason, S., Westerhuis, J.A., van Reenen, M. & Reinecke, C.J. 2018. GC–MS metabolomics reveals multiple dysregulated metabolic pathways following experimental acute alcohol consumption. *Scientific reports*, 8:5775. doi:10.1038/s41598-018-24128-1

Copyright: © Irwin *et al.* 2018. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**RightsLink®****SPRINGER NATURE**

Title: GC-MS-based urinary organic acid profiling reveals multiple dysregulated metabolic pathways following experimental acute alcohol consumption

Author: Cindy Irwin et al

Publication: Scientific Reports

Publisher: Springer Nature

Date: Apr 10, 2018

Copyright © 2018, Springer Nature

Creative Commons

This is an open access article distributed under the terms of the [Creative Commons CC BY](#) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

You are not required to obtain permission to reuse this article.

Are you the [author](#) of this Springer Nature article?

To order reprints of this content, please contact Springer Nature by e-mail at reprintswarehouse@springernature.com, and you will be contacted very shortly with a quote.

6.2.3 BMC Bioinformatics

van Reenen, M., Reinecke, C.J., Westerhuis, J.A. and Venter, J.H. Variable selection for binary classification using error rate p-values applied to metabolomics data. 2016. *BMC bioinformatics*, 17(1):33. <https://doi.org/10.1186/s12859-015-0867-7>.

Copyright: © van Reenen *et al.* 2016. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

van Reenen, M., Westerhuis, J.A., Reinecke, C.J. and Venter, J.H. 2017. Metabolomics variable selection and classification in the presence of observations below the detection limit using an extension of ERp. *BMC bioinformatics*, 18(1):83.

Copyright: © van Reenen *et al.* 2017. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

BMC Bioinformatics | Copyright

- Copyright on any open access article in a journal published by BioMed Central is retained by the author(s).
- Authors grant BioMed Central a [license](#) to publish the article and identify itself as the original publisher.
- Authors also grant any third party the right to use the article freely as long as its integrity is maintained and its original authors, citation details and publisher are identified.
- The [Creative Commons Attribution License 4.0](#) formalizes these and other terms and conditions of publishing articles.

In addition to BioMed Central's copyright policy, some journals also follow an Open Data policy and the [Creative Commons CCo 1.0 Public Domain Dedication waiver](#) applies to all published data in these journals. Further information can be found on the individual journals pages.

Where an author is prevented from being the copyright holder (for instance in the case of US government employees or those of Commonwealth governments), minor variations may be required. In such cases the copyright line and license statement in individual articles will be adjusted, for example to state '© 2016 Crown copyright'. Authors requiring a variation of this type should [inform BioMed Central](#) during or immediately after submission of their article. Changes to the copyright line cannot be made after publication of an article.

Exceptions to copyright policy

Our [policy pages](#) provide details concerning copyright and licensing for articles which were previously published under policies that are different from the above. For instance, occasionally BioMed Central may co-publish articles jointly with other publishers, and different licensing conditions may then apply. In all such cases, however, access to these articles is free from fees or any other access restrictions.

Information specifically regarding permissions and reprints can be found [here](#). Please [contact us](#) if there are questions.

6.3 Journals' Instructions to Authors

6.3.1 PloS one

Article title: PLoS one: accelerating the publication of peer-reviewed science

Website title: Journals.plos.org

URL: <http://journals.plos.org/plosone/s/submission-guidelines>

6.3.2 Scientific Reports

Article title: Submission guidelines | Scientific Reports

Website title: Nature.com

URL: <https://www.nature.com/srep/publish/guidelines>

6.3.3 BMC Bioinformatics

Article title: BMC Bioinformatics

Website title: BMC Bioinformatics

URL: <https://bmcbioinformatics.biomedcentral.com/submission-guidelines/preparing-your-manuscript/methodology-article>