Comparative Study of Neural Networks and Design of Experiments to the Classification of HIV status

WILBERT SIBANDA

Student Number: 21935009

- BSc (Life Sciences) University of Witwatersrand, Johannesburg
- BSc (Med) Hons Pharmacology, University of Cape Town
- MSc (Med) Pharmacy University of the Witwatersrand, Johannesburg

Thesis submitted for the degree of Doctor of Philosophy in Information Technology (IT) at the Vaal Triangle Campus of the North-West University

Promoter: **Prof. Philip Pretorius**
School of Information Technology
Vaal Triangle campus
North-West university
South Africa

Abstract

This research addresses the novel application of design of experiment, artificial neural networks and logistic regression to study the effect of demographic characteristics on the risk of acquiring HIV infection among the antenatal clinic attendees in South Africa.

The annual antenatal HIV survey is the only major national indicator for HIV prevalence in South Africa. This is a vital technique to understand the changes in the HIV epidemic over time. The annual antenatal clinic data contains the following demographic characteristics for each pregnant woman; age (herein called mother's age), partner's age (herein father's age), population group (race), level of education, gravidity (number of pregnancies), parity (number of children born), HIV and syphilis status.

This project applied a screening design of experiment technique to rank the effects of individual demographic characteristics on the risk of acquiring an HIV infection. There are a various screening design techniques such as fractional or full factorial and Plackett-Burman designs. In this work, a two-level fractional factorial design was selected for the purposes of screening. In addition to screening designs, this project employed response surface methodologies (RSM) to estimate interaction and quadratic effects of demographic characteristics using a central composite face-centered and a Box-Behnken design.

Furthermore, this research presents the novel application of multi-layer perceptrons (MLP) neural networks to model the demographic characteristics of antenatal clinic attendees. A review report was produced to study the application of neural networks to modeling HIV/AIDS around the world. The latter report is important to enhance our understanding of the extent to which neural networks have been applied to study the HIV/AIDS pandemic.

Finally, a binary logistic regression technique was employed to benchmark the results obtained by the design of experiments and neural networks methodologies.

The two-level fractional factorial design demonstrated that HIV prevalence was highly sensitive to changes in the mother's age (15-55 years) and level of her education (Grades 0-13). The central composite face centered and Box-Behnken designs employed to study the individual and interaction effects of demographic characteristics on the spread of HIV in South Africa, demonstrated that HIV status of an antenatal clinic attendee was highly sensitive to changes in pregnant mother's age and her educational level. In addition, the interaction of the mother's age with other demographic characteristics was also found to be an important determinant of the risk of acquiring an HIV infection. Furthermore, the central

composite face centered and Box-Behnken designs illustrated that, individually the pregnant mother's parity and her partner's age had no marked effect on her HIV status. However, the pregnant woman's parity and her male partner's age did show marked effects on her HIV status in "two way interactions with other demographic characteristics".

The multilayer perceptron (MLP) sensitivity test also showed that the age of the pregnant woman had the greatest effect on the risk of acquiring an HIV infection, while her gravidity and syphilis status had the lowest effects. The outcome of the MLP modeling produced the same results obtained by the screening and response surface methodologies.

The binary logistic regression technique was compared with a Box-Behnken design to further elucidate the differential effects of demographic characteristics on the risk of acquiring HIV amongst pregnant women. The two methodologies indicated that the age of the pregnant woman and her level of education had the most profound effects on her risk of acquiring an HIV infection. To facilitate the comparison of the performance of the classifiers used in this study, a receiver operating characteristics (ROC) curve was applied. Theoretically, an ROC analysis provides tools to select optimal models and to discard suboptimal ones independent from the cost context or the classification distribution. SAS Enterprise Miner™ was employed to develop the required receiver-of-characteristics (ROC) curves.

To validate the results obtained by the above classification methodologies, a credit scoring add-on in SAS Enterprise Miner™ was used to build binary target scorecards comprised of HIV positive and negative datasets for probability determination. The process involved grouping variables using weights-of-evidence (WOE), prior to performing a logistic regression to produce predicted probabilities. The process of creating bins for the scorecard enables the study of the inherent relationship between demographic characteristics and an individual's HIV status. This technique increases the understanding of the risk ranking ability of the scorecard method, while offering an added advantage of being predictive.

Keywords:
Factorial, Central composite face-centered, Box-Behnken, multilayer perceptron, binary logistic regression, HIV, demographic characteristics.

Declaration

I, Wilbert Sibanda, hereby declare that the thesis entitled **'Comparative Study of the novel application of design of experiments (DOE) and Neural Networks (NN) to the classification of HIV status of antenatal clinic attendees in South Africa'** is my work. No plagiarism has taken place and due acknowledgements and references were given.

Wilbert Sibanda **15/11/2013**

Acknowledgements

I would like to extend special thanks to the following people and/or institutions for their contributions towards this project.

- My beloved wife Cathrine (***Katie***) and my children, Lorraine (***Masi***), Janice (***Mute-Mute***), Njabulo (***Draad***) and Vuyo (***Pumpkin***).
- Prof. Philip Pretorius (my PhD supervisor)
- Medical Research Council of South Africa for the doctoral funding
- South African Centre for Epidemiological Modeling and Analysis (SACEMA), for the doctoral funding and constant workshops that sharpened my skills in epidemiological modeling.
- North-West university (Vaal Triangle campus) for graduate funding

Dedicated to the Memory of my Late Parents

Mr. and Mrs. Ores Leonard Sibanda

Table of Contents

List of Tables

List of Figures

List of Scientific Publications in Peer-reviewed journals

Response surface modeling and optimization to elucidate the differential effects of demographic characteristics on HIV prevalence in South Africa.

Journal: 2012 Computers and Industrial Engineering 42
Publisher: CIE & SAIIE
Date of Publication: 2012

Artificial neural networks- A review of applications of neural networks in the modeling of HIV epidemic.

Journal: International Journal of Computer Applications (0975-8887)
Date of Publication: 2012.

Novel application of multi-layer perceptrons (MLP) neural networks to model HIV.

Journal: International Journal of Computer Applications (0975-8887)
Date of Publication: 2011

Application of two-level fractional factorial design to determine and optimize the effect of demographic characteristics on HIV prevalence.

Journal: International Journal of Computer Applications (0975-8887)
Date of Publication: 2011

Comparative Study of the Application of Box-Behnken Designs (BBD) and Binary Logistic Regression (BLR) to study the effect of demographic characteristics on HIV risk in South Africa.

Journal: International Journal of Applied Medical Sciences

Application of Central Composite Face-Centered (CCF) and Box-Behnken Designs (BBD) to study the effect of demographic characteristics on HIV risk in South Africa.

Journal: Network Modeling and Analysis in Health Informatics and Bioinformatics

Application of ROC curves to compare neural networks, logistic regression and decision trees for modelling the causal relationship between demographic characteristics and the risk of acquiring infection using antenatal seroprevalence data.
Abstract accepted: 42nd Annual Conference of the Operations Research Society of South Africa

Development and validation of an HIV risk scorecard model.
Submitted: International Symposium on Network Enabled Health Informatics, Biomedicine and Bioinformatics (HI-BI-BI 2013), Niagara Falls, Canada