Annexure G:

Development and validation of an HIV risk scorecard model

Development and validation of an HIV risk scorecard model

Wilbert Sibanda*
DST/NWU Preclinical Platform
Faculty of Health Sciences, North-West University
Potchefstroom campus, South Africa
E-mail: Wilbert.sibanda@nwu.ac.za
*Corresponding author

Philip Pretorius
School of Information Technology
Vaal Triangle campus
North-West University, South Africa
E-mail: Philip.pretorius@nwu.ac.za

***Abstract*—This research paper covers the development of an HIV risk scorecard using SAS Enterprise Miner™. The HIV risk scorecard was developed using the 2007 South African annual antenatal HIV and syphilis seroprevalence data. Antenatal data contains various demographic characteristics for each pregnant woman, such as pregnant woman's age, male sexual partner's age, race, level of education, gravidity, parity, HIV and syphilis status. The purpose of this research was to use a scorecard to rank the effects of the demographic characteristics on influencing an individual's risk of acquiring an HIV infection, not the probability of being sick. The project encompassed the selection of the data sample, classing, selection of demographic characteristics, fitting of a regression model, generation of weights-of-evidence (WOE), calculation of information values (IVs), creation and validation of an HIV risk scorecard. The educational level and syphilis status of the pregnant women produced information values below 0.05 and were rejected from inclusion in the final HIV risk scorecard. Based on their respective information values, the following four demographic characteristics of the pregnant women were found to be of medium predictive strength and thus included in the final HIV risk scorecard; age, age of male sexual partner, gravidity and parity. The age of the pregnant woman had the highest information value and Gini coefficient. The HIV risk scorecard showed that the risk of contracting an HIV infection increased gradually up to the age of 30 years for females and 34 years old for their male sexual partners. Thereafter, the risk decreased gradually towards the age of 45.**

Keywords*—*IV; HIV; WOE

I. Introduction

Over the past years, the insurance industry has grown to appreciate the need to insure HIV positive individuals [4]. South Africa is one of a very few number of countries where HIV positive individuals have access to life insurance policies. This comes as no surprise as South Africa has one of the highest levels of HIV infection in the world. A significant number of these HIV positive individuals are gainfully employed in different business sectors within the Republic. With the extensive roll-out of anti-retroviral drugs and the subsequent improvement in life expectancy rate to 60 years [2], it is imperative for the insurance companies to develop life cover products for HIV positive citizens. The number of HIV positive individuals enrolled for anti-retroviral therapy (ART) in South Africa by the year 2011, had increased to 1,8 million [1]. With the above information in mind, it is important to develop a scorecard that can help insurance companies, commercial banks and other financial institutions to predict the likelihood of an individual becoming HIV positive in order to tailor-make the most appropriate financial packages and insurance policies. The same scorecard can be useful for government policy makers in predicting the likely direction of the HIV epidemic and thus develop sound intervention strategies that focus on the specific age-groups that are at most risk.

The following are examples of scorecards that have been successfully developed in life sciences;

- *Diabetes scorecard*- This scorecard was developed by the International Diabetes Federation, for measuring government's commitment of diabetes,
- *Heart disease scorecard*- This scorecard was developed by the American Heart Association (1),
- *Huntington's disease scorecard*- The scorecard was developed to provide drugs and supplements for Huntington's disease (3),
- *Health literacy scorecard*- The scorecard was developed to demonstrate the variables that contribute to ill-health, chronic disease and early death (8).

Prior to this project, the researchers used the following statistical modeling techniques to rank the effects of demographic characteristics on the risk of acquiring an HIV infection, two-level fractional factorial design (10), response surface methodology (9), binary logistic regression (11) and multilayer perceptrons (12). The above methodologies showed that the mother's age had the most influential effect on an individual's risk of acquiring an HIV infection.

The aim of this paper is to demonstrate that a binned variable scorecard can provide as much risk ranking as any other regression based model, with the advantages associated with the simplicity of the scorecards.

The advantages of the HIV scorecard include:

- The binning process allows for the evaluation of the relationship between each demographic characteristic and the HIV risk,
- The binning process enables the risk ranking of demographic characteristics,
- The binning process also allows the researcher to adjust risk relationships based on established epidemiological knowledge, by assigning higher or lower weights-of-evidence (WOE),
- The format of the scorecard is easy to understand, explain and use.

II. Building HIV Risk Scorecard Model

The 2007 antenatal HIV seroprevalence data was used for this research. The database consisted of 32 611 records with seven variables. Of the 32 611 records, 23 165 individuals were HIV negative and 9 446 HIV positive. The antenatal HIV data was collected from pregnant women attending antenatal clinics throughout the nine provinces of South Africa, and the seven demographic characteristics studied were the pregnant woman's age, her educational level, gravidity, parity, male partner's age, syphilis status and HIV status.

The research steps were as follows;

1. HIV status was presented as a binary response with an 0 and 1 representing an HIV negative and positive status respectively,
2. The Interactive Grouping Node in SAS Enterprise Miner™, was used to develop the weights-of-evidence (WOE) for each demographic characteristic,
3. The demographic characteristics with the highest WOE values were binned based on their strength, with high negative WOE values implying high risk and high positive values corresponding to low risk of acquiring an HIV infection,
4. The formula for the calculation of the WOE values in SAS Enterprise Miner™ was based on the standard formula below;

$$WOE_{bin} = \ln\left(\frac{\text{Proportion of HIV negative individuals in bin}}{\text{Proportion of HIV positive individuals in bin}}\right) \qquad \text{(i)}$$

5. The WOE values were used to determine the information values (IV) and based on the IVs, the demographic characteristics with low IV less than 0.05 were discarded. The term IV stands for the ability to separate high risk from low risk to aid the selection of demographic characteristics to be included in the final scorecard.
6. The formula for the calculation of the information values was

$$IV = \sum_{i=1}^{n} \left(\text{Proportion HIVnegative} - \text{Proportion HIVpositive} * \ln\frac{\text{Proportion HIVnegative}}{\text{Proportion HIVpositive}}\right) \qquad \text{(ii)}$$

where n is the number of attributes for a demographic characteristics

7. Siddiqi recommended the classification of variables in terms of IVs as shown in Table 1.

Table 1: Information values

Information value	Meaning
≤0.02	unpredictive
0.02-0.1	weak
0.1-0.3	medium
≥0.3	strong

The WOE analyzes the predictive power of a demographic characteristic in relation to an individual's HIV status. The information value on the other hand assesses the overall predictive power of the demographic characteristic being considered, and thus can be used for comparing the predictive power among different demographic characteristics.

III. Output

Fig. 1 shows the plot of the weights-of-evidence (WOE) for each age-group of the pregnant woman. In this situation, the WOE measures the strength of an age-group in differentiating between HIV negative and positive individuals at each group level. Therefore negative values show that a particular grouping is isolating a higher proportion of HIV positive individuals than their HIV negative counterparts. This therefore implies that individuals with negative WOEs present a greater risk for HIV infection.

For pregnant mothers, the age group 28 to 31 years was found to be at risk of acquiring an HIV infection. As expected, the younger women less than 18 years were least at risk of infection, probably due to less sexual involvement. In addition, the HIV risk seemed to decrease after the age of 32 years. The latter situation was attributed to possible stability in marriage, less risky sexual practices, psychological maturity and awareness of the HIV pandemic.

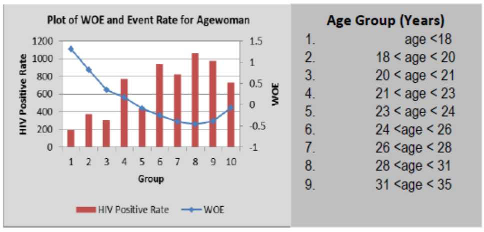

Fig. 1: Plot of weight-of-evidence and HIV positive rate against pregnant woman's age

Fig. 2 shows that as the WOE values decrease, the HIV positive rate for male sexual partners increases gradually to a maximum at the age of 33 years. At this age, males are found to be at most risk of acquiring an HIV infection, thereafter the risk decrease as the age increases to towards the age of 50.

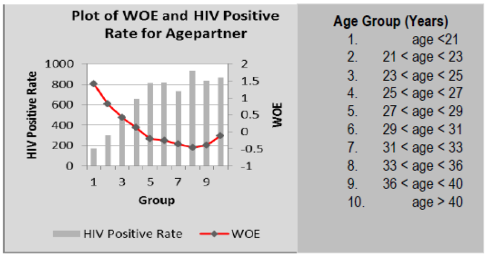

Fig. 2: Plot of weight-of-evidence and HIV positive rate against male sexual partner's age

The plot of the changes in WOE with an increase in gravidity (Fig. 3) indicates that the women experiencing their third pregnancy are at most risk of acquiring an HIV infection. Once more, as expected women presenting at antenatal clinics for their first pregnancy were found to least at risk. These are presumably, the younger women debuting in sexual practice and not yet exposed to multiple partners and the possibility of being infected. On the other hand, Fig. 4 shows that the women presenting to antenatal clinics having given birth to two surviving children are most at risk of acquiring an HIV infection.

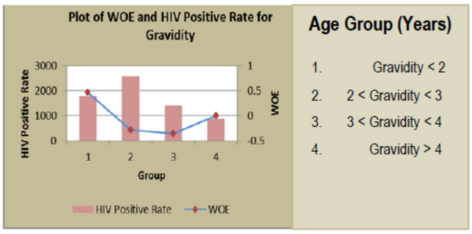

Fig. 3: Plot of weight-of-evidence and HIV positive rate against changes in gravidity

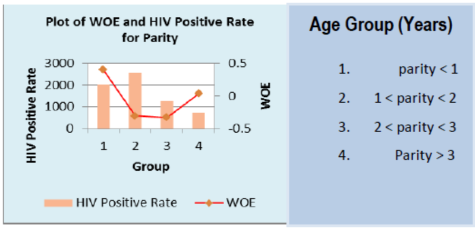

Fig. 4: Plot of weight-of-evidence and HIV positive rate against changes in parity

IV. Binning Based on WOE and Choice of Demographic Characteristics with Adequate Predictive Power

The Interactive Grouping node in SAS Enterprise Miner™ was used to analyze the risk ranking ability of each demographic characteristic. Attributes with low information values (IV), less than 0.05 were discarded. These were the educational level and syphilis status of the pregnant woman. Based on Siddiqi information value classification guideline, the pregnant woman's age, male sexual partner's age, gravidity and parity were binned as these demographic characteristics presented medium predictive tendencies. The age of the pregnant woman had the most predictive information value (IV) followed by her male sexual partner's age with the parity characteristic ranking last.

Table 2: Binning of demographic characteristics

Variable	Expected Role	Original Gini	Original Information value (IV)	Gini Statistic
Agepartner	input	24.349	0.243	24.349
Agewoman	input	24.909	0.249	24.409
Gravidity	input	18.634	0.126	18.634
Parity	input	17.224	0.109	17.224
Education	**rejected**	**0.245**	**0.00**	**0.245**
Syphilis	**rejected**	**0.00**	**0.00**	**0.00**

V. Development of a Scorecard Model Using a Logistic Regression

The scorecard node in SAS Enterprise Miner™ was used to fit a logistic regression model and generation of a scorecard. A stepwise selection method was used to eliminate unnecessary characteristics. The final scorecard had four demographic characteristics namely, age of the pregnant woman, age of the male sexual partner, gravidity and parity. The selected demographic characteristics are shown in the scorecard in Fig. 5, together with their respective scores and associated HIV positive rate.

Table 3: Scorecard regression coefficients

Parameter	DF	Estimate
Intercept	1	-0.8994
WOE_Age_partner	1	-0.5376
WOE_Age_woman	1	-0.4776
WOE_gravidity	1	-0.3540
WOE_parity	1	0.0326

Table 3 shows the values of the regression coefficients generated by the logistic regression. The products of the weight-of-evidence (WOE) groups and the regression coefficients are used to develop the scorecard points shown in Fig. 5.

Scorecard				
Demographic Characteristics	**Age Group**	**Age Group**	**Scorecard Points**	**HIV rate**
Age_partner	age_partner < 21	1	50	172
	21<age_partner<23	2	41	297
	23<age_partner<25	3	35	495
	25<age_partner<27	4	30	660
	27<age_partner<29	5	25	815
	29<age_partner<31	6	24	820
	31<age_partner<33	7	23	730
	33<age_partner<36	8	21	937
	36<age_partner<40	9	22	835
	40<age_partner	10	27	865
Age_woman	Age_woman<18	1	46	195
	18<age_woman<20	2	40	370
	20<age_woman<21	3	33	304
	21<age_woman<23	4	31	775
	23<age_woman<24	5	27	441
	24<age_woman<26	6	25	938
	26<age_woman<28	7	23	822
	28<age_woman<31	8	22	1061
	31<age_woman<35	9	23	977
	35<age_woman	10	27	728
Gravidity	Gravidity<2	1	33	1780
	2<gravidity<3	2	25	2570
	3<gravidity<4	3	25	1389
	4<gravidity	4	28	866
Parity	Parity<1	1	28	2024
	1<parity<2	2	29	2589
	2<parity<3	3	29	1263
	3<parity	4	28	735

Fig.5: Final scorecard

Score Bin	HIV positive count	Marginal HIV rate	Average Predicted Probability
Training Dataset			
100	4000	40	0.40
105	2500	40	0.35
110	1500	35	0.32
115	1000	29	0.28
120	800	25	0.25
125	500	23	0.22
130	500	20	0.18
135	400	15	0.15
140	350	8	0.13
Validation dataset			
100	3000	40	0.4
105	1500	40	0.35
110	1100	35	0.32
115	700	29	0.28
120	500	28	0.25
125	450	22	0.22
130	400	20	0.18
135	350	17	0.15
140	300	20	0.13

Fig. 6: Gains table

Fig. 6 shows close agreement between marginal HIV positive rates and average predicted probabilities. For an example, a score of 115 provided a marginal HIV rate of 29% and the corresponding average predicted probability of 0.28.

VI. VALIDATION OF THE SCORECARD

The HIV risk scorecard was validated using a 25% hold-out sample. Many fit statistics were used for scorecard validation such as cumulative percent response, cumulative percent of responses captured, receiver-of-characteristics (ROC) curve, empirical odds and Kolmogorov-Smirnov plots.

A. Cumulative Percent response curve

This lift curve shows the predictive power of the scorecard model, and provides a score that represents the likeliness of a pregnant woman being HIV positive. The scores are sorted into ten deciles, from left to right along the horizontal axis, with the highest scoring ten percent of individuals on the left edge of the chart.

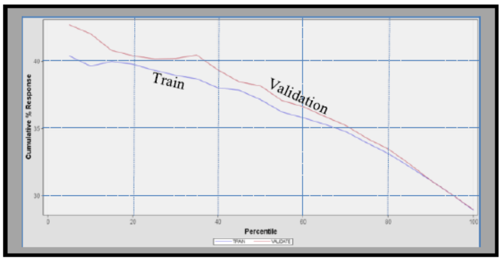

Fig. 7: Cumulative percentage response chart

After training and validation, the scorecard model is able to predict the HIV status of a pregnant woman with an accuracy of 43% for the top 10 % of the scores. The accuracy drops to 41% for the top 20% of the scores.

B. *Cumulative percent of responses captured*

Another useful chart compares the cumulative percent of responses captured as each decile is added to the target. In this project, the top 20% deciles captured about 30% of the HIV positive individuals, compared to a random baseline where two deciles captured 20% of the HIV positive individuals and thus emphasizing the advantage of targeting.

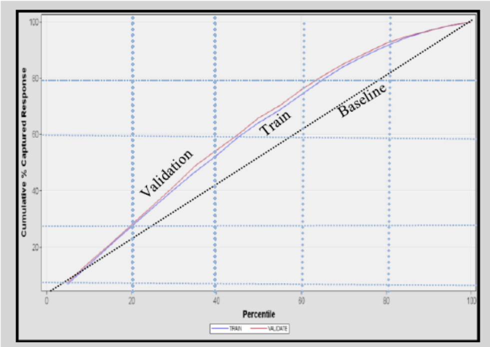

Fig. 8: Cumulative percent captured response

C. *Receiver-of-characteristics (ROC) curves*

The ROC chart is a graphical plot which shows the performance of the scorecard binary classification model as its discrimination threshold is changed. It is therefore generated by plotting the fraction of true HIV positive individuals out of the positives (sensitivity) against the fraction of false positives out of the negatives (100-specificity), at various threshold settings.

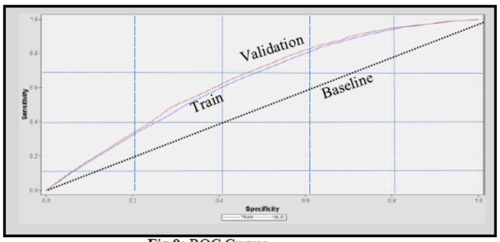

Fig.9: ROC Curves

Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. The area under the ROC curve (AUC) is a measure of how well a parameter can distinguish between two diagnostic groups, such as HIV positive and HIV negative. Therefore the ROC curve is used as a test to distinguish between HIV positive and HIV negative individuals. The closer the ROC curve is to the upper left corner, the higher the overall accuracy of the test (9).

D. *Empirical Odds plot*

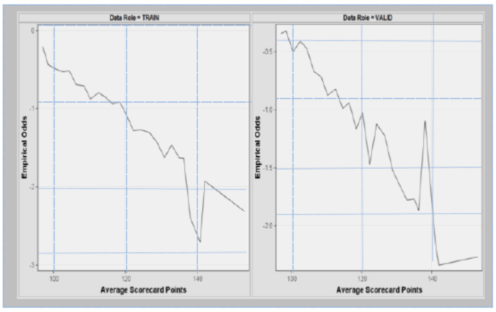

Fig.10: Empirical odds plot

This is used to evaluate the calibration of the scorecard. The chart plots the actual odds values as they are found in the validation data against scorecard. The chart is overlaid with a chart of the values that are predicted by the scorecard. The chart therefore determines those score bands where the scorecard is, or is not sufficiently accurate. The odds are calculated as the logarithm of the number of HIV positive divided by the number of HIV negative individuals for each scorecard bucket range. Thus, a steep negative slope implies that the HIV negative individuals tend to get higher scores than the HIV positive individuals. As the scorecard points increase, so does the number of HIV negative individuals in each score bucket.

E. *Kolmogorov-Smirnov plot*

The Kolmogorov-Smirnov plot shows the Kolmogorov-Smirnov statistics plotted against scorecard cut-off values. Kolmogorov-Smirnov statistic is the maximum distance between the empirical distribution functions for the HIV negative and HIV positive individuals. The difference is plotted for all cut-offs in the Kolmogorov-Smirnov plot. The weakness of reporting only the maximum difference between the curves is that it provides only a measure of vertical separation at one cut-off value, but not overall cut-off values. According to Fig. 10, the best cut-off is about 110 where the K-S score is maximum. Therefore at a cut-off value of 110, the scorecard best distinguishes between HIV negative and HIV positive individuals.

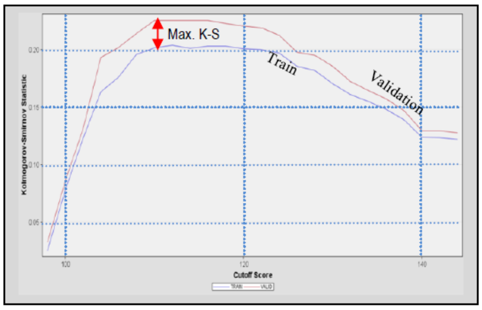

Fig.11: Kolmogorov-Smirnov plot

VI. DISCUSSION OF RESULTS AND CONCLUSION

The information values based on weights-of-evidence (WOE) ranked the contribution of the different demographic characteristics towards predisposing an individual to an HIV infection as follow; age of woman > age of partner > gravidity > parity. The pregnant woman's educational level and syphilis status had information values less than 0.05 and were thus rejected from inclusion in the scorecard. These results were in agreement with those found through the use of a two-level fractional factorial design, response surface methodology and the binary logistic regression model that clearly indicated that the pregnant woman's age had the most significant effect on the risk of acquiring an HIV infection.

The plot of scorecard points clearly shows that the risk of HIV increases with the age of the pregnant woman and her male sexual partner. The risk group at most risk is indicated by the lowest scorecard points corresponding to 28 to 31 years old for the females and 33 to 36 years old for their male sexual partners. Females and males under the ages of 18 and 21 years respectively score the greatest points, and the possible reason being the delay in sexual debut. Equally individuals over the age of 35 years old seem to be less at risk as indicated by the gradual increase in scorecard points. The latter trend could be attributed to the fact that this age group is consisted of mature individuals who are in steady sexual relationships, do not practice unsafe sex and have acquired sufficient knowledge to safe-guard themselves from possible HIV infection.

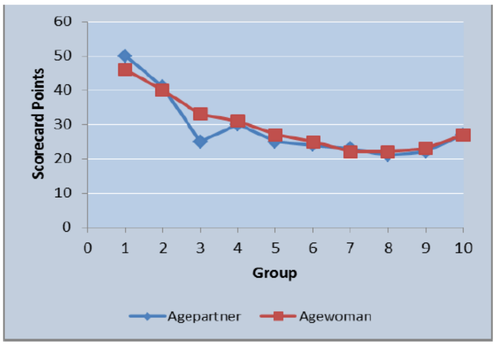

Fig.11: Plot of scorecard points

In conclusion, it is suffice to mention that the use of weights-of-evidence and risk factors provides an excellent starting point for screening individuals for different insurance policies, pension schemes and possible loans from financial institutions. In addition, this information can be of vital importance for planning intervention strategies to curb the spread of HIV within communities.

VII. ACKNOWLEDGEMENTS

Wilbert Sibanda acknowledges funding from South African Centre for Epidemiological Modeling (SACEMA), Medical Research Council (MRC) and North-West University. Special thanks to Cathrine Tlaleng Sibanda and the National Department of Health (South Africa) for the antenatal seroprevalence data.

VIII. REFERENCES

1. http://www.adena.org/files/resources/AHS%20Heart.Assessment_vE%20MM.pdf
2. http://www.idf.org/global-diabetes-scorecard
3. http://www.stanford.edu/group/hopes/cgi-bin/wordpress/2012/03/huntingtons-disease-scorecard/
4. IRIN News Service. Southern Africa: Life insurance for HIV-positive people, at a price. IRIN news service (the humanitarian news and analysis of the UN Office for the Co-ordination of Affairs).
5. Johnson L.F. (2012). Access to antiretroviral treatment in South Africa, 2004-2011. South African Journal of HIV Medicine, vol. 13, No. 1.
6. Mayosi B.M. (2012). Health in South Africa: Changes and Challenges since 2009. The Lancet, Vol. 380, Issue 9858, Pages 2029-2043.
7. Naeem Siddiqi. Credit risk scorecards: Developing and implementing intelligent credit scoring. SAS Institute Inc., John Wiley. 2006.
8. Scott C. Ratzan. A Scorecard for Chronic Disease: Advancing Health Literacy Globally Among Key Audiences. Johnson and Johnson.

9. Sibanda, W., Pretorius, P., & Grobler, A. (2012, August). Response surface modeling and optimization to elucidate the differential effects of demographic characteristics on HIV prevalence in South Africa. In Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on (pp. 818-826). IEEE.

10. Sibanda, Wilbert, and Philip Pretorius. "Application of Two-level Fractional Factorial Design to Determine and Optimize the Effect of Demographic Characteristics on HIV Prevalence using the 2006 South African Annual Antenatal HIV and Syphilis Seroprevalence data." International Journal of Computer Applications (0975-8887) Volume 35 (2011).

11. Sibanda, Wilbert, and Philip Pretorius. "Comparative Study of the Application of Box Behnken Design (BBD) and Binary Logistic Regression (BLR) to Study the Effect of Demographic Characteristics on HIV Risk in South Africa." Journal of Applied Medical Sciences 1, no. 2 (2012): 15-40.

12. Sibanda, Wilbert, and Philip Pretorius. "Novel Application of Multi-Layer Perceptrons (MLP) Neural Networks to Model HIV in South Africa using Seroprevalence Data from Antenatal Clinicsl." International Journal of Computer Applications (0975–8887) 35, no. 5 (2011).

13. Zweig, Mark H., and Gregory Campbell. "Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine." Clinical chemistry 39, no. 4 (1993): 561-577.