Weight of Evidence Model: Application in Tracking Changes in HIV Risk Profile Using 10 Year Annual Antenatal HIV Seroprevalence Data

Wilbert Sibanda and Philip Pretorius

***Abstract*—A weight-of-evidence model based on antenatal HIV seroprevalence data is explored to study the effect of demographic characteristics on the risk of acquiring an HIV infection amongst pregnant women in South Africa. Antenatal data obtained from each pregnant woman contains a wealth of information in the form of demographic characteristics. In this research we use weights-of-evidence models (WOE) and information values (IV) as measures of the risk of acquiring an HIV infection to monitor changes in HIV risk over a period of 10 years from 2001 to 2010. The study demonstrated that the risk of acquiring an HIV infection amongst pregnant women in South Africa was higher for the younger women below the age of 28 during the early years of 2001 to 2005. However, during the subsequent years of 2006 to 2010, the risk dropped amongst the younger women with the simultaneous increase amongst the older women over the age of 28. Married women were found to be least at risk of acquiring an HIV infection, while widowed women were observed to be most at risk.**

***Index Terms*—Weight-of-evidence, information value, HIV, antenatal data.**

I. Introduction

The fundamental motive of research in epidemiological sciences is to gather enough data to provide a basis for both short and long term sound decisions. In South Africa, the annual antenatal HIV survey is the only existing national surveillance for determining HIV prevalence and is therefore a vitally important tool to track the geographic and temporal trends of the epidemic [1]. The National Department of Health in South Africa has been conducting antenatal surveys since 1990 [2].

Antenatal clinic data contains the following demographic characteristics for each clinic attendee; pregnant woman's age, marital status, race, level of education, gravidity, parity, name of clinic, HIV and syphilis results.

This study aims to develop weights-of-evidence (WOE) and information values (IV) for each demographic characteristic for each year from 2001 to 2010, in order to observe changes in these parameters as measures of the risk of being infected with HIV. This research will hopefully shade light on the dynamics of HIV risk over time. Our previous research using different statistical modeling techniques, such as factorial design , multilayer perceptrons [3] and response surface methodologies [4] found the age of the pregnant woman to be highly linked to the risk of acquiring an HIV infection amongst antenatal clinic attendees. The above results were confirmed by our recent research that involved the development and validation of an HIV risk scorecard model [5], based on 2007 South African annual antenatal HIV seroprevalence data. The entire dataset for this research is shown in Table I.

Manuscript received February 15, 2014; revised April 15, 2014. This work was supported by the National Research Foundation of South Africa (Grant no. 86946). Any opinion, findings and conclusions or recommendations expressed in this material are those of the author (s) and therefore the NRF does not accept any liability in regard thereto.

The authors are with the School of Information Technology, North-West University, Vaal Triangle campus, Hendrik van Eck Boulevard, Vanderbijlpark, South Africa 1900 (e-mail: Wilbert.sibanda@nwu.ac.za, Philip.pretorius@nwu.ac.za).

II. Data-Sets

TABLE I: Research Data Set

Year	HIV rate (%)	Number of subjects
2001	24.8	16734
2002	26.5	16586
2003	27.9	16637
2004	29.5	16057
2005	30.2	16498
2006	29.1	32990
2007	29.4	33585
2008	29.3	33670
2009	29.4	32861
2010	30.2	32225

Table I shows the total number of pregnant women that were included in the antenatal HIV seroprevalence survey from the year 2000 to 2010. The table further shows the corresponding HIV prevalence rates for each year. From 2001 to 2005, the National Department of Health of South Africa included on average 16 000 pregnant women in the study. The sample size was increased to over 32 000 from 2006 to 2010. A simple inspection of Table I indicates that the HIV prevalence rates increased gradually from 2001 to 2010.

III. Theory of Weight-of-Evidence (WOE)

Weight-of-evidence (WOE) is a quantitative method for integrating data to back a hypothesis. The method was originally advanced to assist in determining the probability of an individual being diagnosed with a disease on the basis of observable symptoms or the absence of symptoms [6]. The

size of the weights was dependent on the relationship between symptom and the pattern of disease in a large group of patients. The weights were then applied to estimate the likely probability that a new individual would succumb to the disease, on the basis of the presence or absence of symptoms. The weight-of-evidence model is similar to the traditional multiple regression techniques due to the fact that the approach also entails the estimation of response variable and a number of predictor variables.

The weight-of-evidence method is a Bayesian approach presented in a log-linear form, using the prior probability of occurrence of an event like HIV status of a pregnant woman.

IV. Development of Weights-of-Evidence (WOE)

To calculate the contribution rate of each demographic characteristic on the HIV risk, weights-of-evidence (WOE) were developed as shown in equations 1 and 2.

$$WOE_{(HIV+ve)} = \left(\frac{nM_1}{nM_1 + nM_2}\right) / \left(\frac{nM_3}{nM_3 + nM_4}\right) \quad (1)$$

$$WOE_{(HIV-ve)} = \left(\frac{nM_2}{nM_1 + nM_2}\right) / \left(\frac{nM_4}{nM_3 + nM_4}\right) \quad (2)$$

In the above equations, $WOE_{(HIV+ve)}$ stands for the weight-of-evidence that HIV positive individuals exist, while $WOE_{(HIV-ve)}$ indicates the weight-of-evidence that HIV positive individuals do not exist. *nM* refers to the number of HIV positive or negative individuals within a given level of the demographic characteristic. Therefore the derivation of WOE is based on conditional probability.

V. Information Value Theory

The information value measures the overall predictive power of a demographic characteristic and thus can be used to compare the predictive strengths of different demographic characteristics. It is therefore a vital tool for variable selection during model construction.

$$IV = \sum_i \left(DistHIV_{-ve} - DistHIV_{+ve}\right) \times \ln\left(\frac{DistrHIV_{-ve}}{DistrHIV_{+ve}}\right) \quad (3)$$

TABLE II: Guidelines for Selection of IV [7]

Information value	Predictive power
≤ 0.02	Useless for prediction
0.02-0.1	Weak predictor
0.1-0.3	Medium predictor
0.3-0.5	Strong predictor
>0.5	Suspicious or too good to be true

The WOE values are used to determine the IV values and the demographic characteristics with low IV less than 0.05 should be discarded as they are not predictive, as shown in Table II.

VI. Results

A. Mother's Age

1) Changes in HIV rate across age-groups of women

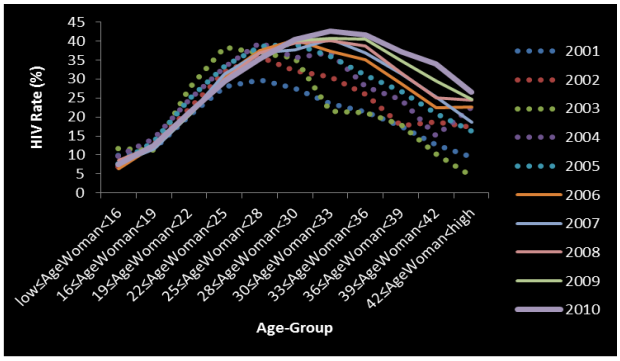

Fig. 1. HIV rate (%) as a function of women's ages.

During the years 2001 to 2005, the younger women aged between 16 and 25 years were found to exhibit higher HIV infection rates compared to the same age groups during the latter years of 2006 to 2010, as shown in Fig. 1. However, the infection rates change after the age of 28, where the latter years of 2006 to 2010 were seen to have higher HIV infection rates compared to earlier years of 2001 to 2005.

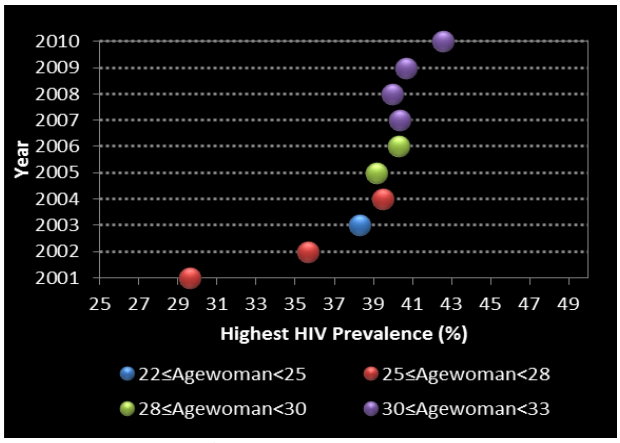

Fig. 2. Plot of highest HIV rate (%) against years.

Fig. 2 confirms the results obtained in Fig. 1 that the highest HIV infection rates are observed in the older woman during the latter years 2006 to 2010, compared to the earlier years of 2001 to 2005. This phenomenon could be attributed to *generational effect* of HIV infection, where the same pregnant women that were HIV infected during their younger years are observed to be HIV positive in later years as they visit antenatal clinics for their subsequent pregnancies. This could also mean that the younger women are becoming more educated about HIV infection and are thus taking necessary steps to avoid possible infection. The above trend could also be signaling that the epidemic is gradually waning down.

The plot of weights-of evidence (Fig. 3), confirms that during the earlier years of 2001 and 2002, higher risk of HIV was observed amongst pregnant women compared to subsequent years of 2009-2010. In general, the more negative

the WOE values, the higher the risk of HIV infection. Fig. 4 also shows that the older women between the ages of 30 and 33 years exhibit higher risk of HIV infection during the latter years of 2005 to 2010. This once more confirms the *generational effect* of HIV infection.

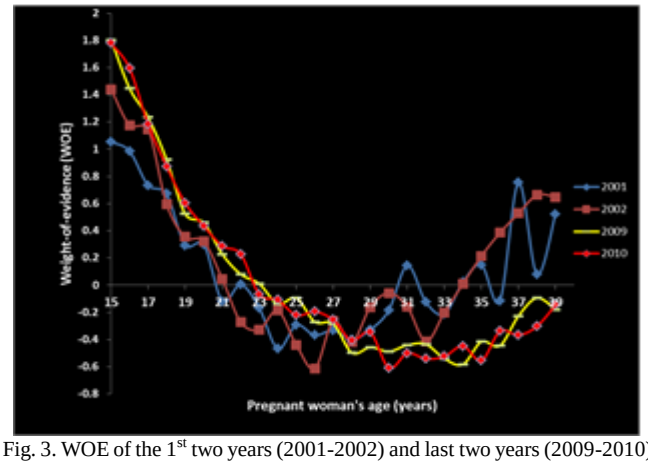

Fig. 3. WOE of the 1st two years (2001-2002) and last two years (2009-2010) as a function of pregnant women's age-groups.

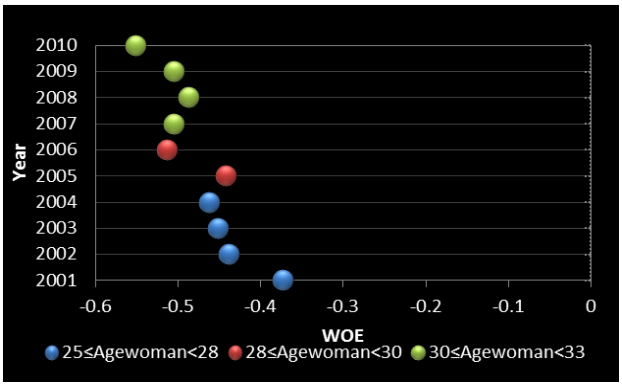

Fig. 4. Plot of highest WOE values for each year.

In order to fully understand the changes in WOE amongst pregnant women, a further investigation was conducted to assess whether the WOE changed within the same age-group of pregnant women over the ten years in the study. The results (Fig. 5) indicated that 15-year-old women were consistently less at risk of HIV infection compared to the 20-year-old women, who in turn exhibited lower risk levels compared to 25-year-old women. The negative WOE values for the 25-year-old women strongly indicated an elevated level of HIV risk. This once again confirmed that the risk of HIV infection increased with an increase in the age of the pregnant woman.

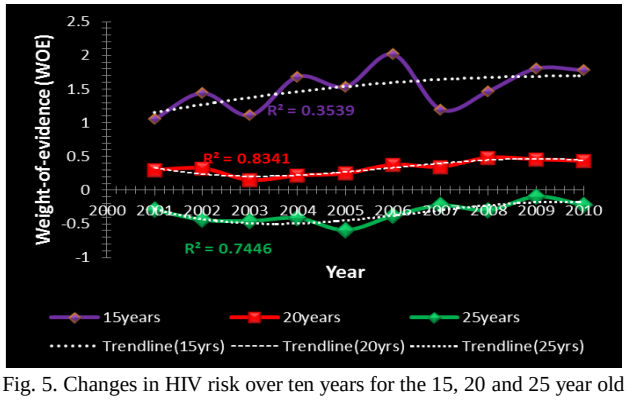

Fig. 5. Changes in HIV risk over ten years for the 15, 20 and 25 year old antenatal clinic attendees.

B. *Partner's Age*

1) *Changes in HIV rate across male age-groups*

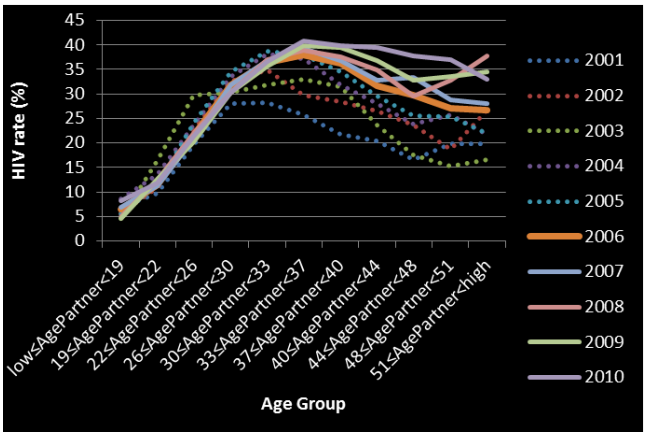

Fig. 6. HIV rate (%) as a function of male partner's ages.

In general, during the earlier years of 2001 to 2005, the HIV infection rate was higher for men below the age of 26 compared to the same age-group during the subsequent years of 2006 to 2010. However, after the age of 33, the latter years of 2006 to 2010 exhibit higher levels of HIV prevalence rates as shown in Fig. 6.

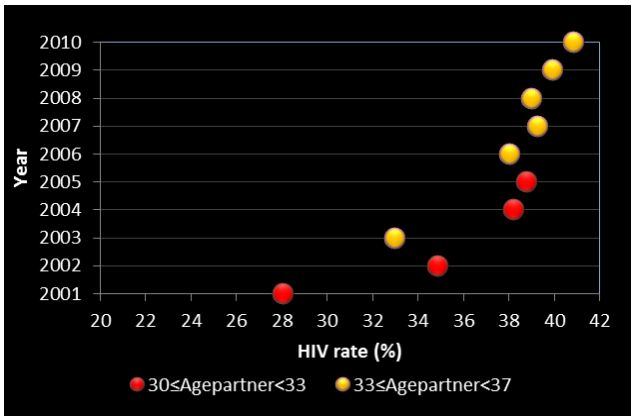

Fig. 7. Plot of highest HIV rate (%) against years.

The plot of the highest HIV prevalence rates (Fig. 7) shows that between the years 2006 and 2010, the highest HIV rates are observed in the older male age-groups between the ages of 33 and 37 years, compared to the earlier years of 2001 to 2005 where the highest HIV prevalence rates are observed between the ages of 30 to 33, further illustrating the possible *generational effect* of HIV infection.

2) *Changes in WOE across male age-groups*

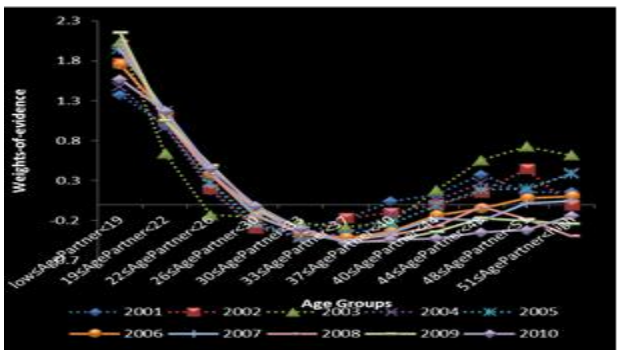

Fig. 8. Plot of WOE as a function of male partner's age-groups.

The weights-of-evidence for the male sexual partners demonstrated that the risk of HIV infection was higher amongst males below the age of 33 years during the study period 2001 to 2005, as shown in Fig. 9. Thereafter, the HIV risk gradually decreased within this age-group. However, during the subsequent period 2006 to 2010, males older than 33 years demonstrated higher levels of HIV infection.

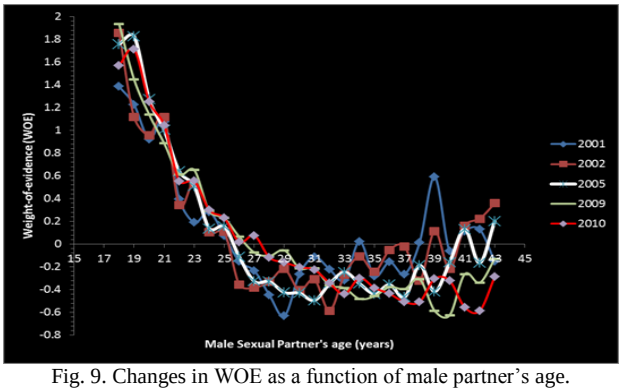

Fig. 9. Changes in WOE as a function of male partner's age.

Fig. 10 is used to illustrate that during the earlier years of 2001 and 2005, younger males (between the ages of 30 and 33) were more at risk of acquiring an HIV infection, while during the latter years of 2009 to 2010, older males (between the ages of 33 and 37) were observed to be more at risk of HIV infection. This situation also confirmed the *generational effect* of HIV infection within the male sexual partners.

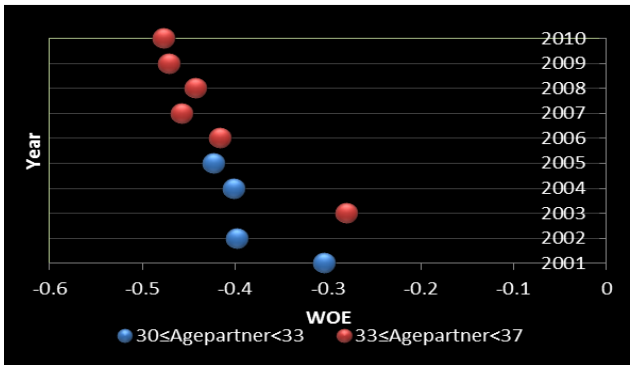

Fig. 10. Plot of highest WOE values for each year.

C. Education

1) Changes in HIV rate across educational levels

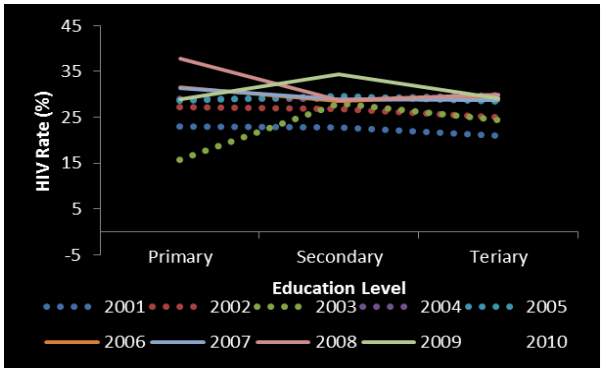

Fig. 11. HIV rate (%) as a function of educational level.

For primary education, the earlier years of 2001 to 2005 exhibited lower HIV infection rates for pregnant women between the ages of 15 and 30 years, compared to the latter years of 2006 to 2010 that showed significantly higher HIV prevalence rates ranging between 30 and 38%. The trend however, indicated that individuals with tertiary education had a decreased HIV infection rate compared to their counterparts with a primary education.

2) Changes in WOE values across educational levels

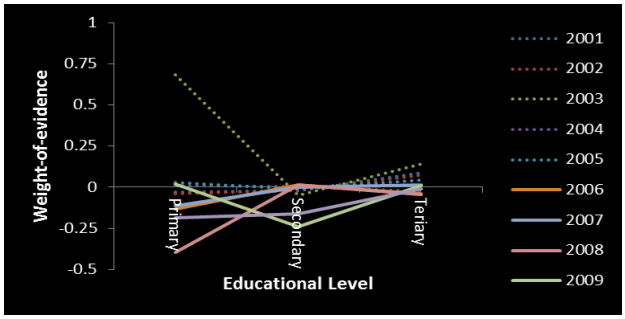

Fig. 12. Plot of highest WOE values for each year.

The WOE (Fig. 12) plot of individuals with primary education showed that between the years 2006 and 2010, there was a higher risk of HIV infection compared to the individuals with the same level of education between the years 2001 and 2005. However, that risk was considerably reduced for individuals with tertiary education.

D. Gravidity

1) Changes in HIV rate across gravidity levels

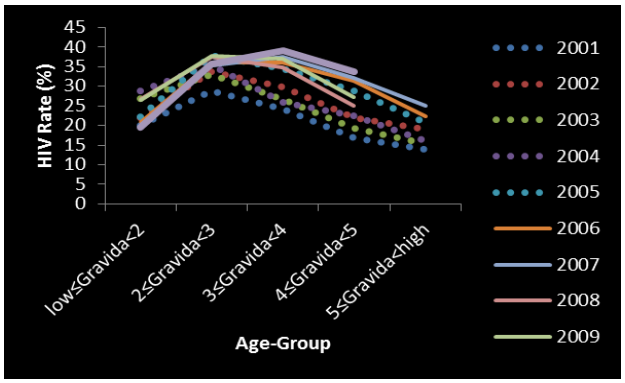

Fig. 13. HIV rate (%) as a function of gravidity.

The HIV rate plot (Fig. 13) for the period 2001 and 2010 indicates that the highest infection rates were observed amongst pregnant women presenting to antenatal clinics for their second pregnancy. The peak of HIV infection during second pregnancy was observed during the year 2009 at 39.75%. However, for all years under the study, the HIV infection rates were observed to decrease after the second pregnancy. Pregnant women presenting to antenatal clinics for their third and subsequent pregnancy, exhibited lower HIV infection rates during the period 2001 to 2005 compared to their counterparts during the period 2006 to 2010. The latter situation could be attributed to the fact that individuals with more than two pregnancies were likely to be older women and thus confirmed the *generational effect* that was observed with the pregnant women's ages.

The WOE plot (Fig. 14) confirmed that during the earlier years of 2001 to 2005, the risk of acquiring an HIV infection was observed to be higher for pregnant women presenting to antenatal clinics for their first or second pregnancy. However,

during the period 2006 to 2010, the higher risk of HIV infection was observed amongst pregnant women presenting to antenatal clinics for their third and subsequent pregnancies. This means that women with more than two pregnancies were most at risk of HIV infection.

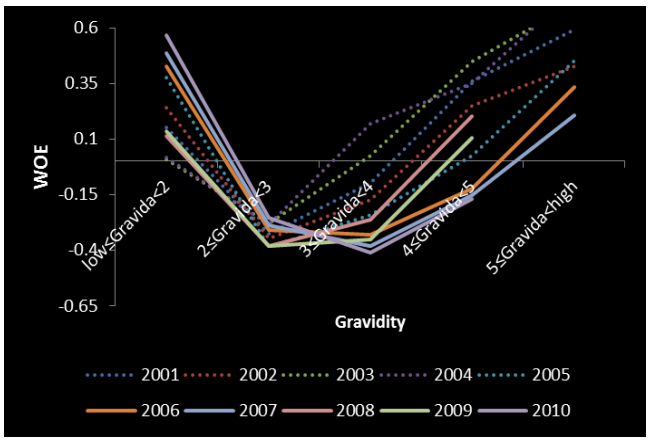

Fig. 14. Plot of highest WOE values for each year.

E. Syphilis

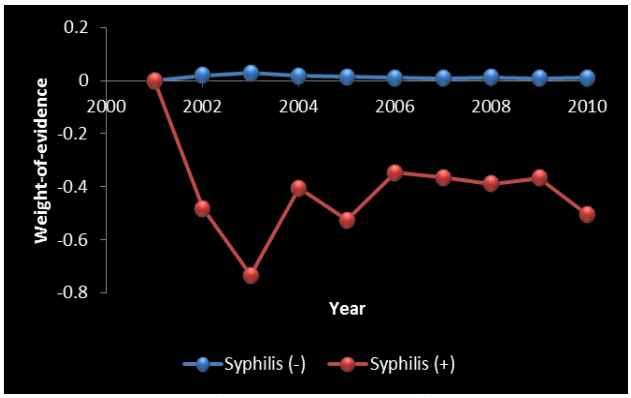

Fig. 15. Plot of Syphilis WOE values for each year.

The WOE plot (Fig. 15) showed that consistently for the years 2001 to 2010, women who were infected with syphilis exhibited an elevated risk of HIV infection compared to their counterparts without syphilis infection. This confirms the already known fact in literature that any sexually transmitted infection such as syphilis tends to enhance an individual's risk of acquiring an HIV infection.

F. Marital Status

1) Changes in HIV rate with marital status

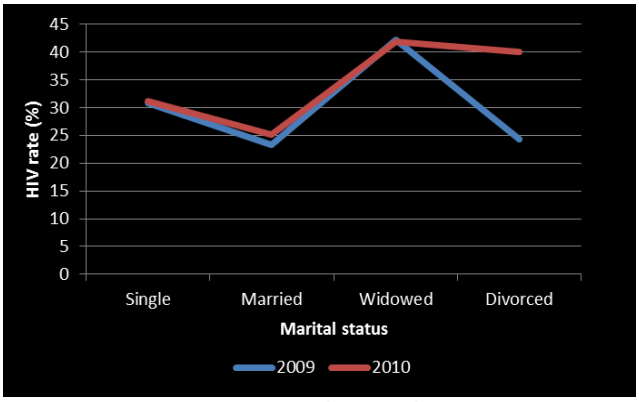

Fig. 16. HIV rate (%) as a function of marital status.

This study also demonstrated that widowed pregnant women had the highest risk of acquiring an HIV infection (Fig. 16). This could mean that these women were widowed as a result of their partners dying of HIV/AIDS. The lowest infection rates were observed amongst married women. These results were confirmed by the WOE plots, as shown in Fig. 17.

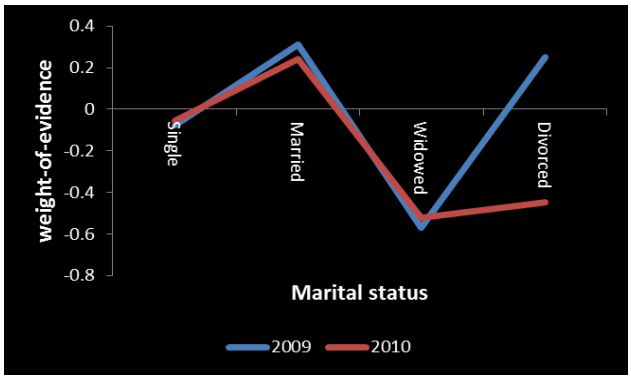

Fig. 17. Plot of marital status WOE values.

G. Comparison of HIV Risk Predictive Strengths of Demographic Characteristics Using Information Values

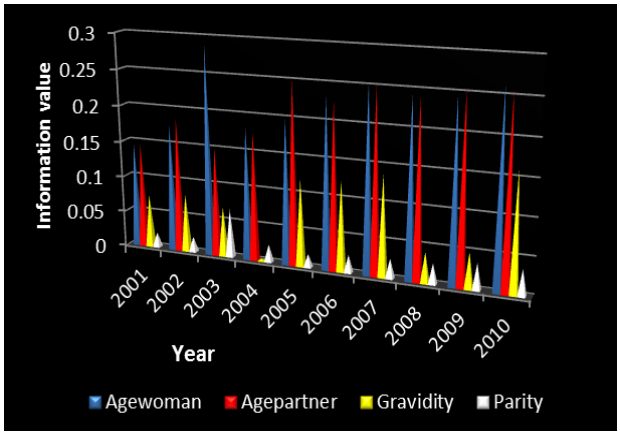

Fig. 18. Information values of demographic characteristics.

The plot of information values (Fig. 18) over the 10 year period from 2001 to 2010 shows that the ages of the pregnant women and their male sexual partners had medium predictive strength with regards to predicting the risk of acquiring an HIV infection amongst pregnant women attending antenatal clinics in South Africa. Based on Siddiqi information value classification method, gravidity and parity had unpredictive to weakly predictive strengths.

VII. Conclusion

The WOE model demonstrated that the risk of acquiring an HIV infection amongst pregnant women in South Africa was higher for the younger women below the age of 28 during the early years of 2001 to 2005. However, during the subsequent years of 2006 to 2010, the risk dropped amongst the younger women with the simultaneous increase amongst the older women over the age of 28. A similar observation was made for the age of the male sexual partners, though the HIV risk was lower for males younger than 33 years for the period 2001 to 2005. Thereafter, the risk decreased in the latter age-groups

followed by an increase in HIV risk amongst males older than 33, during the period 2006 to 2010. This phenomenon was termed *generational effect* of HIV infection and was attributed to a possible moving out of the epidemic within the reproductive population. Pregnant women with a primary (elementary) education were observed to be most at risk of HIV infection compared to their counterparts with a tertiary education. Pregnant women presenting to antenatal clinics with their second pregnancy were found to be most at risk of HIV infection. Syphilis infection was found to enhance the risk of acquiring an HIV infection. Married women were found to be least at risk of acquiring an HIV infection, while widowed women were observed to be most at risk. The researchers postulated that the women were possibly widowed as a result of the partners dying of HIV/AIDS.

The research findings are of fundamental importance to policy makers and health workers in planning sound intervention strategies to curb the spread of HIV within vulnerable populations in South Africa.

ACKNOWLEDGMENT

Wilbert Sibanda acknowledges funding from the South African Centre for Epidemiological Modeling (SACEMA), Medical Research Council (MRC) and North-West University. This work is based in part on the research supported by the National Research Foundation of South Africa (Grant no. 86946). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author (s) and therefore the NRF does not accept any liability in regard thereto.

REFERENCES

[1] W. Sibanda and P. Pretorius, "Application of two-level 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*, vol. 35, no. 12, pp. 15-20, 2011.

[2] Department of Health, "National antenatal sentinel HIV and syphilis prevalence survey in South Africa," 2010.

[3] W. Sibanda and P. Pretorius, "Novel application of multilayer perceptrons (MLP) neural networks to model HIV in South Africa using seroprevalence data from antenatal clinics," *International Journal of Computer Applications*, vol. 35, no. 5, pp. 26-31, 2011.

[4] W. Sibanda and P. Pretorius, "Response surface modeling and optimization to elucidate the differential effects of demographic characteristics on HIV prevalence in South Africa," in *Proc. 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, Instabul, 2012, pp. 818-826.

[5] W. Sibanda and P. Pretorius, "Development and validation of an HIV risk scorecard model," in *Proc. 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, Niagara, 2013, pp. 916-922, 2013.

[6] H. Wang, G. Cai, and Q. Cheng, "Data integration using weights of evidence model: applications in mapping mineral resource potentials," in *Proc. Symposium on Geospatial Theory, Processing and Applications*, Ottawa, 2002.

[7] N. Siddiqi, "Credit risk scorecard: Developing and implementing intelligent credit scoring," SAS Institute Inc., John Wiley, 2006.

Wilbert Sibanda became an associate member of the Pharmaceutical Society of South Africa. The author holds the B.Sc. degree in human physiology, in University of the Witwatersrand, Johannesburg, B.Sc. degree in med. hons (pharmacology) from University of Capetown), M.Sc. degree in med. (pharmacy) in University of the Witwatersrand, Johannesburg, and Ph.D. degree in statistical modeling of antenatal HIV data from North-West University, South Africa). The author is involved in statistical modeling of HIV antenatal data in South Africa.

He is a subject specialist (researcher) at the School of Information Technology, North-West University, Vaal Triangle campus, Hendrik van Eck Boulevard, Vanderbijlpark, South Africa. He has published extensively in the field of statistical modeling of HIV antenatal data

Dr. Sibanda chaired a special session at the International Computers and Industrial Engineering (CIE 42) Conference in Capetown in 2012 and Machine Learning and Databases Session at the 2013 IEEE International Conference on Bioinformatics and Biomedicine, at Tongji University, Shanghai, China. Dr. Sibanda is a reviewer of the International Journal of Computational Biology, Informatics and Control (IJCBIC). Dr. Sibanda sits on the Editorial Board of the Journal of Proteomics, Genomics and Bioinformatics Studies.