Rwanda Journal ISSN 2305-2678 (Print); ISSN 2305-5944 (Online)

DOI : http://dx.doi.org/10.4314/rj.v1i1.2F

A Complex Survey Data Analysis of Tb Mortality in South Africa

J. L. Murorunkwere, Mwambi H.

University of Kwazulu-Natal, School of Statistics and Actuarial Science, Pietermaritzburg, South Africa
Correspondent author: Mulaisioo87@yahoo.fr

Abstract

Many countries in the world record annual summary statistics such as economic indicators like Gross Domestic Product (GDP) and vital statistics for example the number of births and deaths. In this paper we focus on mortality data from various causes including Tuberculosis (TB). TB is an infectious disease caused by bacteria called Mycobacterium tuberculosis. It is the main cause of death in the world among all infectious diseases (Herchline and Amorosa, 2010). An additional complexity is that HIV/AIDS acts as a catalyst to the occurrence of TB. People infected with mycobacterium tuberculosis alone have an approximately 10% life time risk of developing active TB, compared to 60% or more in persons co-infected with HIV and mycobacterium tuberculosis (Vaidynathan and Singh, 2003). In 2006, South Africa was ranked seventh highest by the World Health Organization (WHO, 2009) among the 22 TB high burden countries in the world and fourth highest in Africa.

The research work in this presentation uses the 2007 Statistics South Africa (STATSSA) data on TB as the primary cause of death to build statistical models that can be used to investigate factors associated with death due to TB. Logistic regression and generalized linear models (GLM) will be used to assess the effect of some risk factors or predictors to the probability of deaths associated with TB. This study will be guided by a theoretical approach to understanding factors associated with TB death. Of the 615312 deceased, (89%) died from natural death, (2%) were stillborn and (9%) from non-natural death possibly accidents, murder, suicide. Among those who died from natural death and disease, (12%) died of TB.

Keywords: *TB mortality, prevalence, HIV incidence, TB/HIV co-infection, survey data analysis, logistic regression model, and STATSSA South Africa.*

1. Introduction

Many countries in the world record annual summary statistics such as economic indicators (example Gross Domestic Product: GDP) and vital statistics (example number of births and deaths). In particular, Statistics South Africa (STATSSA) collects annual data on nationwide number of deaths and associated causes. Tuberculosis (tubercle bacillus- TB) is an infectious disease caused by bacteria called Mycobacterium tuberculosis. These bacteria attack mainly the lungs (pulmonary TB), but also at lower extent other parts of the body such as the central nervous system, circulatory system, and the skeletal system (Khaled, 2008).

TB is the main cause of death in the world among all infectious diseases (Herchline and Amorosa, 2010). TB is classified as latent when it is not yet causing illness or active when illness has already been developed. HIV/AIDS acts as catalyst to the occurrence of TB; hence it can dramatically increase the proportion of active TB cases. A study done in India by Vaidyanathan and Singh (2003) revealed that people infected with mycobacterium tuberculosis alone have an approximately 10% life time risk of developing active TB, compared to 60% or more in persons co-infected with HIV and mycobacterium tuberculosis.

The research work in this study uses the 2007 Statistics South Africa (STATSSA) data on TB as the primary cause of death to build statistical models that can be used to investigate factors associated with death due to TB.

According to Singer (1997) TB problem in South Africa TB tends to affect the poorer populations, who have historically suffered a low standard of health care. In 2006, South Africa was ranked seventh highest by the WHO among the 22 TB high burden countries in the world; and fourth highest in Africa; with an estimated incidence of all TB cases of 940 per 100,000 in the population (WHO, 2008).

According to the report published by Williams and Dye in 2003, HIV/AIDS has dramatically increased the incidence of TB in Sub-Saharan Africa where up to 60% of TB patients are co-infected with HIV and each year 200,000 TB deaths are attributed to HIV co-infection.

However, as published by World Health Organization (WHO), in their report on *Global Tuberculosis Control: Surveillance, Planning, Financing,* in 2003, 30% of people in Sub-Saharan Africa are latently infected with *Mycobacterium Tuberculosis* and the rapid spread of HIV.

In South Africa, more than 16% of the populations are infected with HIV, and 1000 people die from AIDS-related diseases each day, and two-thirds of those with HIV also suffer from TB, because of their weakened immune systems (AMREF, 2008).

2. Theoretical Framework

This study will be guided by a theoretical approach to understanding factors associated with TB death. Such factors can be defined into the following figure:

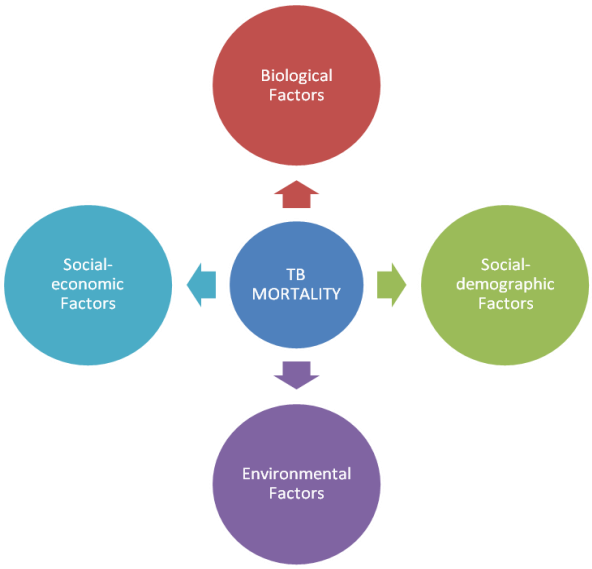

Figure 1: Factors associated with TB mortality

3. Methodology

The data used in this study is registration and records survey data on deaths from various causes gathered by Statistics South Africa in 2007. Our special interest is on deaths due to TB and HIV.

Exploratory analysis is performed using graphical displays and some basic summary statistics in the form of tables. Logistic regression and generalized linear models (GLM) will be used to assess the effect of some risk factors or predictors to the number of deaths associated with TB. Statistical modeling and analysis will be done using STATA software.

4. Objectives

The study aims to understand factors that can be used to explain TB mortality in South Africa. The work will be concerned with statistical methods that can be best used to model these associations through the following specific objectives:

i) To evaluate the proportion of deaths associated with TB/HIV in South Africa
ii) To review regression modeling for relating a binary namely death due to TB to a number of predictor variables including HIV co-infection.
iii) To investigate the factors associated with TB mortality in South Africa.

5. Results

The data sourced from Statistics South Africa consist of 615312 deaths from various causes in the year 2007. As a preliminary exploratory analysis, the use of tools such as cross tabulations and graphical displays will guide in understanding important relationships.

Results from such an exploratory analysis will assist in building a more formal statistical model to understand the relationship between key predictor variables and the response variable. Our interest in the current work is death due to tuberculosis (TB) and HIV. The synergy between TB and HIV has attracted a huge interest in recent times.

However, in this study, the author most importantly considered some variables such as age group, sex, marital status, and education level namely those which have potential significant effect on TB death and HIV death defined as the presence or absence of the disease.

of the 615312 deceased people, (89%) died from natural death and disease, specific conditions (2%) were stillborn and (9%) died from non-natural death (possibly accidents, murder, suicide).

Among the 546917 who died from natural death and disease, (12%) died of TB. The percentage of TB deaths among males is 11.18%, P <0.001. The percentage of deaths among 0-15 years old is 2.29%, 16.49% for 16-30 years old, 19.05% for 31-45, 12% for 46-60 years old. It shows that death due to TB appear to be in younger age groups (16-30 years, 31-45 years, 46-60 years) than older people, P < 0.001.

Table 1 indicates that death due to TB is higher for single individuals with a percentage of 12.56%. Curiously, the results indicate that non-educated individuals have lower rate of death by TB than people with some level of education. Nonetheless those with university or tertiary education have lower rate of TB death than those with primary and secondary education.

Table 1 Percentage of TB and NON TB deaths, With P-values for Chi-Square test, According to selected Demographic and Social characteristics

	HIV+	TB	N
Demographic characteristics			
Age group	P<0.001	P<0.001	
0-15	1.16	2.29	86111
16-30	3.88	16.49	85939
31-45	4.32	19.05	157694
46-60	1.94	11.9	114469
61-75	0.28	4.28	93158
76-90	0.05	1.47	66109
>90	0.48	2.41	11832
Sex	P<0.001	P<0.001	
Male	1.95	11.18	314138
Female	2.53	9.95	299933
Other	1.85	6.53	1241

Social Characteristics			
Education level		P<0.001	
None	1.24	4.43	123671
Primary Education	2.56	14.42	118851
Secondary Education	2.75	13.66	69752
University/Tech	1.38	6.89	8315
Other	2.41	10.97	294723
Marital status	P<0.001	P<0.001	
Single	2.65	12.56	309066
Civil marriage	1.28	6.93	73904
Living as married	1.61	10.22	18904
Widowed	0.63	4.08	48876
Religious law marriage	1.15	7.62	13668
Divorced	1.15	6.99	8979
Customary marriage	1.43	10.3	38101
Other	3.05	11.17	103813

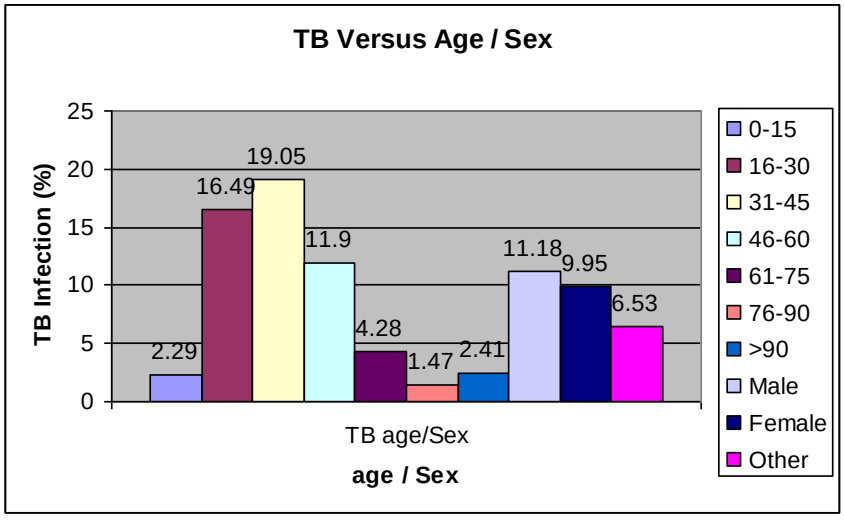

Figure 2 TB versus Age group and Sex

Logistic Regression Model (LRM)

The logistic regression model (LRM) is a special case of generalized linear models. The logistic regression model will be discussed because it will be the main application tool in analysis of the mortality data in the study.

The logistic regression model is a member of generalized linear models used to model binary data. Consider n independent observations y_i of a binary random variable Y_i taking values 1 for success and 0 for failure.

The probability of response $p = P(Y = y \mid \mathrm{x}_1, x_2, \cdots, x_p)$ is said to follows a logistic distribution if

$$p(\mathbf{x}) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)} \quad (1)$$

or in terms of the logit function as

$$\mathrm{logit}(\mathrm{p}(\mathrm{x})) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \quad (2)$$

where $\beta_0, \beta_1, \beta_2, \cdots \beta_p$ are unknown model parameters to be estimated (Agresti, 2002, p.182). The predictor variables $\mathrm{x}_1, x_2, \cdots, x_p$ can be continuous (example, age) or categorical (example, sex, marital status). The interpretation of the parameters $\beta_0, \beta_1, \beta_2, \cdots \beta_p$ are interpreted as log odds ratios with respect to the reference level of the factor variable under consideration.

Odds ratios

For the logistic regression model given by (1), many researchers prefer reporting odds ratios than the direct model parameter $\hat{\beta}_j$, $j = 1,2,\ldots, p$. In general, in the case of a binomial distribution with probability of success p, the odds of a success is defined as

$$O = \frac{\text{prob of success}}{\text{prob of failure}} = \frac{p}{1-p}.$$

For two probabilities of success p_1 and p_2, the ratio of the associated odds O_1 and O_2 is called odds ratio and is given by

$$\psi = \frac{O_1}{O_2} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$ (Agresti, 2002, p.44).

Cluster Survey Logistic Regression Model (CSLRM)

Logistic regression models used to analyze data from the complex sampling designs will be called survey logistic regression models in this study, to distinguish between them from ordinary logistic regression models discussed above. Survey logistic regression models follow the same theory as ordinary logistic regression models. The exception is that they account for the complexity of survey designs. When data are from simple random sampling, the survey logistic regression model and the ordinary logistic regression model are identical

In order to concisely define the model consider the problem of disease prevalence in epidemiology.
Let $\pi_{ijh} = p(y_{ijh} = 1)$ be the probability that the disease is present and $1-\pi_{ijh} = p(y_{ijh} = 0)$ that it is not present $(i = 1.2,\ldots, m_{hj}; j = 1,2,\ldots, n_h; h = 1,2,\ldots, \mathrm{H})$ in the i^{th} observation or individual within the j^{th} primary sampling unit (PSU) nested within the h^{th} stratum $(i = 1.2,\ldots, m_{hj}; j = 1,2,\ldots, n_h; h = 1,2,\ldots, \mathrm{H})$. In this case the log-likelihood function is given by

$$l(\beta; y) = \sum_{h=1}^{H}\sum_{j=1}^{n_h}\sum_{i=1}^{m_{hj}}\left\{ y_{ijh}\log\left(\frac{\pi_{ijh}}{1-\pi_{ijh}}\right) - \log\left(\frac{1}{1-\pi_{ijh}}\right)\right\} \qquad (19)$$

Thus in general the survey logistic regression model is given by

$$\mathrm{logit}(\pi_{ijh}) = \mathrm{X}'_{ijh}\beta, \quad i = 1,2,\ldots, m_{hj}; j = 1,2,\ldots, n_h; h = 1,2,\ldots, \mathrm{H}$$

Where X_{ijh} is the row of the design matrix corresponding to the characteristics of the i^{th} observation in the j^{th} PSU within h^{th} stratum, and β is a vector of unknown parameters of the model.

In fitting the models, TB status was the response variable. Results for both simple and cluster survey logistic regression models are presented in Table 2. Notice that the estimated coefficients are the same from both procedures, but standard errors produced by logistic regression are relatively small compared to those from the survey logistic regression.

Table 2: Logistic Regression

	Simple Logistic Regression				Cluster Survey Logistic Regression			
Variable	**OR (Std.Err)**	**95% CI**		**p-value**	**OR (Std.Err)**	**95% CI**		**p-value**
Age group								
0-15	REF							
16-30	8.406(0.206)	8.0118	8.8209	< 0.001	8.406(0.735)	6.8986	10.244	<0.001
31-45	10.02(0.237)	9.5683	10.497	< 0.001	10.022(1.035)	7.9326	12.662	<0.001
46-60	5.749(0.141)	5.4796	6.0324	< 0.001	5.749(0.632)	4.4839	7.372	<0.001
61-75	1.902(0.053)	1.8009	2.0093	< 0.001	1.902(0.231)	1.4449	2.5042	<0.001
76-90	0.64(0.025)	0.5886	0.6872	< 0.001	0.636(0.078)	0.481	0.841	0.005
>90	1.051(0.67)	0.9268	1.1917	0.439	1.051(0.179)	0.7148	1.545	0.777
Sex								
Male	REF							
Female	0.878(0.007)	0.8642	0.8928	<0.001	0.878(0.015)	0.8449	0.9132	<0.001
Other	0.5548(0.06)	0.4428	0.6952	<0.001	0.555(0.106)	0.3597	0.8559	0.013
Marital status								
Single	REF							
Civil marriage	0.519(0.008)	0.5033	0.5348	<0.001	0.519(0.063)	0.3942	0.6829	<0.001
Living as married	0.793(0.019)	0.7557	0.8322	<0.001	0.793(0.075)	0.6397	0.9830	0.037
Widowed	0.296(0.007)	0.2827	0.3099	<0.001	0.296(0.044)	0.2106	0.4161	<0.001
Religious law marriage	0.575(0.019)	0.5390	0.6126	<0.001	0.575(0.049)	0.4742	0.6963	<0.001
Divorced	0.524(0.022)	0.4825	0.5682	<0.001	0.5236(0.09)	0.3429	0.7994	0.007
Customary marriage	0.799(0.014)	0.7719	0.8274	<0.001	0.799(0.119)	0.5691	1.1224	0.17
Other	0.875(0.009)	0.8563	0.8945	<0.001	0.875(0.120)	0.6411	1.1952	0.359
Education level								
None	REF							
Primary Education	3.6345(0.058)	3.5219	3.7534	<0.001	3.6345(0.213)	3.3828	4.3511	<0.001
Secondary Education	3.4343(0.06)	3.2879	3.5345	<0.001	3.4343(0.304)	2.7906	4.3771	<0.001
University/ Tech	1.5972(0.072)	1.4611	1.7459	<0.001	1.5972(0.221)	1.1673	2.1853	0.008
Other	2.6599(0.039)	2.5827	2.7394	<0.001	2.6599(0.223)	2.1999	3.2161	<0.001

Discussion and Conclusion

The exploratory analysis carried out in this study indicates that TB incidence is higher among males than females; the reason is that males tend to work in more TB conducive environments than females. One possible working environment is that males work in mines more than females where shafts in mines are poorly ventilated and therefore facilitating very easy spread of TB bacteria. Migrant

mine workers carry the bacteria back home during holidays and spread it to their surrounding areas.

The preliminary results on TB death data indicate that TB prevalence seems to be higher among younger individuals. The reason is possibly due to the fact that younger individuals are increasingly becoming more vulnerable due to co-infections with HIV. Given TB is one of the opportunistic infection among HIV infected individuals may explain this correlation.

Exploratory analysis also suggests that people with low level of education are more TB infected. Those who live in informal settlements and working for crowded environments including households and those who work in crowded environments such as in factories where there is a lot of pollution tend to die of TB than other living and working condition.

The exploratory analysis also indicates that HIV is more prevalent among females than males. The reason is that females are exposed to sexual abuse, rape and commercial sex activities for survival which expose them to HIV infection. The prevalence in young individuals could be due to the fact that they are more sexually active and inexperienced which lead them to be at higher risk of HIV infection. The exploratory analysis shows that the level of education is important in explaining the risk of HIV infection. Individuals with lower education levels tend to be less informed about the risks of HIV. Low levels of education, poverty, overcrowding and unemployment are much associated with the less knowledge about HIV/AIDS.

TB and HIV are linked and people with TB that are infected with HIV have increased risk of dying from TB than HIV negative ones. Chart 1 and chart 2 shows that the risk of TB infection is higher among individuals infected with HIV compared to those who are HIV negative. People who are HIV positive are at higher risk of TB infection. The observed probability of dying of TB given HIV positive is 24% compared to 10% for HIV negative. On the other hand, the chance of being HIV positive is higher among individuals who died of TB than among those who did not die of TB.

Table 3 Two-way table showing the joint distribution of TB deaths by HIV deaths.

Variable	Category	TB	No TB	Total
HIV status	**HIV negative**	61734 (10)	539860 (90)	601594
	HIV positive	3318 (24)	10400 (76)	13718
	Total	65052	550260	615312

The table shows that 24% were reported to have died due to co-infection while 10% died of TB but not with HIV. The results also shows that individuals die of other causes of death (non TB) while infected with HIV (76%).

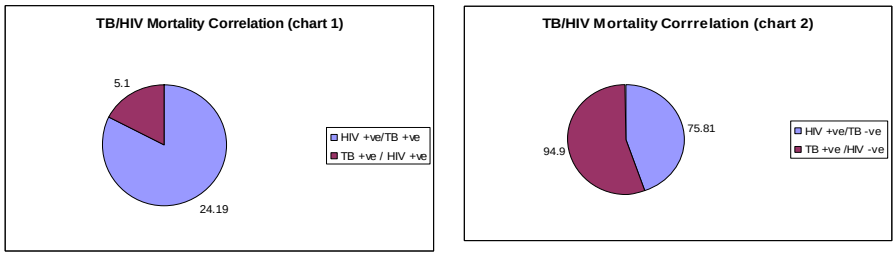

Figure 3: The joint distribution of TB deaths by HIV deaths

The above figures shows that the study of the joint dynamics of HIV and TB present formidable mathematical challenges due to the fact that the models of transmission are quite indistinct. Furthermore, HIV activates TB and an individual who dies of TB could have been co-infected with HIV. Here the risk of TB and HIV infection give 24%.

Finally, this study was able to quantify factors related to TB and co-infection with HIV and such results will help to guide decisions on how to mitigate the problem.

The major limitation of the study is the data which is very large and could not allow analysis at the level of individual members; therefore policy makers and further researchers should focus more on individual level, for example TB and HIV co-infected individuals. In addition to that, analysis at the individual level might give more insight into the disease than analysis at the general level.

Reference

1. African Medical and Research Foundation (AMREF). (2008). *TB and HIV Control in South Africa*.http://www.amref.org/whatwedo/tb and hiv control in South Africa [24 August2008].
2. coinfection in an urban area of hyperendemicity. Clin Infect Dis. 2010 May 15; 50 suppl 3:S208-14.
3. Agresti, A. (2002). *Introduction to categorical data analysis.* John Wiley.
4. Cock, D., & Chaisson, R. E. (1999). *International Journal of Tuberculosis and Lung Disease* , 3, 457.
5. Cohen, J. (2002). *Science.* Retrieved MAY 20, 2011, from Therapies:Confronting the limits of success: http://aidscience.org/science/2320.html
6. Collins, T. F. (1981). Applied epidemiology and logic in tuberculosis control. *South Africa Medical journal* , 61, 566-9.
7. Corbett, E. L., & al, e. (2003). Archives of Internal Medicine. 163, 1009.
8. Dye, C. (2006). Global epidemiology of tuberculosis. *Lancet* , 367 (3): 938-940.
9. Getahun, H., Harrington, M., O'Brien, R., & 16, P. N. (2007, Junuary). Diagnosis of smear-negative pulmonary tuberculosis in people with HIV infection or AIDS in resource-constrained settings. *informing urgent policy changes* , 2042-2049.
10. Herchline, T., & Amorosa, J. K. (2010, October 4). *Tuberculosis.* Retrieved August 2, 2010, from emedicine: http://emedicine.medscape. com/article/230802-print
11. Khaled, K. M. (2008). *Tuberculosis (TB) Progress Toward Millennium Development Goals (MDGS) and DOTS in Who Eastern Mediterranean Region (EMR), MPH Thesis.* Atlanta: Georgia State University.
12. World Health Organization. (2003). *"Global Tuberculosis Control: Surveillance, planning financing (World Health Report) " Tech. Report No. WHO/CDS/TB/2001.287.* Geneva: World Health Organization.
13. World Health Organization (2009). Global tuberculosis control: a short update to the 2009 report. Available on HYPERLINK

"http://www.who.int/tb/publications/global_report/2009/update/en/index.html"
14. World Health Organization, (2000). *Anti-tuberculosis Drug Resistance in the World, World Health Organization Report no. 2, WHO/CDS/TB/ 2000.278.* Geneva: WHO.
15. Raviglione, M. C., & Pio, A. (2002). Evolution of WHO policies for tuberculosis control,1948-2001. *Science direct* , 775-780.
16. Raviglione, M. C., Harries, A. D., Msiska, R., Wilkinson, D., & Nunn, P. (1997).
17. Tuberculosis and HIV:Current status in Africa. *AIDS Supplement 11* , S115-S123.
18. Singer, C. (1997). *TB in South Africa: The People's Plague.* Pretoria: National Department of Health.
19. Tan, D., Upshur, R. E., & Ford, N. (Apr 1 2003). *BMC International Health and Human Rights 3, 2.*
20. UNAIDS. (2003). *Report on the global HIV/AIDS epidemic Tech. Report No. UNAIDS/02.26E* . UNAIDS.
21. Uriz, J., Reparaz, J., Castiello, J., & Sola, J. (2007). Tuberculosis in patients with HIV infection. *An sist sanit Navar* , 30 Suppl 2:131-142.
22. Vaidyanathan, P. S., & Singh, S. (2003). TB-HIV co-infection in India. *NTI Bulletin* , 39, 3&4, 11-18.

Williams, B. G., & C Dye (2003). *Antiretroviral Drugs for Tuberculosis Control in the Era of HIV/AIDS.* Geneva: Sciencexpress