Tuberculosis risk factors in South Africa, 2008 to 2017: A Generalised Estimating Equations approach

Hilda Dhlakama [1,2], Siaka Lougue [2], Henry Godwell Mwambi [2]

(1) Department of Statistics, University of Johannesburg, Doornfontein Campus, Corner Siemert and Beit Street, Johannesburg 2094, South Africa.
(2) School of Mathematics, Statistics and Computer Sciences, University of KwaZulu-Natal Westville Campus, Private Bag X54001, Durban 4000, South Africa.

CORRESPONDING AUTHOR: *Hilda Dhlakama, Department of Statistics, University of Johannesburg, Doornfontein Campus, Corner Siemert and Beit Street, Johannesburg 2094, South Africa, Tel (+27115596021), E-mail: hildad@uj.ac.za*

DOI: ***10-2427/13227***
Accepted on April 21, 2020

ABSTRACT

Background: Although, death due to tuberculosis has been on the decline. In 2016, 124 000 people died of tuberculosis in South Africa and the disease was declared the leading cause of death by Statistics South Africa. Continued efforts to use research to create a nation free of tuberculosis are underway.
Methods: A repeated measures investigation was performed with the aim of identifying the persistent predictors and the long-term patterns of tuberculosis infection in South Africa for the period 2008 to 2017. The most suitable Generalised Estimating Equations that describe the population average probability of infection over time were applied to a sample of respondents taken from the National Income Dynamics Survey data, wave 1 to wave 5. The response variable was binary with the outcome of interest being the respondents that self-reported to have been diagnosed with tuberculosis. To improve estimation efficiency, the best working correlation matrix for this data was selected.
Results: We used a sample of 8510 individuals followed for five waves, of these, 3.7%, 2.54%, 4.15%, 5.72% and 5.99% for waves 1, 2, 3, 4 and 5 respectively, reported to have been diagnosed with tuberculosis. Findings revealed that the independent working correlation matrix with the model-based standard error estimates gave the most robust results for the repeated measures tuberculosis data in South Africa. Furthermore, over the years, the average probability of being diagnosed with tuberculosis was positively associated with being single, male, middle-aged (30-59 years), black African, unemployed, smoking, lower education levels, lack of regular exercise, asthma, suffering from other diseases, lack of access to improved sanitation, lower household income and expenditure.
Conclusion: The probabilities of tuberculosis infection are independent within individuals over time. The inequalities in socioeconomic status in South Africa caused the poor to be more at risk of tuberculosis over time from 2008 to 2017.

Key words: Tuberculosis, Generalised Estimating Equations, South Africa, socioeconomic status

INTRODUCTION

Tuberculosis (TB), caused by a bacteria called Mycobacterium tuberculosis that is spread through inhalation and mainly affects the lungs, can remain dormant in the latent infection stage. Since the World Health Organisation (WHO) recommended the TB control strategy known as the Directly Observed Treatment Short-

course (DOTS), there has been continued and improved efforts to implement the treatment [1]. In South Africa, the TB incidence rate has been on the decline since 2009 when it was at its peak. The WHO reports that from 2000 to 2015, 49 million lives were saved through early diagnosis and effective treatment [2]. However, an estimated 10 million people developed TB disease in 2017 globally, where about 90% were adults (15 and above) and an estimated 3% of them were from South Africa. Although TB is curable and preventable through the continued efforts by WHO to effectively diagnose and treat the disease, it is still a major public health problem in South Africa. The aim is to "bring down the global incidence from more than 1000 per million population in 2015 to less than100 per million by 2035" [1,2].

This research was aimed at identifying the persistent predictors of the long term patterns of tuberculosis infection in South Africa with the use of Generalised Estimating Equations (GEE) and was motivated by the burden of TB in South Africa. Although most Statistical disease modelling is based on clinical trials, observational studies have also been useful for users to identify new interventions to curb the TB disease. A couple of survey studies have explored the TB modelling in South Africa but very few included the repeated measures component. This research is also in line with the Stop TB partnership's "Zero TB initiative" whose purpose is to create "islands of elimination" by identifying communities at risk and recommend models of intervention which is also the National TB Control Programme National Strategic Plan (NTCP) objective of reducing TB in SA and the globally [1,3]. In this regard, [4,5] used transmission models to determine if these targets were reachable in the bid to eradicate TB in South Africa.

MATERIALS AND METHODS

The NIDS Data

The NIDS data is a nationally representative sample survey which started wave 1 in 2008 with 28,000 individuals from 7 300 households across South Africa. NIDS has been tracking their lives every two years since 2008. For Wave 5 (2017), they added about 2775 respondents due to attrition of white, Indian/Asian and high-income respondents. The NIDS data is available free on a public domain (http://www.nids.uct.ac.za/nids-data/data-access). A detailed description of this data is given elsewhere[6]. The NIDS data have repeated observations (waves) on five-time points, Wave 1 to Wave 5 data collected in 2008, 2010, 2012, 2014/2015 and 2017 respectively. These repeated measures are assumed to be correlated within individuals over time. The GEEs were applied to the repeated measures to assist study the variations among the different waves and their influence on self-reported TB infections. GEEs will account for the correlations in the repeated responses.

Generalised Estimating Equations (GEEs)

Developed by [7] for the case of correlated data, the GEE models are an extension of Generalised Linear Models (GLM). In addition to binary and count outcomes, GEE models are also applicable to continuous outcomes[7–9]. Also known as marginal models, GEE models regress the dependent variable(Y) on the explanatory variable(X) and the intra-subject dependence. The mean response thus depends on the covariates only and not on previous responses or random effects. The GEE approach's basic feature is that only the marginal distribution of a subject's dependent vector at each time point needs to be specified and there is no need to specify the joint distribution[10]. Considering the GEE approach, let $\mathbf{Y}=(Y_{it})$ response for each subject i, measured at different time points ($t=1,2,...,n_i$) denotes the outcome vector for subjects $(i=1,2,.....,N)$ and $\mathbf{X}=(X_1,X_{2,...,}X_p)$ be a $n_i \times p$ matrix of explanatory/covariate variables for subject i.

GEE is expressed in the form of a GLM but with an extension.

The linear predictor $\eta_{it}=x_{it}^T\beta$ where x_{it} is the covariate vector for subject i at time t.

The link function, $g(\mu_i)=\eta_{it}=x_i^T\beta=E(y_i)$.

For a binary outcome, the link function is logit for a binary outcome with two outcomes, 1 for success and 0 for failure: $g(\mu_i)=logit(\eta_{it})=\log\left[\frac{P(y_i=1)}{1-P(y_i=1)}\right]$ where $P(y_i=1)$ is the probability that a subject self-report to have been TB diagnosed. The parameter estimates are interpreted as the odds that a patient was diagnosed with TB.

The variance is then described as the function of the mean, $V(\mu_{it})=\phi v(\mu_{it})$ where ϕ is the scale parameter that determines the dispersion, $v(\mu_{ij})$ is known variance function.

Lastly, the working correlation structure R_i, for the repeated measures, with dimension $n_i \times n_i$. R_i is assumed to depend on a vector of association parameters α.

The most common working correlation matrices are the unstructured, independent, exchangeable and Autoregressive AR(1)[11].

The marginal covariance matrix, $V_i=\mathrm{Var}(\mathbf{Y}_i)$ involves the nuisance parameters α, is defined as[7]"abstract":"Abstract. This paper proposes an extension of generalized linear models to the analysis of longitudinal data.

$V_i(\beta,\alpha,\phi)=\phi A_i^{1/2}(\beta)R_i(\alpha)A_i^{1/2}(\beta)$, where $R_i(\alpha)$ a n x n symmetric correlation matrix is a "working" correlation structure for a single subject Y_i and $A_i=diag\{a''(\theta_{it}\}$

The score equations for a multivariate marginal model

$Y_i = N(X_i\beta, V_i)$ are given by

$$S(\beta) = \sum_i \sum_t \frac{\partial \mu_{it}}{\partial \beta} v_{it}^{-1}(y_{it} - \mu_{it}(\beta)) = \sum_{i=1}^{N} D_i[V_i(\alpha)]^{-1}(y_i - \mu_i(\beta)) = 0$$

Where:

D_i is a $n_i \times p$ matrix with $(it)^{th}$ elements $\frac{\partial \mu_{it}}{\partial \beta}$

y_i and μ_i are n_i-vectors with elements y_{it} and μ_{it}

When we have univariate GLM, then the quasi-likelihood estimating equation has the form,

$$\sum_i \left(\frac{\delta \mu_i}{\delta \beta}\right)^T = v(\mu_i)^{-1}[y_i - \mu_i(\beta)] = 0$$

The analogue of this in multivariate is the GEE given by[7]

$$\sum_i D_i^T V_i^T [y_i - \mu_i(\beta)] = 0$$

Where $\mu_i = (\mu_{i1}, \ldots\ldots, \mu_{ini})^T$ denotes the mean vector where $\mu_i = E(Y_i)$

GEE is solved by iterating between the solution for β and the method for estimating α as a function of β and this procedure is repeated until convergence [11]. The full specification of the joint distribution is not required, therefore the likelihood tests are not applicable to compare GEE models since they're estimated using quasi-likelihood procedures. Although GEEs are flexible, for average model accuracy, the covariates and the appropriate working correlation matrix need to be selected carefully.

For likelihood-based methods, model selection is done using the Akaike Information Criterion (AIC). For non-likelihood based methods like GEEs, model selection can be done by use of QIC as proposed by [12] there seem to be few model-selection criteria available in GEE. The well-known Akaike Information Criterion (AIC. The Quasi-likelihood under the Independence model Criterion (QIC) is used for selecting the appropriate working correlation matrix and is given by:

$$QIC(R_i) = -2Q(\hat{B}(R_i); \mathbf{I}, D) + 2trace(\hat{\Omega}_\mathbf{I} V)$$

Where Q is the quasi-likelihood, $\hat{\beta} = \hat{\beta}(R_i)$ is the GEE estimator obtained using any general working correlation structure R_i. D is a function of $D = D(\beta)$, V is the robust or sandwich covariance estimator and Ω_I can be estimated by its empirical estimator

$$\hat{\Omega}_I = \frac{-\partial^2 Q(\beta | I, D)}{\partial \beta \partial \beta^{'}|_{\beta=\hat{\beta}}}$$

The working correlation structure that yields the smallest QIC is the best set of GEEs that estimate β more efficiently.

Sample and variables

In this study, the focus was on risk factors of self-reported TB on adults above the age of fifteen, over the years from 2008 to 2017 in South Africa. The response outcome of interest was those that reported having been TB diagnosed. The independent variables used were as follows: marital status: single, age: 15-29, 30-44, 45-59, 60+, gender: male, race: African, education: none, primary school, secondary school, tertiary, home language, isiZulu, Afrikaans, other, employment status: employed, Perceived health: good, regular exercise: Yes, suffer from other disease: yes, diagnosed with Asthma: yes, diagnosed with diabetes: yes, heavy smoker: yes, heavy smoker: yes more than 20 cigars daily, access to improved sanitation: yes and household overcrowded: yes, 4 or more people per room.

Due to dropouts (for various reasons) over the years and the addition of respondents to the survey, not all respondents appear in all the five waves. To get a better picture of the dynamics of TB infection, we considered the individual adults that were in wave 1 and participated in all the five waves. Our sample was comprised of 8510 adult individuals that were followed for all the five waves. The GEEs were the most suitable approach for this research since the response was not a single measurement per subject but a profile of repeated measurements of the same response within the subject. We did simple logistic regression on the variables of interest for all the waves in STATA version 14, separately to determine the most influential risk factors of TB. The selected variables were then modelled using GEEs as main effects to determine the most significant effects. The analysis was done in SAS enterprise guide 6.1. We used the four common working correlation matrices for wave 5 data. These are namely: Independent, exchangeable, unstructured and Autoregressive AR(1). We opted no to use the m-dependent correlation matrix as "AR(1) structure is preferable over banded correlation structures and m-dependent correlation structures are not biologically plausible" [13] We used QIC and QICu to select the best working correlation matrix through the QIC criterion. The model selected to be the final model was fitted using the correlation matrix of choice and the model-based and empirical standard error estimates were used to compare the robustness of results.

RESULTS

The respondents aged 15 and above in wave 1, 2, 3, 4 and 5 were 17 102, 18 569, 21 272, 23 937 and 25 419 respectively. However, we used a sample of 8510 and their demographic factors are shown in Table 1 below.

The percentages of people who reported to have been diagnosed with TB were 3.7%, 2.54%, 4.15%, 5.72%

TABLE 1. Demographic factors

		WAVE 1	WAVE 2	WAVE 3	WAVE 4	WAVE 5
Code	**Variable**	**n(%)**	**n(%)**	**n(%)**	**n(%)**	**n(%)**
TB diagnosed		270 (4.5)	216(3.62)	353(5.28)	487 (5.72)	510 (5.99)
Gender						
1	Male	3168(37.23)				
2	Female	5342(62.77)				
Race						
1	African	7033(83.11)				
2	Other	1437(16.89)				
Language						
1	IsiZulu	2701(31.74)				
2	Afrikaans	1308(15.37)				
3	Other	4501(52.89)				
Marital Status						
1	Married/Living together	3060(35.96)	3091(36.32)	3091(36.32)	3138(36.87)	3267(38.39)
2	Single	5420(63.69	5401(63.47)	5416(63.64)	5367(67.07)	5238(61.55)
Age						
1	15-29	3576(42.03)	3182(37.39)	2891(33.97)	2368(27.83)	1848(21.72)
2	30-44	2252(26.46)	2289(26.93)	2305(27.09)	2472(29.05)	2644(31.07)
3	45-59	1748(20.94)	1888(22.19)	1968(23.13)	2046(24.04)	2119(24.90)
4	60+	934(10.98)	1153(13.53)	1346(15.82)	1624(19.04)	1899(22.31)
Highest education						
1	None	1076(12.64)	1100(12.93)	1111(13.06)	1054(12.39)	1058(12.43)
2	Primary	2147(25.23)	1943(22.83)	1903(22.36)	1868(21.95)	1789(21.02)
3	Secondary	4634(54.45)	4667(54.84)	4609(54.67)	4460(52.41)	4387(51.55)
4	Tertiary	645(7.58)	799(939)	886(0.41)	1124(13.21)	1241(14.58)

and 5.99% for waves 1, 2, 3, 4 and 5 respectively. In order to select the appropriate working correlation matrix, the response variable TB diagnosed yes/no was modelled on the selected variables known to be risk factors of TB.

Results of logistic regression by wave for these variables are shown in Table 2.

In wave 1, employment status, perceived health, exercise, asthma, heavy smoker, and overcrowding were significant. For wave 2, perceived health, smoking, sanitation, other diseases, and asthma were associated with TB. In wave 3, marital status, race, education, perceived health, heavy smoking, other diseases, asthma, household income, and expenditure. TB was associated with marital status, race, language, perceived health, other diseases, asthma, heavy smoking and household income in wave 4 whereas, in wave 5, TB determinants were gender, other diseases, asthma, heavy smoking, perceived health, sanitation, and household income.

The results for a comparison of the working correlation matrices are shown in Table 3:

The algorithm for all of the models converged. Comparing the QICs, the Independent working correlation matrix is the most suitable one to use for this data since it has the lowest QIC.

For the final model, we used the independent working correlation matrix with model-based and empirical-based standard errors. The comparison of results is shown in Table 4.

The parameter estimates for the two models are the same but the model-based standard errors are smaller and more robust than the empirical standard errors. The confidence intervals and the p-values follow suit.

On the model-based standard errors, waves 1 to 3 was significantly different from wave 5 as far as TB infections were concerned whereas wave 4 was not. Over time, the average probability of being diagnosed with TB for the single was more than their married counterparts, odds ratio 1.37. Individuals aged 15-29 were not significantly different from those aged 60 and above over time as far as contacting TB was concerned. Those aged

TABLE 2. Logistic regression results by wave

	Wave 1		Wave 2		Wave 3		Wave 4		Wave 5	
	O.R	P>z	O.R	P>z	O.R	P>z	O.R	P>z	O.R	P>z
Marital status	1.000	0.998	1.100	0.573	1.345	0.025*	1.294	0.021*	1.118	0.299
Age	1.053	0.542	0.960	0.656	0.965	0.609	0.989	0.858	1.043	0.481
Gender	0.819	0.211	0.972	0.874	1.054	0.711	0.975	0.830	0.715	0.004*
Race	0.901	0.626	0.821	0.381	1.406	0.045*	1.570	0.002*	1.071	0.637
Education	0.901	0.313	0.907	0.392	0.765	0.002*	0.945	0.447	0.992	0.909
Language	0.885	0.105	0.973	0.753	0.898	0.116	0.855	0.007*	1.005	0.931
Employment status	0.679	0.011*	0.922	0.654	0.907	0.465	1.040	0.731	1.224	0.074
Perceived health	2.555	0.000*	3.789	0.000*	1.992	0.000*	1.986	0.000*	1.491	0.001*
Exercise	1.464	0.042*	1.427	0.083	0.992	0.954	1.132	0.324	1.287	0.046*
Other Disease	0.705	0.072	0.474	0.006*	0.249	0.000*	0.415	0.000*	0.335	0.000*
Asthma	0.483	0.003*	0.454	0.006*	0.548	0.007*	0.623	0.016*	0.462	0.000*
Diabetes	1.035	0.920	0.946	0.869	0.924	0.740	1.173	0.444	1.484	0.074
Heavy smoker	1.942	0.000*	2.956	0.000*	1.957	0.000*	1.933	0.000*	1.825	0.000*
Sanitation	1.036	0.826	1.560	0.010*	1.174	0.251	0.921	0.490	1.297	0.020*
Household income	1.000	0.120	1.000	0.883	1.000	0.029*	1.000	0.002*	1.000	0.000*
Household expenditure	1.000	0.583	1.000	0.173	1.000	0.006*	1.000	0.371	1.000	0.416
Overcrowded	0.556	0.024*	1.059	0.887	0.755	0.261	0.915	0.688	0.776	0.312
_cons	0.268	0.336	0.015	0.009	0.776	0.821	0.100	0.021	0.278	0.198

O.R: odds ratio
b. P>z: p-value
c.: significant at 0.05*

30-44 and 45-59 years were 2.15 and 1.77 times more likely to be diagnosed with TB respectively, compared to their counterparts aged 60 and above. Males were 23% more likely to be diagnosed with TB compared to females, odds ratio $e^{0.21}$ =1.23. The black African race, with an odds ratio of $e^{0.37}$ =1.45, was 45% more likely to be diagnosed with TB than other races. Having no education or high school education was not significantly different from having tertiary education. Those with Primary school education only were 1.58 times likely to be diagnosed with TB than those with tertiary education over time. With time, isiZulu and Afrikaans speaking people were more likely to be diagnosed with TB than people who speak other languages. The average rate of TB diagnosis was higher among the unemployed than the employed with an odds ratio of 1.16. The average probability of infection for individuals who perceived their health as not good was twice as much (odds ratio 2.08) as those who reported being in good health status. Individuals who exercised regularly had a lesser average probability of being diagnosed with TB over time (odds ratio $e^{-0.15}$=0.86). There was no association over time between TB and diabetes, whereas those diagnosed with Asthma had a higher average probability of being diagnosed with TB (odds ratio $e^{0.66}$ =1.93), compared to their counterparts who were not diagnosed with asthma. Those who were diagnosed with any other disease except the mentioned ones were 2.29 times more likely (odds ratio $e^{0.83}$ =2.29) to be diagnosed with TB over time. Over time, with an odds ratio of 1.8, those who smoke less than 20 cigarettes per day are 80% more likely to be diagnosed with TB over time compared to their counterparts who smoke more than 20 cigarettes per day. The mean probability of contracting TB is approximately 1.1 times more for individuals who belong to a household with no improved sanitation than their counterparts with improved sanitation. Household income and expenditure were both significant, p –values 0.00 and 0.02 respectively. However, they both had estimated coefficients of near-zero implying odds ratio was 1.

DISCUSSION

We formulated the GEE approach for the full model under various working correlation assumptions to analyse the probable performance in relation to the selected covariates. We identified the most suitable working

TABLE 3. Comparison of working correlation matrices

	Independent		Unstructured		Exchangeable		Autoregressive	
	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.
Wave	0.252	0.020	0.223	0.020	0.240	0.019	0.213	0.022
Marital status	0.150	0.079	0.069	0.066	0.148	0.070	0.160	0.074
Age	0.012	0.044	0.066	0.039	0.005	0.040	0.028	0.042
Gender	-0.109	0.092	-0.258	0.089	-0.177	0.090	-0.192	0.099
Race	0.163	0.117	0.030	0.113	0.122	0.113	0.186	0.125
Education	-0.098	0.061	-0.122	0.049	-0.124	0.052	-0.126	0.056
Language	-0.083	0.046	-0.053	0.047	-0.071	0.045	-0.053	0.051
Employment status	-0.034	0.073	0.028	0.059	-0.003	0.059	-0.015	0.063
Perceived health	0.730	0.073	0.349	0.060	0.517	0.062	0.471	0.064
Exercise	0.190	0.069	0.057	0.052	0.071	0.055	0.088	0.057
Other Disease	-0.964	0.092	-0.403	0.084	-0.559	0.087	-0.552	0.092
Asthma	-0.674	0.136	-0.578	0.129	-0.603	0.126	-0.563	0.141
Diabetes	0.136	0.152	0.013	0.100	0.005	0.120	0.006	0.132
Heavy smoker	0.706	0.081	0.340	0.068	0.492	0.069	0.457	0.070
Sanitation	0.135	0.085	0.109	0.068	0.115	0.071	0.070	0.078
Household income	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Household expenditure	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Overcrowded	-0.263	0.137	-0.059	0.129	-0.116	0.123	-0.111	0.123
_cons	-2.406	0.702	-2.451	0.585	-2.485	0.606	-2.508	0.679
Trace	35.580		27.010		28.160		33.380	
QIC	12760.000		12979.000		12843.000		12869.000	
QIC_u	12726.000		12963.000		12824.000		12841.000	

Coeff: estimated coefficient
std err: standard error

correlation matrix and the persistent risk factors of TB in South Africa over the years from 2008 to 2017. The independent correlation matrix was the most suitable working correlation matrix to best describe the TB scenario over time in South Africa. Since this data was cluster set at an individual level, the use of an independent correlation matrix means that within an individual, the probabilities of infection over time are independent but infections are correlated within waves. Though the empirical and model-based standard errors gave similar parameter estimates, the model-based standard errors gave more robust results for this TB data. The change in TB infection diagnosis rate over time was seen to be associated with time (wave), marital status, age, gender, race, education, employment status, perceived health, exercise, asthma, suffering from other diseases, heavy smoking, income, expenditure, and sanitation. There was no significant difference in TB infections between waves 4 and 5, this means that our sample did not show a difference in TB incidence for the years 2014/2015 and 2017. Over time, higher average probabilities of being diagnosed with TB were associated with single, male, middle-aged (30-59 years), black African, unemployed, with primary education, exercise regularly, diagnosed with asthma, suffering from other disease and have no access to improved sanitation.

A number of studies, some of which are referred to in this study, state that TB is a disease of poverty. This link was demonstrated by [14] yet there have been few analyses of the social determinants of tuberculosis, particularly in high-burden settings. We conducted a multilevel analysis of self-reported tuberculosis disease in a nationally representative sample of South Africans based on the 1998 Demographic and Health Survey (DHS using a potential causal pathway for low income and TB. In their study, they used the following variables, Individual: Age, sex, education, race, smoker, alcoholism, Body Mass Index (BMI), employment, urban residence, number of adults per bedroom, affordability of meals and household asset score. In their study on TB in Western Cape South Africa [15], identified the risk factors for infection as

TABLE 4. Empirical versus Model-Based Standard error estimates

		EMPIRICAL STANDARD ERROR ESTIMATES					MODEL BASED STANDARD ERROR ESTIMATES					
Parameter		Est.	S.E	95% Confidence Limits		Pr > |Z|	Est.	S.E	95% Confidence Limits		Z	Pr > |Z|
Intercept		-2.53	0.31	-3.15	-1.92	<.0001	-2.53	0.22	-2.97	-2.10	-11.44	<.0001
Wave	1	-0.84	0.08	-1.01	-0.68	<.0001	-0.84	0.09	-1.01	-0.68	-9.79	<.0001
Wave	2	-0.93	0.09	-1.11	-0.76	<.0001	-0.93	0.09	-1.11	-0.75	-10.15	<.0001
Wave	3	-0.40	0.07	-0.53	-0.27	<.0001	-0.40	0.08	-0.55	-0.25	-5.16	<.0001
Wave	4	-0.09	0.05	-0.19	0.01	0.07	-0.09	0.07	-0.23	0.05	-1.29	0.20
Wave	5	0.00	0.00	0.00	0.00	.	0.00	0.00	0.00	0.00	.	.
Marital status	Married	-0.32	0.08	-0.48	-0.16	<.0001	-0.32	0.06	-0.43	-0.20	-5.49	<.0001
	Single	0.00	0.00	0.00	0.00	.	0.00	0.00	0.00	0.00	.	.
Age	15-29	-0.03	0.16	-0.35	0.29	0.86	-0.03	0.11	-0.24	0.18	-0.26	0.80
	30-44	0.77	0.14	0.49	1.05	<.0001	0.77	0.09	0.59	0.95	8.44	<.0001
	45-59	0.57	0.12	0.32	0.81	<.0001	0.57	0.08	0.40	0.73	6.72	<.0001
	60+	0.00	0.00	0.00	0.00	.	0.00	0.00	0.00	0.00	.	.
Gender	Male	0.21	0.09	0.02	0.39	0.03	0.21	0.06	0.09	0.33	3.35	0.00
	Female	0.00	0.00	0.00	0.00	.	0.00	0.00	0.00	0.00	.	.
Race	African	0.37	0.22	-0.07	0.80	0.10	0.37	0.15	0.07	0.66	2.45	0.01
	Other	0.00	0.00	0.00	0.00	.	0.00	0.00	0.00	0.00	.	.
Education	None	0.26	0.20	-0.13	0.65	0.20	0.26	0.13	-0.01	0.52	1.90	0.06
	Primary	0.46	0.17	0.11	0.80	0.01	0.46	0.12	0.23	0.69	3.87	0.00
	High	0.09	0.15	-0.22	0.39	0.58	0.09	0.11	-0.13	0.30	0.79	0.43
	Tertiary	0.00	0.00	0.00	0.00	.	0.00	0.00	0.00	0.00	.	.
Language	IsiZulu	0.13	0.10	-0.06	0.32	0.18	0.13	0.06	0.01	0.25	2.11	0.03
	Afrikaans	0.69	0.22	0.26	1.11	0.00	0.69	0.15	0.40	0.98	4.65	<.0001
	Other	0.00	0.00	0.00	0.00	.	0.00	0.00	0.00	0.00	.	.
Employed	Yes	-0.15	0.07	-0.29	0.00	0.05	-0.15	0.06	-0.27	-0.03	-2.45	0.01
	None	0.00	0.00	0.00	0.00	.	0.00	0.00	0.00	0.00	.	.
Perceived health: Good	Yes	-0.73	0.07	-0.87	-0.59	<.0001	-0.73	0.06	-0.86	-0.61	-11.45	<.0001
	No	0.00	0.00	0.00	0.00	.	0.00	0.00	0.00	0.00	.	.
Regular exercise	Yes	-0.15	0.07	-0.29	-0.02	0.03	-0.15	0.07	-0.28	-0.02	-2.31	0.02
	No	0.00	0.00	0.00	0.00	.	0.00	0.00	0.00	0.00	.	.
Other disease	Yes	0.83	0.09	0.65	1.01	<.0001	0.83	0.07	0.69	0.97	11.41	<.0001
	No	0.00	0.00	0.00	0.00	.	0.00	0.00	0.00	0.00	.	.
Asthma	Yes	0.66	0.14	0.39	0.93	<.0001	0.66	0.10	0.47	0.86	6.65	<.0001
	No	0.00	0.00	0.00	0.00	.	0.00	0.00	0.00	0.00	.	.
Diabetes	Yes	-0.07	0.15	-0.37	0.22	0.63	-0.07	0.11	-0.29	0.15	-0.64	0.52
	No	0.00	0.00	0.00	0.00	.	0.00	0.00	0.00	0.00	.	.
Heavy smoking	Yes	-0.59	0.08	-0.75	-0.43	<.0001	-0.59	0.07	-0.72	-0.46	-8.91	<.0001
	No	0.00	0.00	0.00	0.00	.	0.00	0.00	0.00	0.00	.	.
Sanitation	Yes	-0.10	0.09	-0.27	0.07	0.25	-0.10	0.06	-0.22	0.02	-1.64	0.10
	No	0.00	0.00	0.00	0.00	.	0.00	0.00	0.00	0.00	.	.
Income		0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00	-2.91	0.00
Expenditure		0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	-2.36	0.02
Overcrowded	Yes	0.21	0.14	-0.06	0.48	0.13	0.21	0.12	-0.02	0.44	1.78	0.08
	No	0.00	0.00	0.00	0.00	.	0.00	0.00	0.00	0.00	.	.
							1.00	.	.	.	.	.

Est: estimate - b. S.E: Standard error - c. Pr>|Z|: P-value

overcrowding, number of infectious cases in the community, malnutrition, alcohol abuse and unemployment. Adherence to medication was also a contributing factor so they recommended that non-medical interventions were key to the success of TB control programmes. On identifying risk factors of TB/HIV in South Africa 2006, those married or have surviving partners were reported to be at lower risk of TB/HIV [16]. Based on a National study, [17] reported Eastern Cape Province the hardest hit by TB and Limpopo the least. They also listed gender, marital status, age groups, poor living conditions, lower socio-economic status, English illiterateness, alcohol use, and lack of secondary/tertiary education as drivers of TB.

These disparities in TB deaths are attributed to socioeconomic factors associated with place of birth, income, education and healthcare access and regional differences [18]. Also, [19] concluded that it is because of their overexposure to poor living conditions in overcrowded places with deficient hygiene, protection, and malnutrition. TB and general health status also depend more on individual risks such as age, sex, migrant status, diabetes, HIV status, marital status, ethnical groups, vagrancy, smoking, alcohol and drug use. Other socioeconomic and environmental risk factors include deprivation, financial instability, and household dwelling conditions. On their cross-sectional data analysis on self-reported TB for a sample in Eastern Cape South Africa, [20] a leaky roof, social capital, unemployment, income also recommended the need to consider "possible benefits of programs that deal with housing and social environments when addressing the spread of TB in economically poor districts".

Our results concur with other researches previously done [14-17] yet there have been few analyses of the social determinants of tuberculosis, particularly in high-burden settings. We conducted a multilevel analysis of self-reported tuberculosis disease in a nationally representative sample of South Africans based on the 1998 Demographic and Health Survey (DHS, even though they were all cross-sectional studies. Non-heavy smoking was also identified as associated with TB maybe because, as reported by [21], once TB patients are diagnosed, most tend to reduce cigarettes intake. Income and expenditure are associated with TB [18,20] but in this study, having an odds ratio of one means that a unit increase in household income or expenditure from year to year has no impact on the average probability of being diagnosed with TB. The WHO reports TB as a threat to diabetic mellitus patients due to their compromised immune system[1]. Although there was a positive relationship between diabetes and TB in this research, there was no significant difference over time between individuals who had been diagnosed with diabetes or not, as far as self-reported TB was concerned. This finding is in contrary to the findings of [22,23], who did cross-sectional clinical research and found an association between TB and diabetes particularly among HIV positive people.

The number one global strategy to eradicate TB starts with early diagnosis and effective treatment [1,3]. However, research has shown that there is a need for social interventions [14,17] yet there have been few analyses of the social determinants of tuberculosis, particularly in high-burden settings. We conducted a multilevel analysis of self-reported tuberculosis disease in a nationally representative sample of South Africans based on the 1998 Demographic and Health Survey (DHS. We recommend further investments in TB screening, detection, and treatment. The same was recommended by [24,25] on chronic respiratory and non-communicable diseases. They reiterated the need for a "comprehensive programme to tackle chronic respiratory diseases" and "integration of non-communicable diseases and TB programs for screening, counselling, and treatment of comorbidities", respectively. We also recommend an awareness campaign that emphasises on the risks of smoking extending to more than the common breathing problems and lung cancer but poses smokers as at double the risk of TB. There is also a need to improve socioeconomic and living conditions for South Africans to help eradicate TB. More research on the synergy between diabetes and TB is required to give a better understanding of this deadly coinfection.

Acknowledgements

No financial support was received for this study.

References

1. WHO. Global strategy and targets for tuberculosis prevention, care and control after 2015. End TB Strategy [Internet]. 2015. Available from: http://www.who.int/tb/publications/global_report/en/
2. WHO | WHO End TB Strategy [Internet]. WHO. [cited 2019 Jan 31]. Available from: https://www.who.int/tb/post2015_strategy/en/
3. Sotgiu G, Falzon D, Castiglia P, Migliori GB, Raviglione M. Tuberculosis and the strategy for the New Millennium: Not simply "more of same." Epidemiol Biostat Public Health. 2014;11:e10116.
4. Knight GM, Dodd PJ, Grant AD, Fielding KL, Churchyard GJ, White RG. Tuberculosis prevention in South Africa. PloS One. 2015;10:e0122514.
5. Knight GM, Gomez GB, Dodd PJ, Dowdy D, Zwerling A, Wells WA, et al. The Impact and Cost-Effectiveness of a Four-Month Regimen for First-Line Treatment of Active Tuberculosis in South Africa. PLOS ONE. 2015;10:e0145796.
6. Leibbrandt M, Woolard I, de Villiers L. Methodology: Report on NIDS Wave. N.i.D.S; 2009 p. 34.
7. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986;73:13–22.
8. Zeger SL, Liang K-Y. Longitudinal Data Analysis for Discrete and Continuous Outcomes. Biometrics. 1986;42:121.
9. Zeger SL, Liang K-Y, Albert PS. Models for Longitudinal Data:

A Generalized Estimating Equation Approach. Biometrics. 1988;44:1049–60.
10. Hedeker D, Gibbons RD. Longitudinal Data Analysis. 1 edition. Hoboken, N.J: Wiley-Interscience; 2006.
11. Models for Discrete Longitudinal Data [Internet]. New York: Springer-Verlag; 2005 [cited 2019 Jan 31]. Available from: http://link.springer.com/10.1007/0-387-28980-1
12. Pan W. Akaike's information criterion in generalized estimating equations. Biometrics. 2001;57:120–5.
13. Gosho M. Criteria to Select a Working Correlation Structure for the Generalized Estimating Equations Method in SAS. J Stat Softw. 2014;57:1–10.
14. 14. Harling G, Ehrlich R, Myer L. The social epidemiology of tuberculosis in South Africa: a multilevel analysis. Soc Sci Med 1982. 2008;66:492–505.
15. 15. Yach D. Tuberculosis in the Western Cape health region of South Africa. Soc Sci Med 1982. 1988;27:683–9.
16. 16. Appunni SS, Blignaut R, Lougue S. TB/HIV risk factors identified from a General Household Survey of South Africa in 2006. SAHARA-J J Soc Asp HIVAIDS. 2014;11:37–41.
17. Dhlakama H, Lougue S. Bayesian Modelling of Tuberculosis Risk Factors in South Africa 2014. Int J Stat Med Res. 2017;6:34–48.
18. Young BN, Rendón A, Rosas-Taraco A, Baker J, Healy M, Gross JM, et al. The effects of socioeconomic status, clinical factors, and genetic ancestry on pulmonary tuberculosis disease in northeastern Mexico. PloS One. 2014;9:e94303.
19. Stoesslé P, González-Salazar F, Santos-Guzmán J, Sánchez-González N. Risk Factors and Current Health-Seeking Patterns of Migrants in Northeastern Mexico: Healthcare Needs for a Socially Vulnerable Population. Front Public Health [Internet]. 2015 [cited 2019 Jan 31];3. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4526788/
20. Cramm JM, Koolman X, Møller V, Nieboer AP. Socio-economic status and self-reported tuberculosis: a multilevel analysis in a low-income township in the Eastern Cape, South Africa. J Public Health Afr [Internet]. 2011 [cited 2019 Jan 31];2. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5345507/
21. Peltzer K. Correlates of tobacco use among tuberculosis patients in South Africa: A brief report. J Psychol Afr. 2016;26:473–6.
22. Oni T, Berkowitz N, Kubjane M, Goliath R, Levitt NS, Wilkinson RJ. Trilateral overlap of tuberculosis, diabetes and HIV-1 in a high-burden African setting: implications for TB control. Eur Respir J. 2017;50:1700004.
23. Berkowitz N, Okorie A, Goliath R, Levitt N, Wilkinson RJ, Oni T. The prevalence and determinants of active tuberculosis among diabetes patients in Cape Town, South Africa, a high HIV/TB burden setting. Diabetes Res Clin Pract. 2018;138:16–25.
24. Marak B, Kaur P, Rao SR, Selvaraju S. Non-communicable disease comorbidities and risk factors among tuberculosis patients, Meghalaya, India. Indian J Tuberc. 2016;63:123–5.
25. Viswanathan K, Rakesh PS, Balakrishnan S, Shanavas A, Dharman V. Prevalence of chronic respiratory diseases from a rural area in Kerala, southern India. Indian J Tuberc. 2018;65:48–51.

*