THE SOCIAL EPIDEMIOLOGY OF TUBERCULOSIS IN SOUTH AFRICA: A MULTILEVEL ANALYSIS

by Guy Harling

HRLGUY001

Submitted to the University of Cape Town in partial fulfilment of the requirements for the degree Master of Public Health

Faculty of Health Sciences

UNIVERSITY OF CAPE TOWN

August 2006

Supervisor:
Dr Landon Myer
School of Public Health & Family Medicine
University of Cape Town

DECLARATION

I, Guy Harling, hereby declare that the work on which this dissertation is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university.

I empower the university to reproduce for the purpose of research either the whole or any portions of the contents in any manner whatsoever.

Signature:…….

Date:…….

Abstract

Tuberculosis has long been considered a disease of poverty but there has been little research into the pathways through which low socio-economic status leads to increased risk of disease. This study reviews the existing literature on risk factors for tuberculosis disease with a particular focus on those variables that reflect the social setting in which an individual lives. It then conducts a multilevel analysis of South African data from the 1998 South African Demographic and Health Survey and the 1996 national census to evaluate individual-, household- and community-level risk factors for tuberculosis disease using a hierarchical regression model. In analyses adjusting for non-socio-economic individual characteristics, the study finds that having low education, being unemployed, living in a household with a low level of wealth, and living in a community with high levels of income inequality are each independently associated with an increased risk of having ever been diagnosed with tuberculosis. The study suggests that tuberculosis prevention and treatment policy might benefit from an expansion to include programs to aid those living in low socio-economic conditions.

Acknowledgements

I would like to thank my advisors, Professor Rodney Ehrlich and particularly Dr. Landon Myer, for their guidance and patience throughout the writing of this study, and two anonymous reviewers for their thoughtful comments.

I would also like to thank the Medical Research Council, in particular Dr. Debbie Bradshaw, for providing the South African Demographic and Health Survey data on which this dissertation is largely based.

Finally, I would like to thank Katie Doolan for her advice and comments at various points in the process of writing this paper, including data sourcing, methodological discussions, commenting on a draft of the final paper, and for originally introducing me to this project.

Table of Contents

1. Introduction 1
2. Aim and Objectives 3
3. Literature Review 4
3.1. Tuberculosis Epidemiology 4
3.1.1. Risk factors for tuberculosis infection 4
3.1.2. Biomedical risk factors for tuberculosis disease 5
3.2. Tuberculosis in South Africa 8
3.3. Social Epidemiology 10
3.4. Socio-Economic Status and Health 13
3.4.1. Income poverty 13
3.4.2. Income inequality 14
3.4.3. Employment 16
3.4.4. Education 16
3.4.5. Deprivation 17
3.5. Socio-Economic Status and Tuberculosis 19
3.5.1. Socio-economic risk factors in low tuberculosis incidence settings 19
3.5.2. Socio-economic risk factors in high tuberculosis incidence settings 20
3.6. Literature Summary 21
4. Methods 22
4.1. The 1998 South African Demographic and Health Survey 22
4.1.1. Survey methodology 22
4.1.2. Variable selection and manipulation 23
4.2. The 1996 South African National Census 27
4.2.1. Census methodology 27
4.2.2. Variable selection and manipulation 27
4.3. Analytic Methodology 33
4.3.1. Data weighting 33
4.3.2. Descriptive statistics and bivariate analysis 33
4.3.3. Multilevel modelling 34
4.3.4. Multivariate analysis 36
4.3.5. Model checking 36
5. Results 37
5.1. Descriptive Statistics 37
5.1.1. Tuberculosis 37
5.1.2. Biomedical, behavioural & social characteristics 37
5.1.3. Socio-economic status 39
5.2. Bivariate Analysis 40
5.2.1. Geography 40
5.2.2. Biomedical, social & behavioural characteristics 42
5.2.3. Individual- and household-level SES 45
5.2.4. Magisterial district-level absolute poverty and unemployment 45
5.2.5. Magisterial district-level relative poverty 46
5.3. Multivariate Models 47
5.3.1. Non-SES risk factors 47
5.3.2. Individual and household SES risk factors 48

5.3.3. Magisterial district income poverty and employment status....56
5.3.4. Magisterial district income inequality....56
5.4. Model Checking....61
6. Discussion....63
6.1. Key Findings....64
6.1.1. Individual- and household-level SES risk factors....64
6.1.2. Community-level SES risk factors....67
6.1.3. Non-SES risk factors....68
6.2. Potential Explanations for Findings....69
6.2.1. Geographic effects....69
6.2.2. Compositional and contextual effects....70
6.2.3. Risk factors in high tuberculosis prevalence settings....75
6.3. Potential Strengths and Weaknesses of the Analysis....76
6.3.1. Study design....76
6.3.2. Analytic design....80
6.3.3. HIV....84
6.3.4. Study validity....85
6.4. Next Steps....87
6.4.1. Research recommendations....87
6.4.2. Social epidemiology....88
6.4.3. Policy recommendations....90
6.5. Conclusion....93

Table of Tables

Table 1: Independent Variables from the South African Demographic and Health Survey 1998 25
Table 2: Monthly Bands and Imputed Values for Census Income Data 28
Table 3: Descriptive Statistics for SADHS Variables 38
Table 4: Descriptive Statistics for Census-Based Group-Level SES Measures 40
Table 5: Weighted Tuberculosis Rates by Geographic Strata 41
Table 6: Bivariate Relationships between Non-SES Variables and Tuberculosis 43
Table 7: Bivariate Relationships between Tuberculosis and SES Variables 46
Table 8: Multivariate Relationships between Non-SES Variables and Tuberculosis 48
Table 9: Multivariate Relationships between Individual-Level SES and Tuberculosis 49
Table 10: Multivariate Relationships between Household-Level SES and Tuberculosis 50
Table 11: Multivariate Relationships between Household-Based, MD-Level Headcount Income Poverty and Tuberculosis 52
Table 12: Multivariate Relationships between Household-Based, MD-Level Unemployment Rate and Tuberculosis 54
Table 13: Multivariate Relationships between Household-Based, MD-Level Income Inequality (Robin Hood index) and Tuberculosis 57
Table 14: Multivariate Relationships between Household-Based, MD-Level Income Inequality (Gini Coefficient) and Tuberculosis 59
Table 15: Characteristics of Individuals with Outlying Pearson or Deviance Residual Values 61
Table 16: Characteristics of MDs with Outlying Cook's Distance Values 62

Table of Figures

Figure 1: Potential Causal Pathways for Low Income and Tuberculosis 11
Figure 2: An Example of a Lorenz Curve 31
Figure 3: Adjusted Odds Ratios and 95% CIs for Asset Score Quintiles and Tuberculosis 65
Figure 4: Models of the Causal Pathways between SES Variables and Tuberculosis 67
Figure 5: Potential Causal Mechanisms for Community-level Income Inequality 72
Figure 6: Adjusted Odds Ratios and 95% CIs for Income Inequality Measures and Lifetime Prevalent Tuberculosis 74
Figure 7: Adjusted Odds Ratios and 95% CIs for Income Inequality Measures and Incident Tuberculosis 74
Figure 8: Potential Effect Pathways for Diagnostic Bias 79
Figure 9: Potential Causal Pathways for HIV 85

Abbreviations

BMI	Body Mass Index
CI	Confidence Interval
DHS	Demographic and Health Survey
DOTS	Directly Observed Treatment, Short-course
EA	Enumerator Area
GLLAMM	Generalised Linear Latent and Mixed Models
HIV	Human Immunodeficiency Virus
IQR	Inter-Quartile Range
MD	Magisterial District
MRC	Medical Research Council
OR	Odds Ratio
PCA	Principal Component Analysis
PSU	Primary Sampling Unit
PYO	Person-Years of Observation
SADHS	South African Demographic and Health Survey
TB	Tuberculosis
UK	United Kingdom
US	United States

1. Introduction

Tuberculosis (TB) is the world's leading curable cause of infectious disease mortality, directly resulting in more than 1.8 million deaths each year (World Health Organisation 1999). Concerted efforts using effective chemotherapeutic drugs in the decades following World War II led many to believe that tuberculosis was slowly being eradicated. More recently, however, the rise of multi-drug resistant strains of tuberculosis and the emergence of HIV-TB co-infection, possibly exacerbated by reduced access to healthcare due to social and economic disruption, has led to a resurgence of the disease (Gandy & Zumla 2002).

Tuberculosis is a significant public health problem in South Africa. The illness has been a public health concern since the beginning of the twentieth century (Packard 1987). Notification rates peaked in the 1950s, but the subsequent decline in reported disease until the early 1990s is likely to have been an artefact of Apartheid disease notification policies rather than diminished case rates. The emergence of HIV infection, and its interaction with tuberculosis, has led South Africa to have the second highest reported incidence rate of tuberculosis in the world (Maher & Raviglione 2005). As a result it was estimated that in 2001 one in eleven deaths nationally were due to tuberculosis (Groenewald et al. 2005). Given the burden of tuberculosis disease in South Africa, research based on local data is needed to delineate risk factors for disease, and to suggest interventions to reduce the burden.

Tuberculosis is traditionally known as a 'disease of poverty'. Many expressions of poverty, for example overcrowding and malnutrition, are widely acknowledged to be risk factors for the disease. Furthermore, there is a long history within epidemiology for considering tuberculosis to be directly associated with socio-economic factors worldwide (Dubos & Dubos 1992; McKeown 1976; Gandy & Zumla 2003). Evidence from South Africa suggests that the links between tuberculosis and socio-economic factors are particularly strong in this country (Packard 1987). The infectious nature of tuberculosis makes it likely that risk of disease is not only a function of an individual's personal characteristics, but also those of their community. This consideration of an

individual's social setting in determining their risk of disease is captured in the growing field of social epidemiology (Berkman & Kawachi 2000).

This study uses a combination of South African data sources to examine the degree and nature of the link between socio-economic status (SES) and tuberculosis in the country in the late 1990s. Examining individual- and group-level characteristics in parallel, it attempts to identify which form(s) of SES are associated with tuberculosis disease, after taking account of other accepted risk-factors for the disease.

2. Aim and Objectives

The aim of this study is to examine the impact of poverty and income inequality on the incidence and prevalence of tuberculosis disease among adults in South Africa, after taking into account measured individual biomedical and behavioural factors. To achieve this, the objectives of the study are:

1. To provide descriptive statistics on tuberculosis disease in South Africa, including incidence and prevalence by geographic region and individual demographic characteristics.
2. To examine the association between individual- and household-level biological, social and behavioural factors commonly associated with tuberculosis disease.
3. To explore the association between prevalent and incident tuberculosis and:
    - household- and community-level poverty based on assets,
    - community-level poverty based on income,
    - individual- and community-level poverty based on unemployment.
4. To measure the importance of income inequality at the community level on tuberculosis, over and above the factors explored under objective 3.

3. Literature Review

3.1. Tuberculosis Epidemiology

In order to develop tuberculosis disease in an individual the causative bacillus, *Mycobacterium tuberculosis*, must first enter the host and then replicate to a level sufficient to cause illness. It is estimated that one third of the world's population is infected with *M. tuberculosis* (Dye et al. 1999). Progression from infection to disease occurs either soon after infection due to an internal inflammation-infection process – primary illness – or at some later date due to a weakened immune system – reactivation illness. The lifetime risk of each is around five percent in an immunocompetent individual (Coberly & Chaisson 2001). In the presence of an immunosuppressive agent, such as chemotherapy or HIV, the risk is one to two orders of magnitude higher.

3.1.1. Risk factors for tuberculosis infection

As an airborne infectious agent, tuberculosis infection is most likely when an individual spends time in close proximity to an infectious source. Thus key risk factors for infection are the existence of a person with active tuberculosis disease (the index case) living in the same house or institution; the infectivity of the index case's disease; and the degree of overcrowding of the setting (Coberly & Chaisson 2001).

Measuring these risk factors is complex and usually requires the use of proxy characteristics of an individual. The degree of contact, both duration and proximity, that an individual has with infectious persons is likely to be patterned by their social activities, and these in turn will be stratified by age, sex and ethnicity. As a result, demographic characteristics are often presented as risk factors for infection (Strebel & Seager 1991).

Furthermore, some variables are risk factors both for tuberculosis infection and for disease. For example, household crowding is a direct risk factor for infection due to increased proximity to other individuals, but is also a proxy for SES and therefore often presented as a risk factor for disease (Cantwell et al. 1998; Bennett et al. 2001; Munch et

al. 2003; Gustafson et al. 2004). There is therefore potential for infection and disease risk factors to be conflated, particularly in low-incidence settings.[1]

3.1.2. Biomedical risk factors for tuberculosis disease

Factors that raise an individual's risk of tuberculosis are those things which reduce their capacity to defend themselves against the mycobacterium. These are typically divided into those factors not amenable to intervention (essential characteristics) and those which are (contingent characteristics) (Strebel & Seager 1991; King 2003).

The former category includes the temporal characteristic – time since infection – and host characteristics such as age, sex, genetics and ethnicity. The latter category contains host and environmental characteristics that lower the host's immunity, these include: (i) past history of disease affecting the lung; (ii) present history of illness that, either by its natural history or due to its treatment, suppresses the immune system; (iii) exposure to harmful substances; and (iv) an absence of factors promotive of good health.

3.1.2.1 Risk factors not amenable to intervention

The risk of becoming actively ill with tuberculosis is highest in the first two years following infection, falling to a constant background level after five years. It is likely that this pattern reflects the lifecycle of tuberculosis – after this time a person has either succumbed to infection or successfully forced the bacillus into dormancy (Coberly & Chaisson 2001).

The effects of age and sex on risk of disease in adults are not simple. Historically, and in the more developed world today, the perceived risk of developing disease declined from birth to age ten, then rose to peak in early adulthood and fell thereafter. This, however, is a cross-sectional picture and reflects the fact that each successive birth cohort has been at lower risk than its predecessor. In reality, the risk of tuberculosis for each cohort falls through their lifetime, not least because those most susceptible to disease tend to die young (Coberly & Chaisson 2001).

[1] This may arise because only certain sections of the population are at risk for infection in low-prevalence settings, and thus also for disease. Disease therefore appears to be patterned by what are actually infection risk factors. In high-prevalence settings, where almost all people are infected with tuberculosis, disease will only be patterned by disease risk factors, since there is no variation in infection.

The effect of sex on risk of disease is modified by age. Men under 15 or over 45 report higher rates of disease than women of the same age, but the pattern is inverted for those of childbearing age. These results have been seen in several settings, including Africa, but are not found in every case (Holmes et al. 1998; Glynn et al. 2000). It has been argued that this may reflect different patterns of health-seeking behaviour, with women being less likely to access care, perhaps due to a higher risk of social stigma (Thorson & Diwan 2003). Active case-finding often finds more sex-neutral infection and disease rates.

The relevance of racial or ethnic background as a marker for tuberculosis disease remains contested. In England the variation in tuberculosis rates in several major cities was found to be explicable predominantly as a function of the proportion of the population born in the Indian subcontinent and Africa (Bennett et al. 2001; Tocque et al. 1999; Parslow et al. 2001). The authors attributed this to immigration from high-incidence areas and the resultant increased contact with infected and infectious individuals. A study of tuberculosis in Hong Kong found a similar association between immigration from mainland China and disease (Leung et al. 2004). In the United States (US) some studies have found a positive association between African-American or Hispanic race and tuberculosis disease rates (Barr et al. 2001; Cantwell et al. 1994), but recent evidence suggests that this too is largely due to a higher exposure to other risk factors for tuberculosis, including exposure to recent immigrants from higher-incidence settings (Cantwell et al. 1998; Acevedo-Garcia 2001).

African epidemiological studies of tuberculosis rarely consider race as an explanatory variable. One study in West Africa found being a racial minority to be a risk factor for disease (Gustafson et al. 2004). This may support an argument that migrant or minority status is a risk factor, rather than any particular ethnicity, through an etiology of raised stress levels or low SES (King 2003).

The balance of evidence seems to suggest that racial background is not a causal risk factor for tuberculosis, although it may well be a mediator, or risk marker, in many settings. At a group level, no genetic predisposition to tuberculosis has been shown for any race or ethnicity, although there is evidence that genetic predisposition exists at the

individual level (Coberly & Chaisson 2001). As a result it is likely that changes in the social environment of ethnic or racial groupings will change their tuberculosis risk, making the risks generally associated with race modifiable.

3.1.2.2 Risk factors amenable to intervention

The above risk factors are inherent to an individual at birth or are a function of passing time; their impact on risk of disease cannot therefore be directly reduced. There are, however, other characteristics that are risk factors for tuberculosis disease, and for which risk-reducing interventions are possible.

Past history of tuberculosis disease is predictive of future disease. This is due to a higher likelihood of both being infected with *M. tuberculosis* and to having healed fibrotic lesions which are conducive to the development of active disease. There is also evidence for silicosis being a predisposing factor in the development of tuberculosis disease, putting miners at particular risk of illness. This effect is even more pronounced among those with HIV infection (Corbett et al. 2000).

HIV infection is a significant risk factor for tuberculosis disease. HIV raises the risk of an infected individual developing active disease from five percent in a lifetime to between five and ten percent per year, a risk that rises as HIV disease progresses (Raviglione et al. 1997; Badri, Wilson & Wood 2002). It is estimated that in sub-Saharan Africa 31 percent of all tuberculosis cases and 34 percent of all deaths are attributable to HIV (Corbett et al. 2003). The impact of HIV is so large in high-incidence settings that it changes other previously-observed epidemiological patterns. A study of tuberculosis cases in Malawi between 1998 and 2001 found two-thirds of patients to be HIV-infected. In this population the normal difference in reported cases between men and women over the age of 45 was inverted, with women being more likely to have disease than men (Glynn et al. 2004).

There is qualified evidence for alcohol and tobacco use raising an individual's risk of developing active tuberculosis. A case-control study in China found no association between smoking or tobacco use and tuberculosis when measured individually, but did find an effect when both substances were jointly consumed (Dong et al. 2001). A review of studies on the relationship between tobacco and tuberculosis found both an

exposure and a dose-response relationship between smoking and tuberculosis disease (Maurya et al. 2002). Evidence suggests that heavy drinkers have a raised risk of tuberculosis. This may be due to it directly exacerbating tuberculosis (Mason et al. 2004), as has been seen in mice, or due to the general weakening of the immune system that excessive alcohol consumption can cause (Gamble et al. 2006).

It is often stated that diabetes mellitus is a risk factor for tuberculosis disease, perhaps due to its immunosuppressive effect (refs). Evidence for this relationship is sparse, but a case-control study in the United States found patients hospitalised for tuberculosis to have a raised crude risk of also having diabetes (Palos-Méndez et al. 1997). A second case-control study in Indonesia found a stronger relationship in populations with similar socio-economic characteristics (Alisjahbana et al. 2006). As both studies note, the nature of their analysis makes causal connections impossible to isolate. A population-based analysis in Veracruz, Mexico found those diagnosed with tuberculosis to be almost seven times more likely to have been previously diagnosed with diabetes than other residents in the area (Ponce-de-Leon et al. 2004). Again, the associations reported are crude.

A causal connection between malnutrition and tuberculosis remains uncertain. Adequate nutrition is crucial in fighting any infection, and malnutrition as measured through low body mass index has been associated with tuberculosis in several settings (Schwenk & Macallan 2000). Given the impact of crowding and poverty on tuberculosis (see section 3.4 below), and their high correlation with poor nutrition, it is hard to precisely measure the size of this effect. This problem is exacerbated by the bi-directionality of any such relationship – tuberculosis disease is characterised by rapid weight loss.

3.2. Tuberculosis in South Africa

South Africa is estimated to currently have the second highest incidence rate – 558 cases per 100,000 population per year – and the eighth highest absolute number of new cases – 250,000 per year – of tuberculosis in the world (Maher & Raviglione 2005; World Health Organisation 2005). While the quality and completeness of case reporting undoubtedly rose following the advent of fully-representative government in 1994,

tuberculosis disease notification rates to the World Health Organisation rose from around 200 per 100,000 population per year in the 1980s to 505 in 2003. The proportion of this rise due to increased illness related to HIV infection and the proportion due to improved case-detection is unclear, but by 2000 50% of tuberculosis cases were estimated to also be HIV-positive (Corbett et al. 2003).

Prior to 1980, reported incident tuberculosis rates had declined by half from a peak of almost 400 cases per 100,000 population per year in the early 1960s, having risen rapidly to this level following the end of World War II (Wulfsohn 1985). While it has been suggested that this fall was due to the waning of tuberculosis disease (Strebel & Seager 1991), others have argued that it was an artefact of decreased surveillance among those at highest risk, particularly African populations in rural areas (Packard 1987). This lack of reporting was most pronounced in the "independent homelands" created within the borders of South Africa, which were excluded from national statistics in the early 1980s. Analysis of urban tuberculosis notification, where health treatment and reporting infrastructure was strongest, suggests that incidence rates remained extremely high among African and Coloured communities, with rates of 2-3000 per 100,000 population per year not uncommon (Andersson 1990). It is therefore likely that tuberculosis rates in South Africa fell rather less than notification rates suggested between 1965 and 1995.

This argument is supported by population surveys carried out during this period which suggest that infection and disease rates were extremely high. Infection prevalence studies conducted among children aged five to nine in the late 1970s found tuberculosis infection rates among African populations averaged 13.9%, among Coloureds 8.6%, among Indians 2.8% and among Whites 1.1%.[2] By the age of 15 this had risen to 33.1% among African populations. A study of the prevalence of tuberculosis disease run in the decade from 1974 found rates to be as high as 3800 cases per 100,000 persons tested

[2] Under the apartheid system of the National Party government in South Africa between 1948 and 1990 all citizens were assigned to one of four 'population groups' or racial categories. These groups were African (or Black), Indian (or Asian), White and Coloured (those of mixed African, Indian and/or White descent). In line with other South African literature, and without implying that these terms have genetic validity, they will be used throughout this paper. The most common reason for continuing to use these categories is to measure the impact of interventions on reducing historical inequalities.

among rural African men. Other studies found tuberculosis prevalence to vary between 960 and 2200 culture-positive cases per 100,000 population in the 1980s (Strebel & Seager 1991).

The reasons for tuberculosis rates being so high among the African, and to some extent Coloured, population today appear to be a combination of historical employment and treatment access patterns, and current socio-economic realities. As Packard outlines in his seminal work on the subject, African men from across Southern Africa were drawn, and within South Africa economically coerced, to work in mines, particularly around Johannesburg, beginning in the late nineteenth century (Packard 1987). Working in settings where risks such as silicosis, long exposure to actively ill persons in a damp, humid environment, poor nutrition and considerable physical stress were daily realities, tuberculosis rapidly became endemic in such populations. Sick workers lost their jobs and due to the pass system, which made it illegal for most African persons without employment to remain in an urban area, had to return to the rural environments from which they had been brought. In these places the actively ill infected friends and family, and as a result tuberculosis infection was endemic in the African population in South Africa by the 1970s (Andersson 1990).

These employment patterns are likely to have led to the high rates of infection seen in the African population, but there is no evidence that the determinants of active disease in South Africa are any different from those seen in other countries. Furthermore, while both infection and disease rates for tuberculosis differ widely by population group, survival rates remain similar. This suggests that although groups are differentially at risk for infection, the same is not true for disease progression (Strebel & Seager 1991).

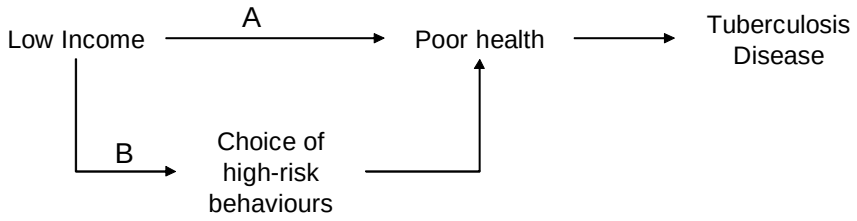
3.3. Social Epidemiology

Social epidemiology is the study of the distribution and determinants of health in populations analysed by individuals' and groups' social conditions (Berkman & Kawachi 2000). It begins from the premise that an individual's risk of becoming ill, and of recovering from such illness, is not determined solely by their biological characteristics but also their place in society. If this premise is true, failure to take social

conditions into account in studying determinants of health may lead to a misdiagnosis of the problem and thus a failure to change those factors which are the root causes of ill health. Studying such hypotheses may be as simple as considering the level of financial resources available to individuals (income or wealth), or as complex as analysing the nature and structure of the society in which they live.

Ill-health can be proximally caused by social conditions (Figure 1, pathway A). For example, having no income may mean an individual is unable to buy food and thus suffers from malnutrition. It is more common, however, for social factors to play a more distal role in the causation of disease (Figure 1, pathway B), often because they limit choices, or lead individuals to make decisions that are detrimental to their health. For example, having no income in a society which does not allow for female empowerment may push a woman towards working as a sex worker, placing her at raised risk of violence, sexually transmitted infections or unwanted pregnancies.

Figure 1: Potential Causal Pathways for Low Income and Tuberculosis



Although the study of social epidemiology may focus only on individual-level social characteristics, a frequent construction used in the field is the group-level explanator of an individual-level outcome. The idea has its roots in Emile Durkheim's concept of a 'social fact', a phenomenon that is not simply the amalgamation of individual characteristics, but is formed by the rules and structures of society as a whole (Schwartz & Diez-Roux 2001). Studying such phenomena requires an investigation of an individual's context, rather than just their personal characteristics. A comprehensive social epidemiology approach will often require methods that consider both individual- and group-level factors, since it is likely that both personal and societal factors will impact the health of an individual.

These group-level factors can be divided, roughly, into two categories. Those factors that are the sum of the characteristics of individuals in a community are referred to as compositional factors; those that reflect emergent aspects of the social or physical environment – closer to Durkheim's 'social facts' – are called contextual factors (MacIntyre & Ellaway 2000). An alternative definition is to be consider compositional factors as being agency-related, in that they rely on the actions and intentions of individuals, and to consider contextual factors as being structure-related, in that they are socially constructed attributes of groups (Veenstra 2005). It is possible for a health outcome to be due to a combination of compositional and contextual factors, and even to interactions between factors in the two groups. In practice it is often difficult to distinguish between compositional and contextual factors, since many compositional attributes are at least partially a result of the pre-existing context in which they arise (MacIntyre et al. 2002).

A wide range of group-level factors has been used to measure SES. These have included aggregated values of individual-level data, such as the mean income or percentage unemployed in a population, emergent values that have no individual-level analogy, such as the degree of income inequality of a group, and non-person-based measures, such as level of public service provision (MacIntyre et al. 2002). It is hoped that each measure will capture a different aspect of the social setting under consideration.

In order to simultaneously consider both group and individual-level risk factors for a disease, a multilevel approach is needed (Diez-Roux & Aiello 2005). Individual-level studies frequently use realisations of an individual's social standing and context such as relative personal income or wealth, education and employment status. Ecologic studies, in contrast, use only group-level variables to explain group-level outcomes. Alone, each of these approaches to understanding would fail to take account of the other, but a multilevel model can account both for personal standing and group context.

One strength of multilevel analysis is the capacity to evaluate whether a variable has a compositional, contextual or mixed effect on the health outcome of interest by including both individual- and group-level realisations of the risk factor in the analysis. If only the

individual-level measure is significant then the group-level measure adds no information and the effect is purely compositional. If only the contextual measure is significant then the individual's characteristic is unimportant and the effect is purely due to the community-level aspect of the variable. If both are significant, then both effects play a role.

3.4. Socio-Economic Status and Health

Socio-economic status refers to either the absolute or relative position of an individual in society. A wide range of measures are believed to reflect SES, each contributing a facet to the overall picture of an individual's socio-economic position (Braveman et al. 2005; Lynch & Kaplan 2000). These measures fall into two overarching categories: first, utility-based – measuring an individual's command over resources, such as income or consumption; second, capability-based – measuring the outcomes achieved with these resources, such as life expectancy or literacy (Booysen 2001). This study focuses on the former, within which resource use may be measured directly (e.g. expenditure or consumption patterns, assets owned), or indirectly through what financial resources are available for such use (e.g. income, wealth). All of these measures have been found to be associated with, although not always causally connected to, poor health.

3.4.1. Income poverty

Poverty has been described as "being deficient and deprived materially, socially and emotionally", excluding those affected both from resource consumption and societal interaction (Benatar 2003, p223). Income poverty is one aspect of overall poverty, and is undoubtedly associated with poor health. Those living in poor countries, on the average, have lower life expectancy, and higher rates of morbidity while alive, than those living in richer countries (Preston 1975, Mathers et al. 2004). Moreover, within a given country, the rich tend to live longer than the poor. Research in the US found that the risk of death in the decade following age 50 to be more than two times higher for those with annual household incomes under $5,000 than for those with incomes over $50,000 (Deaton 2003). Similar gradients were seen in other studies in the US, in Canada and in Holland (Marmot 2002, van Lenthe et al. 2004). The relationship

between income and health appears to be stronger the lower the level of aggregation used in developed nations – using area, state or national measures of income gives a poorer correlation than individual or household measures.

Evidence on the relationship between poverty and health from lower income settings is limited, however a study of the impact of providing old-age pensions to otherwise impoverished families in South Africa suggested a considerable improvement in health resulted (Case 2002). This suggests that poverty may play a role in health outcomes in settings where it is a common condition, rather than the exception as it is in richer nations.

3.4.2. Income inequality

Measures of income poverty provide information on the absolute condition of the poorest segments of society, but they cannot speak to the conditions of those living outside of poverty, or the relationship between these groups. Measures of income inequality within a society can consider all members of society. Furthermore, they are explicitly group-level phenomena – inequality cannot exist at the individual level. A range of theoretical explanations for a link between income inequality and health have been suggested, including inequality reflecting lower levels of investment in health services and other public goods, higher levels of psychosocial stress or reduced social capital (Kawachi & Blakely 2001).

At the level of countries, there is some evidence that high inequality nations have lower life expectancy, higher infant mortality and lower age-at-death than more equal ones (Kawachi 2000). There is also, however, evidence that these studies suffer from limited comparability due to varying data collection methodologies between nations and their results – particularly those relating to non-infant mortality – have been contested (Deaton 2003).

Within the US, studies at the state and metropolitan levels have found an association between mortality and income inequality, even after adjusting for poverty rates (Kaplan et al. 1996, Kennedy et al. 1996, Lynch et al. 1998). A broad review considering studies of mortality, morbidity and health behaviour outcomes with at least one individual-level and one community-level explanatory factor reported that 23 of 25 studies reviewed

found at least one significant relationship between a measure of social environment and a measure of health status (Pickett & Pearl 2001).

A more focused review, looking only at studies that used an explicitly multilevel methodology to consider the confounding effect of individual income on the relationship between income inequality and health showed seven of the ten studies included to have found an independent effect for income inequality (Kawachi & Blakely 2001). While such relationships may be partially confounded by non-SES risk factors, a study controlling for a range of individual-level risk factors and median state-level income found a robust relationship between income inequality and self-reported health status in the US (Subramanian & Kawachi 2004).

However, another review of health and income inequality argued that no causal link between income inequality *per se* and poor health has been proven, but rather that the relationship demonstrated is one of inequality acting as a confounder, perhaps of a true relationship between psychosocial stress and illness (Deaton 2003). The precise relationship between income inequality and health outcomes is important since only by understanding this can one efficiently target policy change. If a causal link between income inequality and health is lacking, as Deaton argues, reducing inequality may lead to health improvements, but only as a second-order effect via the true causal link.

Given that these studies were all conducted in the more developed world, their relevance in poorer settings may be questioned: in situations of widespread absolute poverty, inequality may be of only secondary importance (Kawachi & Blakely 2001). It is therefore interesting to note that more than one study has found the impact of inequality to be strongest amongst the poorest parts of those populations studied, those closest to being in a comparable position to the majority of those living in South Africa (Kawachi 2000, Lochner et al. 2001). The only multilevel study to date to consider a less-developed nation – Chile – also found a significant relationship between income inequality as measured by the Gini coefficient and self-reported poor health (Subramanian et al. 2003). The effect appeared to be protective for those with least inequality.

3.4.3. Employment

Evidence exists for an inverse association between unemployment and a wide range of health outcomes (Jin et al. 1995; Dooley et al. 1996; Mathers & Schofield 1998). Of all the measures of SES, however, employment is the most prone to concerns regarding causality, since there is clear evidence that ill-health often causes unemployment rather than *vice versa*. Some longitudinal studies attempt to adjust for this by selecting a population that is either exogenously been made unemployed (the closing-factory method), or about to enter the workforce (students leaving school) or through panel population surveys (Dooley et al. 1996). These methods are rarely employed, but those studies that have used these approaches have found robust relationships between unemployment and poor health.

At the aggregate level, in addition to these individual-level, compositional effects, there is also the possibility of a contextual effect of high unemployment rates through their impact on the community – either physically or psychologically. Interaction between the two levels is also possible.

3.4.4. Education

In contrast to employment status, measures of education cannot suffer from the problem of reverse causality among adults: once an individual has finished studying, educational achievement is invariant to changes in health status. This does not, however, entirely remove concerns about reverse causation. If adult health outcomes are determined in part by childhood, or even antenatal, health, as suggested by the lifecourse epidemiology literature (Barker 1994), then there is the potential for an individual's health to determine the amount of schooling they receive – a frequently sick child is likely to attend less school, and have poorer attainment, than a healthy one (Behrman 1996).

Bearing this in mind, there is mixed evidence that those with higher educational achievement levels have higher life expectancy and lower morbidity levels both in more and less developed country settings (Krieger et al. 1997). One US study found education to be the strongest individual socio-economic predictor of health outcomes (Winkleby et al. 1992), but another found it to be less important than other measures, such as occupational class (Davey Smith et al. 1998).

In South Africa skewed educational opportunities provided by the Apartheid system prior to 1994 lead to differential relationships between education and health among race groups (Case 2002). As a result, racial group status is likely to act as a confounder, or even an effect modifier, of any relationship in this country and will need to be taken into account in this study. This problem is similar to that experienced for race and sex in other countries, where those who are otherwise similar earn average salaries significantly lower than if they were white and male (Krieger et al. 1997).

3.4.5. Deprivation

While a link between income and health may exist, it seems more reasonable that an individual's capacity to overcome health (and other) challenges is correlated with the total stock of assets rather than the current flow. This is particularly true in countries where health insurance, whether public or private, does not operate widely and thus the individual equivalent – savings and fungible assets – is crucial. Comprehensive measures of wealth and access to goods and services are also likely to take into account many or all of the aspects of socio-economic position raised above (Lynch & Kaplan 2000). A widely used class of measures considering wealth are deprivation and asset indices (Krieger et al. 1997). These attempt to measure whether households are in possession of certain assets that are considered standard for survival in a given social environment. This has the drawback that they are not easily comparable between societies.

In the past two decades researchers in the United Kingdom (UK) have created several deprivation indices that represent area-level access to goods and services. The most commonly used is the Townsend index, which contains factors such as unemployment, lack of car and house ownership and overcrowding. This and other indices have been found to be positively associated with both population health and with mortality (Carstairs 1995; Gordon 1995; Ben-Shlomo et al. 1996). In the US a measure of material hardship, based on the Townsend index, and a broad-based index formed through principal component analysis (PCA) were found to be associated with cardio-vascular disease risk factors and all-cause mortality respectively (Cubbin et al. 2001;

Singh & Stahpush 2002).[3] The latter study contained variables believed to reflect both absolute and relative deprivation. The Care Need Index, a measure of neighbourhood social position used in Sweden has been found to be linked to poor health behaviours, long-term illness, cardiovascular disease and mortality (Malmström et al. 1999; Sundquist et al. 1999; Malmström et al. 2001; Sundquist et al. 2004).

Several of the indicators included in the above indices are unlikely to be relevant in poorer countries. A methodology for creating asset indices based on questions in the widely-used Demographic and Health Surveys is increasingly being used as model in less developed countries including South Africa (Filmer & Pritchett 2001; Booysen 2002). Using PCA this methodology creates an asset index based on those combinations of binary variables which most clearly differentiate different socio-economic groupings (divided by income, geography, etc.) in a given country's data.

Unlike the US index this asset index is based on all available data in the survey, rather than a rigorous evaluation of each item's theoretical and empirical relevance. Also unlike measures used in more developed countries the DHS indices were created at the household, rather than the area, level. This allows for a more nuanced analysis of the level at which deprivation may affect health. The use of a deprivation-based, rather than an income-based, measure of SES is likely to be more powerful in a nation such as South Africa, where between a third and a half of all those seeking work are unable to find it and many communities have a majority of households that report no income (Myer et al. 2004).

[3] Factor Analysis (FA) is a methodology that seeks to uncover the latent structure in a dataset without using a specific dependent variable. A factor is a combination of variables that have strong explanatory power in the dataset. Principal Component Analysis is a form of FA that creates combinations of variables that jointly account for both common and unique variance in the dataset. It first creates a *factor* that includes the maximum variance from the overall dataset, then removes those variables comprising this *factor*, and repeats the process. The proportion of the original variance explained by the *factors* is referred to as the factor loading, and will decline with successive *factors*. The final output is a reduced number of explanatory variables, equal to the number of *factors* believed to be important for explaining the variability of the dataset (Kim & Mueller 1978).

3.5. Socio-Economic Status and Tuberculosis

As an infectious disease, tuberculosis is by its nature a social illness. This is reflected in risk factors raised above such as overcrowding. It additionally has a long history of being associated with poor socio-economic conditions, both in South Africa (Strebel & Seager 1991; Packard 1987) and internationally (Dubos & Dubos 1992, Gandy 2003; van Helden 2003). This may be the result of poor living conditions exacerbating other risk factors, or due to them limiting access to care (Ho 2004; Gandy 2002; Castro 2003). It therefore seems natural to consider the social context of persons affected by the disease. SES is, however, often left out of an analysis of risk factors for tuberculosis disease due to its distal nature in the causative pathway. While poor nutrition or a history of mine work is often included in biomedical studies of the disease, the causal agents of such exposures are frequently ignored.

3.5.1. Socio-economic risk factors in low tuberculosis incidence settings

Numerous studies in the US and UK have studied the ecologic impact of selected socio-economic measures on tuberculosis at various area levels. None of these studies adjusted for other known, non-SES risk factors for the disease, but they found tuberculosis to be associated with high levels of publicly provided housing, free school meals and poverty, low levels of education and car ownership, and high scores on the Jarman, Carstairs and Townsend deprivation indices (Spence et al. 1993; Parslow et al. 2001; Krieger et al. 2003). One US study also found an unadjusted bivariate association between the Gini coefficient measure of inequality and tuberculosis.

A UK study of the predictive power of the constituent elements of the Jarman index in Liverpool found the only meaningful SES-based element to be the unemployment rate (Toque et al. 1999). Even this fell from being the strongest predictor of tuberculosis in the early 1980s to insignificance by the early 1990s. This may be related to the finding that the only significant predictor of change in tuberculosis rates in London between 1982 and 1991 was the change in the employment rate (Mangtani et al. 1995).

The potential importance of including non-SES variables in an analysis of the impact of SES risk factors is illustrated by two ecologic studies from Hong Kong. Bivariate analyses of SES measures and tuberculosis found both low educational achievement and

high poverty in a given area to be associated with higher disease rates (Chan-Yeung et al. 2005). A multivariate analysis of the same tuberculosis dataset from an overlapping period in the same geographic area, however, found that notification rates were not associated with measures of SES, including education and income, after adjusting for immigration and marriage rates and the proportion of the population living in rooms/bedsits (Leung et al. 2004).

In lower income settings socio-economic factors appear to play a stronger role. A case-control study in Estonia found tuberculosis cases to be significantly more likely to have minimal or no income, be unemployed and live in shared residence (Tekkel et al. 2002). The study also considered non-SES risk factors but did not conduct a multivariate analysis of the data. An ecologic study of risk factors in Olinda, Brazil found that two different measures of deprivation, one created through factor analysis of census variables, were both significantly associated with tuberculosis in bivariate analyses (Souza et al. 2000).

3.5.2. Socio-economic risk factors in high tuberculosis incidence settings

In settings with low tuberculosis incidence those infected with the mycobacterium and affected by the disease are among the most deprived and isolated members of society. In higher incidence settings the disease is generalised to much of the population. As a result, risks for infection, and hence for disease, may be different.

Although unemployment is classically considered a risk factor for illness, it is possible that in a high-incidence setting this is offset by a raised risk of coming into contact with an infectious case for employed persons. This latter effect is likely to be magnified by the raised risk of employed persons being infected with HIV in high-HIV-incidence settings. A cohort study in rural northern Malawi found non-farmers and those with better housing to have higher risk of tuberculosis; this may have been related to those of higher SES also having higher risk of HIV (Glynn et al. 2000). In a lower HIV-prevalence setting farming and skilled manual labour were positive risk factors for tuberculosis disease (Lienhardt et al. 2005). On the other hand, South African ecological research on socio-economic risk factors, conducted in the low-SES Western

Cape areas of Ravensmead and Uitsig, found tuberculosis rates to be significantly positively correlated with the unemployment rate in an area (Munch et al. 2003).

3.6. Literature Summary

There is a large international body of literature suggesting a wide range of risk factors for tuberculosis disease, both intrinsic to individuals and arising from their social environment. Each of these should be considered in a South African context. Previous research in this country also suggests that the impact of Apartheid-era policies, both in restricting access to services and forming particular patterns of employment, may well have shaped South African risk factors into an unusual form.

Almost every measure of SES available has been inversely associated with health states and outcomes. Furthermore, many of them have been found to be risk factors for tuberculosis in high income settings. These measures of SES have included both individual- and group-level characteristics, the latter being both compositional and contextual in nature. There remains, however, a dearth of research on the relationship between SES and tuberculosis in low-income, high income-inequality and high tuberculosis-burden settings.

4. Methods

This study is a secondary analysis of the associations between self-reported tuberculosis disease and various personal, household and community-level variables believed to affect the risk of developing tuberculosis disease. The dataset was formed from two existing cross-sectional surveys, the 1998 South African Demographic and Health Survey and the 1996 South African census.

4.1. The 1998 South African Demographic and Health Survey

4.1.1. Survey methodology

The Demographic and Health Survey (DHS) series is an international programme of representative national surveys that has to date been conducted in 70 countries worldwide. The aim of the programme is to create a panel of data on population, health and nutrition in developing countries. By using regular repetitions of surveys and standardised methodologies, comparisons across space and time are possible. In South Africa the first DHS survey was conducted in 1998 as a collaboration between Measure DHS+, the international co-ordinating organisation, and the South African Medical Research Council (MRC). Funding was provided primarily by the South African Department of Health with assistance from Macro-International and USAID (*South African Demographic and Health Survey* 1999).

The 1998 South African Demographic and Health Survey (SADHS) consisted of three questionnaires, covering general household health, individual adult health and women's health. The first questionnaire was asked of one person in every household selected (n=12,860); the second to all adults aged 15 or older in every other household selected (n=14,928); the third to all women between 15 and 49 identified in all selected households (n=12,327). The response rate for eligible persons varied between 92 and 97 percent.

The SADHS involved a two-stage sampling process. A sampling frame was formed by dividing the country into the 86,200 enumerator areas (EAs) created for the 1996 census. Initially, the country was stratified by province and urban versus rural residence to

create 18 strata. At the request of the provincial government, the Eastern Cape was further stratified by its five health regions to give a total of 26 strata.

Among the aims of the SADHS were to provide accurate estimates of a range of health indicators for each of: (a) the four population groups; and (b) the nine provinces. The former of these created difficulties due to the small proportion and highly clustered nature of persons classified as Asian in the country. In order to ensure robust estimates, particularly of Asian women, the areas where they were most likely to reside were oversampled; this led to an oversampling of urban Gauteng (by < 1%) and urban KwaZulu-Natal (by 57%). The latter aim required an oversampling of the smallest provinces, it was achieved by sampling an equal number of individuals in each province, with the exceptions mentioned above for the Eastern Cape, Gauteng and KwaZulu-Natal. As a result of these manipulations the SADHS requires weighting factors for accurate analysis.

Having determined the appropriate strata sample sizes, the first stage of the sampling process involved the selection of an appropriate number of EAs, the primary sampling unit (PSU), for each stratum. This was done by probability proportional to size, based on the number of households (or alternatively census visiting points) in each EA. Any EA that contained no households (e.g. prisons or hostels) was excluded. The second stage of the sampling process was to select ten visiting points for each urban EA (n=690) and 20 visiting points for each rural EA (n=282) selected at the first stage. Visiting points were selected by systematic sampling from a sampling frame based on delimited maps (where available) or on-the-ground observation. As a result of this process, the samples were approximately self-weighting within each of the 26 strata under the assumption that household size does not vary systematically between EAs.

4.1.2. Variable selection and manipulation

In creating a dataset for this study all persons completing the adult health questionnaire (n=13,826) were considered as the population base. Relevant variables from the household questionnaire were appended to this dataset. The dependent variables in the analysis were derived from two questions relating to an individual's history of diagnosed tuberculosis. These questions were:

- Has a doctor or nurse or staff member at a clinic or a hospital told you that you had or have tuberculosis?
- If "yes", when was the first time that you were told that you had tuberculosis:
    - In the last twelve months
    - More than a year ago?

From these questions one categorical outcome variable was created to represent those who had been diagnosed with tuberculosis in the past twelve months and another to represent those who had ever been diagnosed with tuberculosis, including those diagnosed within the past twelve months. In this study these variables are referred to as **incident tuberculosis** and **lifetime prevalent tuberculosis**, respectively. While the former measure is more likely to be causally connected with the contemporaneous measures of potentially explanatory variables collected in the SADHS, the small number of positive observations raises the likelihood that such a measure will lack the statistical power necessary to distinguish true associations. The latter measure, which is highly correlated with the former but has more positive observations, should provide greater statistical power.

A number of independent, explanatory variables were used from the household and adult health questionnaires to cover the range of risk factors outlined in section 3.1 (Table 1). The methods used to create selected measures are detailed below.

4.1.2.1.Education

Educational achievement was categorised according to the highest level of schooling completed – primary, middle, secondary and tertiary – for exploratory analysis. For the multivariate analysis it was reclassified by years of schooling completed as given in the SADHS dataset.

4.1.2.2.Alcohol Abuse

The CAGE questionnaire is an internationally validated method for assessing an individual's level of alcohol dependency and risk of abuse. It has four questions:

- Have you tried repeatedly, and without success, to **C**ut down on the amount of alcohol you drink?
- Have people **A**nnoyed you by criticising your drinking habits?
- Have you ever felt bad or **G**uilty about your drinking?
- Do you need to have a drink as an **E**ye-opener in the morning (to steady your nerves or get rid of a hangover)?

Answering in the affirmative to more than one of these questions is considered an indicator of someone being 'at risk' of alcoholism and has been validated in a rural South African setting (Claasen 1999).

Table 1: Independent Variables from the South African Demographic and Health Survey 1998

Level	Household Questionnaire	Adult Health Questionnaire	No. of respondents
Individual	Age at last birthday		13,826
	Sex		13,826
	Highest education level achieved		13,720
		Race group	13,801
		Worked for payment in the past year	13,811
		Ever worked underground in a mine	13,733
		Ever smoked 100 cigarettes	13,816
		Ever drunk alcohol	13,796
		CAGE questionnaire	13,826
		Body Mass Index	13,546
		Had TB more than 12 months ago	13,771
Household	Urban residence		13,826
	Number of adults per bedroom*		13,675
	Affordability of meals		13,626
	Household asset score		13,561
All Data Available			13,043

Source: *South African Demographic and Health Survey* 1999.

4.1.2.3.Body Mass Index

The Body Mass Index (BMI) is measured as an individual's weight in kilograms divided by the square of their height in metres and aims to provide an indication of the health of

an individual. For this study an individual was classified in a binary fashion depending on whether their BMI was under or over 18.5, a common threshold for considering someone dangerously underweight. This created a proxy measure of malnutrition.

4.1.2.4.Adults per bedroom

This measure of household crowding was derived by dividing the number of adults in household by the number of rooms used for sleeping in the residence.

4.1.2.5.Affordability of Meals

There is a question in the 1998 SADHS asking how often the household goes hungry due to not being able to afford food. The responses are ordinal in the form: often, sometimes, seldom, never. The responses were included as a second proxy measure of malnutrition and also as a measure of SES.

4.1.2.6.Household asset score

The household asset index used was created by researchers at the Medical Research Council based on data in the SADHS (Booysen 2002). It used a factor analysis of nine questions broken down into 55 binary variables. These covered each household's:

- main source of drinking water;
- type of toilet facility;
- fuel used for cooking and heating;
- number of rooms used for sleeping;
- main material of floor and walls;
- affordability of meals; and
- ownership of specific assets.[4]

The factor analysis process identified 14 variables that were of particular explanatory power, based around three core factors. The three factors appeared to reflect the urban

[4] It should be noted that two of these questions were included as explanatory variables elsewhere in this study. Concerns as to overspecification can be allayed for the number of rooms used for sleeping, which was not used in the final asset index. Affordability of meals is included in the index as a binary term comparing those who never go hungry to all others, but the binary term alone has a correlation coefficient of -0.54 with the overall index suggesting it may individually provide additional information.

non-poor, rural poor and all poor persons respectively. Comparison of the 14 variable index to the 55 variable measure showed the smaller index to be both internally coherent (differentiating well between different classes of the population) and robust. It also showed consistency with alternative measures of poverty in South Africa based on income, expenditure and other asset measures. The combined index was used to divide all households into five ordered quintiles.

4.2. The 1996 South African National Census

4.2.1. Census methodology

The 1996 South African census was the first census carried out nationwide in South Africa following the extension of the franchise to all adults in 1994. The census was conducted in October and November 1996 with a post-enumeration survey undertaken a month subsequently to allow for miscounting due to persons absent being from all visiting points during the main census period (Statistics South Africa 2000b).

In order to organise the census South Africa was divided into 86,200 EAs, as mentioned in section 4.1, each consisting of 100 to 250 households. These EAs were then grouped together according to their situation within the 345 Magisterial Districts (MDs), which were the basic administrative unit of South Africa in 1996 (Statistics South Africa 1998).

For the purposes of this study a 10 percent random sample of the national census was utilised (South African Data Archive 2003). This was necessary since Statistics South Africa protects the privacy of South African residents by not making individual-level datasets available to the general public. The 10 percent sample dataset consists of two sets of data: the first is a systematic 10 percent sample of households recorded in the census (n=846,478); the second a combination of all persons living in the selected households plus an independent 10 percent sample of those living in non-household settings (n=3,621,201).

4.2.2. Variable selection and manipulation

The primary reason for using data from the 1996 census was to create representative, income-based measures of SES at higher levels of aggregation than the household,

specifically at the community level. Given the available geographical information from both the SADHS and the 10 percent sample of the census, the lowest level at which data could be matched was the MD.

The relevant variables selected from the census questionnaire were those asking respondents to provide details of their income and recent employment history. Responses to the income question were provided in bands, beginning at no income, and then rising in increasing increments. In order to calculate household incomes and measures of inequality it was necessary to assign each band a point value. Following Statistics South Africa's methodology this was set as the logarithmic mean of the maximum and minimum values in each band, with the exception of the lowest two and the highest one income bands (Statistics South Africa, n.d.). The bands and imputed values are given in Table 2. As a result of these assumptions, results arising from these data should be treated as indicative rather than precise predictors.

Table 2: Monthly Bands and Imputed Values for Census Income Data

Monthly Income Band	Imputed Point Values
None	None
R1 - R200	R133
R201 - R500	R350
R501 - R1000	R707
R1001 - R1500	R1225
R1501 - R2500	R1936
R2501 - R3500	R2958
R3501 - R4500	R3969
R4501 - R6000	R5196
R6001 - R8000	R6928
R8001 - R11,000	R9381
R11,001 - R16,000	R12,266
R16,001 - R30,000	R21,909
R30,001 or more	R60,000

Source: Statistics South Africa, n.d.

The value of each SES measure and for mobility for each of the 287 MDs contained in the SADHS sample were computed from the census sample and then imputed to each individual in the SADHS sample based on their MD of residence. Due to the

complexity of some of the calculations, particularly for the inequality measures, the use of sample weights was not possible. In order to maintain comparability between measures, weights were therefore not used in any analysis including census variables.

4.2.2.1.Measures of income poverty

Income poverty refers to the amount of money a person or household claims over a period of time and can be defined two ways: relatively or absolutely (Woolard & Leibbrandt 1999). In the former case the poverty line, below which one is considered to be poor, is set as a proportion of some measure of societal income, such as the mean or median value. In the latter case the line is set in terms of what is considered a minimum amount on which it is possible to live. The United Nations uses figures of this nature, particularly US$1 or US$2 per day, sometimes adjusted for the cost of living in a given country. Such figures often take into account minimal calorific values needed for meaningful living and other essential goods such as shelter and warmth. An alternative approach is to define the poorest *p* percent of the population as poor, in which case the proportion of persons who are poor never changes.

The selection of different methods can have a significant impact on who is considered poor – a comparison of 1993 South African measures of poverty found the poverty line to range from R105 per capita to R301.70 per adult equivalent per month (Woolard & Leibbrandt 1999). Figures diverge further if the reduced requirements of children and the economies of scale that may be gained from providing for multiple persons are taken into account. Evidence from the above study, however, found that different specifications of poverty lines made little difference to who was defined as poor.

In this study the simplest measure of poverty was used, the headcount ratio. This is the proportion of a given population that falls below a fixed poverty line. While this suffers from a simplification of poverty into a dichotomous variable, which does not appear to reflect the linear trend in health measured against income in other settings (Marmot 2002), it is the only measure that is easily computable from the income information collected by the census. As an alternative measure of compositional MD-level SES, unemployment rates were also calculated.

Four measures of income poverty were created from the datasets. These were formed by counting the proportion of all persons or households in the sample resident within each MD who fell into the following categories:

P1. Persons reporting income of less than R500 per month;

P2. Households reporting income per adult-equivalent of less than R500 per month;[5]

U1. Persons of working age unemployed by the expanded definition;[6]

U2. Households in which no-one was employed.

4.2.2.2.Measures of income inequality

The first measure of income inequality used was the Gini coefficient, which considers how inequitably the income of a population is distributed by measuring the proportion of total income earned by each individual. It is defined as the mean of the absolute differences between every observation and each other observation, standardised to a [0,1] scale. If all observations are ranked from poorest to richest this can be computed by the following formula:

$$\text{Gini} = \frac{2}{n^2 \bar{x}} \sum_{i=1}^{n} i(x_i - \bar{x})$$

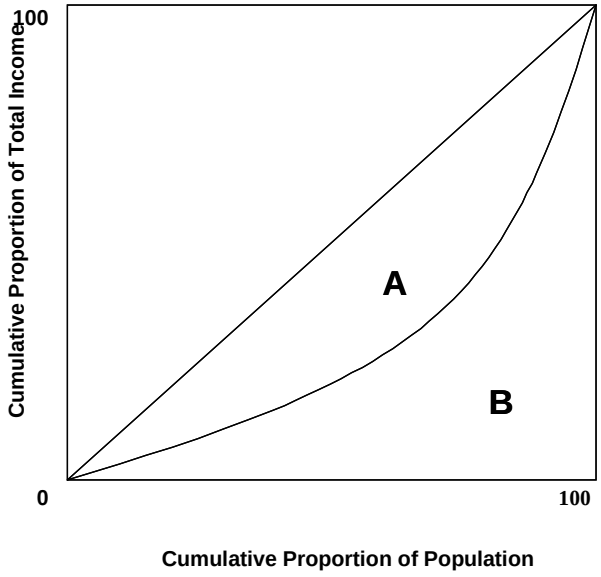
where n is the total population, x_i the value for observation i and $\bar{x}$ the mean value of all observations.

This is equivalent to plotting a curve of these ranked individuals with the cumulative proportion of the population on the x-axis and their cumulative proportion of the whole population's income on the y-axis (this is a Lorenz curve, the lower line in Figure 2). The closer the curve is to a 45-degree line, the more equally the population's income is distributed. The Gini coefficient can be calculated as the ratio of area A to areas A plus

[5] The adult-equivalent size of a household was calculated by adding the number of adults to half the number of children. No account was taken of potential economies of scale within households.
[6] The official definition required that a person: (a) had not worked in the past week; (b) was available to begin work within the next week; (c) had sought work in the past four weeks. The expanded definition required only (a) and (b). The 1996 census responses used were that an individual was either "Unemployed and looking for work" or "Unemployed, not looking for work, but would accept work".

B in Figure 2. A Gini coefficient of 0 indicates that everyone in the population has the same income; a coefficient of 1 indicates that one person has all the income. Therefore, the higher a Gini coefficient a population has, the more unequally income is distributed within it.

Figure 2: An Example of a Lorenz Curve



The second measure of inequality used was the Robin Hood index, a method of calculating what must be taken from the rich and given to the poor in order to equalise incomes. The method defines rich as those with income above the mean and poor as those below this figure. This is most often computed by ranking a population by income, dividing it into ordered deciles based on this ranking and then determining which deciles earn less than 10 percent of the total income in the population. The sum of the income earned by in all such deciles is computed as a proportion of the population's income and subtracted from the proportion of the population that such deciles constitute. The result of this figure is the Robin Hood index for an area: the percentage of all income that must be taken from those deciles earning more than 10% of the population's income, and given to the those deciles earning less than 10% of the

population's income, in order that each decile has exactly 10% of the total. Normally, the Robin Hood index is higher in more unequal populations.

These two measures provide different information and have both advantages and disadvantages. The Gini coefficient has the benefit of meeting various theoretical guidelines for measures of inequality, but the disadvantages of not being easily decomposable into subpopulations and of not being particularly intuitively comprehensible.[7] The Robin Hood measure is not generated from any statistical theory and does not have all the properties preferred of inequality measures, but is considered to be easier to interpret intuitively. Past research in the US suggests that the relationship between income inequality and mortality is not particularly sensitive to the measure of income used (Kawachi & Kennedy 1997). This study may help to determine whether the same is true in South Africa for tuberculosis.

The Robin Hood index (R1, R2, etc) and the Gini coefficient (G1, G2, etc) were each calculated at the MD-level across four populations:

1. The entire population of persons;
2. All households;
3. Persons in the expanded workforce;
4. Households in which someone in the expanded workforce resided.

4.2.2.3.Mobility

A secondary reason for using the 1996 census was to generate a measure of the social mobility of each community. For this purpose a variable was created that measured the proportion of all persons in the sample resident within each MD who had moved home in the past five years.

[7] In general inequality theory four properties are desired of measures of inequality (Dalton 1920):
- Dalton's transfer principle: a transfer from a richer to a poorer person should reduce inequality;
- Population principle: differences in population size should not affect inequality;
- Relative income principle: changes in absolute income levels should not affect inequality;
- Anonymity principle: variation in who is rich and who is poor should not affect inequality

While the Gini coefficient meets all these criteria the Robin Hood index is insensitive to all transfers within those persons above, or below, the mean level.

4.3. Analytic Methodology[8]

4.3.1. Data weighting

The MRC developed sample weightings at the individual and household level for the SADHS to account for the non-simple nature of the survey. In this analysis data were weighted for the adult health survey at the individual level, since this was the core dataset. The survey weights were identical within each of the 18 strata (by province and urban/rural location).[9]

4.3.2. Descriptive statistics and bivariate analysis

The relevant variables from the household questionnaire of the SADHS and the census were merged using the adult health questionnaire to form a single dataset. Data analysis was conducted using Stata version 9.0 (StataCorp; College Station, TX). True categorical data were described using proportions and 95 percent confidence intervals (CIs); those based on underlying continuous variables (quintile measures) were described using means and 95 percent CIs. Continuous variables were reported using means and 95 percent CIs or medians and interquartile ranges (IQR), adjusted for the form of the underlying survey. Given the expectation that some of the socio-economic measures might be very highly correlated with one another, a preliminary examination of potential multicollinearity was performed using correlation coefficients.[10]

The data were first described univariately in crude form and adjusted for the survey design in the case of variables derived from the SADHS. The SES measures derived from the census were described crudely at the MD level and as characteristics of the SADHS sample, weighted for the latter's survey design. Bivariate relationships were then explored between the two outcome measures – incident and lifetime prevalent tuberculosis – and each independent variable using logistic regression (which provide

[8] The 95% level of confidence was used through this study to evaluate statistical significance.
[9] The five rural and five urban strata of the Eastern Cape health districts were merged since the sampling strategy used for each was identical.
[10] Using a simple correlation coefficient is not valid for inference in the presence of non-normally distributed data, which much of this dataset was. However, given the small differences seen between this measure and a Spearman Rank coefficient in a few test cases, it was felt that this would provide sufficient guidance at this preliminary stage.

odds ratios (OR)) and rate ratios. The geographical distribution of tuberculosis cases was explored.

4.3.3. Multilevel modelling

As mentioned in Section 3.2, in order to ascertain the independent impact of risk factors acting at different levels of analysis, it is necessary to jointly enter terms into a regression equation (Snijders & Bosker 1999; Bryk & Raudenbush 1992). In a standard ordinary least squares regression an outcome is modelled as the result of various dependent variables and a random error term:

$$Y_i = \beta_0 + \beta_k X_{ik} + \varepsilon_i , \qquad \varepsilon_i \sim N(0,\sigma^2) \qquad (1)$$

where there are N observations (i=1,…,N), Y is the outcome, X is a matrix of K (k=1,…,K) potentially explanatory variables and the error term is normally distributed with zero mean and constant variance. In the case where each observation is a member of a group, and some additional explanatory variables exist only at the group level, this equation can be expanded to:

$$Y_{ij} = \beta_0 + \beta_k X_{ijk} + \beta_p Z_{jp} + \varepsilon_{ij} , \qquad \varepsilon_{ij} \sim N(0,\sigma^2) \qquad (2)$$

where Z is a matrix of these P group level variables (p=1,…,P) and there are J groups (j=1,…,J). Here both individual and group effects exist, but there is no allowance for the fact that group membership of individual i may have an impact on the way in which explanatory variable X_k affects the outcome. If an individual's characteristics are believed to be in any way affected by the group of which they are a member, then the regression model used needs to take account of this. One method of doing this involves modelling both levels jointly:

$$Y_{ij} = \beta_{0j} + \beta_{jk} X_{ijk} + \varepsilon_{ij} , \qquad \varepsilon_{ij} \sim N(0, \sigma_a^2) \qquad (3.1)$$

$$\beta_{0j} = \gamma_{00} + \gamma_{0kp} Z_{jp} + u_{0j} \qquad u_{0j} \sim N(0, \sigma_b^2) \qquad (3.2)$$

$$\beta_{kj} = \gamma_{k0} + \gamma_{jkp} Z_{jp} + u_{ij} \qquad u_{0j} \sim N(0,\sigma_c^2) \qquad (3.3)$$

Here both the intercept term β_0 and the coefficient terms β_k are modelled as a combination of a fixed component, a component related to group-level outcomes and a

random component. Fitting these three equations together and multiplying out, a multilevel, mixed-effects model can be reached:

$$Y_{ij} = \gamma_{00} + \gamma_{k0}X_{ijk} + \gamma_{0kp}Z_{jp} + \gamma_{jkp}Z_{jp}X_{ijk} + u_{0j} + u_{ij}X_{ijk} + \varepsilon_{ij} \quad (4)$$

In equation (4) there are two categories of effects. The first four, fixed effects are non-stochastic in that they are directly predictable from the explanatory variables. The last three, random effects are stochastic as they depend on a frequency distribution. The term u_{0j} allows for the intercept term in regression to vary by group while the term $u_{ij}X_{ijk}$ allows for the slope coefficient term to vary by group also. The term ε_{ij} is the standard random error term.

The multilevel, or hierarchical, linear model in equation (4) is a more flexible form of the standard, single-level model in equation (1) or the restricted two-level model in equation (2). By allowing for a more complex structure it is possible to take account of effects that might otherwise be hidden.

The use of multilevel modelling in public health is a relatively new phenomenon. It has, however, flourished in the field of social epidemiology, since its methodology is particularly appropriate to the study of the context in which the health of individuals is formed, allowing the simultaneous analysis of individual- and group-level risk factors, including their interactions (Diez-Roux 2000). Unlike single-level modelling the multilevel offers the possibility of analysing what proportion of the variation in health outcomes is due to individual-level risk factors and what proportion to membership of a given community (Merlo 2003). Two reviews of the use of multilevel studies in social epidemiology, looking at the relationship between SES and health outcomes have found a number of studies with varying levels of theoretical and empirical rigour (Pickett & Pearl 2001; Subramanian & Kawchi 2004).

This study was an analysis of data at three levels, allowing for clustering at a fourth level. Individuals were modelled to be nested within households, which were then nested within communities, represented by MDs. Clustering arose from the stratified nature of the primary data source, the SADHS, as described in section 4.1, but was simplified to the nine provinces since several MDs cross rural/urban boundaries. Given

this complex data structure, the analysis was simplified by including only group-level intercepts (u_{0j}) and the error term (ε_{ij}) as random-effects in the modelling process.

4.3.4. Multivariate analysis

All multivariate models were run as three-level multilevel models as described above, using the Generalised Linear Latent and Mixed Models (GLLAMM) add-in for Stata (Rabe-Hesketh et al. 2004), allowing for weightings as given in the SADHS dataset. All explanatory variables found to be associated with either dependent variable in bivariate analyses were included in the multivariate analysis with the exception of ever having consumed alcohol, since it was very highly correlated with a CAGE score greater than 1.

A preliminary analysis was conducted using all non-SES variables. Thereafter SES variables of rising aggregation (individual, household, MD) were modelled, first in an otherwise empty model, then adding all SES variables of lower aggregation and finally adding all non-SES variables. Given the high degree of correlation between the various MD-level SES measures these were not included jointly in any model.

4.3.5. Model checking

Due to the complexity of the model form, some simple model checks were not possible in this analysis. The final model form was, however, checked for multicollinearity among the explanatory variables by inspection of correlation coefficients. Outlying and potentially influential observations were plotted and examined to ensure that results were not being driven by a small number of non-representative observations. The influence of each MD on the model was considered using the Cook's Distance measure.

The variables in the final model were also modelled using an alternate formulation, where the third level of the model was the EA rather than the MD, allowing for clustering to be applied as intended in the survey, by province and urban/rural status, rather than just by province as in the main analysis. This was not used in the main analysis because it had the disadvantage of not including the MD as a level in the model. This technique yielded results that were similar to those seen in the main analysis, and thus for reasons of brevity only the main results are presented below.

5. Results

5.1. Descriptive Statistics

The initial dataset was composed of 13,826 observations. These comprised 18 strata and 966 clusters (EAs). 287 of the 345 MDs in the country were represented within the sample, with between three and 442 observations per MD.

5.1.1. Tuberculosis

In a crude analysis of the data there were 71 cases of tuberculosis reported as having been diagnosed in the previous 12 months. This was a rate of 516 cases per 100,000 person-years of observation (PYO). Allowing for the survey design effect this rate fell to 412 cases per 100,000 PYO (95% CI: 311-544) (Table 3). Three hundred and eighty-four individuals reported ever having been diagnosed with tuberculosis. This was a rate of 2788 cases per 100,000 persons surveyed. Allowing for the survey design effect this rate fell to 2362 per 100,000 persons, (95% CI: 2058-2709).

5.1.2. Biomedical, behavioural & social characteristics

The survey sample was distributed with declining frequency across the range of 15 to 95 years old, with a median age of 35 (IQR: 23-51). Almost three-fifths of the sample (59%) was female. These proportions were close to national figures. Three-quarters of the sample were identified as African, in line with the national population. Coloured persons were intentionally over-, and White persons under-, sampled.

Almost five percent of those interviewed had worked on a mine, of which two-thirds had worked on a gold mine. Thirty-two of the 71 incident cases (49%) reported having previously been diagnosed with the disease. Twenty-seven percent of respondents had ever smoked 100 or more cigarettes and 39% reported having ever drunk alcohol. One fifth of all respondents (48% of all persons who said they had ever drunk alcohol) gave two or more affirmative responses to the CAGE questionnaire. Ten percent of those interviewed were dangerously underweight for their height, with a BMI of less than 18.5.

Table 3: Descriptive Statistics for SADHS Variables

	Crude	Adjusted for survey design	
	%	%	95% CI
Diagnosed with tuberculosis in past year	0.52	0.41	(0.31 - 0.54)
Diagnosed with tuberculosis ever	2.79	2.06	(2.36 - 2.71)
Age			
15-29 years	39.0	39.1	(38.0 - 40.2)
30-44 years	26.7	27.5	(26.5 - 28.5)
45 years and over	34.3	33.4	(32.2 - 34.6)
Female	58.4	59.0	(58.0 - 60.0)
Race group			
African	75.8	76.3	(73.9 - 78.5)
Coloured	12.9	10.5	(9.1 - 12.1)
White	8.0	9.7	(8.1 - 11.4)
Asian	3.4	3.6	(2.7 - 4.7)
Urban Residence	56.1	62.0	(60.6 - 63.3)
Worked for payment in past 12 months	33.7	35.6	(34.2 - 37.1)
Education level reached			
No school year completed	13.8	12.4	(11.5 - 13.3)
Primary (1-5 years)	14.4	13.6	(12.8 - 14.5)
Secondary (6-11 years)	53.1	53.1	(51.8 - 54.5)
Secondary complete and above	18.6	20.9	(19.7 - 22.1)
Ever worked in a mine	4.94	4.20	(3.74 - 4.72)
Ever worked in a gold mine	3.30	2.72	(2.39 - 3.08)
Diagnosed with TB > 12 months ago	2.24	1.83	(1.57 - 2.14)
Ever smoked 100 cigarettes or more	26.7	26.1	(24.9 - 27.3)
Ever drunk alcohol	38.8	39.0	(37.7 - 40.3)
CAGE score of two or greater	18.5	17.1	(16.1 - 18.1)
BMI below 18.5	9.68	8.51	(7.94 - 9.12)
Number of adults per bedroom (mean)	1.47	1.50	(1.46 - 1.54)
Meals missed by household due to lack of funds			
Never	47.8	51.4	(49.1 - 53.7)
Seldom	5.1	4.9	(4.2 - 5.8)
Sometimes	34.4	32.4	(30.5 - 34.4)
Often	12.7	11.3	(10.1 - 12.6)
Household asset score quintile			
1 (poorest)	16.5	12.7	(11.4 - 14.2)
2	20.4	19.8	(18.1 - 21.7)
3	21.3	20.2	(18.6 - 21.9)
4	22.0	23.3	(21.2 - 25.6)
5 (richest)	19.8	23.9	(21.8 - 26.2)
Mobility of population in MD (mean) [a]	26.5	27.3	(26.5 - 28.0)

Figures are percentages unless otherwise stated. Relevant sample sizes given in Table 1.

[a] Taken from the 10% sample of the census.

The median level of crowding within respondents' houses was 1.33 adults per bedroom (IQR: 1-2), the mean 1.47. The households of over half of those questioned had gone without meals due to a lack of funds, and of these almost a quarter had done so often. Among those MDs included in the sample, an average of 26% of persons in the census sample had moved home within the past five years.

5.1.3. Socio-economic status

Only one-third of those questioned had worked for payment in the past twelve months (Table 3). One in seven (14%) respondents had completed no years of schooling, but the great majority (73%) of the sample had completed at least one year of secondary schooling. The respondents to the questionnaire were, on average, of higher asset-owning class than the population as a whole. Although the poorest and the richest quintile were undersampled in a crude analysis, after adjusting for sampling method the lowest two classes were underrepresented and a rising participation trend was seen across the five quintiles.

A sizable proportion of persons in South Africa were unable to obtain employment despite wanting it, and many individuals and households were absolutely poor. Table 4 presents the median values of the 287 MD-level observations for each measure in the left-hand columns, and the SAHDS-weighted mean values of the 13,826 imputed individual-level observations for each measure in the right-hand ones. The mean proportion of persons earning less than R500 per month in SADHS respondents' MDs was almost three-quarters, but this figure was far lower for households. The mean unemployment rate amongst those available for work was almost a quarter (24%), while an average of almost two-fifths (38%) of households had no-one in employment.

Both the Robin Hood and Gini measures of income inequality had extremely high average values. The average value of the Robin Hood index across MDs was 68.9% when calculated across all persons, and 48.6% when calculated across all households.[11]

[11] That is, 68.9% (or 48.6%) of all income in the MD must be taken from those deciles which each accounted for more than 10% of total income in the MD, and redistributed to all other deciles, in order for these two groups each have total income proportional to their population size after the redistribution.

More restrictive definitions of the eligible population lowered these figures and the trends were similar when the observations were based on weighted individuals in the SADHS sample. The Gini coefficients measured at the MD level were higher than those reported nationally in any country worldwide; although direct comparisons should be made with caution due to the imputed nature of income values in this dataset (see section 4.2.2). The average Gini coefficient measured across all persons was 87% and across all households it was 68%. These figures fell as the eligible population was restricted.

Table 4: Descriptive Statistics for Census-Based Group-Level SES Measures

	Magisterial Districts (n=287)		Weighted individuals (n=13,826)	
	Median	(IQR)	Mean	(95% CI)
Headcount ratio [a]				
P1: Persons <R500 per month	83.8	(71.8 - 90.3)	74.4	(73.7 - 75.1)
P2: Households <R500 per adult-equivalent per month	16.5	(10.9 - 24.5)	17.3	(16.8 - 17.8)
Unemployment rate [a]				
U1: Persons expanded rate	23.1	(17.4 - 27.9)	23.7	(23.2 - 24.1)
U2: Households, no employment	35.4	(24.7 - 58.2)	38.4	(37.4 - 39.4)
Robin Hood index [b]				
R1: All persons	68.9	(64.8 - 74.9)	67.9	(67.5 - 68.3)
R2: Persons in the expanded workforce	56.5	(49.9 - 62.7)	53.2	(52.7 - 53.7)
R3: All households	48.6	(45.0 - 52.1)	46.1	(45.7 - 46.5)
R4: Households with expanded workforce member	48.1	(43.9 - 52.3)	45.3	(45.0 - 45.7)
Gini coefficient [b]				
G1: All persons	87.1	(83.5 - 90.7)	84.9	(84.6 - 85.2)
G2: Persons in the expanded workforce	74.1	(68.6 - 79.4)	70.9	(70.5 - 71.3)
G3: All households	67.6	(63.9 - 71.3)	65.5	(65.1 - 65.8)
G4: Households with expanded workforce member	66.7	(62.5 - 71.2)	64.4	(64.0 - 64.8)

[a] Figures are the proportion of individuals in each MD falling into the categories shown.

[b] Figures are the value of the index/coefficient as described in section 4.2.2.2 above.

5.2. Bivariate Analysis

5.2.1. Geography

Incident and lifetime prevalent cases of tuberculosis were non-randomly distributed geographically. At the lowest level, two households had two incident cases of disease in the past year; for lifetime prevalent tuberculosis there were 18 clusters of two cases

within a single household and one cluster of three. At the MD level the number of lifetime prevalent cases ranged from zero (n=129) to 25, while incident tuberculosis cases included one cluster of 11 cases, one of 4, one of 3 and five of 2.

Table 5: Weighted Tuberculosis Rates by Geographic Strata

	Number of respondents	Incident TB per 100,000 PYO		Lifetime prevalent TB	
	(n=13,771)	Rate	95% CI	%	95% CI
Western Cape					
Urban	1006	397	(152 - 1040)	2.29	(1.50 - 3.48)
Rural	139	724	(105 - 5118)	6.48	(2.39 - 17.56)
Eastern Cape					
Urban	1311	1,220	(675 - 2208)	4.35	(3.11 - 6.07)
Rural	2056	924	(616 - 1385)	5.45	(4.38 - 6.77)
Northern Cape					
Urban	910	442	(167 - 1140)	2.20	(1.29 - 3.75)
Rural	347	284	(0 - 2051)	5.19	(3.06 - 8.77)
Free State					
Urban	852	352	(115 - 1073)	1.64	(0.92 - 2.94)
Rural	342	586	(152 - 2233)	3.51	(1.82 - 6.75)
KwaZulu-Natal					
Urban	1468	477	(210 - 1087)	1.98	(1.22 - 3.19)
Rural	571	525	(176 - 1563)	3.33	(1.74 - 9.86)
North West					
Urban	458	0		0.87	(0.22 - 3.43)
Rural	775	516	(111 - 2397)	1.55	(0.81 - 2.97)
Gauteng					
Urban	1088	92	(13 - 651)	1.38	(0.86 - 2.21)
Rural	20	0		0	
Mpumalanga					
Urban	438	0		2.28	(1.24 - 4.20)
Rural	798	500	(197 - 1271)	1.75	(0.96 - 3.20)
Limpopo					
Urban	195	0		2.57	(1.07 - 6.17)
Rural	997	200	(51 - 791)	1.10	(0.66 - 1.86)
Design-adjusted Pearson's test for independence:					
F (7.50, 7108.20)		1.33 (p=0.23)			
F (8.96, 8494.23)				3.93 (p=0.001)	

Table 5 displays a breakdown of tuberculosis rates by the eighteen strata of the dataset, adjusted for the SADHS sampling methodology. Although the numbers of cases per cell were low in the column for incident tuberculosis, and thus the confidence intervals are

wide, the highest rates of reported disease appeared to be in the Eastern Cape where the overall rate was over 1000 cases per 100,000 PYO. Other areas with high rates include urban Northern Cape, rural Free State and KwaZulu-Natal, and rural Western Cape, although the sample size here was particularly small.

Lifetime prevalent tuberculosis was similarly geographically patterned. The highest rates of lifetime disease were found in the rural Western Cape, Northern Cape, Free State and KwaZulu-Natal, and throughout the Eastern Cape. The lowest rates of disease were found in Gauteng, North West province and rural Limpopo. The overall pattern of lifetime disease did not appear to be randomly distributed across the 18 strata ($p<0.001$).

Those living in rural locations were approximately 50% more likely both to have incident or lifetime prevalent tuberculosis, although this was only significant in the latter case ($p=0.009$).

5.2.2. Biomedical, social & behavioural characteristics

No statistically significant relationships between age, sex and incident tuberculosis were seen (Table 6). Relative to the reference category, women aged under 30, women aged 30 to 44 years old were least likely to have incident tuberculosis (184 per 100,000 PYO) while men aged 30 to 44 were most likely to have been recently diagnosed (829 per 100,000 PYO).

Due to differential time at risk for disease, interpretation of the relationships between age, sex and lifetime prevalent tuberculosis is difficult. Relative to women aged under 30, women aged over 45 and men aged over 30 were at statistically significantly higher risk of having ever had tuberculosis. Relative to women of the same age, men aged 30 to 45 (OR: 2.13; 95% CI: 1.38-3.28) and aged over 45 (OR: 1.85; 95% CI: 1.32-2.60) were more likely to have had tuberculosis.

African and Coloured respondents appeared to be at raised risk of tuberculosis, compared to White and Asian individuals. African and Coloured persons had a similar risk of having incident tuberculosis (468 and 480 per 100,000 respectively) while that for Asians was far lower (203 per 100,000) and no White persons reported having been diagnosed in the past year.

Table 6: Bivariate Relationships between Non-SES Variables and Tuberculosis

`	Incident TB		Lifetime Prevalent TB	
	Odds Ratio	95% CI	Odds Ratio	95% CI
Female				
15-29 years	1	(reference)	1	(reference)
30-44 years	0.46	(0.15 - 1.40)	1.45	(0.89 - 2.36)
45 years and over	1.17	(0.53 - 2.58)	1.95	(1.27 - 2.97)
Male				
15-29 years	0.86	(0.34 - 2.22)	0.61	(0.33 - 1.13)
30-44 years	2.08	(0.91 - 4.76)	3.07	(1.92 - 4.92)
45 years and over	0.90	(0.38 - 2.14)	3.61	(2.30 - 5.66)
Race group				
African	1	(reference)	1	(reference)
Coloured	0.98	(0.42 - 2.30)	1.64	(1.15 - 2.32)
Asian	0.43	(0.16 - 1.39)	0.47	(0.06 - 3.06)
White	0		0.45	(0.21 - 0.93)
Urban residence	0.66	(0.37 - 1.15)	0.69	(0.52 - 0.91)
Worked for payment in past 12 months	0.57	(0.33 - 0.99)	0.65	(0.50 - 0.84)
Education level reached				
No school year completed	1	(reference)	1	(reference)
Primary (1-5 years)	2.14	(1.00 - 4.61)	1.09	(0.79 - 1.51)
Secondary (6-11 years)	0.91	(0.45 - 1.83)	0.54	(0.40 - 0.72)
Secondary complete and above	0.29	(0.08 - 1.03)	0.25	(0.15 - 0.43)
Ever worked in a mine				
No	1	(reference)	1	(reference)
Yes, non-gold mine	0		1.45	(0.55 - 3.84)
Yes, gold mine	2.39	(0.94 - 6.08)	2.71	(1.80 - 4.10)
Diagnosed with TB > 12 months ago	46.0	(26.1 - 81.1)		
Ever smoked 100 cigarettes	2.28	(1.31 - 3.97)	2.36	(1.86 - 2.99)
Ever drunk alcohol	1.69	(0.98 - 2.88)	2.45	(1.89 - 3.17)
CAGE score of two or greater	2.90	(1.64 - 5.12)	3.42	(2.69 - 4.36)
BMI below 18.5	4.54	(2.55 - 8.09)	3.16	(2.37 - 4.20)
Adults per bedroom	1.25	(1.03 – 1.53)	1.05	(0.93 - 1.19)
Meals missed due to lack of funds				
Never	1	(reference)	1	(reference)
Seldom	3.63	(1.05 - 12.54)	1.62	(0.80 - 3.24)
Sometimes	2.05	(1.10 - 3.80)	2.23	(1.66 - 3.01)
Often	2.92	(1.34 - 6.35)	2.42	(1.68 - 3.49)
Household asset score quintiles				
1 (poorest)	1.77	(0.83 - 3.76)	2.28	(1.58 - 3.31)
2	0.91	(0.42 - 1.97)	1.40	(0.89 - 2.20)
3	1	(reference)	1	(reference)
4	0.79	(0.34 - 1.85)	1.16	(0.77 - 1.72)
5 (richest)	0.14	(0.03 - 0.65)	0.65	(0.40 - 1.05)
Mobility within Magisterial District	0.36	(0.04 – 2.47)	0.34	(0.13 – 0.91)

When White and Asian persons were combined into a single category, to raise the discriminatory power of the analysis, their risk of having incident tuberculosis was one-ninth that of Africans (OR: 0.12, p=0.034). The only difference between incident and lifetime prevalent tuberculosis risk was that Coloured persons were significantly more likely to have ever been diagnosed with tuberculosis than African persons (p=0.006), despite having similar median sample ages (Coloured: 36; African: 34).

Those who had ever worked on a gold mine had an incident tuberculosis rate of 940 per 100,000 PYO compared to 396 per 100,000 PYO among those who had never worked on a mine; the difference was not statistically significant. This ratio was slightly higher for lifetime prevalent tuberculosis and was statistically significant (p<0.001). A past history of tuberculosis was an extremely strong predictor of incident tuberculosis; the rate of incident tuberculosis among those who reported a previous episode was 9,814 per 100,000 PYO compared to 234 per 100,000 PYO for the rest of the sample.

Ever having smoked 100 cigarettes, or having ever drunk alcohol, were both significantly associated with reporting incident and lifetime prevalent tuberculosis. A stronger predictor of both outcomes than having ever drunk alcohol was giving two or more positive responses on to the CAGE questionnaire. Persons with a BMI of less than 18.5 were almost five times as likely to have incident tuberculosis as the rest of the sample, and more than three times as likely to have lifetime prevalent tuberculosis.

A one-person increase in the number of adults per bedroom was significantly associated with a 25% rise in risk of incident tuberculosis (p=0.028) in the bivariate analysis, but was not significantly associated with lifetime prevalent disease (p=0.46).

Compared to those who had never had to miss meals, all other persons were at significantly raised risk of having incident tuberculosis and all bar those who ‘seldom’ had to miss meals were significantly more likely to have ever had tuberculosis. There was a trend across the four categories for both outcomes (p<0.001). Collapsing the variable into a binary measure of having ever had to miss meals gave a significant predictor of both incident (OR: 2.41, p=0.004) and lifetime prevalent (OR: 2.21, p<0.001) disease.

A higher degree of mobility within the MD in which an SADHS respondent was living was associated with an increased risk of either outcome. In the case of incident tuberculosis this was not significant (p=0.28), but in the case of lifetime prevalent disease it was (p=0.032).

5.2.3. Individual- and household-level SES

Recent employment was significantly associated with reduced levels of both incident (p=0.046) and lifetime prevalent (p=0.001) tuberculosis (Table 6).

Incident tuberculosis was positively associated with having one to five years of schooling and inversely associated with having more than five years of schooling. A non-parametric test for trend across the four schooling categories was negative and significant (p<0.001). The same pattern was seen for lifetime prevalent tuberculosis, except that primary schooling was not significantly associated with disease and the negative trend was stronger (p<0.001).

Based on the asset index, the wealthiest quintile of individuals significantly less likely to have experienced incident tuberculosis (p=0.012) and the poorest quintile significantly more likely to have experienced lifetime prevalent tuberculosis (p<0.001), compared to the third quintile. The differences between the middle three quintiles were not significant for either outcome. There was an overall negative trend across wealth for both outcomes (p<0.001).

5.2.4. Magisterial district-level absolute poverty and unemployment

In bivariate analyses both measures of MD-level headcount poverty were positively associated with tuberculosis disease (Table 7). For incident tuberculosis the individual-based measure was most strongly associated, with a 10% increase in the MD measure of this being associated with a 26.6% increased risk of disease (p=0.049). For lifetime prevalent tuberculosis individual-based poverty was again significantly associated with the outcome, but household-based poverty was also associated, with a 10% increase in mean MD poverty being associated with a 19.7% rise in risk of tuberculosis (p=0.003).

The unemployment rate of the MD in which a respondent was resident was inversely, but not significantly, associated with incident and lifetime prevalent tuberculosis. The

proportion of households with no working member was, however, positively associated with both outcomes, significantly in the case of lifetime prevalent disease. A 10% rise in the proportion of households in an MD without an employed member was associated with a 9.5% rise in risk of incident tuberculosis for a resident of that MD.

Table 7: Bivariate Relationships between Tuberculosis and SES Variables

	Incident TB		Lifetime prevalent TB	
	Odds Ratio	95% CI	Odds Ratio	95% CI
Headcount ratio				
P1: Persons <R500 per month	1.26	(1.00 - 1.58)	1.17	(1.06 - 1.30)
P2: Households <R500 per adult-equivalent per month	1.16	(0.87 - 1.56)	1.20	(1.05 - 1.37)
Unemployment rate				
U1: Persons expanded rate	0.90	(0.58 - 1.39)	0.96	(0.79 - 1.17)
U2: Households, no employment	1.09	(0.98 - 1.23)	1.09	(1.03 - 1.15)
Robin Hood index				
R1: All persons	1.12	(0.76 - 1.66)	1.23	(10 - 1.51)
R2: Persons in the expanded workforce	1.22	(0.93 - 1.61)	1.19	(1.04 - 1.36)
R3: All households	1.32	(0.78 - 2.23)	1.60	(1.23 - 2.09)
R4: Households with expanded workforce member	1.43	(0.90 - 2.25)	1.60	(1.27 - 2.04)
Gini coefficient				
G1: All persons	1.44	(0.85 - 2.43)	1.42	(1.10 - 1.83)
G2: Persons in the expanded workforce	1.37	(0.97 - 1.92)	1.29	(1.09 - 1.53)
G3: All households	1.37	(0.86 - 2.17)	1.50	(1.17 - 1.91)
G4: Households with expanded workforce member	1.40	(0.91 - 2.16)	1.47	(1.18 - 1.83)

Odds ratios are for a 10% change in the relevant SES variable.

5.2.5. Magisterial district-level relative poverty

All measures of the Robin Hood index were positively, but not significantly, associated with incident tuberculosis. The Robin Hood indices calculated across individuals had a similar level of association with lifetime prevalent tuberculosis as they did with incident disease, but were more significant. For example, when calculated from all persons (R1), a 10% increase in the index was associated with a 22.7% rise in risk of lifetime incident tuberculosis.

The Robin Hood indices calculated across households were very much more strongly associated with lifetime tuberculosis than were those calculated across individuals, with a 10% rise in either index being associated with a 60.5% rise in risk of tuberculosis.

All four Gini coefficients were positively associated with both incident and lifetime prevalent tuberculosis; although only the lifetime prevalent associations were significant, the point estimates of the measures were broadly similar for both sets of associations. For example, for the broadest definition (G1), a 10% rise in the Gini coefficient in an MD was associated with a 43.6% rise in risk of incident tuberculosis and 41.6% rise in risk of lifetime prevalent disease.

5.3. Multivariate Models

The initial SADHS dataset of 13,826 was reduced to a balanced panel of 13,043 for multivariate analyses in order to allow for consistent comparison between regressions. This excluded two incident and 15 prevalent cases of tuberculosis. Based on exploratory analysis and theory,[12] and in an effort to simplify to the analysis, household-based, MD-level measures of SES are reported in preference to individual-based ones. Results using individual-based measures were not qualitatively different.

5.3.1. Non-SES risk factors

Most of the relationships seen in the multivariate model of non-SES variables and tuberculosis reflected those seen in the bivariate regressions (Table 8).[13] In several cases the size of the point estimate fell, and in some cases associations were no longer statistically significant. Specifically, race category, history of working in a gold mine, rural residence and MD-level mobility ceased to be predictive of lifetime prevalent disease, while missing meals ceased to be predictive of incident tuberculosis.

Consistent non-SES predictors of tuberculosis were smoking, a CAGE score >1 and a low BMI. Additionally being African or Coloured, compared to being White or Asian, or being male were predictive of incident disease, and being older or having missed meals due to lack of funds were statistically significant predictors of lifetime prevalent disease.

[12] A household is often described as a group of people who share resources, those who 'eat from the same pot'. If this assumption holds, then differences in income between households should be more directly related to outcomes of any kind than differences between individuals.

[13] Four variables – ever worked in gold mine, urban residence, number of adults per bedroom and proportion of MD moved in past 5 years – were not significantly associated with tuberculosis in any regressions in section 5.3, but were retained for completeness. They are not shown to simplify the tables.

Table 8: Multivariate Relationships between Non-SES Variables and Tuberculosis

Dependent Variable	Incident	Lifetime Prevalent
Age	1.00 (0.98 - 1.03)	1.03 (1.01 - 1.04)
Male	0.64 (0.42 - 0.98)	1.05 (0.79 - 1.39)
Coloured vs. African	0.81 (0.19 - 3.46)	1.82 (0.66 - 5.02)
White/Asian vs. African	0.14 (0.05 - 0.38)	0.68 (0.25 - 1.82)
Ever worked in gold mine	1.52 (0.41 - 5.70)	1.32 (0.83 - 2.10)
Ever smoked 100 cigarettes or more	2.29 (1.20 - 4.34)	1.65 (1.04 - 2.60)
CAGE score greater than 1	2.01 (1.22 - 3.33)	2.51 (1.88 - 3.35)
BMI under 18.5	4.28 (1.82 - 10.04)	3.03 (2.06 - 4.46)
Urban Residence	0.68 (0.27 - 1.74)	0.83 (0.59 - 1.19)
Adults per bedroom	1.17 (0.85 - 1.60)	1.00 (0.85 - 1.16)
Missed meals due to lack of funds	1.66 (0.72 - 3.83)	2.09 (1.11 - 3.92)
Proportion of MD moved in past 5 yrs	1.36 (0.11 - 17.42)	0.62 (0.16 - 2.48)

Each relationship in this table is adjusted for all other variables shown, and also for: Ever worked in gold mine; Urban Residence; Adults per bedroom; Proportion of MD moved in past 5 years. 95% CIs shown in parentheses.

5.3.2. Individual and household SES risk factors

Individual measures of SES that were predictors of tuberculosis at the bivariate level generally remained predictors when included in multivariate regressions (Table 9). An additional year of education was associated with an 11% reduction in the risk of incident and lifetime prevalent disease, while having worked in the past twelve months was associated with a 22-28% lower risk.[14] These point estimates were stable in the case of

[14] As noted in section 3.4, reverse causality may have a role to play in the result for incident tuberculosis, but such fears may be assuaged by the similarity with the relationship for lifetime prevalent disease, where such causality is only possible in the minority of lifetime prevalent cases that were diagnosed in a year of the survey.

employment, attenuating by approximately 50% in the case of education, when non-SES measures from Table 8 were added to model. Although several of these variables were not statistically significant for incident tuberculosis, the values of the odds ratios were in both cases similar to the, significant, values for lifetime prevalent tuberculosis.

The household-level asset score in quintiles was inversely associated with tuberculosis disease (Table 10). Being in the lowest asset quintile was (non-significantly) associated with a 48% increased risk of incident disease relative to the middle quintile and being in the highest asset quintile was significantly associated with a 75% reduced risk of incident disease, even after adjusting for individual-level SES and non-SES covariates (column 3).

Table 9: Multivariate Relationships between Individual-Level SES and Tuberculosis

Dependent Variable	Incident TB		Lifetime Prevalent TB	
Column number	*1*	*2*	*3*	*4*
Years of Education	0.89 (0.84 - 0.95)	0.94 (0.88 - 1.01)	0.886 (0.864 - 0.908)	0.96 (0.93 - 0.99)
Worked in last 12 months	0.72 (0.44 - 1.18)	0.73 (0.47 - 1.12)	0.781 (0.631 - 0.967)	0.72 (0.52 - 0.98)
Age		1.00 (0.96 - 1.03)		1.02 (1.01 - 1.03)
Male		0.69 (0.46 - 1.02)		1.11 (0.84 - 1.47)
Coloured vs. African		0.85 (0.20 - 3.56)		1.90 (0.70 - 5.15)
White/Asian vs. African		0.19 (0.06 - 0.56)		0.84 (0.32 - 2.22)
Ever smoked 100 cigarettes or more		2.26 (1.20 - 4.24)		1.66 (1.03 - 2.66)
CAGE score greater than 1		2.00 (1.26 - 3.19)		2.52 (1.867 - 3.40)
BMI under 18.5		3.99 (1.72 - 9.24)		2.85 (1.92 - 4.22)
Missed meals due to lack of funds		1.52 (0.66 - 3.48)		1.93 (1.00 - 3.74)

Columns 2 and 4 are also for adjusted for: Ever worked in gold mine; Urban Residence; Adults per bedroom; Proportion of MD moved in past 5 years. 95% CIs shown in parentheses.

Regressions for lifetime prevalent tuberculosis gave similar results for the lowest asset quintile, being significantly associated with a 67% increased risk of disease. Being in the highest asset quintile was, however, only weakly associated with this outcome. In all cases the odds ratios became closer to the null hypothesis of no effect as more explanatory variables were added to the models.

Table 10: Multivariate Relationships between Household-Level SES and Tuberculosis

Dependent Variable	Incident TB		
Column number	*1*	*2*	*3*
Asset score quintile 1 (poorest)	2.04 (0.94 - 4.40)	1.71 (0.84 - 3.47)	1.52 (0.89 - 2.60)
Asset score quintile 2	1.00 (0.43 - 2.30)	0.94 (0.41 - 2.13)	0.85 (0.42 - 1.72)
Asset score quintile 3	1 (reference)	1 (reference)	1 (reference)
Asset score quintile 4	0.70 (0.41 - 1.18)	0.76 (0.44 - 1.30)	0.70 (0.36 - 1.39)
Asset score quintile 5 (richest)	0.14 (0.04 - 0.44)	0.17 (0.06 - 0.54)	0.25 (0.12 - 0.52)
Years of education		0.93 (0.89 - 0.98)	0.96 (0.90 - 1.01)
Worked in last 12 months		0.84 (0.51 - 1.37)	0.73 (0.47 - 1.15)
Age			1.00 (0.97 - 1.03)
Male			0.70 (0.47 - 1.04)
Coloured vs. African			1.10 (0.22 - 5.40)
White/Asian vs. African			0.41 (0.12 - 1.41)
Ever smoked 100 cigarettes or more			2.18 (1.18 - 4.03)
CAGE score greater than 1			1.96 (1.21 - 3.16)
BMI under 18.5			3.94 (1.68 - 9.212)
Missed meals due to lack of funds			1.23 (0.59 - 2.60)

Column 3 is also for adjusted for: Ever worked in gold mine; Urban Residence; Adults per bedroom; Proportion of MD moved in past 5 years. 95% CIs shown in parentheses.

Dependent Variable	Lifetime Prevalent TB		
Column number	*4*	*5*	*6*
Asset score quintile 1 (poorest)	2.40 (1.45 - 3.97)	1.84 (1.11 - 3.07)	1.67 (1.11- 2.53)
Asset score quintile 2	1.29 (0.92 - 1.80)	1.18 (0.84 - 1.65)	1.03 (0.73- 1.45)
Asset score quintile 3	1 (reference)	1 (reference)	1 (reference)
Asset score quintile 4	1.04 (0.55 - 1.98)	1.18 (0.63 - 2.223)	0.93 (0.56 - 1.53)
Asset score quintile 5 (richest)	0.57 (0.28 - 1.18)	0.83 (0.40 - 1.74)	0.81 (0.48 - 1.38)
Years of education		0.90 (0.88 - 0.92)	0.97 (0.93 - 1.00)
Worked in last 12 months		0.82 (0.66 - 1.01)	0.72 (0.53- 0.99)
Age			1.02 (1.01- 1.04)
Male			1.13 (0.86- 1.49)
Coloured vs. African			2.05 (0.77- 5.46)
White/Asian vs. African			0.93 (0.32 - 2.67)
Ever smoked 100 cigarettes or more			1.63 (1.02 - 2.58)
CAGE score greater than 1			2.52 (1.86 - 3.41)
BMI under 18.5			2.84 (1.91 - 4.22)
Missed meals due to lack of funds			1.76 (0.93 - 3.33)

Column 6 is also for adjusted for: Ever worked in gold mine; Urban Residence; Adults per bedroom; Proportion of MD moved in past 5 years. 95% CIs shown in parentheses.

The addition of the asset score to the model attenuated the point estimates of individual education and employment very slightly and had little effect on most non-SES covariates. The asset score did, however, attenuate the point estimates on White/Asian by around half compared to their values in Table 8.

Table 11: Multivariate Relationships between Household-Based, MD-Level Headcount Income Poverty and Tuberculosis

Dependent Variable	Incident TB		
Column	*1*	*2*	*3*
Household headcount quintile 1 (poorest)	1.33 (0.60 - 2.96)	0.74 (0.28 - 1.95)	0.99 (0.47 - 2.06)
Household headcount quintile 2	0.59 (0.16 - 2.16)	0.48 (0.09 - 2.68)	0.50 (0.16 - 1.58)
Household headcount quintile 3	1 (reference)	1 (reference)	1 (reference)
Household headcount quintile 4	0.41 (0.17 - 0.99)	0.58 (0.15 - 2.23)	0.49 (0.15 - 1.58)
Household headcount quintile 5 (richest)	0.83 (0.33 - 2.04)	1.70 (0.30 - 9.54)	1.48 (0.40 - 5.45)
Asset score quintile 1 (poorest)		1.77 (0.81 - 3.85)	1.39 (0.72 - 2.69)
Asset score quintile 2		0.96 (0.41 - 2.24)	0.84 (0.41 - 1.72)
Asset score quintile 3		1 (reference)	1 (reference)
Asset score quintile 4		0.65 (0.35 - 1.20)	0.68 (0.35 - 1.32)
Asset score quintile 5 (richest)		0.13 (0.06 - 0.28)	0.23 (0.13 - 0.42)
Years of education		0.93 (0.88 - 0.98)	0.96 (0.90 - 1.01)
Worked in last 12 months		0.79 (0.46 - 1.37)	0.74 (0.44 - 1.24)
Age			1.00 (0.97 - 1.03)
Male			0.70 (0.47 - 1.04)
Coloured vs. African			0.81 (0.26 - 2.52)
White/Asian vs. African			0.37 (0.12 - 1.20)
Ever smoked 100 cigarettes or more			2.19 (1.18 - 4.05)
CAGE score greater than 1			1.99 (1.23 - 3.21)
BMI under 18.5			4.12 (1.78 - 9.54)
Missed meals due to lack of funds			1.28 (0.59 - 2.74)
Column 3 is also for adjusted for: Ever worked in gold mine; Urban Residence; Adults per			

bedroom; Proportion of MD moved in past 5 years. 95% CIs shown in parentheses.

Dependent Variable	Lifetime Prevalent TB		
Column	*4*	*5*	*6*
Household headcount quintile 1 (poorest)	1.59 (0.79 - 3.21)	0.88 (0.53 - 1.44)	1.11 (0.68 - 1.82)
Household headcount quintile 2	0.94 (0.53 - 1.66)	0.71 (0.38 - 1.30)	0.83 (0.49 - 1.39)
Household headcount quintile 3	1 (reference)	1 (reference)	1 (reference)
Household headcount quintile 4	0.85 (0.63 - 1.15)	0.92 (0.67 - 1.26)	0.89 (0.63- 1.27)
Household headcount quintile 5 (richest)	1.20 (0.58 - 2.49)	1.48 (0.72 - 3.02)	1.13 (0.71- 1.79)
Asset score quintile 1 (poorest)		1.92 (1.17 - 3.16)	1.62 (1.06 - 2.47)
Asset score quintile 2		1.20 (0.87 - 1.66)	1.02 (0.73- 1.44)
Asset score quintile 3		1 (reference)	1 (reference)
Asset score quintile 4		1.09 (0.62 - 1.90)	0.92 (0.56 - 1.51)
Asset score quintile 5 (richest)		0.73 (0.40 - 1.32)	0.79 (0.47 - 1.34)
Years of education		0.90 (0.88 - 0.93)	0.97 (0.93 - 1.00)
Worked in last 12 months		0.80 (0.63 - 1.01)	0.72 (0.52 - 1.00)
Age			1.02 (1.01- 1.03)
Male			1.13 (0.86 - 1.49)
Coloured vs. African			1.92 (0.72 - 5.09)
White/Asian vs. African			0.92 (0.31 - 2.70)
Ever smoked 100 cigarettes or more			1.62 (1.02- 2.59)
CAGE score greater than 1			2.53 (1.86 - 3.43)
BMI under 18.5			2.87 (1.96- 4.22)
Missed meals due to lack of funds			1.75 (0.93 - 3.31)

Column 6 is also for adjusted for: Ever worked in gold mine; Urban Residence; Adults per bedroom; Proportion of MD moved in past 5 years. 95% CIs shown in parentheses.

Table 12: Multivariate Relationships between Household-Based, MD-Level Unemployment Rate and Tuberculosis

Dependent Variable	Incident TB		
Column	*1*	*2*	*3*
Household unemployment quintile 1 (highest)	0.94 (0.48 - 1.83)	0.52 (0.25 - 1.09)	0.89 (0.44 - 1.82)
Household unemployment quintile 2	0.48 (0.18 - 1.28)	0.34 (0.12 - 1.01)	0.49 (0.20 - 1.24)
Household unemployment quintile 3	1 (reference)	1 (reference)	1 (reference)
Household unemployment quintile 4	0.35 (0.10 - 1.22)	0.47 (0.12 - 1.80)	0.51 (0.14 - 1.89)
Household unemployment quintile 5 (lowest)	0.461 (0.22 - 0.96)	0.83 (0.31 - 2.22)	0.83 (0.34 - 1.99)
Asset score quintile 1 (poorest)		1.95 (0.95 - 3.98)	1.53 (0.83 - 2.83)
Asset score quintile 2		1.00 (0.45 - 2.24)	0.88 (0.43 - 1.78)
Asset score quintile 3		1 (reference)	1 (reference)
Asset score quintile 4		0.73 (0.34 - 1.56)	0.71 (0.34 - 1.48)
Asset score quintile 5 (richest)		0.16 (0.06 - 0.43)	0.25 (0.12 - 0.54)
Years of education		0.93 (0.89 - 0.98)	0.96 (0.90 - 1.02)
Worked in last 12 months		0.80 (0.50 - 1.30)	0.74 (0.46 - 1.20)
Age			1.00 (0.97 - 1.03)
Male			0.71 (0.48 - 1.03)
Coloured vs. African			1.08 (0.22 - 5.33)
White/Asian vs. African			0.42 (0.12 - 1.54)
Ever smoked 100 cigarettes or more			2.17 (1.19 - 3.94)
CAGE score greater than 1			1.91 (1.19 - 3.08)
BMI under 18.5			3.88 (1.67 - 9.03)
Missed meals due to lack of funds			1.24 (0.60 - 2.58)

Column 3 is also for adjusted for: Ever worked in gold mine; Urban Residence; Adults per bedroom; Proportion of MD moved in past 5 years. 95% CIs shown in parentheses.

Dependent Variable	Lifetime Prevalent TB		
Column	*4*	*5*	*6*
Household unemployment quintile 1 (highest)	0.96 (0.41 - 2.26)	0.75 (0.39 - 1.45)	1.22 (0.52 - 2.84)
Household unemployment quintile 2	0.71 (0.39 - 1.28)	0.70 (0.41 - 1.20)	0.99 (0.56 - 1.77)
Household unemployment quintile 3	1 (reference)	1 (reference)	1 (reference)
Household unemployment quintile 4	0.40 (0.20 - 0.83)	0.58 (0.31 - 1.07)	0.57 (0.34 - 0.96)
Household unemployment quintile 5 (lowest)	0.63 (0.28 - 1.39)	1.00 (0.53 - 1.89)	0.95 (0.66 - 1.38)
Asset score quintile 1		1.94 (1.18 - 3.21)	1.64 (1.07 - 2.52)
Asset score quintile 2		1.21 (0.87 - 1.70)	1.03 (0.73 - 1.46)
Asset score quintile 3		1 (reference)	1 (reference)
Asset score quintile 4		1.17 (0.66 - 2.07)	0.94 (0.57 - 1.54)
Asset score quintile 5		0.81 (0.43 - 1.54)	0.79 (0.46 - 1.37)
Years of Education		0.90 (0.88 - 0.93)	0.97 (0.93 - 1.00)
Worked in last 12 months		0.81 (0.66 - 1.00)	0.73 (0.53 - 1.01)
Age			1.02 (1.01 - 1.03)
Male			1.14 (0.87 - 1.49)
Coloured vs. African			2.09 (0.80 - 5.48)
White/Asian vs. African			0.96 (0.33 - 2.76)
Ever smoked 100 cigarettes or more			1.64 (1.04 - 2.58)
CAGE score greater than 1			2.51 (1.867 - 3.36)
BMI under 18.5			2.84 (1.92 - 4.22)
Missed meals due to lack of funds			1.75 (0.91 - 3.37)

Column 6 is also for adjusted for: Ever worked in gold mine; Urban Residence; Adults per bedroom; Proportion of MD moved in past 5 years. 95% CIs shown in parentheses.

5.3.3. Magisterial district income poverty and employment status

Regressions containing quintile measures of household-based, MD-level headcount poverty rates (P2) suggested no significant relationship between headcount poverty and tuberculosis illness, particularly after taking other factors into account (Table 11). Furthermore, the inclusion of MD-level measures in columns 3 and 6 changed other point estimates little compared to the equivalent figures in Table 10.

Regressions containing quintile measures of MD-level household unemployment rates (U2) found little association between MD-level unemployment rates and tuberculosis disease, although quintile 4 did appear to be somewhat protective, particularly against lifetime prevalent tuberculosis (Table 12). The inclusion of MD-level unemployment also did not alter the relationship between tuberculosis and any of the existing explanatory variables, including individual-level unemployment.

5.3.4. Magisterial district income inequality

A clear relationship between higher income inequality and tuberculosis disease was seen in regressions containing quintile measures of MD-level, household-based values of the Robin Hood index (R3) (Table 13). Lifetime prevalent tuberculosis was significantly positively associated with living in MDs with a Robin Hood index score in quintiles 1 or 2, relative to those in quintile 3, after adjusting for other variables. The reported odds ratios for these quintiles in regressions for incident tuberculosis were even higher, but not statistically significant.

The relationships reported for regressions containing quintile measures of MD-level, household-based values of the Gini coefficient (G3) were similar to those found for the Robin Hood index (Table 14): those living in MDs in quintiles 1 and 2 again had a raised adjusted risk of tuberculosis, relative to those living in quintile 3. This relationship was statistically significant in the case of lifetime prevalent disease.

Living in MDs in quintile 5, and to a certain extent quintile 4, for both income inequality measures put an individual at increased risk for tuberculosis, relative to living in quintile 3, although this was at no time statistically significant. The inclusion of income inequality measures had little impact on other point estimates in the model. The strongest effect was that of the Robin Hood index on asset score quintile 1, shifting the point estimate from 1.67 to 1.60, a value that remained significantly greater than the null of no relationship.

Table 13: Multivariate Relationships between Household-Based, MD-Level Income Inequality (Robin Hood index) and Tuberculosis

Dependent Variable	Incident TB		
Column	*1*	*2*	*3*
Household Robin Hood quintile 1 (most unequal)	2.59 (0.98 - 6.80)	1.69 (0.65 - 4.37)	2.16 (0.72 - 6.51)
Household Robin Hood quintile 2	2.07 (0.37 - 10.81)	1.68 (0.34 - 8.25)	1.96 (0.39 - 9.91)
Household Robin Hood quintile 3	1 (reference)	1 (reference)	1 (reference)
Household Robin Hood quintile 4	1.30 (0.54 - 3.16)	1.42 (0.56 - 3.55)	1.68 (0.69 - 4.11)
Household Robin Hood quintile 5 (least unequal)	1.56 (0.47 - 5.18)	2.86 (0.82 - 10.09)	2.97 (0.85 - 10.34)
Asset score quintile 1 (poorest)		1.85 (0.80 - 4.25)	1.57 (0.86 - 2.87)
Asset score quintile 2		0.97 (0.42 - 2.24)	0.87 (0.43 - 1.74)
Asset score quintile 3		1 (reference)	1 (reference)
Asset score quintile 4		0.68 (0.39 - 1.18)	0.68 (0.35 - 1.32)
Asset score quintile 5 (richest)		0.15 (0.06 - 0.34)	0.23 (0.12 - 0.45)
Years of Education		0.93 (0.88 - 0.98)	0.96 (0.90 - 1.02)
Worked in last 12 months		0.81 (0.48 - 1.37)	0.73 (0.44 - 1.18)
Age			1.00 (0.97 - 1.03)
Male			0.695 (0.48 - 1.02)
Coloured vs. African			1.01 (0.27 - 3.84)
White/Asian vs. African			0.40 (0.13 - 1.21)
Ever smoked 100 cigarettes or more			2.20 (1.19 - 4.06)
CAGE score greater than 1			1.98 (1.21 - 3.25)
BMI under 18.5			4.02 (1.75 - 9.23)
Missed meals due to lack of funds			1.27 (0.58 - 2.76)

Column 3 is also for adjusted for: Ever worked in gold mine; Urban Residence; Adults per bedroom; Proportion of MD moved in past 5 years. 95% CIs shown in parentheses.

Dependent Variable	Lifetime Prevalent TB		
Column	*4*	*5*	*6*
Household Robin Hood quintile 1 (most unequal)	1.92 (1.14 - 3.24)	1.46 (0.99 - 2.13)	1.69 (1.25 - 2.29)
Household Robin Hood quintile 2	1.39 (0.82 - 2.35)	1.27 (0.83 - 1.94)	1.49 (1.03 - 2.15)
Household Robin Hood quintile 3	1 (reference)	1 (reference)	1 (reference)
Household Robin Hood quintile 4	0.75 (0.48 - 1.18)	0.87 (0.54 - 1.40)	0.86 (0.57 - 1.29)
Household Robin Hood quintile 5 (least unequal)	0.90 (0.46 - 1.78)	1.19 (0.64 - 2.23)	1.14 (0.82 - 1.58)
Asset score quintile 1 (poorest)		1.76 (1.06 - 2.91)	1.60 (1.06 - 2.42)
Asset score quintile 2		1.16 (0.83 - 1.63)	1.02 (0.72 - 1.46)
Asset score quintile 3		1 (reference)	1 (reference)
Asset score quintile 4		1.22 (0.69 - 2.16)	0.96 (0.59 - 1.56)
Asset score quintile 5 (richest)		0.84 (0.45 - 1.58)	0.83 (0.50 - 1.36)
Years of Education		0.90 (0.88 - 0.93)	0.97 (0.93 - 1.00)
Worked in last 12 months		0.82 (0.66 - 1.03)	0.73 (0.53 - 1.00)
Age			1.02 (1.01 - 1.03)
Male			1.15 (0.87 - 1.51)
Coloured vs. African			2.08 (0.77 – 5.65)
White/Asian vs. African			0.94 (0.33 - 2.66)
Ever smoked 100 cigarettes or more			1.63 (1.03 - 2.59)
CAGE score greater than 1			2.53 (1.87 - 3.42)
BMI under 18.5			2.85 (1.93 - 4.20)
Missed meals due to lack of funds			1.77 (0.93 - 3.35)

Column 6 is also for adjusted for: Ever worked in gold mine; Urban Residence; Adults per bedroom; Proportion of MD moved in past 5 years. 95% CIs shown in parentheses.

Table 14: Multivariate Relationships between Household-Based, MD-Level Income Inequality (Gini Coefficient) and Tuberculosis

Dependent Variable	Incident TB		
Column	*1*	*2*	*3*
Household Gini coefficient quintile 1 (most unequal)	1.89 (0.85 - 4.18)	1.16 (0.58 - 2.32)	1.38 (0.70 - 2.72)
Household Gini coefficient quintile 2	2.12 (1.10 - 4.09)	1.61 (0.88 - 2.93)	1.66 (0.79 - 3.48)
Household Gini coefficient quintile 3	1 (reference)	1 (reference)	1 (reference)
Household Gini coefficient quintile 4	1.16 (0.43 - 3.11)	1.35 (0.57 - 3.17)	1.29 (0.58 - 2.88)
Household Gini coefficient quintile 5 (least unequal)	1.19 (0.50 - 2.83)	2.24 (0.81 - 6.21)	1.94 (0.75 - 5.00)
Asset score quintile 1 (poorest)		1.81 (0.80 - 4.10)	1.55 (0.86 - 2.79)
Asset score quintile 2		0.963 (0.41 - 2.28)	0.86 (0.42 - 1.75)
Asset score quintile 3		1 (reference)	1 (reference)
Asset score quintile 4		0.69 (0.39 - 1.23)	0.68 (0.35 - 1.32)
Asset score quintile 5 (richest)		0.15 (0.06 - 0.34)	0.23 (0.12 - 0.45)
Years of Education		0.930 (0.88 - 0.98)	0.95 (0.90 - 1.01)
Worked in last 12 months		0.81 (0.50 - 1.33)	0.73 (0.46 - 1.17)
Age			1.00 (0.97 - 1.03)
Male			0.70 (0.48 - 1.02)
Coloured vs. African			1.03 (0.25 - 4.26)
White/Asian vs. African			0.41 (0.12 - 1.39)
Ever smoked 100 cigarettes or more			2.17 (1.16 - 4.05)
CAGE score greater than 1			1.94 (1.22 - 3.09)
BMI under 18.5			3.98 (1.71 - 9.29)
Missed meals due to lack of funds			1.23 (0.55 - 2.78)

Column 3 is also for adjusted for: Ever worked in gold mine; Urban Residence; Adults per bedroom; Proportion of MD moved in past 5 years. 95% CIs shown in parentheses.

Dependent Variable	Lifetime Prevalent TB		
Column	*4*	*5*	*6*
Household Gini coefficient quintile 1 (most unequal)	1.94 (1.22 - 3.07)	1.35 (0.91 - 2.00)	1.72 (1.30 - 2.28)
Household Gini coefficient quintile 2	2.60 (1.68 - 4.04)	2.22 (1.40 - 3.53)	2.43 (1.41 - 4.19)
Household Gini coefficient quintile 3	1 (reference)	1 (reference)	1 (reference)
Household Gini coefficient quintile 4	1.11 (0.64 - 1.94)	1.28 (0.78 - 2.11)	1.07 (0.68 - 1.69)
Household Gini coefficient quintile 5 (least unequal)	1.16 (0.40 - 3.31)	1.53 (0.54 - 4.36)	1.28 (0.71 - 2.30)
Asset score quintile 1 (poorest)		1.80 (1.09 - 2.98)	1.66 (1.08 - 2.55)
Asset score quintile 2		1.18 (0.84 - 1.64)	1.03 (0.72 - 1.47)
Asset score quintile 3		1 (reference)	1 (reference)
Asset score quintile 4		1.24 (0.68 - 2.25)	0.97 (0.59 - 1.60)
Asset score quintile 5 (richest)		0.86 (0.44 - 1.67)	0.82 (0.50 - 1.35)
Years of Education		0.90 (0.88 - 0.92)	0.97 (0.93 - 1.00)
Worked in last 12 months		0.82 (0.66 - 1.03)	0.73 (0.53 - 1.00)
Age			1.02 (1.01 - 1.03)
Male			1.15 (0.87 - 1.52)
Coloured vs. African			2.16 (0.80 - 5.78)
White/Asian vs. African			0.96 (0.34 - 2.68)
Ever smoked 100 cigarettes or more			1.62 (1.02 - 2.59)
CAGE score greater than 1			2.51 (1.86 - 3.38)
BMI under 18.5			2.86 (1.92 - 4.24)
Missed meals due to lack of funds			1.72 (0.91 - 3.27)

Column 6 is also for adjusted for: Ever worked in gold mine; Urban Residence; Adults per bedroom; Proportion of MD moved in past 5 years. 95% CIs shown in parentheses.

5.4. Model Checking

Based on the regressions reported above, the preferred models from this analysis are those presented in columns 3 and 6 of Table 13 and Table 14. A plot of the correlations between the independent variables in the models suggested no multicollinearity. The Pearson correlation coefficient between the quintiles of MD-level variables used in section 5.3 ranged from 0.58 to 0.87, suggesting them to be highly correlated with one another. Furthermore, each of these four household-based MD-level variables was correlated between 47% and 59% with household-level asset score.[15]

Table 15: Characteristics of Individuals with Outlying Pearson or Deviance Residual Values

Outlier	*1*	*2*	*3*	*4*	*5*
Lifetime Prevalent TB	No	No	No	No	No
Incident TB	No	No	No	No	No
Community					
Headcount rate	5	1	2	3	3
Unemployment	4	2	2	4	3
Gini coefficient	4	2	3	4	2
Robin Hood index	4	2	3	5	3
Household					
Asset score quintile	5	4	1	5	3
Missed meals ever	No	No	Yes	No	No
Individual					
Education	Tertiary	Standard 1	Standard 5	Tertiary	Sub A
Employed in past year	Yes	Yes	No	Yes	Yes
Age	37	18	53	22	35
Sex	Female	Male	Male	Male	Male
Race	White/Asian	African	African	White/Asian	African
Smoked 100 cigarettes ever	No	No	No	No	Yes
CAGE score > 1	No	No	Yes	No	Yes
BMI under 18.5	No	No	No	Yes	No

For all quintile measures, a higher score represents better social conditions.

Plots of Pearson and Deviance residuals suggested that the models were well fitted, with a clear separation in residual values between those with and without tuberculosis. A few

[15] Multicollinearity is generally considered a problem if explanatory variables are more than 90% correlated with one another.

outlying values were observed in these residuals (Table 15). Four of the five strongest outliers appear to be poorly fitted by the model due to the very low-risk combination of risk factors, while the fifth had a high risk for disease but had never been diagnosed with tuberculosis.

Table 16: Characteristics of MDs with Outlying Cook's Distance Values

	Whole Sample	Extreme MDs	
Number of Observations	13,043	414	93
Lifetime Prevalent TB	2.8%	6.0%	1.1%
Incident TB	0.5%	2.7%	1.1%
Community			
Urban	55.8%	Yes	Yes
Headcount rate	3	3	5
Unemployment	3	3	5
Gini coefficient	3	4	5
Robin Hood index	4	4	5
Household			
Asset score quintile	3	4	5
Missed meals ever	52.3%	57.2%	10.8%
Individual			
Years of Education	8	9	11
Employed in past year	33.8%	38.2%	47.3%
Age	35	35	36
Male	41.6%	47.6%	40.8%
African	75.1%	60.4%	45.2%
Coloured	13.2%	21.3%	4.3%
Smoked 100 cigarettes ever	26.9%	29.5%	33.3%
CAGE score > 1	18.5%	26.6%	9.7%
BMI under 18.5	9.8%	9.2%	4.3%

Values are medians or proportions using unweighted data unless otherwise stated.
For all quintile measures, a higher score represents better social conditions.

Similarly to the residual analyses, plots of the Cook's Distance measure for each MD in the four regressions suggested no extreme outlying, and therefore influential, values. The two MDs described in Table 16 had the highest Cook's Distance values. The first had particularly high rates of tuberculosis but only average scores on several of the key risk factors in the model. The second had only one incident case of tuberculosis, but this placed it above the sample average, and the MD had high SES by several measures. None of the values for these measures were far from the main body of observations.

6. Discussion

In the study of tuberculosis there is a tradition of attributing reductions in disease incidence across Europe and North America between 1850 and 1980 to ecological changes in the social circumstances of populations. René and Jean Dubos recognised the importance of social and economic structures of the disease over half a century ago, writing that it was:

> "apparent that the spread of tuberculosis during the nineteenth century was the outcome of the social tragedies that followed in the wake of the industrial revolution, rather than the consequence of city life *per se*." (Dubos & Dubos 1992, p199)

The Dubos' also questioned the benefit of early twentieth-century efforts to eradicate the disease, noting the confounding impact of rising quality of life. This connection was more forcefully made by Thomas McKeown's quantitative analysis of national tuberculosis rates in England and Wales between the eighteenth and twentieth centuries (McKeown 1975). McKeown found tuberculosis disease rates to have been in secular decline long before chemotherapeutic treatment was available, and attributed the fall to improvements in living standards and nutrition, rather than medical or public health interventions. McKeown's methods, and particularly conclusions, have been disputed and perhaps discredited, but his evidence of a decline in disease rates prior to the availability of effective medical treatment remains uncontested (Colgrove 2002).

More recently, quantitative studies of cities and national populations in more developed countries have found ecologic associations between tuberculosis disease rates and a range of socio-economic markers, notably unemployment, low education, income poverty, income inequality and social deprivation (see section 3.5 above). There seems little doubt that tuberculosis is associated with poor socio-economic conditions in low-incidence, high-income settings.

Furthermore, there is considerable evidence that tuberculosis disease is most prevalent in low-income countries. The twenty-two high-incidence countries, as identified by the World Health Organisation had per capita gross domestic products of between US$ 110 and US$ 3,630 in 2004. This compares to US$ 33,547, the average figure for the 30

industrialized nations that comprise the Organisation for Economic Co-operation and Development (World Bank 2004). The rise of HIV in already high-burden African countries that are also amongst the poorest in the world has skewed this comparison yet further in recent years.

This study adds two important dimensions to the existing literature. First, it is the first study of the impact of socio-economic factors on tuberculosis disease in Africa, offering evidence on whether the findings in richer settings are replicated elsewhere. Second, it is the first study to examine the impact of a range of socio-economic factors on tuberculosis disease risk, taking account of a range of commonly-accepted non-SES risk factors. It therefore represents a significant step forwards in understanding the impact of socio-economic factors on the distribution of tuberculosis within societies.

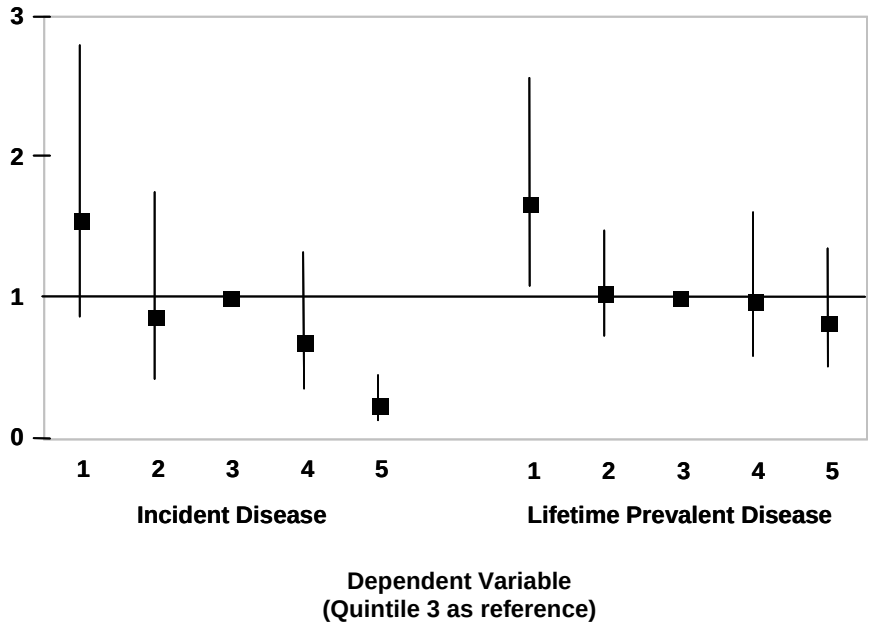
6.1. Key Findings

This study provides evidence on the distribution of tuberculosis in South Africa geographically, by age, by sex and by race group. It furthermore presents a first effort to quantitatively and simultaneously measure the association of tuberculosis with a range of SES measures in a high-incidence setting.

6.1.1. Individual- and household-level SES risk factors

At the individual and household levels, this analysis finds recent employment and more education to be correlated with reduced individual risk of disease, both incident and lifetime prevalent. For lifetime prevalent disease, having been employed in the previous twelve months is associated with a 27% reduced risk of disease (95% CI: 0.53-1.00), and each additional year of education is associated with a 3.5% reduced risk (95% CI: 0.93-1.00) (Table 14, column 6). The respective odds ratios for incident disease were 0.73 (95% CI: 0.46-1.17) and 0.95 (95% CI: 0.90-1.01) (Table 14, column 3). A person who has completed secondary education (12 years) therefore has a 35% reduced risk of lifetime prevalent disease and a 44% reduced risk of incident disease relative to an identical individual who completed no schooling.

Figure 3: Adjusted Odds Ratios and 95% CIs for Asset Score Quintiles and Tuberculosis



The raw household asset score represents a measure of absolute deprivation, since it is a combination of responses on ownership of goods and access to services. When it is divided into quintiles, as in this analysis, it is, however, transformed into a measure of relative deprivation, a household's position relative to others in South Africa.

This study finds relative deprivation to be inversely associated with tuberculosis disease in a linear fashion, falling as asset ownership rises (Figure 3). For example, those living in the most deprived quintile of households have a 66% (95% CI: 7.5%-155%) higher risk of having ever been diagnosed with tuberculosis compared to those in the middle quintile. Similarly, those living in the least deprived quintile of households have a 77% (95% CI: 55%-88%) reduced risk of having been diagnosed with tuberculosis in the past year compared to those in the middle quintile.

The roughly linear inverse relationship between deprivation and tuberculosis is in line with the ecologic findings in high-income settings outlined in section 3.5. This study's result is, however, in contrast to the finding in the closest comparable study to date, from a Brazilian city, where a threshold effect was reported such that the most deprived 40

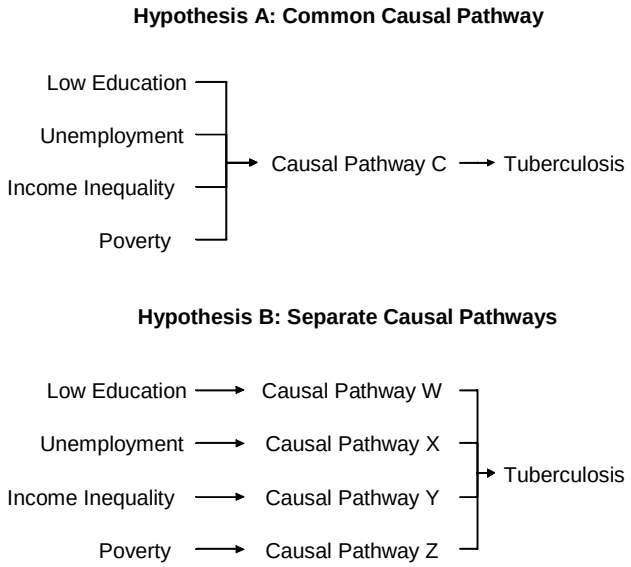
percent of census tracts were at raised risk of tuberculosis relative to the rest of the population (Souza et al. 2000).

A number of studies have previously found area-level deprivation to affect individual-level health, either directly, or via other risk factors for poor health (Reijneveld 1998; Smith et al. 1998; Malmström et al. 1999; Sundquist et al. 1999; Malmström et al. 2001; Sundquist et al. 2004; Petrelli et al. 2006). This may, however, be the first study to consider the impact of household-level deprivation on individual-level health outcomes, independent of area-level SES, and as a result may find nuanced results not seen elsewhere.

Each of the three relationships between individual- and household-level measures of SES and tuberculosis is broadly supportive of the existing literature on individual-level SES and health, finding employment and increased education and decreased deprivation to be promotive of good health. In high-income settings, however, the inclusion of several SES measures as predictors of disease rates often leads to a loss of significance for some of them (Davey-Smith et al. 2004; Holtgrave & Crosby 2004; Chan-Yeung et al. 2005). This suggests a model such as that shown in the upper panel of Figure 4, where all SES measures work through a single causal pathway, and furthermore have high cross-elasticity (i.e. one can be easily substituted for another without changing the effect seen).

This study, however, finds that the inclusion of any one individual- or household-level measure of SES has little effect on the relationship between other SES measures (Table 6, Table 9 & Table 10). Furthermore, the inclusion of community-level measures of SES also has little impact on the lower-level relationships (Table 11 to Table 14). This suggests a model such as that shown in the lower panel of Figure 4, where each SES variable acts through more or less independent causal pathways to cause disease.

Figure 4: Models of the Causal Pathways between SES Variables and Tuberculosis



6.1.2. Community-level SES risk factors

At the MD level, neither headcount poverty nor unemployment rates is found to be associated with tuberculosis disease, after adjusting for lower-level measures of SES. Compared to the middle quintile, the poorest quintile has odds ratios of 0.99 (95% CI: 0.47-2.06) and 1.11 (95% CI: 0.68-1.82) respectively when regressed on incident and lifetime prevalent disease (Table 11). The quintile with highest unemployment has odds ratios of 0.89 (95% CI: 0.44-1.82) and 1.22 (95% CI: 0.52-2.84) for the same regressions (Table 12). Across both sets of regressions, only one quintile (Table 12, regression 6, quintile 4) has an effect significantly different from the null.

Higher MD-level income inequality is, however, significantly associated with raised rates of lifetime prevalent tuberculosis, even after adjusting for household-level poverty in the form of the asset index. Using the Robin Hood index the most income unequal quintile has an odds ratio for lifetime prevalent disease of 1.69 (95% CI: 1.25-2.29), and the second most unequal quintile has an odds ratio of 1.49 (95% CI: 1.03-2.15), compared to the middle quintile (Table 13). The respective figures using the Gini coefficient are 1.72 (95% CI: 1.30-2.28) and 2.43 (95% CI: 1.41-4.19) (Table 14).

6.1.3. Non-SES risk factors

Among the non-SES risk factors, this study finds commonly accepted correlates of tuberculosis disease such as alcohol abuse and cigarette smoking to be risk factors in South Africa. Unlike many other studies however, it does not find a previous history of minework, household overcrowding or race group to be predictive of illness. The differences in these associations may be due to differences in causal pathways between those present in previous studies. Alternatively these findings may reflect the wide range of risk factors included in this study. The latter explanation would suggest that these variables are either mediators in a causal pathway that begins with other risk factors, and thus enter into the regression through these variables (for example SES), or confounders for other causal factors – associated with the causal factors but not directly with the outcome.

This study also finds a crude association between a history of minework, particularly a history of gold mine work, and tuberculosis. This relationship is not, however, robust to the inclusion of other individual-level risk factors. Minework has been associated with a range of pulmonary illnesses which are also associated with raised risk of tuberculosis, including silicosis and bronchitis in South Africa, in the latter case using the same 1998 SADHS dataset as the current study (Corbett et al. 2000; Ehrlich et al. 2004). This disparity in findings may reflect the relatively small number of individuals in the study population who have ever worked in a gold mine (n=452) and the even smaller number of these who had gone on to develop tuberculosis (incident n=6; lifetime prevalent n=38) since the point estimates remained quite sizeable in multivariate analysis (OR=1.52 for incident tuberculosis; OR=1.32 for lifetime prevalent disease in Table 8).

It is perhaps not surprising that household crowding is not significant in this study after adjustment has been made for the various SES measures included. As explained in section 3.1.1, overcrowding is generally believed to be a risk factor for infection due to the close proximity in which it puts individuals, raising the risk of contracting tuberculosis. It is also sometimes believed to be a risk factor for disease since it acts as a marker for low SES. This being so, the observed attenuation of the bivariate association between overcrowding and tuberculosis disease after adjustment for SES is to be

expected. In the bivariate analysis a one-person increase in the number of people sleeping in each bedroom is associated with a statistically significantly increased risk of incident tuberculosis, but not of lifetime prevalent disease. This is congruent with overcrowding being a risk factor for infection since current overcrowding reflects proximity to others. In contrast it is likely to be a poorer predictor of lifetime prevalence, since the level of household crowding is likely to change over an individual's lifetime.

This study finds being labelled White or Asian to be protective for incident tuberculosis relative to African individuals, in bivariate and individual-level multivariate analysis. These associations, however, disappear once household asset levels are included in the analysis and remain absent thereafter. Previous studies in the US, UK and China have found racial background to be associated with risk of tuberculosis disease, most likely because of the proximity of racial and ethnic minorities to immigrants, who are more likely to be infected, and infectious, with tuberculosis (see section 3.1). It seems from this study that the crude relationship seen between race and tuberculosis in South Africa, where tuberculosis infection is widespread, is due to race being correlated with socio-economic factors, rather than factors relating either to racial susceptibility, or the legacy of past policies, such as Apartheid. That is, race acts as a proxy for SES in models which do not contain SES variables, but has no independent effect on disease risk.

6.2. Potential Explanations for Findings

The results seen in this study are likely to reflect a wide range of causal mechanisms. A few potential pathways and issues are explored below.

6.2.1. Geographic effects

A first possible explanation for the community-level SES results seen is that they may reflect something about the geographical distribution of tuberculosis. Although the model assumes observations to be clustered at the provincial level, and includes an explanatory variable for individual urban/rural residence, it is possible that other geographic variation is being measured through these variables. In particular no account

is taken in this data of the difference between urban and peri-urban settlements, except through measures of community SES. All four measures of MD-level SES are significantly, inversely correlated with living in an urban community, although the income inequality measures are less so than the unemployment and poverty measures.

If there is a true association between community-level SES, particularly between poverty or unemployment, and tuberculosis disease rates, the inclusion of the rural/urban residence variable in the regression may lead to confounding – and thus an attenuation – of this effect. The limited change in the point estimate on rural/urban residence when community-level variables are included in the analysis (not shown) suggests that this effect is not an important one in this analysis.

Furthermore, the complex interplay between urban and rural areas linked to mining and other migratory work outlined by Packard cannot be explicitly included in this analysis (Packard 1987). For example, there is no variable to represent the proportion of persons in a community that commute to a major city to work. To the extent that the SES-related results seen here truly reflect this long-distance spatial patterning of disease, they should be considered to be confounded by geographic factors.

6.2.2. Compositional and contextual effects

As explained in section 3.3, there are a number of ways in which group-level results can be interpreted. For those variables with both individual- and group-level forms included in the analysis it is possible to evaluate whether their effects are compositional (the result of individual agency) or contextual (due to the social structure of the setting). Unemployment and income poverty can be evaluated in this way in this study. Group-level variables with no individual-level equivalent must be considered to act in a contextual manner, since they cannot be the result of individual-level characteristics. The measures of income inequality used in this analysis fall into this category.

Both unemployment and income poverty – the latter measured by the asset score index at the household level, appear to display compositional effects. That is, allowing for lower-level variability, community-level unemployment and poverty rates appear to have no effect on tuberculosis rates.

The finding of an association between these measures and tuberculosis suggest a contextual effect of SES on tuberculosis in this setting. Furthermore, this effect is seen over and above individual-level measures of wealth/deprivation and employment. There is thus something about unequal communities in South Africa that puts persons living in them at higher risk of becoming ill with tuberculosis than those living in more equal ones, even after allowing for individual SES characteristics.

There has been limited debate in the literature about the nature and usefulness of concepts of area-level contextual and compositional effects in the relationship between SES and health. A recent review of the concepts, however, noted that there remain problems with the use of these terms (MacIntyre et al. 2002). The authors note that the two are poorly defined with respect to one another due to bi-directional effects – individual-level factors are shaped by community-level ones, and vice versa. Additionally, due to an absence of theory considering how area-level SES affects health, contextual factors are often treated as the residual explanators, rather than as risk factors acting through recognisable channels. Careful definition and use of terms should, however, allow the circumvention of these pitfalls.

6.2.2.1.Contextual mechanisms

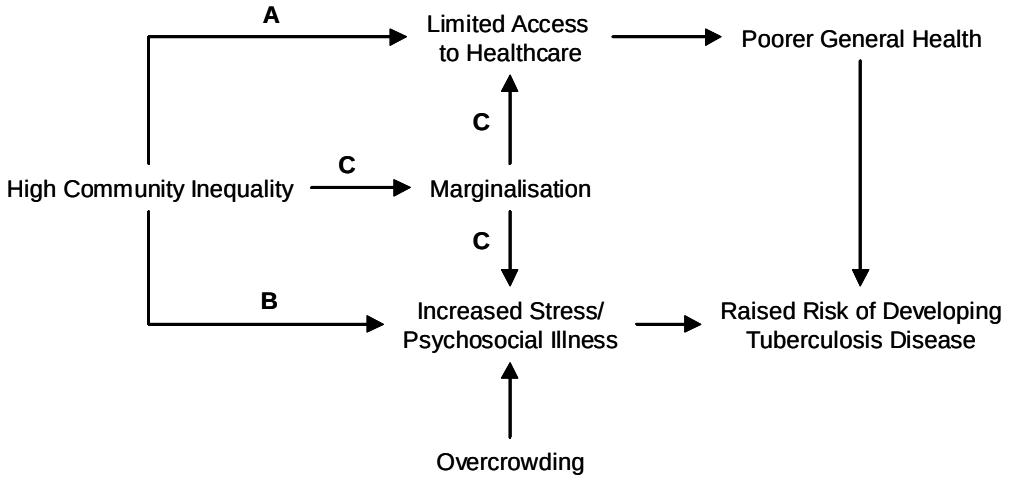
Several causal mechanisms might be occurring to cause this emergent, contextual effect of income inequality apart from any effect linked to poverty (Subramanian et al. 2002). Importantly, each of these hypotheses is consistent with finding that inequality is harmful to the poor, rather than beneficial to the rich, as is seen in section 5.3 where an individual's risk of tuberculosis disease is positively associated with living in the most unequal communities, but not inversely associated with living in the least unequal ones.

First, inequality may affect a community's access to healthcare. More unequal communities may have more limited access to healthcare for their poorest members than similarly poor individuals in more egalitarian settings, in which case poor people in unequal settings would be expected to have worse health than their wealth level would otherwise predict (Figure 5, pathway A). This argument is a facet of the materialist approach to income inequality, which argues that unequal income is an outcome of

historical processes through which resources were skewed towards the rich (Lynch et al. 2000).

An example of how such a situation could occur would be if the existence of private healthcare services masked the absence of accessible healthcare services for the poor. This, in turn, is consistent with the existence of a two-tier health service such as that seen in South Africa: private health insurance, which predominantly uses private healthcare services, covered 18 percent of the South African population in 1995 (Söderlund & Hansl 2000). These services were heavily skewed towards the higher end of the income scale, with 2 percent of those in the lowest income quintile having private insurance compared to 60 percent of those in the highest income quintile.

Figure 5: Potential Causal Mechanisms for Community-level Income Inequality



High income-inequality areas are likely to also have skewed availability of public goods such as infrastructure and services. This may also arise if income inequality leads to a loss of political cohesion across income groups, with the result that poorer individuals or sub-communities have fewer services provided to them (Subramanian & Kawachi 2004).

Second, inequality may affect the structure of communities in which it is prevalent. Highly unequal societies have been hypothesised to suffer from higher levels of psychosocial illness than more egalitarian ones, whether measured at the national or

more local level (Wilkinson & Pickett 2006). It is believed that low self-esteem, and lack of personal agency, raises an individual's risk of, and lowers their capacity to cope with, illness. Furthermore, it is believed that stress lies in the causal pathway between overcrowding and tuberculosis disease (section 3.1). As a result, an increased burden of psychosocial illness, caused by higher inequality, might be expected to add to the burden of tuberculosis disease in unequal communities (Figure 5, pathway B).

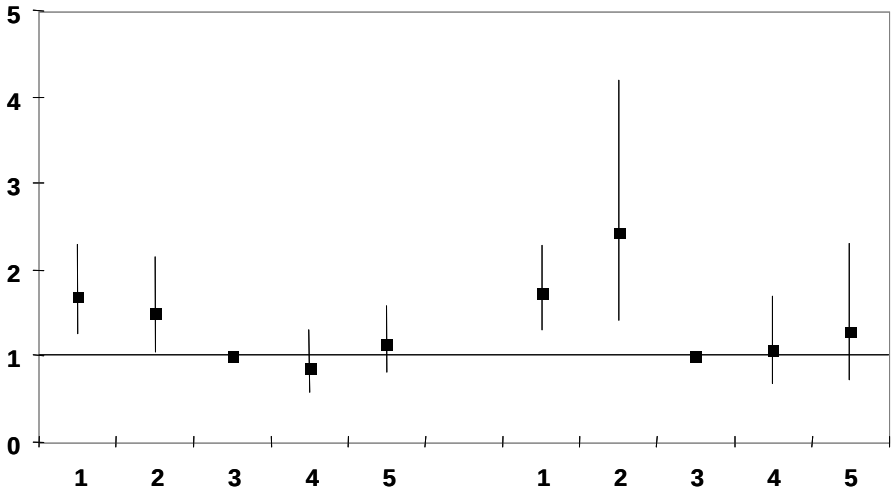
Third, in an unequal setting it is likely that the poorest in society will be marginalised, if unequal societies have weaker bonds between members than more equal ones. Insofar as lower social cohesion is connected to higher levels of social exclusion and lower levels of social capital, it is possible that more unequal areas will see poorer health outcomes (Kawachi & Berkman 2000). Posited mechanisms for this link between social capital and health include poor diffusion of health information or healthy behavioural norms, and lesser capacity to defend against budget cuts or to organise to lobby for the provision of services. This suggests that marginalisation might act as a mediator in either of the two causal mechanisms previously mentioned (Figure 5, pathway C).

The relative importance of materialist and psychosocial explanations for the relationship between high inequality and poor health remain contested (Lynch et al. 2000).

6.2.2.2.Choice of community-level SES measures

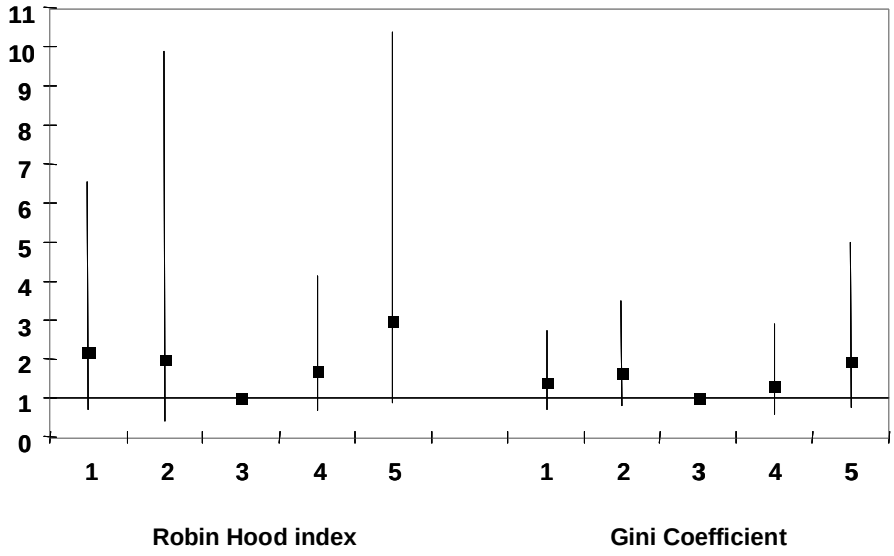
It has generally been accepted in the literature dealing with community-level SES and health outcomes that the exact nature of measures used at the community-level is less important than at the individual-level, since community-level measures tend to be more highly correlated with each other than individual-level ones (Pickett & Pearl 2001). This does not appear to be entirely supported in this analysis, where the measure of SES used – income poverty, unemployment or income inequality – strongly affects the significance and nature of the relationship seen between community-level SES and tuberculosis disease (Table 11 to Table 14). The effect of all measures on lower-level variables was however, similar in its slightness.

Figure 6: Adjusted Odds Ratios and 95% CIs for Income Inequality Measures and Lifetime Prevalent Tuberculosis



(Quintile 3 as reference; higher quintile reflects lower inequality)

Figure 7: Adjusted Odds Ratios and 95% CIs for Income Inequality Measures and Incident Tuberculosis



(Quintile 3 as reference; higher quintile reflects lower inequality)

This study can provide some support to the finding of insensitivity to the income inequality measure used seen in earlier work. Kawachi and Kennedy (1997) found that a range of nine income inequality measures were all highly correlated with one another (absolute correlation coefficients between 0.86-1.00 with the large majority over 0.95) and were all similarly correlated with total mortality (0.38-0.49). Strong relationships remained for all measures after adjusting for median household poverty and income rates. The results for lifetime prevalent disease (Figure 6) show similar relationships between each quintile of income inequality and tuberculosis, supporting Kawachi and Kennedy's conclusions. The inference cannot be so strongly made in the case of incident tuberculosis, however, due to the large size of the confidence intervals seen, but the pattern of point estimates is once again similar (Figure 7).

6.2.3. Risk factors in high tuberculosis prevalence settings

As mentioned in section 3.5, the large majority of research into risk factors for tuberculosis disease has taken place in settings where the prevalence of tuberculosis, both for infection and disease, is lowest – high-income Europe and the United States (Dye et al. 1999). In such settings many risk factors for disease, such as ethnicity, education levels and urban living are highly correlated with contact with immigrant communities from high-tuberculosis-prevalence settings, such as Africa and Asia. As a result, many of these variables may not be true risk factors for tuberculosis disease.

This study provides a useful counterpoint to studies conducted in low-prevalence settings. It finds neither ethnicity nor household overcrowding to be associated with an increased risk of tuberculosis disease, suggesting that these associations seen in higher-income settings may well reflect risk of infection, rather than risk of disease *per se*. In contrast, the finding that that low community-level SES is associated with heightened risk of tuberculosis disease in this study suggests that SES is truly related to tuberculosis disease, rather than simply being a marker for increased contact with highly tuberculosis-infected populations.

6.3. Potential Strengths and Weaknesses of the Analysis

This secondary data analysis is based on two well-validated primary surveys. It further attempts to take into account the variations of these surveys, and adjust for a wide range of potentially influential covariates. Nevertheless, it is important to consider the potential shortcomings in the planning and execution of this work.

6.3.1. Study design

6.3.1.1.Temporal issues

There are a number of aspects of the study design that might have biased the observed results. Several of these relate to the nature of survey data used. First, the SADHS questions in the used to create incident tuberculosis measure do not distinguish between diagnosis of old or recently arisen tuberculosis. Hence a respondent reporting having been diagnosed with tuberculosis within the past year may in fact be reporting the diagnosis of a now-dormant infection that had previously been disease. To the degree that this occurred, it would lead to an overestimate of tuberculosis incidence in this population.

Second, the analysis is based on two cross-sectional datasets, one a survey, the other a random sample of a total population census. As a result it is important to consider the possibility that the reported results suffer from reverse causation due to the outcome preceding the explanatory variable, and that prevalent cases of long duration have been oversampled. The duration of cases is not considered in this study, since both outcome measures were based on date of diagnosis, rather than existence of disease at any point in time.

The possibility of reverse causation was considered as it arose in sections 4 and 5 and is certainly a concern for relationships between individual-level explanatory variables and the outcome variables. In particular, those individuals with a lifetime prevalent tuberculosis diagnosis may well have poorer current SES outcomes due to the illnesses reducing their capacity to attend school or perform work tasks. Less obviously, past tuberculosis illness may lead to an individual living in a lower-SES community, due to a diminished capacity to maintain their residence in better surroundings. In such a case, the causal relationship between low SES and tuberculosis may run from disease to SES.

Even in the case the incident tuberculosis, reverse causation cannot be dismissed for individual- and household-level explanatory variables. For example, early symptoms of tuberculosis may have led to diminished earning capacity or even loss of employment. This in turn will have reduced household asset levels and maybe contributed to poorer socio-behavioural characteristics, such as increased alcohol consumption. Given this potential for reverse causation, causal connections drawn from this study must be tentative.

6.3.1.2.Recall Bias

A source of duration-related bias that cannot so easily be dismissed is that of recall: cases that were of long duration are more likely to be recalled, even if the illness occurred long ago, since they were present for longer in the respondent's life. If present, this bias will have an attenuative effect on any results, meaning the true effect is larger than that reported, so long as the duration of illness is not correlated with any of the explanatory variables. In the case of treated tuberculosis, variation in duration should be limited, since treatment is of a standardised length, although the length of symptomatic disease prior to diagnosis may vary systematically.

A second form of recall bias may have arisen if SADHS respondents misunderstood the question asked of them and told interviewers that they had been diagnosed with tuberculosis disease, when in fact they had been diagnosed with tuberculosis infection. If this were to have been correlated with low levels of education or other SES variables then this would have led the reported odds ratios for SES variables being an overstatement of the true association.

A third type of recall bias may have occurred if respondents intentionally misreported their true tuberculosis diagnosis history, hiding their diagnosis from the interviewer, perhaps due to stigma attached to the illness. The impact of this bias will depend on the correlation of deceit levels with explanatory variables. If this effect was more prevalent among high SES individuals then the reported results would be too strong, while if the effect was more prevalent among deprived persons then the reported results would be an understatement of the true situation.

6.3.1.3.Diagnostic Bias

In a self-reported dataset such as this, the possibility of diagnostic bias must be considered. In order for a respondent to report tuberculosis disease, they must first have become actively ill, second sought to access healthcare services and third succeeded in doing so. If the first step in this process is what the study wishes to capture, variation in the other two steps may lead to the results being biased. To the extent that diagnosis rates diverge from true disease rates, this study will reflect healthcare demand, rather than healthcare need.[16]

As of 2003 it was believed that only 45 percent of worldwide smear-positive tuberculosis cases were being detected by control programmes (Dye et al. 2005). The remaining 55 percent of cases either did not seek to, or succeed in accessing care. If the rate of non-detection, and thus non-diagnosis, is correlated with explanatory variables then the results seen will not be a true reflection of causal pathways. In this context it is encouraging that the tuberculosis detection rate is close to 100 percent according to the World Health Organisation (2005).

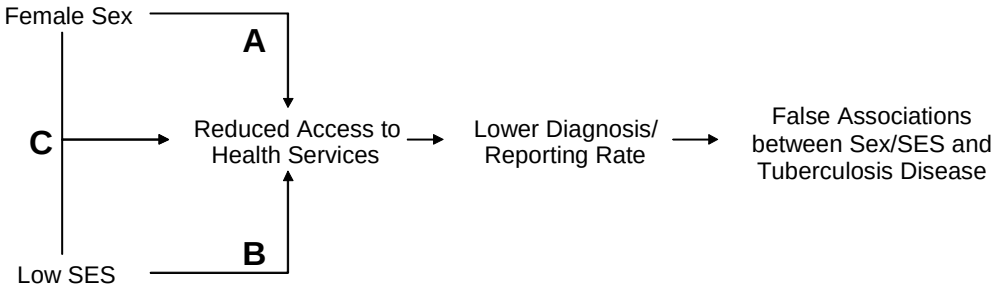
One explanatory variable generally believed to be correlated with diagnosis rates, as mentioned in section 3.1, is sex. It is thought that women are less likely to be diagnosed with tuberculosis disease, even though incidence rates may be sex-neutral, due to differential patterns of access to healthcare. Specifically, women are more likely to access private care or self-medicate than to approach public clinics. It has been hypothesised that this may be partly due to the stigma of the disease (Thorson & Diwan 2003). Other work has, however, suggested that prevalent and diagnosed tuberculosis rates have similar sex-ratios, with the possible exception of Africa (Borgdorff et al. 2000). Although no clear relationship between sex and tuberculosis is seen in this study, a true difference may be being hidden, if reporting rates diverge from true disease rates.

A second variable likely to be correlated with diagnosis rates is SES. As explained in section 6.2, poorer individuals are likely to have limited access to healthcare, even if no direct costs are charged for services. If this is so, low SES individuals will underreport disease, leading to an attenuation of a true inverse association between SES and tuberculosis in the reported results.

[16] Demand is used here in the economic sense: as a need for services that is backed up by the capability to consume them, both through an ability to access and if necessary to pay for them.

These two effects will affect estimates of individual relationships, as shown in Figure 8 below by mechanisms A and B. In addition, if poor women are less likely to access care, since the scarce resources required to do so are reserved for male household members, then there may be an additional interaction effect, mechanism C in Figure 8, between sex and SES in their relationships to reported tuberculosis rates (Ensor & Cooper 2004).

Figure 8: Potential Effect Pathways for Diagnostic Bias



6.3.1.4.Statistical Power

The population-based sampling-frame of the SADHS also presents problems of statistical power when the phenomenon under investigation is relatively rare – even in a high-incidence setting such as South Africa active tuberculosis disease occurs in only one person in 200 each year. The resultant lack of statistical power for the analysis of incident cases of disease limits the study's capacity to differentiate between the absence of a relationship and an inability to discern a true association.

The similarity of the results in section 5 for lifetime prevalent and incident tuberculosis, both in direction and magnitude, offers some evidence for the lack of significance in the case of incident tuberculosis being primarily a result of sample size, rather than a true lack of association. For example, in Table 14 the point estimate for the lowest asset score quintile in column 6 is 1.66 (95% CI: 1.08-2.55), similar to that in column 3 of 1.55 (95% CI: 0.86-2.79). Due to the width of confidence intervals, however, the former

is statistically significant, while the latter is not. Even more strikingly, the values for the protective effect of recent employment in the same table are both 0.73, but for incident tuberculosis the confidence interval runs from 0.46 to 1.02, giving a non-statiscally significant result, while for lifetime prevalent disease the interval stretches from 0.53 to 0.997, thus attaining significance. Graphically this can be seen by comparing Figure 6 and Figure 7. The confidence intervals in the latter figure are such that even though the point estimates are larger in some cases, none of them are statistically significantly.

6.3.2. Analytic design

6.3.2.1.Appropriate community level

A key question in designing any multilevel analysis is determining what the appropriate level of analysis for community-level effects should be. The aim of selecting a community-level is to capture the level at which the relevant causal pathways exist (Diez-Roux 2001). In order for a community effect to be seen, the level used must also display gradients in risk factors. This is likely to happen if the divisions used create economically homogeneous areas, and/or if they follow administrative demarcations (Krieger et al. 2003).

In order to decide which is more important, it is necessary to hypothesise as to the pathways through which the group-level variables (in this study, SES measures) will affect the outcome (tuberculosis). If the link is materialise – material deprivation leads to higher risk of disease – then administrative areas are likely to be most useful construct, since it at these levels that mitigating policies, such as social programmes, will vary. If the link is psychosocial – deprivation leads to psychosocial illness, which in turn leads to harmful behaviours and thus higher risk of disease – then areas that share social ties will be of more interest.

In this analysis the use of MDs was one of convenience as it was the lowest level of aggregation at which the two datasets could be confidently merged. MDs existed to demarcate judicial boundaries, rather than healthcare ones, and as such may not directly reflect variation in healthcare provision. They are, however, of a similar level of aggregation to the 303 municipality-based administrative areas established in 2000 to replace them. MDs therefore roughly equate to municipalities, a level which makes

conceptual sense as a community for variation in healthcare provision (including diagnostic services) and in SES.

Evidence from the US suggests that in that setting using smaller communities produced more robust community-level associations between SES and health. Analysis there suggested little difference in the size or nature of community-level effects when measured at the census tract and census block levels – average populations of 4000 and 1000 persons respectively (Diez-Roux et al. 2001). These areas are one to two orders of magnitude smaller than the MDs used in this study, which had an average population of around 100,000 persons. Another US study suggested that the use of zip codes, average population 30,000, was less consistent in the results arrived at than the census area measures (Krieger et al. 2003). The authors note that this may have been because zip codes are created for delivery of mail, rather than for economic or government policy reasons. While the South African census EA would have provided a community size closer to that of the census block – the average population of each EA was under 500 persons – the two datasets could not be matched on this level.

Empirically, however, evidence for the effect of income inequality on health is far stronger at higher levels of aggregation – it is commonly seen in international studies and those considering the state level in the US, but only intermittently at lower levels of aggregation (Subramanian & Kawachi 2004; Wilkinson & Pickett 2006). This is true both for multilevel and ecologic studies, and may reflect the importance of variations in political mechanisms, which in turn lead to variation in the provision of social goods (Subramanian & Kawachi 2004).

6.3.2.2.Income measurement

There are serious difficulties with the use of income-based socio-economic data in South Africa. First, a very large proportion of individuals report having no income whatsoever – only 38% of working-age adults reported being employed in the 1996 census (Statistics South Africa 2004). Second, only formal-sector earnings are likely to be reported in a census, since it is unlikely that tax is being paid on other income, and thus unlikely that individuals will wish to report them to a government body. Third, as referred to in section 3.4.5, evidence suggests that in low-income settings the stock of

goods (assets) owned may be more relevant to outcomes than the current flow of income.

The difficulty of differentiating between the large number of individuals and household with no, or very little income, is made still more difficult in this study by the fact that South African census income data is gathered in bands. In the census these bands become larger in absolute terms as income rises, but generally smaller in the proportional increase from one band-edge to the next.[17] Imputing a single value to all individuals in a band both decreases the inequality between members of the same band (to nothing) and changes the level of inequality between members of different bands (direction indeterminate) unless individuals are perfectly distributed across the range of each band in every subgroup (i.e. MD). The overall effect, and thus whether the banding has a differential effect on the income inequality results in this study, is unclear.

Since the Gini coefficient (but not the Robin Hood index) is calculated in absolute income differences, the first effect will be systematic in reducing Gini inequality in the most unequal settings, tending to attenuate any relationship seen in these areas. The second effect for all measures, and the first effect for the Robin Hood index, is ambiguous in direction unless something is known about the distribution of incomes within each band. It can therefore only be suggested that this banding may introduce more random variation on the average, which will tend to attenuate true results.

These distortions should not affect the income poverty findings, since the headcount measure did not impute values to individuals, but instead used the bands as given. Given the uncertainty generated by the use of banded data, measures created from the census data for this analysis should be considered indicative of relative poverty and inequality between different areas of South Africa. Point estimates should be interpretable in terms of direction, but the magnitude of any effects found should be treated with great caution.

[17] For example, the fourth band reaches from R501 per month to R1000 per month; the absolute increase is R500, the proportional increase 100%. The absolute Rand increases for all bands are: 200, 300, 500, 500, 1000, 1000, 1000, 1500, 2000, 3000, 5000, 14,000, ∞. The respective percentage increases are: ∞, 150, 100, 50, 67, 40, 29, 33, 33, 38, 45, 88, n/a.

6.3.2.3.Data selection

The data for the SADHS was collected in 1998, while that for the census was collected in 1996. Two concerns arise from this. The first is the comparability of the two datasets, given that they were not collected at the same time. This concern can be allayed by the fact that only rankings of MDs on the various SES measures are brought into the analysis, rather than absolute values. Unless it is believed that a radical change in the nature of MDs occurred between 1996 and 1998 the data should still be applicable to the SADHS dataset. Evidence from the 1996 and 2001 censuses do not suggest any radical change in the economic nature of the country during this period (Statistics South Africa 2004). The labour participation rate (the proportion of working-age persons who are economically active) changed marginally from 57.5% in 1996 to 57.7% in 2001. Census income was collected in different categories in the two censuses, but 25% of working-age adults earned less than R500 per month in 1996 compared to 33% who earned less than R800 in 2001.

The second, and more serious concern, is that the nature of tuberculosis disease and its causes may have changed between 1998 and the present day. Given the rapid rise of HIV-related illness in South Africa over the past decade, it is possible that factors that were related to disease in 1998 are not today. Unfortunately, this DHS dataset is the most recent national dataset available on tuberculosis in South Africa. This concern need not be a shortcoming so long as the context, particularly of the HIV epidemic, is taken into account when interpreting these results.

In any study of an infectious, and thus social, disease such as tuberculosis it is important to consider the widest possible range of potentially explanatory variables in order to avoid omitted variable bias. In this study, these include biomedical characteristics, social habits and socio-economic descriptors of individual respondents, the size and wealth of their households and the socio-economic position of their community, as well as their geographic setting. With the exception of HIV, each significant risk factor previously suggested in the literature has been at least indirectly measured. The potential for omitted variables to be driving the results has thus been minimised.

The converse argument is that too many variables may have been included in this analysis, leading to overadjustment or reduced discriminatory power due to the inclusion of irrelevant data. When considering risk factors that may have a distal impact on the outcome, the inclusion of more proximal factors on the same causal pathway amounts to an overadjustment of the data, and will tend to wash out any true effect of the more distal factor. It has been argued that in studies of area-level SES this occurs when individual-level risk factors such as drinking or smoking are included, since an individual's propensity for such behaviours is likely to be influenced by the nature of their local environment (MacIntyre et al. 2002). This view is, however, discounted by other researchers who see SES and other social factors as both being fundamental causes of health disparities, rather than one being a marker or confounder for the other (Phelan et al. 2004).

While it is possible that point estimates on SES risk factors in this study may be attenuated by such overadjustment, the results seen suggest either that: the impact of the adjustment is only partial (since some community-level effects remain); or that the pathways through which they act are not all contained in the study (i.e. some omitted mediating variables have been left out, leaving community-level effects); or that area-level SES has some more proximal effect on tuberculosis, closer to the view posited by Phelan and her colleagues.

6.3.3. HIV

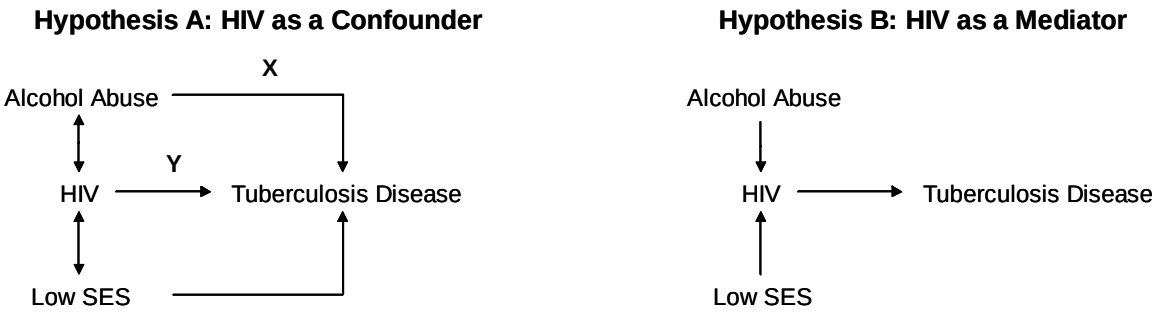
As alluded to above, the most serious data selection concern in this analysis is that it does not take into account the HIV status of the respondents to this study. This important omitted variable means that explanatory variables associated with risk of HIV infection and with risk of tuberculosis disease will appear to be more closely linked to tuberculosis than is actually the case.

In Figure 9, hypothesis A shows such a situation, where the true effect of alcohol abuse should be measured partially through pathway X and partially through pathway Y. In the absence of HIV as an explanatory variable, both effects will be measured through X, overstating the importance of alcohol abuse. Given that HIV is often associated with

low SES, and several of other variables in this analysis, such an explanation of the reported results cannot be dismissed.

The importance of the omission of HIV data is, however, lessened if one considers HIV to be a mediator in the pathway between many of the social factors in this analysis and tuberculosis disease, shown as hypothesis B in Figure 9. This requires that HIV be both a risk-factor for the outcome, which was established earlier, and a result of the exposure, that is, be socially patterned. The literature cited in section 3.4 supports such an assertion for health in general and there is evidence that social deprivation is linked particularly strongly to sexually-transmitted diseases (Brandt 1987). There is additionally a literature suggesting that HIV is a result of both poverty and economic inequality in poor and unequal settings such as Haiti and Peru (Farmer 1999). To the extent that HIV is caused by the social factors included in this model, the inclusion of more distal causes of tuberculosis will take into account this mediation process, reducing the potential for omitted variable bias. Indeed, in such a situation, the inclusion of HIV might be considered an overadjustment of the model.

Figure 9: Potential Causal Pathways for HIV



6.3.4. Study validity

The question of validity addresses both whether the data is representative of the population from which they were drawn (internal validity), and whether they are generalisable to broader populations (external validity). The former is not a concern in this study: a careful analysis of the sample methodology and response rates was

conducted as part of the SADHS process, and appropriate weightings were then used in this analysis to ensure national representivity (*South African Demographic and Health Survey* 1999), while the very large sample drawn from the census (each MD had an average of 10,000 datapoints) make significant deviations from population values unlikely. As a result, this study should be representative of the South African situation in the late 1990s.

The issue of generalisability is more complex. It is important to note that South Africa is middle-income nation with an unusually high level of both income inequality and HIV infection. Given that South Africa is almost unique in this combination of important tuberculosis risk factors, and that it is not clear which, if any, interactions of these factors drives the results seen in this study, care must be taken in extrapolating these results outside of South Africa.

Extrapolation of the income inequality results should be reasonable to other high-inequality settings. Reviews of the impact of income-inequality on disease have previously found significant results only in countries with higher levels of inequality (Brazil, Chile, Taiwan, US) as opposed to more egalitarian settings (Canada, Denmark, Japan, New Zealand, Sweden) (Subramanian et al. 2002, Subramanian & Kawachi 2004).

The Chilean finding that income inequality appeared to have a threshold effect for communities with a Gini coefficient above 0.40 suggests that income inequality-health effects may only apply broadly in countries with higher levels of inequality. As explained above, however, it seems likely that the presence of HIV will act as a magnifier for income-inequality measures, so results may additionally not transfer well to unequal countries with low HIV prevalence.

The relevance of these measures to high-income, low tuberculosis incidence nations is not clear. The dynamics of an infectious disease differ greatly depending on the baseline prevalence, and while the results on the importance of income-inequality support work in the US, the benefit of direct comparison is unclear. Similarly, South Africa is almost an order of magnitude richer (per capita) than many nations in Africa: in settings where

the disparities in income seen in this setting are not present due to widespread absolute poverty, it is hard to see the applicability of the SES-related results found in this study.

6.4. Next Steps

6.4.1. Research recommendations

This study is the first of which the author is aware jointly considering biomedical, social and socio-economic risk factors for tuberculosis in sub-Saharan Africa. It is also the first study to use a multilevel methodology to analyse SES risk factors for tuberculosis disease. It is therefore important to treat these results as tentative, and to attempt to replicate them in other datasets and settings. This is particularly true given the fragility of the incident tuberculosis data, with only 71 cases. One method for strengthening the credibility of the reported results would be to conduct longitudinal studies of health outcomes in South Africa which also collect data on economic variables. An alternative approach would be to collect tuberculosis and other health outcome data in exiting, repeated economic surveys such as the General Household Survey (Statistics South Africa 2006).

One weaknesses of this study was the need to impute Rand values to individual and household income as reported in bands in the census. It would of considerable use to have a nationally representative study of area-level deprivation and SES in South Africa. This would require more detailed questioning, and thus more resources, than the census questionnaire, but could be done on a sample of the population, rather than all persons. Its benefit would stretch well beyond tuberculosis research, or even health research, to provide a useful measure of community SES for researchers in other social science fields, such as education and economic development.

The omission of HIV status from this study is an important one. Although it may not affect the reported results for the reasons mentioned above, it would be extremely useful to replicate this analysis with either individual- or group-level HIV infection markers. A survey which included anonymous, linked HIV testing might be able to provide additional income on this question; so might the linking of individual-level data to community-level HIV information. Neither is available at present in South Africa, but

anonymous HIV testing has been built into recent DHS studies elsewhere in the world (Mishra et al. 2006).

This study has also only touched on one aspect of the random-effects modelling system, by including random intercepts for households and communities (Merlo et al. 2005a). It might be of interest to consider the possibility of random slopes for higher model levels, that is that the community-level SES variables modify the individual-level associations seen (Merlo et al. 2005b). It might also be interesting to consider the proportions of the variance in outcomes due to different levels of the model, that is how much individual-level variation can explained by household and/or community clustering of tuberculosis (Merlo et al. 2005c). A further possibility would be to consider whether the effects of community SES are differentially felt in different populations, such as the old, different racial groupings or those with low education or assets. Such a study was conducted in the US for self-rated health, but found few significant effects (Subramanian & Kawachi 2006). A final possible extension to the methodologies used here would be to investigate the possibility of quadratic effects for age, years of education and other polychotomous categorical variables. This would allow for the possibility of non-linear relationships between these variables and risk of tuberculosis disease.

6.4.2. Social epidemiology

The field of Social Epidemiology is a young one, having barely existed 30 years ago; the first textbook in the subject was published in 2000 (Berkman & Kawachi 2000). As a result much of the field remains fluid and methodologies are still being organised. There can be little doubt that the social and socio-economic environment, whether considered at the macro (national), meso (community) or micro (household/individual) level, in which individuals' lives affects both their health status and their capacity to manage their health. The challenge for practitioners in the field is to approach their enquiry in both a theoretically and methodologically sound manner, in particular considering the following issues.

First, there is a need for greater theoretical rigour. Given the increased computing power of recent times, it has become far easier to conduct empirical studies. These studies have often involved the inclusion of as many variables as are available, at as

many levels as possible. If enough regressions are run on enough variables, random chance will mean some of the relationships will be statistically significant. In order to avoid the frequent reporting of potentially spurious findings as true causal associations, it is crucial to follow the advice of those in the field who demand that potential causal pathways, and the group-level(s) at which these pathways are believed to act, are hypothesised *a priori*, rather than mined for (Pickett & Pearl 2001; Diez-Roux 2001).

The importance of purely contextual phenomena remains underemphasised. At present most SES measures used are based on the scaling up of individual concepts – employment, education and income. Integral measures are those that can only be measured at the community level and are therefore by definition contextual; they are, however, rarely included in social epidemiology studies (Pickett & Pearl 2001). They include the level of social service provision, the number of hospital beds or the number of community groups in an area. Ignoring such measures excludes a range of potential, emergent causal effects.

It is also important to simultaneously consider a range of levels at which SES might affect health; many studies consider the impact of higher-level variables without first considering the impact of individual-level ones. As a recent review of multilevel studies suggested, those studies that included only one (or no) individual-level measure of SES are more likely to find a significant result for any group-level SES measures used, since the latter will act as partial proxies for the former in their absence (Pickett & Pearl 2001).

Second, there is a need for increased statistical rigour. Since social epidemiology attempts to connect an individual to their environment, there will always be a need for both individual and higher-level factors to be considered. In a statistical context this may be done through purely individual- or purely group-level regressions, but both have flaws. Using ecologic regressions to determine individual-level risks is open to the ecologic fallacy: that relationships seen at a population level reflect those seen at the individual level, when in fact no such individual-level relationship exists. Using individual-level regressions, on the other hand is open to the individualist fallacy: that an individual's health outcomes are a function only of their personal characteristics. The

combination of individual- and group-level factors in a single model, using simple ordinary least squares regression is open to bias due to the non-independence of these explanatory variables. The use of hierarchical models can overcome this, but must be used with care, and a deeper understanding of such models is crucial to the development of the field (Merlo 2003).

A second statistical matter rarely considered is that of endogeneity. It is commonly assumed that health outcomes are the result of socio-economic variables rather than considering it likely, at least to some extent, that social and socio-economic choices are made contingent on health status. This is particularly strong in the case of employment status, but if the 'Barker hypothesis' – that *in utereo* nutrition is linked to later health outcomes – is accepted, and given the high likelihood that this nutritional status will be associated with material socio-economic conditions, it is possible that reverse causality also arises for all socio-economic variables (Barker 1994).

A more troubling critique raised by Oakes of the use of hierarchical/multilevel models is that there is unavoidable endogeneity within the explanatory variables used (Oakes 2004). Oakes argues first that social stratification leads the causal mechanisms in each community to differ such that they are no longer comparable; and second that any contextual effects in a community are a reflection of the composition of that community, and thus not separately identifiable. The consideration, and management, of endogeneity has been common in economics for some time, through the use of causality tests and instrumental variables techniques that, at a cost to precision, allow the researcher to avoid the possibility of such problems arising (Martens et al. 2006). This does not solve Oakes' first concern, but remains an important avenue of enquiry in social epidemiology.

6.4.3. Policy recommendations

Taking the results of this study seriously would mean making significant changes to tuberculosis control programmes in South Africa. The current approach, following the World Health Organisation's lead, is to focus on Directly Observed Treatment, Short-course (DOTS) provision to all persons passively found to be sputum-smear positive. This has led to a very high detection rate for new smear-positive cases, but cure rates

remain almost 20% below the level needed to reduce the overall burden of disease (World Health Organisation 2005).

The reasons for this low cure rate are likely to be multifaceted, but as the Proje Veye Sante showed in rural Haiti, the impact of income support at a time of illness can be crucial. The clinic ran a trial where 50 individuals were provided $30 a month for their first three months of treatment, plus a travel allowance and a monthly attendance reminder. After one year, all these patients were disease-free and alive. Of a comparison group of 50 patients who did not receive this support, only 24 were disease-free and alive after twelve months (Farmer 1999). This evidence supports the South African government's policy of providing a disability grant to anyone receiving treatment for tuberculosis.

This study, however, found tentative evidence to suggest that poverty may be related not to treatment success, but to initially becoming ill with tuberculosis. In the light of this evidence, it can be argued that the provision of money to those living in poverty may well alleviate some of the government expenditures on tuberculosis treatment by preventing disease from occurring. One method for doing so would be the provision of a basic income grant, a universal payment supported by many in South Africa (Bhorat 2003). At present, no government support payment is available unemployed adults of working age who do not have a symptomatic disease.

The theoretical implication of the finding that income poverty and unemployment have compositional effects on risk of tuberculosis disease is that the targeting of policies to reduce poverty or unemployment, in order to reduce risk of tuberculosis, may be best performed at the individual level, rather than through interventions aimed at raising entire communities out of poverty. The practical impact of such a finding is less clear – reduction of community-level poverty (or unemployment) will require very similar interventions to those needed to reduce individual-level poverty (or unemployment).

A clearer message arising from the findings on SES is that low SES is associated with more tuberculosis disease, rather than high SES being associated with less. In both Figure 3 and Figure 6 the quintiles in which significant results are seen are the poorest and most unequal in the population. This has important implications for the targeting of

policies, suggesting the need to focus on the worst-off in particular, rather than the non-well-off in general.

The finding that even after adjusting for individual poverty, those living in more unequal settings were at raised risk of becoming ill, promotes both the bolstering of the tuberculosis treatment programme in more unequal regions, and the provision of support to the poor in unequal areas to make their communities less unequal.

The relationships highlighted in this study suggest potential benefits from a paradigm shift in tuberculosis treatment from the current focus on tertiary prevention (treating disease) towards secondary prevention (reducing progress from infection to disease). Interventions aimed at reducing poverty and inequality, and possibly at raising education levels and reducing unemployment, should help reduce the level of tuberculosis disease in South Africa. Public health campaigns and taxation policies highlighting the risks of smoking tobacco and alcohol abuse may also have a role to play. Such policies would need to be in addition to treatment of disease, rather than in place of it. As a result of the generalisability issues outlined in section 0, replication of these results in other data is needed prior to making these policy recommendations international in scope.

Evidence that low SES is associated with a wide range of health outcomes is not a new finding. The continuing existence of widespread poverty both within South Africa and worldwide does not necessarily reflect a lack of concern about the health of the poor, but may reflect a lack of capacity to effect change in this area, primarily in the government arena.

This inability may be due to constraints on policy capacity. For example, the general medicalisation of health, if not in the minds of the medical profession then in the minds of politicians and the public, may make the reallocation of resources from healthcare to social programmes unfeasible. Medical interventions tend to have clear, identifiable costs and benefits, whereas poverty alleviation has little immediate impact on health outcomes, despite statisticians assuring the public that the long-term benefits are considerable. Given a fixed total budget then even assuming doctors are willing to accept lower funding for hospitals and clinics in order to raise grants for social services and support payments, patients on waiting lists for operations, and thus politicians

vulnerable to public opinion, may not be wiling to reduce medical spending in order to reduce poverty.

Alternatively, the problem may be a lack of technical capacity. Even if funding is available, it may be that the public-sector lacks the skills to effectively target it to where it would have the greatest impact. Those areas with existing low levels of service are likely to be either geographically remote, difficult to retain staff in or have little political value – hence the fact that they are underprovided in the first place. It is likely that improving service delivery in such areas, probably predominantly rural and peri-urban in South Africa, will have a lower benefit-cost ratio than focusing on more accessible regions.

Finally, even if careful prioritisation of services is organised through central government, there is no certainty that lower levels of administration will have the capacity to carry it out. Furthermore, many of the most appropriate policies for reducing poverty and inequality, such as means-tested benefits, are also those most susceptible to manipulation and corruption. When government is stretched to meet demanding policy goals, and thus oversight is limited, as appears to be the case in South Africa at present, the hard work of efficiently targeting interventions may inconsistently implemented. While universal income support offers an alternative approach to reduce poverty, it will do little to reduce income inequality, which is found to be important in this study.

6.5. Conclusion

This study is a comprehensive analysis of risk factors for tuberculosis in South Africa. It shows that after adjusting for biological characteristics, individuals are at risk of becoming ill with tuberculosis through a range of behavioural, social and socio-economic reasons. It shows, perhaps for the first time, that not only personal and household characteristics are determinants of an individual's risk of falling ill with tuberculosis, but also the nature of the community in which one lives. It finds that in South Africa, being poor in either terms of financial resources or education, is detrimental to your health, and that living in an unequal community exacerbates this. In

order to reduce the burden of tuberculosis in this country it is crucial to improve the SES of the population, not just their health-related behaviour.

References

Acevedo-Garcia D. 2001. Zip Code-Level Risk Factors for Tuberculosis: Neighborhood Environment and Residential Segregation in New Jersey, 1985-1992. *Am J Public Health*, 91(5): 734-41.

Alisjahbana B, Van Crevel R, Sahiratmadja E, Den Heijer M, Maya A, Istriana E, Danusantoso H, Ottenhoff THM, Nelwan RHH, Van Der Meer JWM. 2004. Diabetes Mellitus is Strongly Associated With Tuberculosis in Indonesia. *Int J Tuberc Lung Dis*, 10(6): 696-700.

Andersson N. 1990. Tuberculosis and Social Stratification in South Africa. *Int J Health Serv*, 20(1): 141-65.

Badri M, Wilson D, Wood R. 2002. Effect of Highly Active Antiretroviral Therapy on Incidence of Tuberculosis in South Africa: a Cohort Study. *Lancet*, 359: 2059–64.

Barker DJP. 1994. *Mothers, Babies, and Disease in Later Life*. London: BMJ Publishing.

Barr RG, Diez-Roux AV, Knirsch CA, Pablos-Mendez A. 2001. Neighborhood Poverty and the Resurgence of Tuberculosis in New York City, 1984-1992. *Am J Public Health*, 91(9): 1487-93.

Behrman JR. 1996. The Impact of Health and Nutrition on Education. *World Bank Res Obs*,11 (1): 23-37.

Benatar SR. 2003. Global Poverty and Tuberculosis: Implications for Ethics and Human Rights. In: Gandy M, Zumla A (eds). *The Return of the White Plague: Global Poverty and the 'New' Tuberculosis*. London: Verso.

Bennett J, Pitman R, Jarman B, Innes J, Best N, Alves B, Cook A, Hart D, Coker R. 2001. A Study of the Variation in Tuberculosis Incidence and Possible Influential Variables in Manchester, Liverpool, Birmingham and Cardiff in 1991-1995. *Int J Tuberc Lung Dis*, 5(2): 158-63.

Ben-Shlomo Y, White IR, Marmot M. 1996. Does the Variation in the Socioeconomic Characteristics of an Area Affect Mortality? *BMJ*, 312: 1013-4.

Berkman LF, Kawachi I. 2000. A Historical Framework for Social Epidemiology. In: Berkman, LF, Kawachi I (eds). *Social Epidemiology*. New York, NY: Oxford University Press.

Bhorat H. 2003. *Estimates for Poverty Alleviation in South Africa, with an Application to a Universal Income Grant*. Development Policy Research Unit Working Paper 03/75. Cape Town: University of Cape Town.

Booysen F. 2001. The Measurement of Poverty. In: Bradshaw D, Steyn K (eds). *Poverty and Chronic Diseases in South Africa: Technical Report 2001*. Parrow: Medical Research Council.

Booysen FR. 2002. Using Demographic and Health Surveys to Measure Poverty: an Application to South Africa. *J Stud Econ Econometrics*, 26(3): 53-70.

Borgdorff MW, Nagelkerke NJD, Dye C, Nunn P. 2000. Gender and Tuberculosis: a Comparison of Prevalence Surveys with Notification Data to Explore Sex Differences in Case Detection. *Int J Tuberc Lung Dis*, 4(2): 123-132.

Brandt A. 1987. *No Magic Bullet: A Social History of Venereal Disease in the United States since 1880*. New York: Oxford University Press.

Braveman PA, Cubbin C, Egerter S, Chideya S, Marchi KS, Metzler M, and Posner S. 2005. Socioeconomic Status in Health Research: One Size Does Not Fit All. *J Am Med Assoc*, 294(22): 2879-88.

Bryk AS, Raudenbush SW. 1992. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, CA: SAGE Publications.

Cantwell MF, McKenna MT, McCray E, Onorato IM. 1998. Tuberculosis and Race/Ethnicity in the United States: Impact of Socioeconomic Status. *Am J Respir Crit Care Med*, 157(4): 1016-20.

Cantwell MF, Snider DE Jr, Cauthen GM, Onorato IM. 1994. Epidemiology of Tuberculosis in the United States, 1985 through 1992. *J Am Med Assoc*, 272(7): 535-9.

Carstairs V. 1995. Deprivation Indices: Their Interpretation and Use in Relation to Health. *J Epidemiol Community Health*, 49 (Suppl. 2): S3-8.

Case A. 2002. Health, Income and Economic Development. In: Peskovic B, Stern N (eds). *Annual World Bank Conference on Development Economics* 2001/02. New York, NY: Oxford University Press.

Castro A, Farmer P. 2003. Infectious Disease in Haiti: HIV/AIDS, Tuberculosis and Social Inequalities. *EMBO Rep*, 4 (Suppl. 1): S20-23.

Chan-Yeung M, Yeh AF, Tam CM, Kam KM, Leung CC, Yew WW, Lam CW. 2005. Socio-Demographic and Geographic Indicators and Distribution of Tuberculosis in Hong Kong: a Spatial Analysis. *Int J Tuberc Lung Dis*, 9(12): 1320-6.

Claasen JN. 1999. The benefits of the CAGE as a screening tool for alcoholism in a closed rural South African community. *S Afr Med J*, 89(9): 976-9.

Coberly JS, Chaisson RE. 2001. Tuberculosis. In: Nelson KE, Williams CM, Graham NMH (eds). *Infectious Disease Epidemiology: Theory and Practice*. Gaithersbury, MD: Aspen.

Colgrove J. 2002. The McKeown Thesis: a Historical Controversy and its Enduring Influence. *Am J Public Health*, 92: 725-9.

Corbett EL, Churchyard GJ, Clayton TC, Williams BG, Mulder D, Hayes RJ, De Cock KM. 2000. HIV Infection and Silicosis: the Impact of Two Potent Risk Factors on the Incidence of Mycobacterial Disease in South African Miners. *AIDS*, 14: 2759-68.

Corbett EL, Watt CJ, Walker N, Maher D, Williams BG, Raviglione MC, Dye C. 2003. The Growing Burden of Tuberculosis: Global Trends and Interactions with the HIV Epidemic. *Arch Intern Med*, 163: 1009-21.

Cubbin C, Hadden WC, Winkleby MA. 2001. Neighborhood Context and Cardiovascular Disease Risk Factors: the Contribution of Material Deprivation. *Ethn* Dis, 11(4): 687-700.

Dalton H. 1920. The Measurement of the Inequality of Incomes. *Econ J*, 3: 348-61.

Davey Smith G, Hart C, Hole D, MacKinnon P, Gillis C, Watt G, Blane D, Hawthorne V. 1998. Education and Occupational Social Class: Which is the More Important Indicator of Mortality Risk? *J Epidemiol Community Health*, 52: 153-60.

Deaton A. 2003. Health, Inequality and Economic Development. *J Econ Lit*, 41: 113-58.

Diez-Roux AV. 2000. Multilevel Analysis in Public Health Research. *Annu Rev Public Health*, 21: 171-92.

Diez-Roux AV. 2001. Investigating Neighborhood and Area Effects on Health. *Am J Public Health*, 91(11): 1783-9.

Diez-Roux AV, Aiello AE. Multilevel Analysis of Infectious Diseases. 2005. *J Infect Dis*, 191(Suppl 1): S25-33.

Diez-Roux AV, Kiefe CI, Jacobs DR, Haan M, Jackson SA, Nieto FJ, Paton CC, Schulz R. 2001. Area Characteristics and Individual-Level Socioeconomic Position Indicators in Three Population-Based Epidemiologic Studies. *Ann Epidemiol*, 11: 395-405.

Dong B, Ge N, Zhou Y. 2001. [Smoking and Alcohol Consumption as Risk Factors of Pulmonary Tuberculosis in Chengdu: a Matched Case-Control Study] (Chinese). *Hua Xi Yi Ke Da Xue Xue Bao*, 32(1): 104-6.

Dooley DJ, Fielding L, Levi L. 1996. Health and Unemployment. *Annu Rev Public Health*, 17: 449-64.

Dubos R, Dubos J. 1992. *The White Plague: Tuberculosis, Man, and Society*. 2nd Ed. New Brunswick, NJ: Rutgers University Press.

Dye C, Scheele S, Dolin P, Pathania V, Raviglione, MC. 1999. Global Burden of Tuberculosis: Estimated Incidence, Prevalence, and Mortality by Country. *J Am Med Assoc*, 282 (7): 677-86.

Dye C, Watt CJ, Bleed DM, Hosseini SM, Raviglione MC. 2005. Evolution of Tuberculosis Control and Prospects for Reducing Tuberculosis Incidence, Prevalence, and Deaths Globally. *JAMA*, 293(22): 2767-75.

Ehrlich RI, White N, Norman R, Laubscher R, Steyn K, Lombard C, Bradshaw D. 2004. Predictors of Chronic Bronchitis in South African Adults. *Int J Tuberc Lung Dis*, 8(3): 369-76.

Ensor T, Cooper S. 2004. Overcoming Barriers to Health Service Access: Influencing the Demand Side. *Health Policy Planning*, 19(2): 69-79.

Farmer P. 1999. *Infections and Inequalities: The Modern Plagues*. University of California Press: Berkeley.

Filmer D, Pritchett LH. 2001. Estimating Wealth Effects Without Expenditure Data - or Tears: An Application to Educational Enrollments in States of India. *Demography*, 38 (1): 115-32.

Gamble L, Mason CM, Nelson S. 2006. The Effects of Alcohol on Immunity and Bacterial Infection in the Lung. *Med Mal Infect*, 36(2): 72-7.

Gandy M, Zumla A (eds.). 2003. *The Return of the White Plague: Global Poverty and the 'New' Tuberculosis*. London: Verso.

Gandy M, Zumla A. 2002. The Resurgence of Disease: Social and Historical Perspectives on the 'New' Tuberculosis. *Soc Sci Med*, 55: 385-96.

Glynn JR, Crampin AC, Ngwira BM, Mwaungulu FD, Mwafulirwa DT, Floyd S, Pönnighaus JM, Warndorff DK, Fine PE. 2004. Trends in Tuberculosis and the Influence of HIV Infection in Northern Malawi, 1988-2001. *AIDS*, 18: 1459-63.

Glynn JR, Warndorff KD, Malema SS, Mwinuka V, Pönnighaus JM, Crampin AC, Fine PE. 2000. Tuberculosis: Associations with HIV and Socioeconomic Status in Rural Malawi. *Trans Royal Soc Trop Med Hygiene*, 94: 500-3.

Gordon D. 1995. Census Based Deprivation Indices: Their Weighting and Validation. *J Epidemiol Community Health*, 49 (Suppl. 2): S39-45.

Groenewald P, Nannan N, Bourne D, Laubscher R, Bradshaw D. 2005. Identifying Deaths from AIDS in South Africa. *AIDS*, 19:193–201.

Gustafson P, Gomes VF, Vieira CS, Rabna P, Seng R, Johansson P, Sandström A, Norberg R, Lisse I, Samb B, Aaby P, Nauclér A. 2004. Tuberculosis in Bissau: Incidence and Risk Factors in an Urban Community in Sub-Saharan Africa. *Int J Epidemiol*, 33(1): 163-72.

Ho M-J. 2004. Sociocultural Aspects of Tuberculosis: a Literature Review and a Case Study of Immigrant Tuberculosis. *Soc Sci Med*, 59: 753-62.

Holmes CB, Hausler H, Nunn P. 1998. A Review of Sex Differences in the Epidemiology of Tuberculosis. *Int J Tuberc Lung Dis*, 2(2): 96-104.

Holtgrave DR, Crosby RA. 2004. Social Determinants of Tuberculosis Case Rates in the United States. *Am J Prev Med*, 26(2): 159-62.

Jin RL, Shah CP, Svoboda TJ. 1995. The Impact of Unemployment on Health: a Review of the Evidence. *Can Med Assoc J*, 153(5): 529-40.

Kaplan GA, Pamuk E, Lynch JW, Cohen RG, Balfour JL. 1996. Income Inequality and Mortality in the United States: Analysis of Mortality and Potential Pathways. *Br Med J*, 312: 999-1003

Kawachi I, Berkman LF. 2000. Social Cohesion, Social Capital, and Health. In: Berkman LF, Kawachi I (eds). *Social Epidemiology*. New York, NY: Oxford University Press.

Kawachi I, Blakely TA. 2001. When Economists and Epidemiologists Disagree. *J Health Polit Policy Law*, 26 (3): 533-41.

Kawachi I, Kennedy BP. 1997. The Relationship of Income Inequality to Mortality: Does the Choice of Indicator Matter? *Soc Sci Med*, 45(7): 1121-7.

Kawachi I. 2000. Income Inequality and Health. In: Berkman, LF, Kawachi I (eds). *Social Epidemiology*. New York, NY: Oxford University Press.

Kennedy BP, Kawachi I, Prothrow-Stith D. 1996. Income Distribution and Mortality: Cross-sectional Ecological Study of the Robin Hood Index in the United States. *Br Med J*, 312: 1004-7.

Kim J, Mueller C. 1978. *Introduction to Factor Analysis: What It Is and How to Do It*. Sage University Paper: Quantitative Applications in the Social Sciences 13. Beverley Hills, CA: Sage Publications.

King NB. 2003. Immigration, Race and Geographies of Difference in the Tuberculosis Pandemic. In: Gandy M, Zumla A (eds.). *The Return of the White Plague: Global Poverty and the 'New' Tuberculosis*. London: Verso.

Krieger N, Chen JT, Waterman PD, Rehkopf DM, Subramanian SV. 2003. Race/Ethnicity, Gender, and Monitoring Socioeconomic Gradients in Health: A

Comparison of Area-Based Socioeconomic Measures-The Public Health Disparities Geocoding Project. *Am J Public Health*, 93(10): 1655-71.

Krieger N, Waterman PD, Chen JT, Soobader M-J, Subramanian SV. 2003. Monitoring Socioeconomic Inequalities in Sexually Transmitted Infections, Tuberculosis and Violence: Geocoding and Choice of Area-Based Socioeconomic Measures. *Public Health Rep*, 118: 240-60.

Krieger N, Williams DR, Moss NE. 1997. Measuring Social Class in US Public Health Research: Concepts, Methodologies, and Guidelines. *Annu Rev Public Health*, 18: 341-78.

Leung CC, Yew WW, Tam CM, Chan CK, Chang KC, Law WS, Wong MY, Au KF. 2004. Socio-Economic Factors and Tuberculosis: a District-Based Ecological Analysis in Hong Kong. *Int J Tuberc Lung Dis*, 8(8): 958-64.

Lienhardt C, Fielding K, Sillah JS, Bah B, Gustafson P, Warndorff D, Palayew M, Lisse I, Donkor S, Diallo S, Manneh K, Adegbola R, Asby P, Bah-Sow O, Bennett S, McAdam K. 2005. Investigations of the Risk Factors for Tuberculosis: a Case-Control Study in Three Countries in West Africa. *Int J Epidemiol*, 34: 914-23.

Lochner K, Pamuk ER, Makuc D, Kennedy BP, Kawachi I. 2001. State-level Income Inequality, Household Income and Individual Mortality Risk: a Prospective Multilevel Study. *Am J Public Health*, 91(3): 385-91.

Lynch JW, Davey Smith G, Kaplan GA, House JS. 2000. Income Inequality and Mortality: Importance to Health of Individual Income, Psychosocial Environment, or Material Conditions. *Br Med J*, 329: 1200-4.

Lynch JW, Kaplan GA, Pamuk ER, Cohen RD, Balfour JL, Yen IH. 1998. Income Inequality and Mortality in Metropolitan Areas of the United States. *Am J Public Health*, 88: 1074-80.

Lynch JW, Kaplan GA. 2000. Socioeconomic Position. In: Berkman, LF, Kawachi I (eds). *Social Epidemiology*. New York, NY: Oxford University Press.

MacIntyre S, Ellaway A, Cummins S. 2002. Place Effects on Health: How Can We Conceptualise, Operationalise and Measure Them? *Soc Sci Med*, 55: 125-39.

MacIntyre S, Ellaway A. 2000. Ecological Approaches: Rediscovering the Role of the Physical and Social Environment. In: Berkman, LF, Kawachi I (eds). *Social Epidemiology*. New York, NY: Oxford University Press.

Maher D, Raviglione M. 2005. Global Epidemiology of Tuberculosis. *Clin Chest Med*, 26: 167-82.

Malmström M, Johansson S, Sundquist J. 2001. A Hierarchical Analysis of Long-Term Illness and Mortality in Socially Deprived Areas. *Soc Sci Med*, 53: 265-75.

Malmström M, Sundquist J, Johansson S-E. 1999. Neighborhood Environment and Self-Reported Health Status: A Multilevel Analysis. *Am J Public Health*, 89(8): 1181-86.

Mangtani P, Jolley DJ, Watson JM, Rodrigues LC. 1995. Socioeconomic Deprivation and Notification Rates for Tuberculosis in London during 1982-91. *Br Med J*, 310: 963-6.

Marmot M. 2002. The Influence of Income on Health: Views of an Epidemiologist. *Health Affairs*, 21 (2): 31-46.

Martens EP, Pestman WR, De Boer A, Belitser AV, Klungel OH. 2006. Instrumental Variables: Applications and Limitations. *Epidemiol*, 17: 260-7.

Mason CM, Dobard E, Zhang P, Nelson S. 2004. Alcohol Exacerbates Murine Pulmonary Tuberculosis. *Infect Immun*, 72(5): 2556-63.

Mathers CD, Iburg KM, Salomon JA, Tandon A, Chatterji S, Ustun B, Murray CJ. 2004. Global Patterns of Healthy Life Expectancy in the Year 2002. *BMC Public Health*, 4:66.

Mathers CD, Schofield DJ. 1998. The Health Consequences of Unemployment: the Evidence. *Med J Aust*, 168: 178-82.

Maurya V, Vijayan VK, Shah A. 2002. Smoking and Tuberculosis: an Association Overlooked. *Int J Tuberc Lung Dis*, 6(11): 942-51.

McKeown, T. 1976. *The Modern Rise of Population*. New York, NY: Academic Press.

Merlo J, Chaix B, Yang M, Lynch J, Råstam L. 2005a. A Brief Conceptual Tutorial of Multilevel Analysis in Social Epidemiology: Linking the Statistical Concept of Clustering to the Idea of Contextual Phenomenon. *J Epidemiol Community Health*, 59: 443-9.

Merlo J, Chaix B, Yang M, Lynch J, Råstam L. 2005b. A Brief Conceptual Tutorial of Multilevel Analysis in Social Epidemiology: Investigating Contextual Phenomena in Different Groups of People. *J Epidemiol Community Health*, 59: 729-36.

Merlo J, Chaix B, Yang M, Lynch J, Råstam L. 2005c. A Brief Conceptual Tutorial of Multilevel Analysis in Social Epidemiology: Interpreting Neighbourhood Differences and the Effect of Neighbourhood Characteristics on Individual Health. *J Epidemiol Community Health*, 59: 1022-9.

Merlo J. 2003. Multilevel Analytical Approaches in Social Epidemiology: Measures of Health Variation Compared with Traditional Measures of Association. *J Epidemiol Community Health*, 57: 550-2.

Mishra V, Vaessen M, Boerma JT, Arnold F, Way A, Barrere B, Cross A, Hong R, Sangha J. 2006. HIV Testing in National Population-based Surveys: Evidence from the Demographic and Health Surveys. *Bull World Health Organ*, 84: 537-45.

Munch Z, Van Lill SW, Booysen CN, Zietsman HL, Enarson DA, Beyers N. 2003. Tuberculosis Transmission Patterns in a High-Incidence Area: a Spatial Analysis. *Int J Tuberc Lung Dis*, 7 (3): 271-7.

Myer L, Ehrlich RI, Susser ES. 2004. Social Epidemiology in South Africa. *Epidemiol Rev*, 26: 112-23.

Oakes JM. 2004. The (Mis)estimation of Neighborhood Effects: Causal Inference for a Practical Social Epidemiology. *Soc Sci Med*, 58: 1929-52.

Pablos-Méndez A, Blustein J, Knirsch CA. 1997. The Role of Diabetes Mellitus in the Higher Prevalence of Tuberculosis Among Hispanics. *Am J Public Health*, 87: 574-79.

Packard R. 1989. *White Plague, Black Labor: Tuberculosis and the Political Economy of Health and Disease in South Africa*. Berkeley, CA: University of California Press.

Parslow R, El-Shimy NA, Cundall DB, McKinney PA. 2001. Tuberculosis, Deprivation, and Ethnicity in Leeds, UK, 1982-1997. *Arch Dis Child*, 84(2): 109-13.

Petrelli A, Gnavi R, Marinacci C, Costa G. 2006. Socioeconomic Inequalities in Coronary Heart Disease in Italy: A Multilevel Population-Based Study. *Soc Sci Med*, 63(2): 446-56.

Phelan JC, Link BG, Diez-Roux A, Kawachi I, Levin B. 2004. "Fundamental Causes" of Social Inequalities in Mortality: A Test of the Theory. *J Health Soc Behav*, 45: 265-85.

Pickett KE, Pearl M. 2001. Multilevel Analyses of Neighbourhood Socioeconomic Context and Health Outcomes: A Critical Review. *J Epidemiol Community Health*, 55: 111-22.

Ponce-de-Leon A, Garcia-Garcia MdL, Garcia-Sancho MC, Gomez-Perez FJ, Valdespino-Gomez JL, Olaiz-Fernandez G, Rojas R, Ferreyra-Reyes L, Cano-Arellano B, Bobadilla M, Small PM, Sifuentes-Osornio J. 2004. Tuberculosis and Diabetes in Southern Mexico. *Diabetes Care*, 27: 1584-90.

Preston SH. 1975. The Changing Relation between Mortality and Level of Economic Development. *Population Studies*, 29(2): 231-48.

Rabe-Hesketh S, Pickles A, Skrondal S. 2004. *GLLAMM Manual*. UC Berkeley Division of Biostatistics Working Paper Series: Working Paper 160. Available from URL: http://www.bepress.com/ucbbiostat/paper160 [Accessed 7 August 2006].

Raviglione M, Harries A, Msika R, Wilkinson D, Nunn P. 1997. Tuberculosis and HIV: Current Status in Africa. *AIDS*, 11: S115–S123.

Reijneveld SA. 1998. The Impact of Individual and Area Characteristics on Urban Socioeconomic Differences in Health and Smoking. *Int J Epidemiol*, 27(1): 33-40.

South African Data Archive. 2003. *Census 1996: 10% Sample Codebook*. SADA 0129. Pretoria: South African Data Archive.

South African Demographic and Health Survey 1998: Full Report. 1999. Parrow: Medical Research Council.

Schwartz S, Diez-Roux R. 2001. Commentary: Causes of Incidence and Causes of Cases – a Durkheimian Perspective on Rose. *Int J Epidemiol*, 30: 435-9.

Schwenk A, Macallan DC. 2000. Tuberculosis, Malnutrition and Wasting. *Curr Opin Clin Nutr Metab Care*, 3(4): 285-91.

Singh GK, Stahpush M. 2002. Increasing Inequalities in All-Cause and Cardiovascular Mortality among US Adults Aged 25-64 Years by Area Socioeconomic Status, 1969-1998. *Int J Epidemiol*, 31: 600-13.

Smith GD, Hart C, Watt G, Hole D, Hawthorne V. 1998. Individual Social Class, Area-Based Deprivation, Cardiovascular Disease Risk Factors, and Mortality: the Renfrew and Paisley Study. *J Epidemiol Community Health*, 52(6):399-405.

Snijders TA, Bosker RJ. 1999. *Multilevel Analysis: an Introduction to Basic and Advanced Multilevel Modeling*. London: SAGE Publications.

Söderlund N, Hansl B. 2000. Health Insurance in South Africa: an Empirical Analysis of Trends in Risk-Pooling and Efficiency Following Deregulation. *Health Policy Planning*, 15(4): 378-385.

Souza WV, Ximenes R, Albuquerque MFM Lapa TM, Portugal JL, Lima ML, Martelli CM. 2000. The Use of Socioeconomic Factors in Mapping Tuberculosis Risk Areas in a City in Northeastern Brazil. *Rev Panam Salud Publica* 2000, 8(6): 403-10.

Spence DP, Hotchkiss J, Williams CS, Davies PD.1993. Tuberculosis and Poverty. *Br Med J*, 307: 759-61.

Statistics South Africa. n.d. Census 1996: Meta-Data for the Household File. Available from URL: http://www.statssa.gov.za/census01/census96/HTML/Metadata/Metadta/comm_profiles_dbases_toc.htm [Accessed 7 August 2006]

Statistics South Africa. 1998. *The People of South Africa, Population Census 1996: The Count and How it Was Done*. Report No 03-01-17(1996). Pretoria: Statistics South Africa.

Statistics South Africa. 2000a. *Measuring Poverty in South Africa*. Pretoria: Statistics South Africa.

Statistics South Africa. 2000b. *The People of South Africa, Population Census 1996: Summary Report*. Report No 03-01-12(1996). Pretoria: Statistics South Africa.

Statistics South Africa. 2004. *Primary Tables South Africa: Census '96 and 2001 Compared*. Report No 03-02-04. Pretoria: Statistics South Africa.

Statistics South Africa. 2006. *General Household Survey, July 2005*. Statistical Release P0318. Pretoria: Statistics South Africa.

Strebel PM, Seager R. 1991. Epidemiology of Tuberculosis in South Africa. In Coovadia HM, Benatar SR (eds). *A Century of Tuberculosis: South African Perspectives*. Cape Town: Oxford University Press.

Subramanian AV, Belli P, Kawachi I. 2002. The Macroeconomic Determinants of Health. *Annu Rev Public Health*, 23:287-302.

Subramanian SV, Delgardo I, Jadue L, Vega J, Kawachi I. 2003. Income Inequality and Health: Multilevel Analysis of Chilean Communities. *J Epidemiol Community Health*, 57: 844-8.

Subramanian SV, Kawachi I. 2004. Income Inequality and Health: What Have We Learned So Far? *Epidemiol Rev*, 26: 78-91.

Subramanian SV, Kawachi I. 2006. Whose Health is Affected by Income Inequality? A Multilevel Interaction Analysis of Contemporaneous and Lagged Effects of State Income Inequality on Individual Self-Rated Health in the United States. *Health Place*, 12: 141-56.

Sundquist J, Malmström M, Johansson S-E. 1999. Cardiovascular Risk Factors and the Neighbourhood Environment: A Multilevel Analysis. *Int J Epidemiol*, 28: 841-5.

Sundquist K, Malmström M, Johansson S-E. 2004. Neighbourhood Deprivation and Incidence of Coronary Heart Disease: A Multilevel Study of 2.6 Million Women and Men in Sweden. *J Epidemiol Community Health*, 58: 71-7.

Tekkel M, Rahu M, Loit H-M, Baburin A. 2002. Risk Factors for Pulmonary Tuberculosis in Estonia. *Int J Tuberc Lung Dis*, 6(10): 887-94.

Thorson A, Diwan VK. 2003. Gender and Tuberculosis: A Conceptual Framework for Identifying Gender Inequalities. In: Gandy M, Zumla A (eds). *The Return of the White Plague: Global Poverty and the 'New' Tuberculosis*. London: Verso.

Tocque K, Regan M, Remmington T, Beeching NJ, Jamieson I, Syed Q, Davies PD. 1999. Social Factors Associated with Increases in Tuberculosis Notifications. *Eur Repir J*, 13(3): 541-5.

van Helden PD. 2003. The Economic Divide and Tuberculosis. *EMBO Reports*, 4: S24-8.

van Lenthe FJ, Schrijvers CT, Droomers M, Joung IM, Louwman MJ, Machenbach JP. 2004. Investigating Explanations of Socio-Economic Inequalities in Health. *Eur J Public Health*, 14(1): 63-70.

Veenstra G. 2005. Location, Location, Location: Contextual and Compositional Health Effects of Social Capital in British Columbia, Canada. *Soc Sci Med*, 60(9): 2059-71.

Wilkinson RG, Pickett KE. 2006. Income Inequality and Population Health: A Review and Explanation of the Evidence. *Soc Sci Med*, 62: 168-84.

Winkleby MA, Jatulis DE, Frank E, Fortmann SP. 1992. Socioeconomic Status and Health: How Education, Income, and Occupation Contribute to Risk Factors for Cardiovascular Disease. *Am J Public Health*, 82(6): 816-20.

Woolard I, Leibbrandt M. 1999. *Measuring Poverty in South Africa*. DPRU Working Paper 99/33. Cape Town: Development Policy Research Unit.

World Bank. *World Development Indicators 2004*. 2004. Available from URL: http://devdata.worldbank.org/data-query [Accessed 7 August 2006].

World Health Organisation. 1999. Global Tuberculosis Control: Annual Report 1999. WHO/CDS/CPC/TB/99.259. Geneva: World Health Organisation.

World Health Organisation. 2005. Global Tuberculosis Control: Annual Report 2005. WHO/HTM/TB/2005.349. Geneva: World Health Organisation.

Wulfsohn M. 1985. Epidemiology of Tuberculosis. *Epidemiol Comments*, 12(9): 2-19.