An Econometric Method for Estimating Population Parameters from Non-Random Samples: An Application to Clinical Case Finding

Rulof Burger, Ph.D.
Associate Professor
Dept. of Economics
University of Stellenbosch
rulof@sun.ac.za

Zoë M. McLaren, Ph.D.*
Assistant Professor
School of Public Health
University of Michigan
zmclaren@umich.edu

DRAFT – Please do not cite without permission

This version: April 6, 2017

Abstract

The problem of sample selection complicates the process of drawing inference about populations. Selective sampling arises in many real world situations when agents such as doctors and customs officials search for targets with high values of a characteristic. We propose a new method for estimating population characteristics from these types of selected samples. We develop a model that captures key features of the agent's sampling decision. We use a

* Corresponding author. Both authors contributed equally to this work. Names are ordered alphabetically.

generalized method of moments with instrumental variables and maximum likelihood to estimate the population prevalence of the characteristic of interest and the agents' accuracy in identifying targets. We apply this method to tuberculosis (TB), which is the leading infectious disease cause of death worldwide. We use a national database of TB test data from South Africa to examine testing for multi-drug resistant TB (MDR-TB). Approximately one-quarter of MDR-TB cases were undiagnosed between 2004-2010. The official estimate of 2.5% is therefore too low and MDR-TB prevalence is as high as 3.5%. Signal-to-noise ratios are lower for new patients than for repeat patients. Our approach is widely applicable because of the availability of routinely collected data and abundance of potential instruments. Using routinely collected data to monitor population prevalence can guide evidence-based policy making.

1. Introduction

The problem of sample selection complicates the process of drawing inference about populations when the characteristic of interest is only observed for a non-random subsample of the population. This type of selective sampling frequently arises in many real world situations when agents are tasked with searching for targets with high values of a particular characteristic of interest. For example, doctors decide to test certain patients for diseases based on observed symptoms or risk factors; tax authorities select firms and individuals for audits based partly on the content of tax returns; customs officials investigate suspicious travelers and shipments; police pull over erratic drivers; universities and firms interview candidates with the most promising resumes. In most of these cases, data is passively collected during the process, which provides

an opportunity to draw inference about these populations if the sample selection problem can be addressed. In this study, we propose a new method for estimating population characteristics from a selected sample of observations drawn using the abovementioned "high-value search" sampling mechanism, which is intended to identify population targets with high-values of a particular characteristic, not facilitate the estimation of specific population attributes.

When it is not possible to perfectly discern high-value from low-value targets or to select all targets in the population for testing, agents must triage available resources to test only those targets deemed most likely to have high values of the characteristic of interest. We develop a model that captures key features of the agent's decision making surrounding the sampling of targets from the population. We then use a generalized method of moments and maximum likelihood to estimate (a) the prevalence of the characteristic of interest in the population and (b) the accuracy with which agents are able to identify high-value targets. In addition, we implement an instrumental variables method that leverages exogenous, discontinuous changes in the availability of resources for testing. These types of sample selection adjustments are an attractive alternative to random samples or testing the full population. Performing random sample surveys of the population can provide unbiased estimates, but these types of surveys are infrequently conducted because they entail significant financial and time costs. Testing or selecting all observations in the population is generally neither feasible nor cost-effective.

We apply our method to the problem of estimating disease prevalence. Accurate information on disease prevalence is essential for health policy making so that limited resources can be targeted globally and nationally to improve patient outcomes and maximize overall population health.

We focus on tuberculosis (TB), which kills more than 1.5 million people annually and recently surpassed HIV to become the leading infectious disease cause of death in the world (WHO 2015). Though multi-drug resistant TB (MDR-TB) patients comprised an estimated 5% of TB cases notified in 2014, they accounted for 13% of TB deaths and 20% of TB spending worldwide (WHO 2015). However, only 12% of incident TB cases were tested for MDR-TB.

We use 7 years of data from South Africa's National Health Laboratory Service (NHLS) database, which has wide national coverage and includes individual patient TB test results. Our identification strategy leverages a series of national and local policy changes that increasingly prioritized MDR-TB as an urgent health concern to enable us to obtain precise estimates of the prevalence of MDR-TB. We find that between 16 and 27 percent of all MDR-TB cases were undiagnosed between 2004-2010 in South Africa, which worsens patient outcomes, increases transmission and leads to the development of additional drug resistance. These estimates are validated against results from national surveillance surveys that use random sampling methods.

The contribution of this paper is two-fold. First, we develop a new method for estimating population parameters from non-random samples, which is widely applicable to many settings. It builds on the literature that addresses selection on unobservables, which focuses on two approaches: instrumental variables (see Imbens and Angrist 1994, Angrist and Imbens 1995) or selection models (see Heckman 1976). We demonstrate the ease with which this method can be applied to other contexts because it uses existing data, is low cost and can be quickly scaled up. The HIV literature has demonstrated the importance of using statistical methods to adjust ante-natal care and population prevalence estimates for representativeness (see Sakarovitch et al.

2007, Nyirenda et al. 2010, Bärnighausen et al. 2011, Hogan et al. 2012, Clark and Houle 2014, McGovern et al. 2015), however we develop more rigorous methods for routine data and are the first to apply these types of methods to TB. Second, by applying the model to the case of MDR-TB, we are the first to demonstrate that approximately one-quarter of MDR-TB cases in South Africa went undiagnosed between 2004-2010, which is a significant threat to TB control. Our method is a significant improvement over than the standard World Health Organization method for estimating TB prevalence estimates that performs a crude adjustment to notified TB cases to account for under-detection (Glaziou et al. 2015; See appendix 1 for details). Our method can be employed in low- and middle-income countries to cost-effectively develop guidance for health policy making and ultimately improve population health.

2. Model

2.1 Assumptions

Consider a population characterized by two variables x and y, which have the joint cumulative distribution function $F_{X,Y}(x, y)$ such that $E(y|x)$ is a monotonic function of x, scaled to be non-decreasing. This assumption is reasonable if x is a good predictor of y such as when x is a symptom indicative of y or a risk factor with positive predictive value. It is particularly reasonable in cases where the observable x variable is partly caused by the unobservable value of y, such as if the disease y produces a symptom x. Suppose the analyst is interested in estimating the (unconditional) population mean of y denoted $\mu \equiv E(y)$, but the analyst only has access to a non-random (i.e. selected or purposive) sample. This sample was drawn by agents with the purpose of selecting target observations with the highest expected values of y (unobservable at the time of sampling) based on values of x, which are observable to the agent but not necessarily

to the analyst. Sampling reveals the values of y for the sampled observations, but has an associated cost.

We demonstrate how, providing there is sufficient exogenous variation in the proportion sampled and given specific assumptions about stationarity regarding population characteristics and the agents' sampling algorithms, we can draw inferences about both the sampling mechanism and the distribution of y in the population.

Define θ as the proportion of the population that is sampled, which is determined exogenously to the agent. Define x_0 as the threshold value such that values of x equal or greater than x_0 are sampled: $x \geq x_0 = F_X^{-1}(1-\theta_0)$, where $F_X(x)$ is the cumulative distribution function of the marginal distribution of x: $F_X(x) = F_{X,Y}(x, \infty)$. The analyst knows the value of θ and sampled values of y, but not necessarily the values of x, so is generally convenient to express the expected value of the sampled value of the outcome in terms of θ rather than x: $E(y|\theta \leq \theta_0) = E(y|x \geq x_0)$.

2.2 Identifying the population mean

It is difficult to draw inferences about the population mean, μ, from $E(y|\theta \leq \theta_0)$ because a high value can be indicative of a high μ, but may also reflect an efficient sampling mechanism that identifies a higher share of high-y observations in a low-μ population. The marginal sampling efficiency at a specific sampling share (i.e. the expected value of the outcome for observations that are added to the sample when the sampling share increases infinitesimally) characterizes the nature of the selective sampling and can be leveraged to infer information about the population mean (μ):

$$E(y|\theta_0) \equiv \lim_{\Delta\theta \to 0} E(y|\theta_0 \leq \theta \leq \theta_0 + \Delta\theta) = E(y|\theta \leq \theta_0) + \theta_0 \frac{dE(y|\theta \leq \theta_t)}{d\theta_t}$$

Since θ_0, $E(y|\theta \leq \theta_t)$ and $\frac{dE(y|\theta \leq \theta_t)}{d\theta_t}$ are observable to the analyst, we can also treat $E(y|\theta_0)$ as observable. Exogenous variation in θ across time or subgroups that have the same joint distribution of $F(y, x)$ can be used to identify the marginal sampling efficiency $E(y|\theta_0)$, which can provide a plausible basis for drawing inferences about μ.

2.3 Statistical inference

Provided that we observe outcomes for subsamples of the population that are selected using the same sampling method, it is instructive to observe how $E(y|\theta \leq \theta_t)$ varies with changes in θ_t. If $\frac{dE(y|\theta \leq \theta_t)}{d\theta_t} = 0$ then this implies that $\mu = E(y|\theta \leq \theta_0)$. Intuitively, if an increase in the sample average of the outcome remains unchanged when the sampling share is increased, then this suggests that the sampling occurs in a haphazard way that is unable to effectively distinguish between high-y and low-y observations. Provided that all sampling shares between 0 and 1 are drawn based on a consistent mapping of characteristics x to outcome y, then $E(y|\theta \leq \theta_0)$ provides an unbiased estimate of μ. On the other hand, if we can reject the hypothesis that $\frac{dE(y|\theta \leq \theta_t)}{d\theta_t} = 0$ in favor of $\frac{dE(y|\theta \leq \theta_t)}{d\theta_t} < 0$, then sampling can be concluded to occur based on informative values of x.

The simplest approach to point-identifying μ is to express the observable subgroup means in the general form as a function of θ_t rather than as a joint function $F_{X,Y}(x, y)$:

$$E(y|\theta \leq \theta_t) = \int_{-\infty}^{\infty} \frac{yP(\theta \leq \theta_t|y)f_Y(y)}{\theta_t} dy$$

and then to make distributional assumptions about $P(\theta \leq \theta_t|y)$ and $f_Y(y)$. When the outcome is binary, the conditional probability simplifies to:

$$P(y = 1|\theta \leq \theta_t) = \frac{P(\theta \leq \theta_t|y = 1)\mu}{\theta_t}$$

which only requires assumptions about the distribution of $P(\theta \leq \theta_t|y = 1)$ in order to identify μ. If x is known to be partly determined by the value of the outcome y, then expressing x in error form as $x = \beta_0 + \beta_1 y + e$ where $E(e|y) = 0$ makes it possible to write the observable sample share of the outcome as:

$$P(y = 1|\theta \leq \theta_t) = \frac{P(e \geq F_X^{-1}(1 - \theta_t)\text{-}\beta_0\text{-}\beta_1)\mu}{\theta_t}$$

A distributional assumption for the error term e is therefore all that is required to identify both $\beta_0 + \beta_1$ and μ.

3. Context

TB has been the leading cause of death for over a decade in South Africa but the lack of reliable estimates of local TB prevalence makes it difficult to allocate government resources efficiently (Statistics South Africa 2011). Conventional thinking about estimating MDR-TB prevalence focuses on two avenues that are expensive and logistically complex: increasing the frequency and coverage of TB prevalence studies and expanding access to high-technology testing (e.g. Xpert) so that everyone can be tested for drug resistance (Cohen et al. 2008, Weyer et al. 2013). Theron et al. (2015) calls for better use of existing data to inform tailored responses in the fight against TB, however advances in this area have been slow.

Official guidelines counsel clinicians to screen patients for TB based on symptoms (current cough, weight loss, night sweats, fever) which are non-specific to TB (i.e. a noisy signal) and do not differ for patients with MDR-TB (Department of Health 2013). Before Xpert drug resistance testing was widely available in South Africa, the guidelines indicated that *only* those with the highest risk of MDR-TB should be tested, which included people with TB symptoms who had been previously treated for TB, patients who had failed TB treatment, and those who were known contacts of MDR-TB cases (Department of Health 2013). Clinicians can ascertain a patient's approximate risk of MDR-TB from a medical history and physical exam to determine a patient's MDR-TB risk relative to other patients before ordering laboratory testing. The risk factors are a good but imperfect signal of MDR-TB so many cases could initially go undetected (Kendall et al. 2015).

4. Methods

4.1 Data

We use data from South Africa's National Health Laboratory Service (NHLS) database on TB tests performed on patients aged 16-64 in public health facilities (hospitals and health clinics) for the period January 2004 - September 2010, which includes 8,647,12 patients in 4,764 facilities. Our analysis sample comprises 2,190,780 TB-positive test records, 249,779 of which are tested for MDR-TB. For TB and drug susceptibility testing, the data include the type of test performed, test result, testing facility location, test date and basic patient demographics. About 50.11% of the sample is female and the median age is 37 (std. dev. 11) with an interquartile range of 29-46

years. Race data is poorly recorded in laboratory data however, of those patients with information on race, 96.75% are Black African. We consider TB-positive results from culture testing and smear microscopy. Patient records are linked using unique patient identifiers created by the NHLS. Our dataset spans almost 7 years (27 quarters) of frequent observations, which allows us to observe several sudden policy shifts that affected the inclination to test for MDR-TB. Figure 1 shows the patients who are TB-positive, tested for MDR-TB, and MDR-TB-positive by quarter over time.

In some analyses, we use data only from new patients in order to limit the sample to one diagnostic episode per patient, exclude treatment monitoring tests, and examine the sample without a history of TB testing (which in most cases implies without a history of TB treatment). Ethics approval was obtained from the University of Michigan Institutional Review Board and the University of Cape Town Faculty Ethics in Research Committee.

4.2 Identification strategy

We apply our model to the clinician's decision to perform drug susceptibility testing on a patient with suspected TB. Clinicians search for MDR-TB-positive patients (i.e. high-value targets) because a confirmed diagnosis is required to initiate MDR-TB patients onto treatment. Prior to 2011, guidelines did not recommend testing all TB-positive patients for MDR-TB and MDR-TB testing resources were highly limited. In the absence of resources to test every patient for drug resistance, the clinician's testing decision is based on their knowledge of the patient's risk factor

profile and the policy guidelines in place, subject to having time and resources available for drug resistance testing. Suppose the clinician observes information about the patient's underlying propensity to have MDR-TB in the form of a noisy signal (x) that includes non-specific symptoms such as cough and fever as well as the presence of risk factors. The clinician can use this observed signal to determine a patient's relative likelihood of having MDR-TB conditional on being TB-positive. Clinicians will triage patients and order drug resistance testing for the proportion θ of patients with the highest expected likelihood of having MDR-TB based on the noisy signal.

Variation in the proportion of individuals who can be tested for MDR-TB, θ, is caused by fluctuations in the availability of time, staff, testing materials or other resources required for testing. Discontinuous changes in the availability of resources for or attention towards MDR-TB testing from period to period may occur due to stochastic events (e.g. stockouts, staff strikes, etc.) or institutional changes (e.g. policy changes or new budget allocations). Exogenous changes in θ should not affect the clinician's ability to rank patients (which is based on information and beliefs) or the underlying share of drug-resistant patients in the short run, but will influence MDR-TB prevalence among those tested by shifting the characteristics of the marginal tested patient.

If changes in the proportion of patients tested from period to period are exogenous, we can use the time dummies as instrumental variables to identify the parameters of interest. However, we may be concerned that using variation over time in the MDR-TB testing rate as a source of identifying variation may reflect underlying transmission dynamics of the epidemic that drive the

true prevalence. We therefore also use instruments based on exogenous institutional variation in MDR-TB testing rates due to policy changes. The national policy changes are orthogonal to facility-level variation in the lagged proportion of patients tested and lagged MDR-TB prevalence because their timing is neither determined by clinician decision making nor by deviations from the underlying prevalence trend which were unmeasured during this period. Intuitively, these instruments represent discontinuous changes in θ that cannot, in the short term, be correlated with relatively smooth trends in prevalence or the signal-to-noise ratio which captures the clinician's ability to determine a patient's relative likelihood of having MDR-TB. In addition, the lag between a change in beliefs and the implementation of a new policy supports instrument exogeneity.

Clinicians' beliefs about the relative importance of symptoms and risk factors in predicting MDR-TB is unlikely to have systematically changed during this period because there was little new information on the mapping between risk factors and prevalence. In fact, an Institute of Medicine Forum noted that the true extent of the MDR-TB epidemic was unknown due to limited resources and inadequacies in existing health systems infrastructure (Institute of Medicine 2011). Rising concern about MDR-TB shifts the proportion tested but should not affect patients' relative likelihood of having MDR-TB.

We include the following policy instruments: a conference presentation of an extensively drug resistant TB (XDR) study from South Africa at the Conference on Retroviruses and Opportunistic Infections (CROI) raised concern about drug-resistance (April 2006); the introduction of the National Strategic Plan on HIV, STIs and TB (June 2006) and release of

updated National TB Guidelines (April 2009) brought increased attention and resources to TB; national Kick TB Campaign raises concern about TB (March 2010). We also use an instrument for the facility-level anti-retroviral therapy (ART) availability and one for the local-area-level ART availability to incorporate sub-national exogenous variation in access to AIDS treatment due to the staggered program rollout beginning in July 2004. With the possible exception of the ART rollout, these policy changes are unlikely to change the composition of population of people present at health facilities or affect the ranking ability of clinicians and should therefore serve as valid instruments.

4.3 Estimation

In our empirical analysis we assume that *e* in the equation $x = \beta_0 + \beta_1 y + e$ follows a standard normal distribution *e*~nid(0,1) and we normalize β_0 to 0. This is sufficient to identify the model parameters signal-to-noise ratio $\beta = \frac{\beta_1}{\sigma_e}$ and population mean μ with observable sample variation in $P(y = 1|\theta \leq \theta_t)$ induced by different values of θ_t across time. We estimate the model first assuming that μ is constant over time, and then allowing μ to take a quadratic function of time where the prevalence is restricted to fall within the unit interval:

$$\mu_t = \Phi\left(\mu_0^* + \mu_1^* t + \mu_2^* t^2\right),$$

where Φ() represents the standard normal cumulative distribution function and *t* is the time period. Estimation of this model is complicated by the fact that x follows a mixed normal distribution and its inverse will not generally have a closed-form solution:

$$F_X(x) = \mu\Phi(x\text{-}\beta) + (1-\mu)\Phi(x).$$

The inverse, $F_X^{-1}(1-\theta_t)$, must therefore approximated using either simulations or numerical approximation techniques. We use two estimators to obtain point estimates for our model parameters: a generalized method of moments estimator $F_X^{-1}$and a maximum likelihood estimator. In both cases a Hermite polynomial of degree 7 is used to approximate F_X^{-1}).

4.3.1 Maximum likelihood estimation

The log-likelihood of the model in section 2 can be expressed as

$$\log L(\mu,\beta) = \sum_{i=1}^{N}\left[y_{it}\log\left(\frac{\mu}{\theta_t}\{1-\Phi(F_X^{-1}(1-\theta_t)-\beta)\}\right) + (1-y_{it})\log\left(1-\frac{\mu}{\theta_t}\{1-\Phi(F_X^{-1}(1-\theta_t)-\beta)\}\right)\right]$$

A Gaussian quadrature procedure is used to approximate the function F_X^{-1} with a Hermite polynomial of order 7. Numerical optimization techniques are then applied to find the values of the model parameters that maximize the likelihood function. In estimating the model, we allow MDR-TB prevalence to be either constant over time or to evolve as a quadratic function of time. The starting parameter values are obtained after performing a series of grid searches to obtain the most promising parameter space.

$$F_X^{-1}(1-\theta_t)$$

4.3.1 Generalized method of moments estimation

We can define the model error term (i.e. the difference between observed and predicted outcomes) as

$$u_{it} = y_{it}\text{-}P(y_{it} = 1|\theta \leq \theta_t).$$

The generalized method of moments (GMM) estimator exploits the exogeneity assumption:

$$E(u_{it}|\Omega) = 0$$

where Ω represents all the information available to the clinician at the time of the testing decision. In the sample data this implies that

$$E[z_{it}\{y_{it}\text{-}P(y_{it} = 1|\mu, \beta, \theta_t)\}] = 0$$

where z_{it} is a vector of instrumental variables, the elements of which are believed to be orthogonal to the individual's likelihood of having MDR-TB. If we define the GMM moment function as

$$g_{it}(y_{it}, z_{it}, \mu, \beta, \theta_t) = z_{it}\{y_{it}\text{-}P(y_{it} = 1|\mu, \beta, \theta_t)\}$$

then the generalized method of moments estimator can be expressed as

$$\arg\min_{\mu,\beta} \left(\sum_{t=1}^{T}\sum_{i=1}^{N} g_{it}(y_{it}, z_{it}, \mu, \beta, \theta_t)\right)' W \left(\sum_{t=1}^{T}\sum_{i=1}^{N} g_{it}(y_{it}, z_{it}, \mu, \beta, \theta_t)\right)$$

where W is the weighting matrix. As with the maximimum likelihood estimator, we cannot calculate $P(y_{it} = 1|\mu, \beta, \theta_t)$ analytically, but replace this probability with its approximated counterpart using Gaussian quadrature methods.

F_X^{-1}

4.3.3 Tests of validity and robustness

We estimate MDR-TB prevalence using maximum likelihood as well as generalized method of moments with either time periods or policy changes as instruments. We calculate estimates of the sampling efficiency (i.e. MDR-TB prevalence among those tested) for the sample over the range of the sampling proportion (θ), as well as the marginal sampling efficiency. To evaluate our method's goodness of fit we compare our the time pattern of MDR-TB estimates to the

observed pattern. We also report two pseudo-R^2 measures: (1) the ratio of variances of predicted to observed MDR-TB prevalence and (2) the squared correlation coefficient between predicted and observed MDR-TB prevalence. Asymptotically valid standard errors are calculated using the relevant parameter covariance matrices. GMM estimates are clustered at the quarter –by-facility level.

5. Results

Figure 2 shows that the proportion of TB-positive patients tested for MDR-TB (•) and the proportion that test positive among those tested (▲) are negatively correlated, both in long-run trends and short-term fluctuations. These patterns are consistent with our theoretical model in which clinicians triage TB patients for testing based on the observed likelihood of being MDR-TB. An increase in the tested proportion (θ) implies extending the test to patients deemed less likely to have MDR-TB by the clinicians, in other words the signal-to-noise ratio $\beta>0$ and $\frac{dE(y|\theta\leq\theta_t)}{d\theta_t} < 0$.

Figure 2 also shows that the percent of all TB-positive patients who were diagnosed with MDR-TB (■) tracks the percent of all TB-positive patients who were tested for MDR-TB (•) reasonably well over this period. As more TB-positive patients are tested for MDR-TB, more MDR-TB cases are found (consistent with our model in that θ is a limiting factor and the signal-to-noise ratio $\beta<\infty$) and the share of MDR-TB tested patients who are MDR-TB-positive falls ($\beta>0$).

The percentage of all TB-positive patients (based on smear, culture or PCR) who were tested for MDR-TB (•) was fairly stable between 8-10% from 2004-2006, spiked up at the end of 2006 and again at the end of 2007 before steadily increasing from the end of 2008 to 2010, when it reached 23%. The percentage of all TB-positive patients who tested positive for MDR-TB (■) was stable between 1-1.3% from 2004-2006 and rose up to 2% in 2010. MDR-TB cases as a percentage of all those tested for MDR-TB (▲) was steady at around 12% until late 2007 when it rose to 15% and then steadily declined.

To further investigate the sampling mechanism, Figure 3 plots the sampling efficiency (solid line) as θ varies between approximately 8% and 23% in different quarters of data. As expected, our results show that the prevalence in the selected sample falls when θ rises (i.e. $\frac{dE(y|\theta \leq \theta_t)}{d\theta_t} < 0$). The dashed line represents the marginal sampling efficiency – the expected value of y for observations that are added to the sample when θ increases infinitesimally. It declines up to approximately 18% of the TB+ patients being tested for MDR-TB, and then rises slightly at the highest values of θ.

We estimate that MDR-TB prevalence in South Africa was approximately 3% over this period (Table 1), which is about 0.5 percentage points higher than the 2011 WHO estimate of 2.5% based on case notification rates (WHO 2011). This indicates that between 17 and 30 percent of all MDR-TB cases went undetected during this period. The signal-to-noise ratio (β) is approximately 1; the signal is therefore 50% noise. ML and both versions of GMM-IV methods produce very similar estimates of MDR-TB prevalence (μ) and the signal-to-noise ratio (β) in the full sample, which demonstrates the robustness of our methods. The one exception is the GMM-

IV method with time dummies and clustered standard error, which was as high as 3.42% (Column 3). The standard errors for our estimates are small. The policy change instruments are highly correlated with the testing proportion. In a linear regression of the testing share on the policy instruments, the associated F-test statistic is 200.05 when clustering by quarter and 5.15 when also controlling for a quadratic time trend. Both our pseudo-R^2 measures are approximately 0.70 for most specifications.

When we relax the assumption that MDR-TB prevalence is constant over time, the GMM-IV estimates show that MDR-TB prevalence rose from 5.5% in 2004 to a peak of 5.8% from mid-2005 to mid-2006, and then fell to 4.4% by the end of 2010 (Figure 4, Table 1). The GMM estimates using time dummies and policy changes are very similar, indicating that changes in testing resources over time were fairly exogenous. ML estimates were in general about 0.5 percentage points lower than the GMM estimates. The signal-to-noise ratio ranges between 0.51-0.67. An exponential decay functional form produced substantially lower pseudo-R^2 values than the quadratic and did not restrict prevalence to be between 0 and 1.

Figure 5 provides robust evidence of the validity of our estimation method because the time pattern of MDR-TB prevalence predicted by our model matches the observed MDR-TB prevalence reasonably well in both the long and short run. This shows that the majority of the variation in observed MDR-TB prevalence can be explained by changes in the proportion of patients tested alone. As expected, the match is worse where the observed prevalence has more peaks and troughs (2007-2010).

New patients and patients with a previous test result have different underlying signal-to-noise ratio and estimated MDR-TB prevalence. The ML results show a prevalence of 3.1% for new patients and 5.3% for repeat patients. Though the prevalence survey distinguishes between new and retreatment cases, which we are not able to do in our data, our results for new patients are still close to prevalence survey estimates for new cases (2.1%, 95% CI: 1.5%-2.7%) as are our results for repeat patients compared to retreatment cases (4.6% CI 95%: 3.2%-6.0%) (Centre for Tuberculosis 2016). As expected, the values for β reflect that clinicians have less information upon which to assess the risk profile of the new patients compared to repeat patients. For the ML, β is estimated at 0.59 for new and 1.26 for repeat patients.

6. Discussion

Our results indicate that the assumptions about clinician behavior in our theoretical framework are consistent with the data. Figure 2 shows that clinicians do prioritize testing patients that are more likely to be MDR-TB positive, but that this prioritization is imperfect. Figure 3 shows, as expected, that the sampling efficiency falls as a greater proportion of the population is sampled. Our simple framework is able to match observed patterns in the data very closely (Figure 5), which supports the validity of our approach. The fact that our GMM-IV-policy change estimates differ little from the GMM-IV time dummy estimates provides evidence to support our assumption that the constraint on diagnostic testing resources (θ) changes exogenously over time. The fact that our GMM-IV results change little with the addition of an instrument related to the rollout of ART, which occurred at the facility level and varied geographically and

temporally, provides additional support that μ and β are well-identified. Our results do not exhibit characteristics indicative of weak instruments (Stock, Wright and Yogo 2002).

Our estimates of MDR-TB prevalence are between 16 and 27 percent higher (0.5-1 percentage point) than the WHO 2010 estimate of 2.5% (WHO 2011). Because our data do not have full coverage of KwaZulu-Natal , which likely has the highest MDR-TB burden, our estimates are a lower bound on the true national MDR-TB prevalence. As expected, our results show that MDR-TB prevalence increased following the 2001-02 prevalence study, which found an MDR-TB prevalence of 2.9% overall and 6.6% in the population with a history of TB treatment (Weyer et al. 2007). Extrapolating our GMM-IV-policy change time-varying prevalence results finds an MDR-TB prevalence of 4.5% for 2001-02 which is above the CI of 2.4-3.5% from the prevalence survey, and 3.3% for 2012-2014 which falls within the CI of 2.0-3.6% of the estimate from the 2012-14 prevalence survey (Weyer et al. 2007, Centre for Tuberculosis 2016). In light of these results, it is likely that resource allocation to MDR-TB between 2002-2016 was sub-optimal due to under-estimates of MDR-TB prevalence. Additional resources should be therefore be allocated to the National Tuberculosis Program to increase efforts to control MDR-TB.

From a health policy perspective, high rates of under-detection of MDR-TB highlight the need for additional diagnostic resources and MDR-TB treatment for new cases that are identified. In addition, our new MDR-TB estimates should be used as input parameters for TB modeling studies that inform health policy because MDR-TB prevalence is often highly influential in these models (see Acuna-Villaorduna et al. 2008, Vassall et al. 2011, Meyer-Rath et al. 2012, Dowdy

et al. 2014). Finally, more frequent prevalence surveys are needed to track the evolution of MDR-TB prevalence over time.

6.1 Data Limitations

Our study population is the same as for the TB prevalence studies: individuals who present at a public health facility, are determined to be at risk for TB and have TB testing performed. Both will underestimate the prevalence of TB and MDR-TB in the population to the degree that cases do not present to health facilities, or are overlooked as at-risk by health workers, or due to diagnostic tests not being perfectly sensitive. Though the data have been deduplicated using an algorithm devised by the NHLS, poor patient linking across time may lead to double counting of MDR-TB patients and bias our estimates upwards. If clinicians order drug susceptibility testing only after treatment failure has been observed, then in the data clinicians will appear to have better information (stronger signal value) than they actually do. In the absence of prevalence study benchmarking, this would bias our estimates upwards.

6.2 Conclusion

This study developed a novel econometric method for estimating disease prevalence from routinely collected data. We found that approximately one-quarter of MDR-TB cases in South Africa were undiagnosed between 2004-2010 which contributed to high transmission rates and high TB mortality rates. The empirical evidence supports the validity of our method. These findings demonstrate the need for increased investment in early detection of MDR-TB, such as

the ongoing implementation of Xpert technology, and more effective treatment, such as new antibiotics (WHO 2014).

Our results underscore the importance of continuous surveillance that accounts for under-detection rather than simply relying on notification rates in order to ensure that the health system is diagnosing as many MDR-TB cases as possible. Our method is particularly attractive because it relies solely on existing routine data, which is widely available and inexpensive to collect, and can be run using standard statistical software. In addition, the data requirements for our method are minimal compared to alternative prediction or imputation methods since patient characteristics need not be observed by the analyst. Notably, our method does not require that all patients or a random sample of patients be screened for the disease which makes it a lower-cost alternative to increasing the frequency of prevalence surveys.

Our approach to disease surveillance is simple and adaptable enough to be applied to many infectious and non-infectious diseases in the developing world where prevalence data is lacking. This method can be applied to MDR-TB where access to Xpert drug resistance technology is limited, and to extensively drug resistant TB where Xpert is available but testing for resistance to additional first- and second-line drugs is less common.

Statistical analysis of routinely collected data is an inexpensive and effective way to monitor the prevalence of any number of population characteristics. Our approach is widely applicable because of the minimal data requirements, wide availability of routinely collected data and abundance of policy changes to serve as valid instruments. Our straightforward, yet powerful,

approach can be used to evaluate the effectiveness not only of clinicians, but also of tax authorities, customs officials and law enforcement. Ultimately, using routinely collected data to monitor population prevalence and agent effectiveness is a high-value strategy to guide evidence-based policy making and implementation in resource-limited settings.

Acknowledgements

We thank Sue Candy, Ananta Nanoo, Michelle Potgeiter and Andrew Whitelaw for assistance with the data and helpful comments. We thank Michael Budros, Elizabeth Brouwer, David Ederer, Kathryn Fischer, Alana Sharp and Jifang Zhou for research assistance. We thank Jacob Bor, Florian Marx, Tyler McCormick, Tom Moultrie, Edward Norton, Kate Schnippel, Margaret Triyana, Sean Wasserman, Mark Wilson, seminar audiences at the University of Michigan and Stellenbosch University, and participants at the Union for African Population Studies conference, University of KwaZulu Natal Microeconomic Analysis of South Africa conference and European Workshop on Econometrics and Health Economics and two anonymous referees for helpful feedback. Financial support was provided by the University of Michigan School of Public Health Global Public Health Program, the Center for Global Health and the University of Michigan Health Management and Policy Department McNerney Award.

References

Acuna-Villaorduna, C., Vassall, A., Henostroza, G., Seas, C., Guerra, H., Vasquez, L., ... & Gotuzzo, E. (2008). Cost-effectiveness analysis of introduction of rapid, alternative methods to identify multidrug-resistant tuberculosis in middle-income countries. Clinical Infectious Diseases, 47(4), 487-495.

Angrist, J. D., and G. W. Imbens, "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," Journal of the American Statistical Association 90:430 (1995), 431–442.

Bärnighausen, T., Bor, J., Wandira-Kazibwe, S., & Canning, D. (2011). Correcting HIV prevalence estimates for survey nonparticipation using Heckman-type selection models. Epidemiology, 22(1), 27-35.

Cameron, A.C, and P.K. Trivedi. Microeconometrics: methods and applications. Cambridge university press, 2005.

Centre for Tuberculosis, 2016, South African Tuberculosis Drug Resistance Survey 2012–14, Johannesburg: Centre for Tuberculosis.

Churchyard, G. J., L. D. Mametja, L. Mvusi, N. Ndjeka, A. C. Hesseling, A. Reid, S. Babatunde, and Y. Pillay. (2014) "Tuberculosis control in South Africa: Successes, challenges and recommendations." SAMJ: South African Medical Journal 104(3): 234-248.

Clark, SJ., and Houle B., "Validation, Replication, and Sensitivity Testing of Heckman-Type Selection Models to Adjust Estimates of HIV Prevalence" PLoS ONE 9 (11) e112563. doi: 10.1371/journal.pone.0112563(2015)

Cohen, T, et al. "Challenges in estimating the total burden of drug-resistant tuberculosis." American journal of respiratory and critical care medicine 177.12 (2008): 1302-1306.

Department of Health of South Africa, January 2013, Management of Drug Resistant Tuberculosis Policy Guidelines.

Dowdy, D. W., Houben, R., Cohen, T., Pai, M., Cobelens, F., Vassall, A., ... & White, R. (2014). Impact and cost-effectiveness of current and future tuberculosis diagnostics: the contribution of modelling. The International Journal of Tuberculosis and Lung Disease, 18(9), 1012-1018.

Glaziou P, Sismanidis C, Pretorius C, Timimi H and Floyd K, 2015, Global TB Report 2015: Technical appendix on methods used to estimate the global burden of disease caused by TB, Geneva: World Health Organization.

Greene, W. H. "Econometric Analysis." 6th ed. (2008).

Heckman, J. J., "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," Annals of Economic and Social Measurement 5:4 (1976), 475–492.

Hogan, DR et al. "National HIV prevalence estimates for sub-Saharan Africa: controlling selection bias with Heckman-type selection models" Sexually Transmitted Infection 91.8 (2015) i17-i23.

Imbens, G. W., and J. D. Angrist, "Identification and Estimation of Local Average Treatment Effects," Econometrica 62:2 (1994), 467–475.

Institute of Medicine (US) Forum on Drug Discovery, Development, and Translation; Academy of Science of South Africa. *The Emerging Threat of Drug-Resistant Tuberculosis in Southern Africa: Global and Local Challenges and Solutions: Summary of a Joint Workshop*. Washington (DC): National Academies Press (US); 2011. 2, The Incidence of Drug-Resistant TB in Southern Africa. Available from: https://www.ncbi.nlm.nih.gov/books/NBK55577/

Judd, Kenneth L. Numerical methods in economics. MIT press, 1998.

Karim, S S A, Churchyard G J, Karim Q A, Lawn S D. HIV infection and tuberculosis in South Africa: an urgent need to escalate the public health response. Lancet 2009; 374(9693):921-933.

Kendall, E. A., Fofana, M. O., & Dowdy, D. W. (2015). Burden of transmitted multidrug resistance in epidemics of tuberculosis: a transmission modelling analysis. *The Lancet Respiratory Medicine*, 3(12), 963-972.

Kim, J.Y., A. Shakow, K. Mate, C. Vanderwarker, et al. 2005. “Limited Good and Limited Vision: Multidrug-Resistant Tuberculosis and Global Health Policy,” Social Science and Medicine Journal, 61: 847–859.

Manski & Lerman 1981. Structural Analysis of Discrete Data with Econometric Applications, Editor D. McFadden, Cambridge: M. I. T. Press, 1981.

McFadden, D. 1989. "A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration," Econometrica, 57(5): 995-1026.

McGovern, ME et al “On the Assumption of Bivariate Normality in Selection Models *A Copula Approach Applied to Estimating HIV Prevalence”* Epidemiology 26.2 (2015) 229-237

Medecins Sans Frontieres, Partners in Health and Treatment Action Group. An evaluation of drug-resistant TB treatment scale-up. July 2011.

Meyer-Rath G, Schnippel K, Long L, MacLeod W, Sanne I, et al. (2012) The Impact and Cost of Scaling up GeneXpert MTB/RIF in South Africa. PLoS ONE 7(5): e36966. doi:10.1371/journal.pone.0036966

Nyirenda M, et al. “Adjusting HIV Prevalence for Survey Non-Response Using Mortality Rates: An Application of the Method Using Surveillance Data from Rural South Africa” PLoS ONE 5 (8): e12370,doi: 10.1371/journal/pone.0012370

Phelps, E. S. (1972). The statistical theory of racism and sexism. American Economic Review, 62(4), 659-661.

Sakarovitch, C et al. "Estimating incidence of HIV infection in childbearing age African women using serial prevalence data from antenatal clinics" Statistics in Medicine in Wiley InterScience (2007) 320-335

Statistics South Africa. Mortality and causes of death in South Africa, 2010: findings from death notification. Statistical Release P0309.3. http://www.statssa.gov.za/publications/ p03093/p030932010.pdf Pretoria, South Africa: Statistics South Africa, 2013. Accessed October 2014.

Stock, J. H., Wright, J. H., & Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. Journal of Business & Economic Statistics 20(4).

Theron, G et al. How to eliminate tuberculosis 1 "Data for action: collection and use of local data to end tuberculosis" The Lancet 386.10010 (2015) 2324-2333.

Vassall A, van Kampen S, Sohn H, Michael JS, John KR, et al. (2011) Rapid Diagnosis of Tuberculosis with the Xpert MTB/RIF Assay in High Burden Countries: A Cost-Effectiveness Analysis. PLoS Med 8(11): e1001120. doi:10.1371/journal.pmed.1001120

Weyer, K, et al. "Rapid molecular TB diagnosis: evidence, policy making and global implementation of Xpert MTB/RIF." European Respiratory Journal 42.1 (2013): 252-271.

Weyer, K., Brand, J., Lancaster, J., Levin, J., & Van der Walt, M., 2007, "Determinants of multidrug-resistant tuberculosis in South Africa: results from a national survey." South African Medical Journal 97(11).

World Health Organization, 2011, WHO Report 2011 Global Tuberculosis Control, Geneva: World Health Organization.

World Health Organization, 2013, Multidrug-resistant tuberculosis (MDR-TB), 2013 Update, March 2013.

World Health Organization, 2014, WHO Report 2013 Global Tuberculosis Control, Geneva: World Health Organization.

Appendix 1: WHO TB Estimation Method

In South Africa the WHO estimates MDR-TB prevalence based on "case notification data combined with expert opinion about case detection gaps" (Glaziou et al. 2015). Using the equation below, $I = \frac{f(N)}{1-g}, g \in [0,1)$, where *f(N)* is a simple function of case notifications (i.e. TB cases reported through the national notification system) and *g* is the detection gap, which is obtained asking a small number of TB "experts" who work in TB programs to provide their "educated best guess of the range". The authors note limitations of this approach, including incomplete data and a lack of information on "vested interests" when eliciting expert opinion.

Table 1: Estimated MDR-TB prevalence (μ), quadratic specification coeffcients (μ_0, μ_1 and μ_2) and signal-to-noise ratio (β).

Method:	ML	GMM	GMM	GMM	GMM	ML	GMM	GMM	GMM	GMM
Instruments:		Time	Time	Policies	Policies		Time	Time	Policies	Policies
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
μ	0.0305***	0.0298***	0.0342***	0.0305***	0.0301***					
	(0.0002)	(0.0013)	(0.0004)	(0.0016)	(0.0004)					
μ_0						-2.8758***	-2.8779***	-3.1033***	-2.8395***	-2.8269***
						(0.0031)	(0.1436)	(0.0378)	(0.3725)	(0.0864)
μ_1						0.0224***	0.0241***	0.0256***	0.012**	0.0094***
						(0.0002)	(0.0038)	(0.0010)	(0.0053)	(0.0016)
μ_2						-0.0011***	-0.0012***	-0.0011***	-0.0008***	-0.0007***
						(0.0000)	(0.0002)	(0.0000)	(0.0003)	(0.0001)
β	1.0749***	1.0891***	0.9336***	1.0716***	1.0846***	0.5192***	0.5165***	0.6687***	0.5457**	0.5421***
	(0.0003)	(0.0489)	(0.0117)	(0.0570)	(0.0142)	(0.0036)	(0.1003)	(0.0297)	(0.2638)	(0.0620)
Observations	262,845	262,845	262,853	262,850	262,842	262,845	262,845	262,853	262,850	262,842
Clustering	No	No	Yes	No	Yes	No	No	Yes	No	Yes
Pseudo R^2 #1	0.694	0.699	0.529	0.691	0.703	0.725	0.798	0.841	0.799	0.737
Pseudo R^2 #2	0.695	0.695	0.694	0.695	0.696	0.740	0.741	0.743	0.729	0.724
Log likelihood	-90646.845					-90623.48				
GMM criterion		0.00131	0.01626	0.00044	0.00780		0.00108	0.01410	0.00031	0.00670

Notes: Table presents coefficients and standard errors in parentheses. Sample includes TB-positive patients ages 16-64 in public health facilities from January 2004-September 2010. *** - Significant at the 1% level, ** - 5% level, * - 10% level.

Table 2: Estimated MDR-TB prevalence (μ) and signal-to-noise ratio (β) for new patients and repeat patients.

	ML New patients	ML Repeat patients
μ	0.031***	0.0533***
	(0.0021)	(0.0024)
β	0.5902***	1.2678***
	(0.0565)	(0.0536)
Observations	190,449	72,409
Pseudo R^2 #1	0.264	0.386
Pseudo R^2 #2	0.257	0.380
Log likelihood	-47809.729	-37663.286

Notes: Table presents coefficients and standard errors in parentheses. Sample includes TB-positive patients ages 16-64 in public health facilities from January 2004-September 2010. *** - Significant at the 1% level, ** - 5% level, * - 10% level.

Figure 1: Number of patients who are TB-positive, tested for MDR-TB, and MDR-TB-positive by quarter over time.

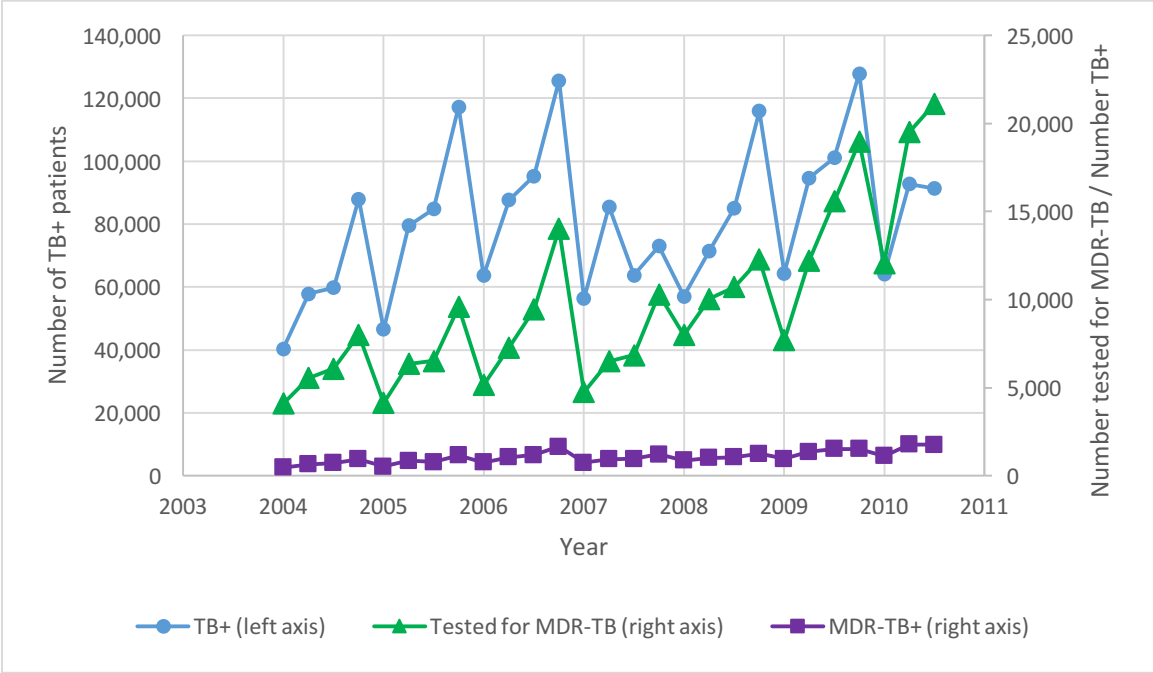

Figure 2: Percent of TB-positive cases tested for MDR-TB, percent of TB-positive cases MDR-TB-positive, and percent of MDR-TB-tested cases MDR-TB-positive from National Health Laboratory Service data (scaled to two Y-axes to show how the testing rate and testing-positive rate track reasonably well over time).

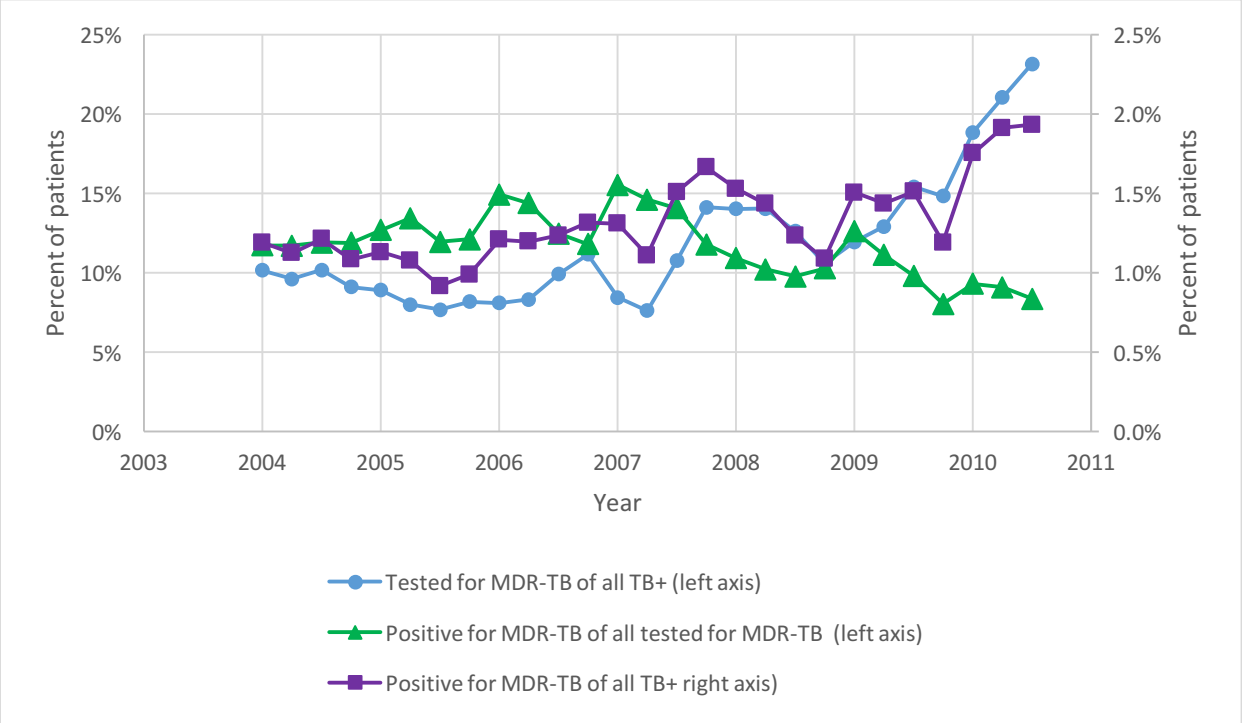

Figure 3: Sampling efficiency and marginal sampling efficiency at different proportions of TB+ patients being tested for MDR-TB (θ).

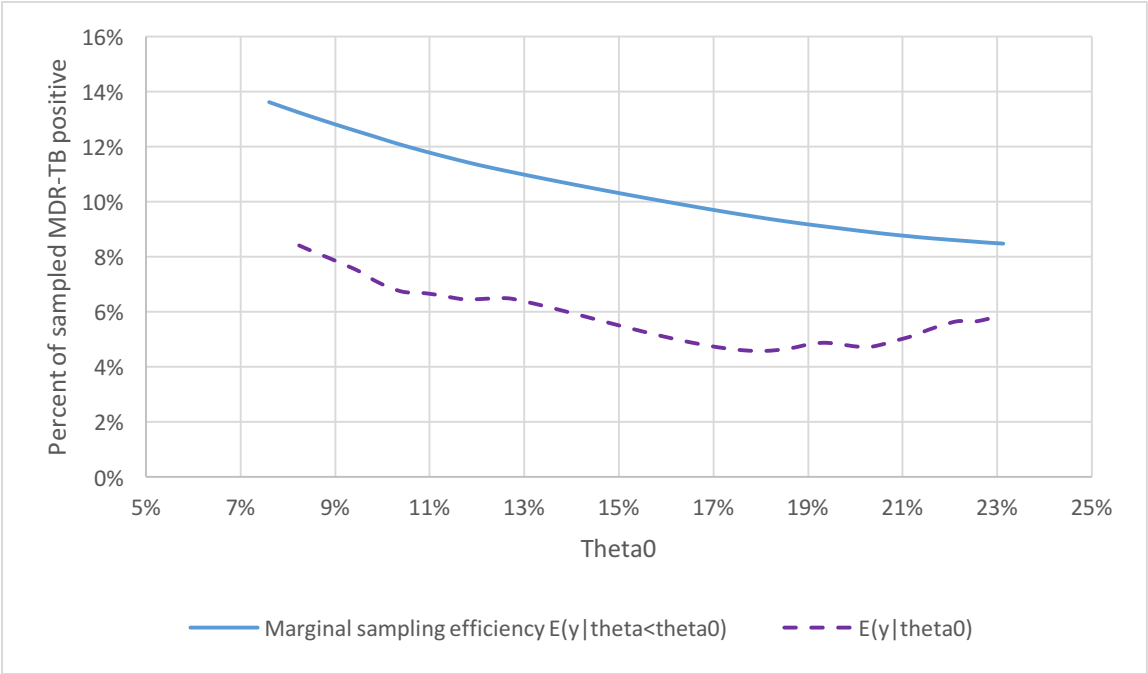

Figure 4: Predicted time trends in MDR-TB prevalence (%) in South Africa estimated from NHLS data using generalized method of moments and maximum likelihood estimation.

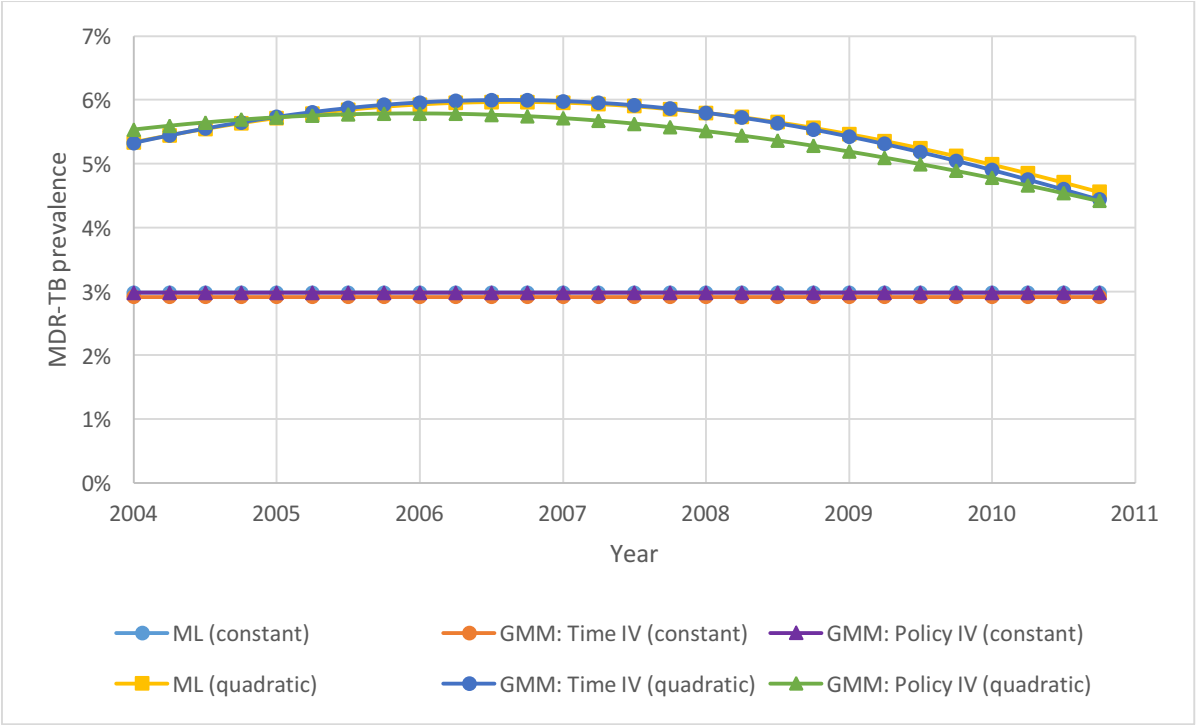

Figure 5: Observed MDR-TB prevalence over time in NHLS data compared to MDR-TB prevalence estimated from maximum likelihood estimation.