A New Econometric Method for Estimating Disease Prevalence: An Application to Multi-Drug Resistant Tuberculosis

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

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

DRAFT – Please do not cite without permission
This version: June 1, 2016

Abstract

Introduction

Accurate information on disease prevalence is needed to target limited health resources in order to maximize overall population health. Applying rigorous econometric methods to routinely collected data can produce accurate estimates of disease prevalence and under-detection rates at a fraction of the cost of alternatives such as prevalence surveys or universal diagnostic testing. Such estimates are valuable in developing countries to inform evidence-based health policy.

Methods

We develop a simple framework with minimal assumptions to capture key features of clinical decision making surrounding diagnostic testing in resource limited settings. When it is infeasible to test every at-risk patient, clinicians must triage available resources to test those deemed most likely to have the disease. We use standard econometric estimation methods and iterative numerical optimization techniques to estimate (a) disease prevalence and (b) the accuracy with which clinicians triage patients for testing. We implement an instrumental variables approach using national and local policy changes that exogenously shift the available resources for diagnostic testing as instruments. We apply this method to tuberculosis (TB), which recently surpassed HIV as the leading infectious disease cause of death in the world. We use a national database of TB test data from South Africa, which includes over 11 million patients, to examine diagnostic testing for multi-drug resistant TB (MDR-TB).

Results

The predictions from our model closely match observed patterns in the data. We find that at least one-quarter of MDR-TB cases were undiagnosed between 2004-2011. Our estimates show that the official World Health Organization estimate of 2.5% based on notification rates is too low, and MDR-TB prevalence in South Africa could be as high as 3.29 - 3.37%. Noise-to-signal ratios in MDR-TB detection estimated in our model enable the identification of areas where clinicians do a poor job of sorting patients by MDR-TB risk prior to testing.

Discussion

In the case of MDR-TB there is a need for greater investment in early detection and more effective treatment. Our method of identifying areas with high MDR-TB under-detection rates, which was heretofore unmeasured and contributes to high transmission rates, provides clinicians and policy makers with a formidable new tool for targeting efforts to control TB. This method should be deployed in countries such as India, China and Russia, which together account for over 50% of MDR-TB cases worldwide, as well as applied to other infectious and non-infectious diseases where prevalence data is lacking.

1. Introduction

1.1 Background

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. Prevalence surveys can provide accurate estimates of disease prevalence, but they are infrequently conducted because they entail significant financial and time costs. Unadjusted estimates from routinely collected health care data are likely to be biased because clinicians generally only perform diagnostic testing on patients who appear at-risk for the disease. Testing all patients for every disease is neither feasible nor cost-effective.

In this study, we develop an innovative approach to estimating disease prevalence that accounts for under-detection, provides continuous surveillance and relies solely on existing routinely-collected data from diagnostic laboratories. Notably, our method does not require that all patients or a random sample of patients be screened for the disease. Instead, it applies rigorous statistical (econometric) methods to routinely collected data, which is readily available but acutely underused, to produce low-cost, unbiased estimates of disease prevalence.

Our framework may be applied to a broad range of health conditions including hypertension, HIV, malaria, and anti-microbial resistance among others. In this study, we apply the method to tuberculosis (TB), which kills more than 1.5 million people annually and recently surpassed HIV as 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, they accounted for 13% of TB deaths and 20% of TB spending worldwide in 2014 (WHO 2015). Early and accurate diagnosis of MDR-TB is therefore critical. However, only 12% of incident TB cases were tested for MDR-TB. The under-detection of MDR-TB drives the development of new forms of drug resistance such as extensively (XDR) and totally drug-resistant TB worldwide (Klopper et al. 2013).

We develop a simple framework with minimal assumptions to capture key features of clinical decision making surrounding MDR-TB testing. In resource limited settings, resources are not available to test every TB-positive patient for MDR-TB. Clinicians must therefore triage available resources to test those deemed most likely to have MDR-TB. We use standard econometric estimation methods and iterative numerical optimization techniques to estimate (a) the prevalence of MDR-TB and (b) the accuracy with which clinicians triage for MDR-TB testing. Our method leverages the same principle that underlies regression discontinuity analysis – the contrast between slow-changing MDR-TB prevalence and sudden (discontinuous) changes in the likelihood that MDR-TB cases are diagnosed due to the availability of testing resources (e.g. funding, testing materials, human resources, lab capacity).

We find that approximately one-quarter of all MDR-TB cases were undiagnosed between 2004-2011 in South Africa, which worsens patient outcomes, increases transmission and leads to the development of additional drug resistance. This straightforward, yet powerful, approach to disease surveillance is a viable strategy for identifying localities and patient groups with high disease burdens. It is simple and adaptable enough to be applied to many infectious and non-infectious diseases in the developing world where prevalence data is lacking.

The contribution of this paper is three-fold. First, we develop a new method for estimating disease prevalence that is widely applicable to many diseases. 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, 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 find that approximately one-quarter of MDR-TB cases in South Africa went undiagnosed between 2004-2011, which is a significant threat to TB control. Third, we demonstrate the ease with which this method can be applied to other diseases and contexts because it uses existing data, is low cost and can be quickly scaled up. Our method can be employed in low- and middle-income countries to cost-effectively develop guidance for health policy making to ultimately improve population health.

1.2 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). This is especially important for MDR-TB which accounts for 2.5-3.5% of all TB patients in South Africa, but consumes about 50% of the TB budget (WHO 2011; Pooran et al. 2012). Treatment success rates in South Africa are close to 45% compared to 79% for drug susceptible TB (WHO 2014).

Conventional thinking about estimating MDR-TB prevalence focuses on increasing the frequency and coverage of TB prevalence studies and expanding access to rapid molecular tests such as Xpert for drug resistance testing (Cohen et al. 2008, Weyer et al. 2013). However, both avenues are expensive and logistically complex. 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. For over a decade there have been calls for more MDR-TB prevalence studies yet only 8.3% of the population in the 27 high MDR-TB burden countries live in an area where at least two accurate population surveillance data points are available to estimate trends (Cohen et al. 2014).

Official guidelines counsel clinicians to screen patients for TB based on symptoms (current cough, weight loss, night sweats, fever) which are indistinguishable from MDR-TB (Department of Health 2013). Before Xpert was widely available in South Africa, the guidelines indicated that persons 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 were at highest risk of MDR-TB and therefore should be tested (Department of Health 2013). Clinicians can ascertain a patient's approximate risk of MDR-TB from a medical history and physical exam to rank patients on MDR-TB risk before ordering laboratory testing. The risk factors are a good but imperfect signal of MDR-TB so many cases could initially go undetected. Previously, MDR-TB was primarily due to acquisition through incorrect or incomplete treatment, however recent evidence shows that most incident MDR-TB cases are due to transmission instead, which makes risk factors worse predictors (Kendall et al. 2015). Resource shortages such as stockouts of test materials or drugs, long lab wait times or heavy clinical workload may prevent some at-risk patients from being tested for MDR-TB.

2. Methods

2.1 Data

We use data from the 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 - December 2011, which includes over 11 million patients. Our analysis sample comprises 2,271,538 TB-positive test records from 2,520,337 unique patients in 5,122 health facilities (359,174 of which are tested for MDR). 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. We consider TB-positive results from culture testing and smear microscopy as well as scanty positives of 3 or more acid fast bacilli (AFB) per 100 immersion fields. Patient records are linked using unique patient identifiers created by the NHLS. Our dataset spans 8 years of frequent observations, which allows us to observe several sudden policy shifts that affected the inclination to test for MDR. Since this variation forms the basis for our identification strategy, the long time-series dimension makes it especially well-suited for this analysis.

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). We exclude 89,032 records (3.9% of the sample) that were missing gender data from the gender sub-group analysis. We also exclude data from KwaZulu-Natal from the provincial analyses because only a small fraction of tests performed in that province have been electronically captured. Ethics approval was obtained from the University of Michigan Institutional Review Board and the University of Cape Town Faculty Ethics in Research Committee.

2.2 Theory model

Using minimal assumptions, we develop a simple theoretical model of the clinician's decision to perform drug susceptibility testing on a patient with suspected TB. 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 (that includes risk factors such as HIV status and non-specific symptoms such as cough and fever) denoted:

$$s \equiv a + \sigma u$$

where s is the noisy signal (observable to the clinician but not the econometrician), a is the true likelihood of MDR-TB, u is noise, and, and σ is therefore the noise-to-signal ratio (i.e. a measure of accuracy of the observed signal). This framework follows the labor economics literature on screening (see for example Phelps 1972) and assumes both a and u are normally and independently distributed on the unit interval *(0,1)*. Furthermore, suppose the clinician does not know the actual prevalence of MDR-TB in the population, μ, or the true likelihood of each patient to have MDR-TB (a) but can use the observed signal to rank (triage) individuals from most to least likely of having MDR-TB (scaled to the unit interval) conditional on having TB:

$$q \equiv \Phi\left(\frac{s}{\sqrt{1+\sigma^2}}\right)$$

where Φ is the normal cumulative distribution function (CDF), so that *q~U(0,1).*

The proportion of individuals who can be tested for MDR-TB, θ, is exogenously determined by institutional factors (such as the XDR-TB outbreak in 2006, changes in national testing guidelines, and the availability of test materials and human resources) and is therefore uncorrelated with μ. 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. We define *c* as the binary variable representing whether a patient is tested for MDR-TB:

$$c \equiv 1(q > 1 - \theta)$$

Furthermore, we define *d* as the binary variable representing whether the patient actually has MDR-TB:

$$d \equiv 1(\Phi(a) > 1 - \mu)$$

Finally, *y* is the binary variable representing whether a patient tests positive for MDR-TB:

$$y \equiv cd$$

For the sake of notational convenience, we define $\tilde{\theta} \equiv \Phi^{-1}(1-\theta)$ and $\tilde{\mu} \equiv \Phi^{-1}(1-\mu)$ so we can rewrite the outcomes as:

$$c = 1\left(\frac{s}{\sqrt{1+\sigma^2}} > \tilde{\theta}\right)$$

$$d = 1(a > \tilde{\mu})$$

The proportion of individuals who test positive for MDR-TB is therefore

$$\begin{aligned} P(y=1) = P(c=1 \cap d=1) &= P(c=1|d=1)P(d=1) \\ &= P\left(\frac{a+\sigma u}{\sqrt{1+\sigma^2}} > \Phi^{-1}(1-\theta)|a > \Phi^{-1}(1-\mu)\right)\mu \\ &= \Phi\left(\frac{\Phi^{-1}(1-\mu) - \sqrt{1+\sigma^2}\Phi^{-1}(1-\theta)}{\sigma}\right)\mu \end{aligned}$$

2.3 Identification

From the routinely collected patient data the econometrician can observe *(c,y)*, but not *(d,a,u)*. In a sufficiently large random sample of routinely collected data, we can therefore obtain reliable estimates of the proportion of TB cases tested for MDR-TB, θ, and the proportion of those testing positive for MDR-TB, *P(y=1)*, but not of the "true" underlying proportion with MDR-TB, μ. This would be difficult with cross-sectional data, since any observable combination of θ and *P(y=1)* in the population could have been produced by infinite combinations of (μ,σ). For example, in our sample 13.86% of patients are tested for MDR-TB and 1.43% test positive. These outcomes are consistent with a health system in which clinicians perfectly observe the patients propensity to have MDR-TB $(\sigma=0)$ and the underlying MDR-TB prevalence in the population is μ=1.43%. However, it is also consistent with a system in which the clinicians observe only noise $(\sigma\rightarrow\infty)$ and the actual MDR-TB prevalence in the population is μ=1.43%/13.86%=10.32%.

Suppose that due to exogenous institutional variation in θ (due to a national policy change or the 2006 XDR-TB outbreak, for example) there is an increase in the proportion of individuals who get tested for MDR-TB over time. This naturally has implications for the type of individual who gets tested, but should have no effect on the clinician's ability to rank patients or on the underlying share of drug-resistant patients.

Let us consider what the two extreme cases predict for the change in the share of individuals that tests positive for MDR-TB. If the signal observed by the clinicians consists only of noise ($\sigma \rightarrow \infty$) then any increase in θ_t should lead to a proportional increase in the share of individuals who test positive, $P(y=1)$. On the other hand, if there is no noise in the signal ($\sigma=0$) then the increase in θ_t should have no effect on this share. If an increase in the share of tested individuals leads to a smaller than proportional increase in the share of individuals who test positive (which is the case in our data) then this suggests intermediate values of σ and μ.

The assumption of exogenous changes in the proportion of patients tested over time implies a set of moment conditions that can be used to identify the parameters of interest. In this case we can use the period dummy variables as instrumental variables. However, we may be concerned that identifying off changes over time may reflect underlying transmission dynamics of the epidemic that drive the true prevalence. We therefore introduce instruments based on exogenous institutional variation in MDR-TB testing rates. 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. Intuitively, these instruments represent discontinuous changes in θ that cannot, in the short term, be correlated with relatively smooth trends in prevalence or the noise-to-signal ratio.

We include the following policy changes: MDR-TB surveillance study results reported (Jan 2002); national anti-retroviral therapy (ART) for AIDS rollout begins (July 2004); WHO declares TB an emergency in Africa (August 2005); first poster on XDR-TB presented at Conference on Retroviruses and Opportunistic Infections (CROI) (February 2006); South African government National Strategic Plan released (January 2007); clinical guidelines require one rather than three negative smears for smear-negative diagnosis (January 2009). We also use the facility-level and local-area-level availability of ART as instruments to incorporate sub-national exogenous variation. With the exception of the ART rollout, these policy changes are highly 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.

2.4 Estimation

2.4.1 Simulated maximum likelihood

In our empirical analysis we use a simulated maximum likelihood estimator (SMLE) to estimate the model parameters. In period *t* each individual is either tested for MDR-TB or not, where the former occurs with probability

$$P(c_{it} = 1) = P(q_{it} > 1 - \theta_t) = P\left(\frac{s_{it}}{\sqrt{1+\sigma^2}} > \Phi^{-1}(1-\theta)\right) = \theta_t$$

Furthermore, the likelihood that an individual tests positive for MDR-TB in period *t* is

$$P(y_{it} = 1) = P\left(a_{it} + \sigma u_{it} > \sqrt{1+\sigma^2}\Phi^{-1}(1-\theta) | a_{it} > \Phi^{-1}(1-\mu)\right)\mu$$

Using MLE to estimate the model parameters is achieved by finding the values of *(μ,σ)* that maximize the likelihood function

$$L\,(\mu,\sigma) = \prod\nolimits_{i=1}^{N}\prod\nolimits_{t=1}^{T}\left(P(y_{it} = 1|\mu,\sigma,\theta_t)\right)^{d_{it}}\left(1 - P(y_{it} = 1|\mu,\sigma,\theta_t)\right)^{1-d_{it}}$$

or the log-likelihood function, which we use for our estimation

$$
\begin{aligned}
l\,(\mu,\sigma) &= \log L\,(\mu,\sigma) \\
&= \sum\nolimits_{t=1}^{T} \sum\nolimits_{i=1}^{N} \{d_{it} \log\big(P(y_{it} = 1|\mu,\sigma,\theta_t)\big) \\
&\quad + (1-d_{it}) \log\big(1 - P(y_{it} = 1|\mu,\sigma,\theta_t)\big)\}
\end{aligned}
$$

where $P(y = 1|\mu,\sigma,\theta_t) = \Phi\left(\frac{\Phi^{-1}(1-\mu)-\sqrt{1+\sigma^2}\Phi^{-1}(1-\theta_t)}{\sigma}\right)\mu.$ Because this probability contains a double integral that cannot be calculated analytically using standard software, we must approximate it using simulations (Judd 1998: 291).

We draw trials of $K \times T$ values (where in our estimates $K = 1$ million and $T = 31$ quarters) of a and u so that both variables are *nid(0,1)*. For any set of trial parameter values (μ,σ), observable θ_t and randomly drawn values of (a_{kt},u_{kt}) we can calculate values for c_{kt}, d_{kt}, and y_{kt} as follows:

$$c_{kt}(\sigma,\theta_t) \equiv 1\left(\Phi\left(\frac{a_{kt} + \sigma u_{kt}}{\sqrt{1+\sigma^2}}\right) > 1 - \theta_t\right)$$

$$d_{kt}(\mu) = 1(\Phi(a_{kt}) > 1 - \mu)$$

$$y_{kt}(\mu,\sigma,\theta_t) = c_{kt}(\sigma,\theta_t) d_{kt}(\mu)$$

We then use these values to approximate $P(y = 1|\mu,\sigma,\theta_t)$ from our auxiliary model as

$$\tilde{P}(y = 1|\mu,\sigma,\theta_t) = \frac{1}{K} \sum\nolimits_{k=1}^{K} y_{kt}(\mu,\sigma,\theta_t)$$

Plugging these values into our likelihood function produces the simulated likelihood function

$$
\begin{aligned}
\tilde{l}\,(\mu,\sigma) = \sum\nolimits_{t=1}^{T} \sum\nolimits_{i=1}^{N} &\left\{d_{it} \log\left(\tilde{P}(y = 1|\mu,\sigma,\theta_t)\right)\right. \\
&\left. + (1-d_{it}) \log\left(1 - \tilde{P}(y = 1|\mu,\sigma,\theta_t)\right)\right\}
\end{aligned}
$$

Theoretically, we can obtain a simulated likelihood function that is arbitrarily close to the actual likelihood function by choosing a sufficiently large number of simulated draws, *K*. In practice, because our outcome variable is discrete, the likelihood function is not a continuous function of the model parameters. Therefore small changes in model parameters may have no effect on the simulated values of y_{kt} and will leave the simulated likelihood function unchanged. This makes using numerical optimization techniques difficult because they rely on small changes in the likelihood to find the parameters that maximize the function.

This is the well-documented problem of simulating choice probabilities with a binary Accept-Reject (AR) simulator (Manski & Lerman 1981). We follow the solution proposed by McFadden (1989) of smoothing the discrete individual outcomes with a logit function. We therefore generate the outcomes c_{kt} and d_{kt} as:

$$c_{kt}(\sigma,\theta_t) \equiv \frac{\exp\left(\dfrac{\dfrac{a_{kt}+\sigma u_{kt}}{\sqrt{1+\sigma^2}}-\Phi^{-1}(1-\theta_t)}{\lambda}\right)}{1+\exp\left(\dfrac{\dfrac{a_{kt}+\sigma u_{kt}}{\sqrt{1+\sigma^2}}-\Phi^{-1}(1-\theta_t)}{\lambda}\right)}$$

$$d_{kt}(\mu) = \frac{\exp\left(\dfrac{a_{kt}-\Phi^{-1}(1-\mu)}{\lambda}\right)}{1+\exp\left(\dfrac{a_{kt}-\Phi^{-1}(1-\mu)}{\lambda}\right)}$$

where the degree of smoothing is determined by the value of λ. High values of λ produce a simulated likelihood surface that is very smooth, which makes numerical optimization easier, but also produces worse approximations to the likelihood function and hence potentially biased parameter estimates. As $\lambda \rightarrow 0$ the logit-smoothed AR simulator approaches the binary AR simulator and better approximates the likelihood function, but also has greater difficulties in finding the optimal parameter values. In our empirical application we set this value to λ=0.005.

According to our theory model the parameters should be restricted so that σ ∈[0,∞) and μ ∈[0,1], which can be applied by expressing the likelihood function in terms of the transformed unrestricted parameters $(\tilde{\mu}, \tilde{\sigma}) = \left(\log \frac{\mu}{1-\mu}, \log \sigma\right)$. We can then obtain point estimates of the parameters of interest by performing the inverse transformation on these parameters. Standard errors are calculated via the delta method.

2.4.2 Method of simulated moments

Another estimator that can be used to estimate the parameters of our auxiliary model is the method of simulated moments (MSM). If we define the model error term as

$$\varepsilon \equiv y - P(y = 1|\mu, \sigma, \theta)$$

then the exogeneity assumption implies that

$$E(\varepsilon|\mathbf{\Omega}) = 0$$

where **Ω** represents all the information available to the clinician at the time of the testing decision. This implies that

$$E[\boldsymbol{z}_{it}\{y_{it} - P(y_{it} = 1|\mu, \sigma, \theta_t)\}] = \mathbf{0}$$

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

$$g_{it}(y_{it}, \boldsymbol{z}_{it}, \mu, \sigma, \theta_t) = \boldsymbol{z}_{it}\{y_{it} - P(y_{it} = 1|\mu, \sigma, \theta_t)\}$$

then the generalized method of moments (GMM) estimator can be expressed as

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

where $\boldsymbol{W}$ is the weighting matrix. We cannot calculate $P(y_{it} = 1|\mu,\sigma,\theta_t)$ analytically, but replacing this probability with its simulated counterpart in the GMM estimator allows us to estimate the model parameters with the method of simulated moments.

2.5 Time Trends

To calculate the time trends we estimate the following equation:

$\mu = \Phi(\mu_0* + \mu_1*t + \mu_2*t^2)$,

where μ is the prevalence of MDR-TB, t is the time period and μ_k are the time coefficients and $\Phi()$ represents the standard normal cumulative distribution function.

3. Results

3.1 Descriptive Statistics

Figure 1 shows that the proportion of TB-positive patients tested for MDR-TB and the proportion that test positive are negatively correlated, both in long-run trends and short-term fluctuations. The patterns in the data 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 (i.e. σ is less than infinity). The percentage of all TB-positive patients (based on smear, culture or PCR) who were tested for MDR-TB was fairly stable at around 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 2011, when it reached 27%. The percentage of all TB-positive patients who tested positive for MDR-TB was stable around 1% from 2004-2006 and rose above 2% in 2010. MDR-TB cases as a percentage of all those tested for MDR-TB was steady at around 12% until 2007 when it rose to 15% and then steadily declined.

Figure 2 rescales the percent of all TB-positive patients who were diagnosed with MDR-TB to show that it tracks the percent of all TB-positive patients who were tested for MDR-TB reasonably well, especially after 2009. The data are consistent with our model in that σ (the noise to signal ratio) is neither zero nor infinity. As more TB-positive patients are tested for MDR-TB, more MDR-TB cases are found (consistent with θ being a limiting factor and σ being less than infinity) and the share of MDR-TB tested patients who are MDR-TB-positive falls (σ>0).

3.2 Estimation Results

We estimate that MDR-TB prevalence in South Africa could be as high as 3.29 - 3.37% (Table 1) which is approximately 0.8 percentage points higher than the 2011 WHO estimate of 2.5% based on notification rates (WHO 2011). This indicates that approximately one-quarter of all MDR-TB cases went undetected during this period. The standard errors for our estimates are small. We are most confident in estimates that are very similar between the two methods (MLE and MSM-IV). The noise-to-signal ratio in the clinician-observed signal of risk factors is estimated to range between 2.12-2.15, in other words the standard deviation of noise is about twice as large as the standard deviation of the signal (which is normalized to be 1). MLE and MSM-IV methods produce very similar estimates of MDR prevalence (μ) and the noise-to-signal ratio (σ) in the full sample.

Figure 3 provides 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 MDR-TB prevalence can be explained by changes in the proportion of patients tested (θ) alone. The match is worse where the observed prevalence has more peaks and troughs (2006-2010).

New patients and patients with a previous test result have different underlying noise-to-signal ratio and estimated MDR-TB prevalence. While the MLE results show a prevalence of 5.65% for new and 4.69% for repeat patients, the MSM-IV results of 2.21% for new and 6.29% for repeat patients are very close to the notification rates for new (1.8%) and retreatment patients (6.7%), respectively (Weyer et al. 2007). As expected, the values for σ reflect that clinicians have less information upon which to assess the risk profile of the new patients compared to repeat patients. For the MLE, σ is estimated at 16.24 for new and 1.23 for repeat patients, and for MSM-IV it is 2.25 compared to 1.90.

When we relax the assumption that MDR-TB prevalence is constant over time, we find that MDR-TB prevalence decreases between 2004 and 2011 from 4.0% to 3.6% for the linear trend ($\sigma = 2.40$) and 4.2% to 3.7% for the quadratic trend ($\sigma = 2.56$) (Figure 4, Table 2).

The subgroup analysis shows that MDR-TB prevalence for men is estimated to be 3.0-3.4% for men and slightly higher at 3.1-3.5% for women, with similar values of σ for both sexes (Table 3). Both methods show the highest MDR-TB prevalence for patients 30-40 of 3.2% (MLE) and 3.5% (MSM-IV), while the MLE shows lower prevalence for patients 20-30 and 40-50 and the MSM-IV shows similar for patients 20-50. Patients 50-60 make up 18% of patients and are shown to have MDR-TB prevalence between 3.2-4.3%. However, the unusually high noise to signal ratio for the new patients in the MLE estimation suggests that the estimate of MDR prevalence (μ) for this age group is suspect because the two parameters are jointly estimated. Both methods show that patient populations tested in hospitals have lower estimated MDR-TB prevalence (2.3-2.9%) than those tested in smaller health clinics (5.0-5.5%).

Both methods showed the highest MDR-TB prevalence in Eastern Cape and Northern Cape. The MLE method produced an especially high prevalence of 6.4% in Northwest Province, however the estimate of σ is very large which suggests that this estimate is elevated. Limpopo, Free State and Western Cape consistently have the lowest estimated prevalence of MDR-TB.

4 Discussion

Our results indicate that the assumptions about clinician behavior in our theoretical framework are consistent with the data. Figures 1 and 2 show that clinicians do prioritize testing patients that are more likely to be MDR-TB positive, but that this prioritization is imperfect. Our simple framework is able to match observed patterns in the data very closely (Figure 3). The fact that our MSM-IV estimates differ little from the MLE estimates provides evidence to support our assumption that the constraint on diagnostic testing resources (θ) changes exogenously over time. Our results do not exhibit characteristics indicative of weak instruments – large standard errors or sensitivity to changes in the sample – therefore concerns about bias in the estimates due to an influence of local MDR-TB prevalence on MDR-TB testing rates appear unfounded (Stock, Wright and Yogo 2002). The fact that our MSM-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 σ are well-identified. This method can be further validated with applications to other data sources.

Our estimates of MDR-TB prevalence are at least 33% higher (0.8 percentage points) than the WHO 2010 estimate of 2.5% (WHO 2011) which is based on an adjustment to notification rates. 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 are also higher than results from the 2001 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). The forthcoming results from the most recent surveillance study may be higher than anticipated. In light of these results, additional resources should be allocated to the National Tuberculosis Program to increase efforts to control MDR-TB.

Our subgroup analyses found no substantial differences in MDR-TB prevalence between genders or age groups. Estimated MDR-TB prevalence was almost twice as high for patients tested in clinics and health centers compared to those tested at hospitals. This raises the question of whether smaller (low volume) health facilities have sufficient access to diagnostic resources to identify patients at risk for MDR-TB and test them promptly.

The provincial estimates show similar patterns to the 2001 drug resistance prevalence study with the Eastern Cape and North West province having higher prevalence while the Western Cape and Free State have lower prevalence (Weyer et al. 2007). The two exceptions are Limpopo, that had high prevalence in the 2001 study but low prevalence in our results and Mpumalanga, that had relatively high prevalence in the 2001 study but were in the middle of the range in our results. These two provinces are also the only ones with estimates of σ that fall outside of the 2.0-2.23 range. Limpopo's noise to signal ratio of 2.35 suggests that clinicians experience slightly more difficulties in accurately ranking patients based on MDR-TB risk. Mpumalanga, on the other hand, has a low value of σ (1.67) which indicates that patients are not tested for MDR-TB until after treatment failure has been observed, which likely contributes to the high MDR-TB prevalence we observe for this province. Perniciously, this type of wait-and-see approach is "neither effective nor benign" because it risks patient health and leads to further propagation of drug resistance (Kim et al. 2005).

Though HIV-positive status was found to be a risk factor for MDR-TB in the 2001 surveillance study, the provincial patterns of MDR-TB in our data do not appear to be closely associated with provincial HIV prevalence (Weyer et al. 2007). Other province-specific factors such as early detection of treatment failure or high treatment success rates for drug-susceptible TB may have more influence on MDR-TB prevalence than HIV prevalence per se.

4.1 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.

4.2 Conclusion

This study developed a novel econometric method for estimating disease prevalence from routinely collected data. We found that approximately 25% of MDR-TB cases in South Africa were undiagnosed between 2004-2011 which contributed to high transmission rates and high TB mortality rates. Our analysis also identified areas where MDR-TB cases were undetected due to inaccurate clinician triage of patients into MDR-TB testing rather than due to a lack of testing resources. Identifying areas with high under-detection rates, which was heretofore unmeasured, provides clinicians and policy makers with a formidable new tool for targeting efforts to control TB.

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). In particular, additional diagnostic resources should be allocated to areas with low noise-to-signal estimates, which suggest that patients may not be tested for drug resistance unless they are at very high risk or have recently experienced treatment failure. A heightened index of suspicion for MDR-TB among patients could effectively identify more MDR-TB cases to not only control the spread of MDR-TB but also curtail future human and financial costs of TB. Our method can be applied to MDR-TB where access to Xpert 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.

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. The current MDR-TB budget allocation is therefore likely to be insufficient. 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. Prevalence surveys and rigorous statistical analysis of routine data are complements rather than substitutes: recently completed MDR-TB prevalence studies can serve to further calibrate methods such as ours, and estimates from the analysis of routinely collected data can inform the design of prevalence studies to maximize precision and minimize cost.

Statistical analysis of diagnostic test results from routinely collected data is an economical and effective way to monitor disease prevalence and guide the targeting of resources to control TB. It uses existing data, which is inexpensive to collect and widely available, and can easily be scaled up to the national level. Countries such as India, China and Russia, which together account for over 50% of MDR-TB cases worldwide could benefit from the deployment of this method to target investments in MDR-TB diagnosis, such as the rollout of new diagnostic technologies or new second-line TB drugs. Routine statistical analysis results can also function as an early warning system for outbreaks, especially if they are able to discern deviations from the prevalence trends over time. Ultimately, using routinely collected data to monitor population prevalence is a viable, low-cost, high-value strategy to guide evidence-based health 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 and University of KwaZulu Natal Microeconomic Analysis of South Africa conference 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.

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

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.

Cohen, T, et al. "On the spread and control of MDR-TB epidemics: An examination of trends in anti-tuberculosis drug resistance surveillance data.” Drug Resistance Updates 17.4 (2014): 105-123.

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.

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

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

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.

Klopper, M et al. "Emergence and Spread of Extensively and Totally Drug-Resistant Tuberculosis, South Africa" Emerging Infectious Diseases 19.3 (2013) 449-455

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.

Pooran A, Pieterson E, Davids M, Theron G, Dheda K (2013) What is the Cost of Diagnosis and Management of Drug Resistant Tuberculosis in South Africa? PLoS ONE 8(1): e54587. doi:10.1371/journal.pone.0054587.

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.

World Health Organization, 2015 Indicators of diagnosis, notification and treatment of multidrug-resistant TB, by region or country and year. http://www.who.int/tb/country/en/, Accessed 12 May 2015.

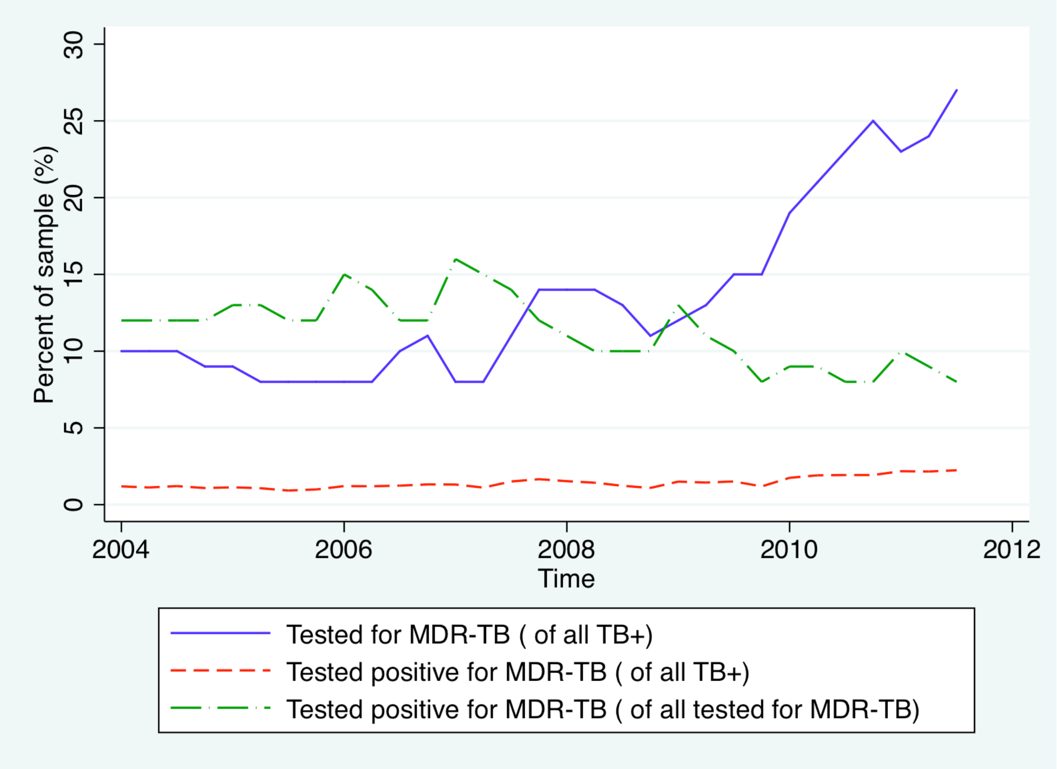

Figure 1: Percent of TB-positive cases tested for MDR-TB, percent of TB-positive cases MDR-TB-positive, and percent of MDR-TB-tested cases MDR-TB-positive from National Health Laboratory Service data.

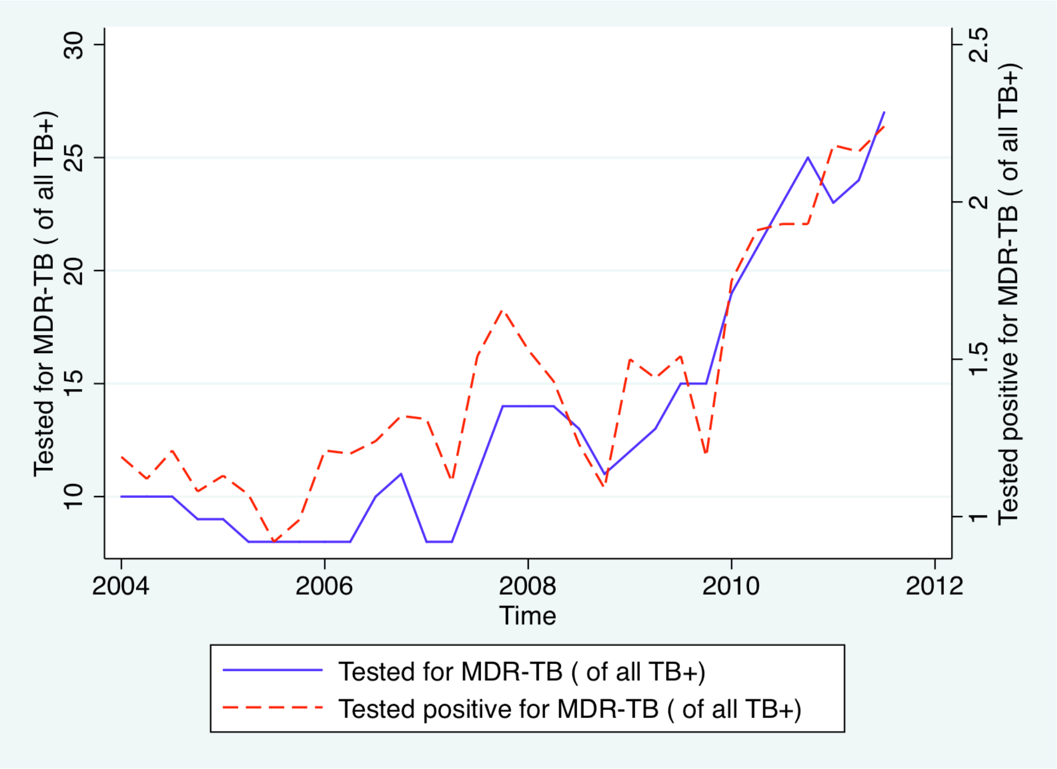

Figure 2: Percent of TB-positive cases tested for MDR-TB and percent of TB-positive cases MDR-TB-positive from National Health Laboratory Service data (scaled to two Y-axes to show how the testing rate and testing-positive rate track reasonably well over time).

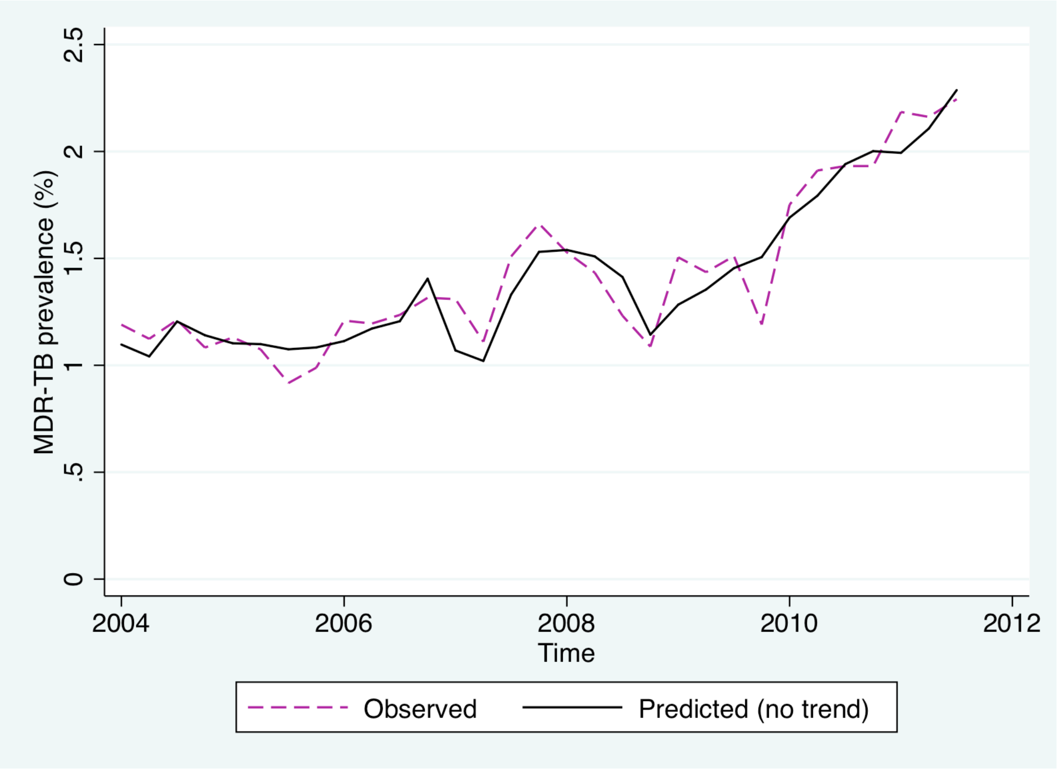

Figure 3: Observed MDR-TB prevalence over time in NHLS data compared to MDR-TB prevalence estimated from simulated maximum likelihood estimation.

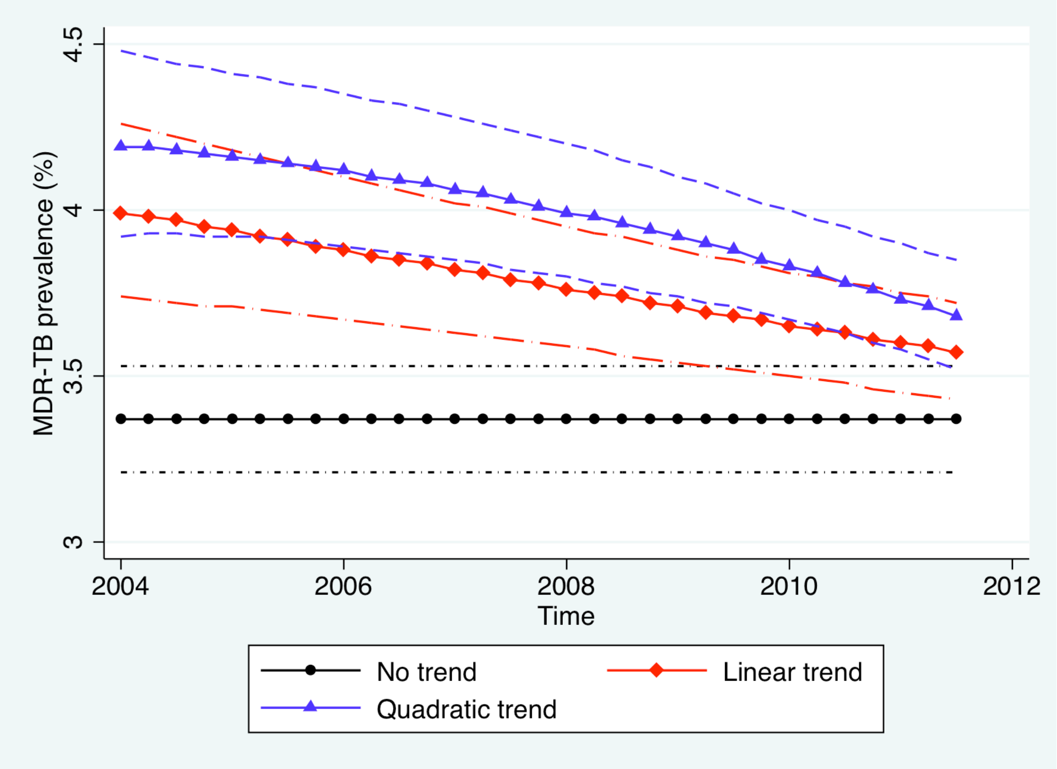

Figure 4: Predicted time trends in MDR-TB prevalence (%) in South Africa estimated from NHLS data using simulated maximum likelihood estimation. Dotted lines represent 95% confidence intervals.

Table 1: Estimated MDR-TB prevalence (μ) and noise to signal ratio (σ).

	SMLE		MSM-IV	
	Noise to signal ratio (σ)	Estimated prevalence (μ)	Noise to signal ratio (σ)	Estimated prevalence (μ)
Whole sample	2.119***	0.0337***	2.154***	0.0329***
	(0.057)	(0.0008)	(0.087)	(0.0011)
Split sample				
Repeat patients	1.227***	0.0469***	1.9***	0.0629***
	(0.067)	(0.002)	(0.169)	(0.0045)
New patients	16.235***	0.0565***	2.254***	0.0221***
	(5.727)	(0.0033)	(0.206)	(0.0016)

Notes: Table presents coefficients and standard errors. Sample includes TB-positive patients ages 16-64 in public health facilities from January 2004-December 2011.
*N = 2,565,951. New patients defined as within three months of first TB test in data. SMLE = Simulated maximum likelihood estimation. MSM-IV = Method of simulated moments – Instrumental variables. *** - Significant at the 1% level, ** - 5% level, * - 10% level.*

Table 2: Estimated MDR-TB prevalence (*μ)* and noise to signal ratio (σ) over time from simulated maximum likelihood estimation.

	Noise to signal ratio(*s*)	Prevalence index coefficient (μ_0*)	*Time coefficient* *(*μ_1**)*	*Time squared coefficient* (μ_2*)
No trend	2.12***	-1.83***		
	(0.06)	(0.01)		
Linear trend	2.40***	-1.75***	-0.0017***	
	(0.07)	(0.02)	(0.0004)	
Quadratic trend	2.56***	-1.73***	-0.0006	-0.00004
	(0.08)	(0.02)	(0.0012)	0.00004)

Notes: Table presents coefficients and standard errors. Sample includes TB-positive patients ages 16-64 in public health facilities from January 2004-December 2011.
*N = 2,565,951. *** - Significant at the 1% level, ** - 5% level, * - 10% level.*

Table 3: Estimated MDR-TB prevalence (μ) and noise to signal ratio (σ) by location type, gender, age and province.

	SMLE		MSM-IV	
	Noise to signal ratio (σ)	Estimated prevalence (μ)	Noise to signal ratio (σ)	Estimated prevalence (μ)
Gender				
Male	1.787***	0.0302***	2.119***	0.0335***
	(0.110)	(0.0015)	(0.187)	(0.0023)
Female	1.709***	0.0312***	2.102***	0.0352***
	(0.115)	(0.0017)	(0.187)	(0.0024)
Location type				
Hospital	1.562***	0.0227***	2.169***	0.029***
	(0.081)	(0.0010)	(0.2)	(0.0022)
Clinic	2.34***	0.0554***	2.012***	0.0498***
	(0.153)	(0.0023)	(0.136)	(0.0022)
Age group				
20-29	1.62***	0.0287***	2.107***	0.0345***
	(0.152)	(0.0021)	(0.256)	(0.0033)
30-39	1.8***	0.0322***	2.106***	0.0348***
	(0.111)	(0.0015)	(0.199)	(0.0025)
40-49	1.664***	0.0299***	2.106***	0.0347***
	(0.147)	(0.0021)	(0.264)	(0.0034)
50-59	3.057***	0.0431***	2.141***	0.0316***
	(0.645)	(0.0063)	(0.449)	(0.0054)
Province				
Eastern Cape	5.859***	0.0774***	2.058***	0.0431***
	(1.065)	(0.0051)	(0.248)	(0.0037)
Free State	1.769***	0.0214***	2.229***	0.0243***
	(0.334)	(0.0035)	(0.735)	(0.0067)
Gauteng	1.221***	0.025***	2.091***	0.0354***
	(0.082)	(0.0012)	(0.245)	(0.0031)
Limpopo	1.45***	0.0158***	2.352**	0.017***
	(0.462)	(0.0044)	(1.121)	(0.0068)
Mpumalanga	2.105***	0.0405***	1.673***	0.0352***
	(0.366)	(0.0055)	(0.305)	(0.005)
North West	8.341*	0.0637***	2.216**	0.0255**
	(6.469)	(0.0170)	(1.252)	(0.0135)
Northern Cape	2.578***	0.0465***	2.091***	0.0392***
	(0.330)	(0.0037)	(0.248)	(0.0032)
Western Cape	1.249***	0.0225***	2.124***	0.033***

	(0.097)	(0.0012)	(0.262)	(0.0031)

*Notes: Table presents coefficients and standard errors. Sample includes TB-positive patients ages 16-64 in public health facilities from January 2004-December 2011. KwaZulu-Natal omitted from provincial analysis due to data limitations. N = 2,565,951. SMLE = Simulated maximum likelihood estimation. MSM-IV = Method of simulated moments – Instrumental variables. *** - Significant at the 1% level, ** - 5% level.*