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

Zoë McLaren, Assistant Professor
School of Public Health, University of Michigan

Rulof Burger, Associate Professor
Dept. of Econ, Stellenbosch University, South Africa

April 2017
Paper download at zoemclaren.com
Forthcoming in *Health Economics*

Motivation

How are MDR-TB prevalence rates currently determined?

- Surveillance study
  - Accurate but infrequent
- Notification rates
  - Number of reported cases likely underestimate
- WHO crude adjustment to notification rates
  - Based on expert opinion (Glaziou et al. 2015)

Our contribution

- Develop a new econometric method for estimating population means from a selected sample
  - Identifies under-detection
  - Minimal data requirements: routine data
  - Minimal assumptions
  - Low cost, easy real-world implementation
- Useful for monitoring rare and emerging diseases
- We estimate that 16 to 26 % of all multi-drug resistant TB cases in South Africa were undiagnosed 2004-2011

Foundation for method

- Methods to address sample selection on unobserved characteristics in economics literature
  - Instrumental variables (see Imbens and Angrist 1994, Angrist and Imbens 1995)
  - Bivariate normal style selection models (see Heckman 1976)
- HIV lit: adjusting survey estimates for representativeness
  - Interviewer random effects (McGovern et al. 2015)
  - Heckman-type selection models (Barnighausen et al. 2011, Hogan et al. 2012, Clark and Houle 2014)
  - Adjusting for survey non-response using mortality rates (Nyirenda et al. 2010)

Context: MDR-TB

Multi-drug resistant TB

- Resistant to the two first-line TB drugs
- Indistinguishable from drug-susceptible
- Only 12% of new TB cases tested for MDR globally
- MDR patients comprise:
  - Less than 5% of TB cases
  - 13% of TB deaths
  - 20% of TB spending

Guidelines for MDR testing

- Clinicians observe risk factors
  - TB history
  - Weakened immune system
  - High risk of exposure: prisoners, miners, health workers
- Risk factors are imperfect predictors of MDR
- Cannot test everyone for all possible forms of drug resistance
- Test the proportion of patients θ with highest likelihood of MDR based on observed signal x

Theoretical Model

General model

- Suppose we have a population for which we want to know the mean of outcome y
- We have a routine sample with observations selected to maximize value of y
- Characteristic *x* is observed
- Mapping of *x* to unobservable *y* is not fully known
- Determining value of *y* has associated cost

Key features of clinical decision making

- Patients must be tested before MDR treatment
- Too few resources to test every patient
- Clinician observes noisy signal about patient's likelihood of MDR-TB
- Testing resources determined exogenously:
  - Test materials, lab capacity
  - Funding
  - Clinician awareness and training

Key features of clinical decision making

- Clinician will test patients deemed most likely to have drug-resistance until resources are exhausted
- We assume clinician beliefs about mapping between risk factors and MDR+ is *consistent* within time periods
- Clinicians do not know actual MDR-TB prevalence

Methods

Identification strategy

- Use plausibly exogenous variation in threshold θ to draw inferences about
  - Distribution of y in the population
  - Sampling mechanism
    - Clinician's ability to predict MDR+ based on observable signal x
- Assume consistent beliefs about mapping of x to y
- Regression discontinuity intuition
  - Relax constraint on testing resources

Identification strategy

- Sample means at observed threshold proportion tested θ_0 informative about unobservable x_0:

$$E(y|\theta \leq \theta_0) = E(y|x \geq x_0)$$

- With a binary y, conditional expectation simplifies to:

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

where μ is population prevalence

- Rewriting relationship between y and x in error form:

$$x = \beta_0 + \beta_1 y + e$$

- Normalize β_0 to zero

Estimation

- Rewrite conditional expectation:

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

- Assume error term e follows standard normal distribution
- No closed-form (analytical) solution so we use numerical approximation techniques
- Estimate parameter values using maximum likelihood and generalized method of moments
- Grid searches to find most promising parameter space

Policy changes as instrumental variables

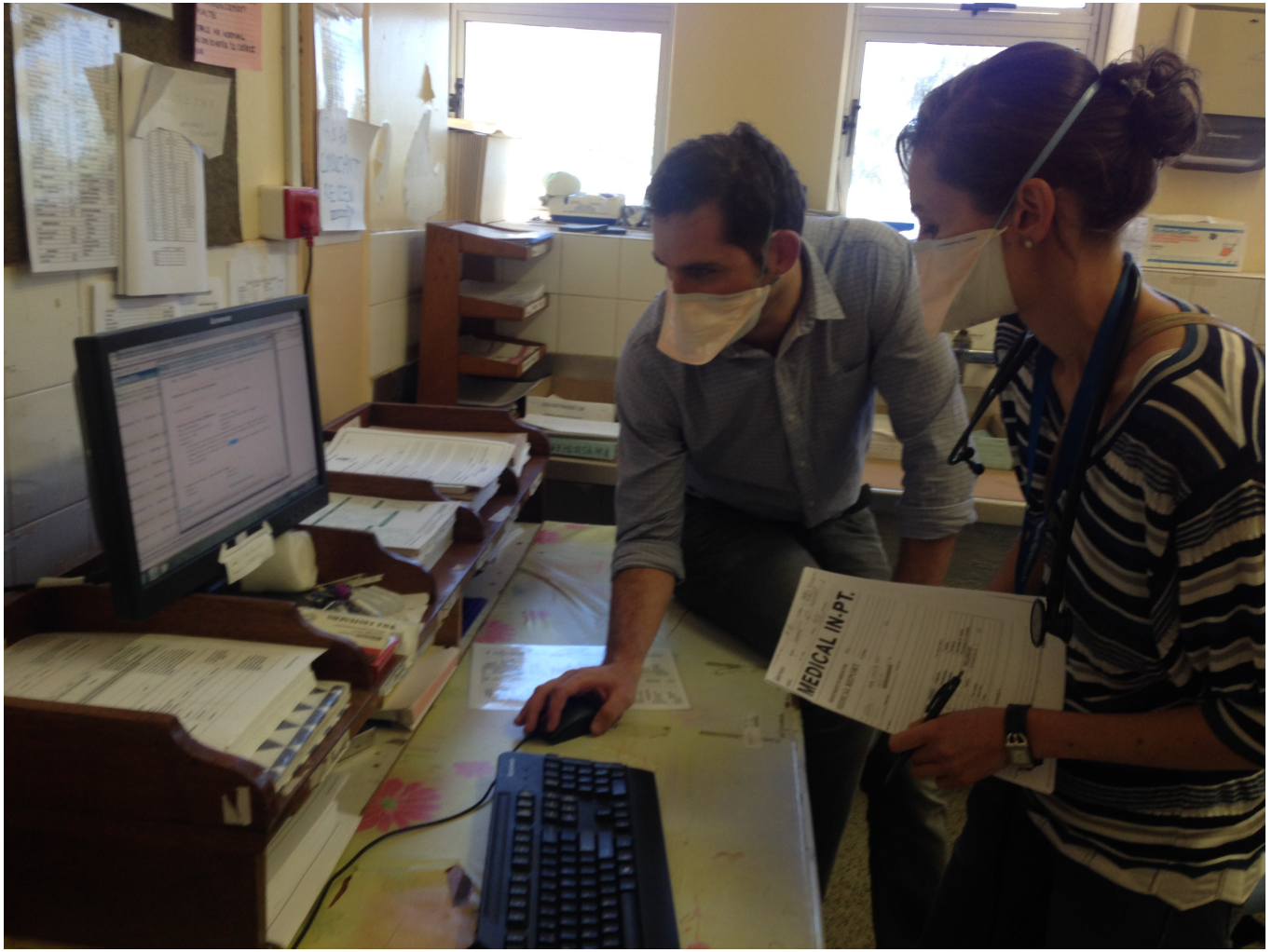
- Exogenous discontinuous changes in testing resources (θ)
  - Time period
  - XDR paper presented at international conference
  - National strategic plan introduced
- Should not affect clinician's understanding of risk factors or the underlying rate of MDR-TB

Data

Data

- National Health Laboratory Service database
- Laboratory database of test results for ~90% of all *suspected* TB cases
  - Jan 2004 – Sept 2010 (Prior to Xpert)
  - 2,190,780 TB+ test results
  - 8,647,12 patients
  - 4,764 health facilities

Accessing NHLS TB database

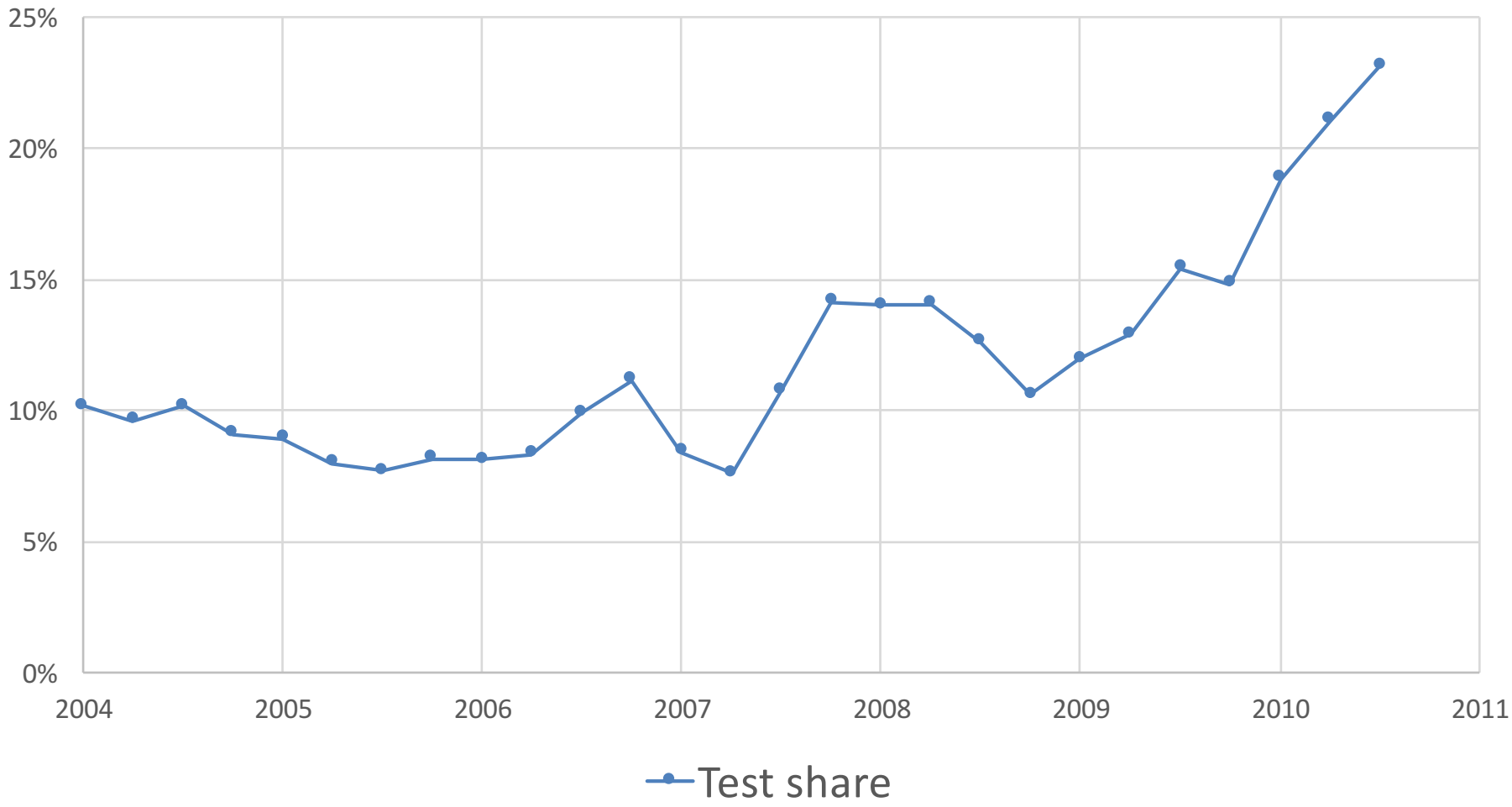


Data

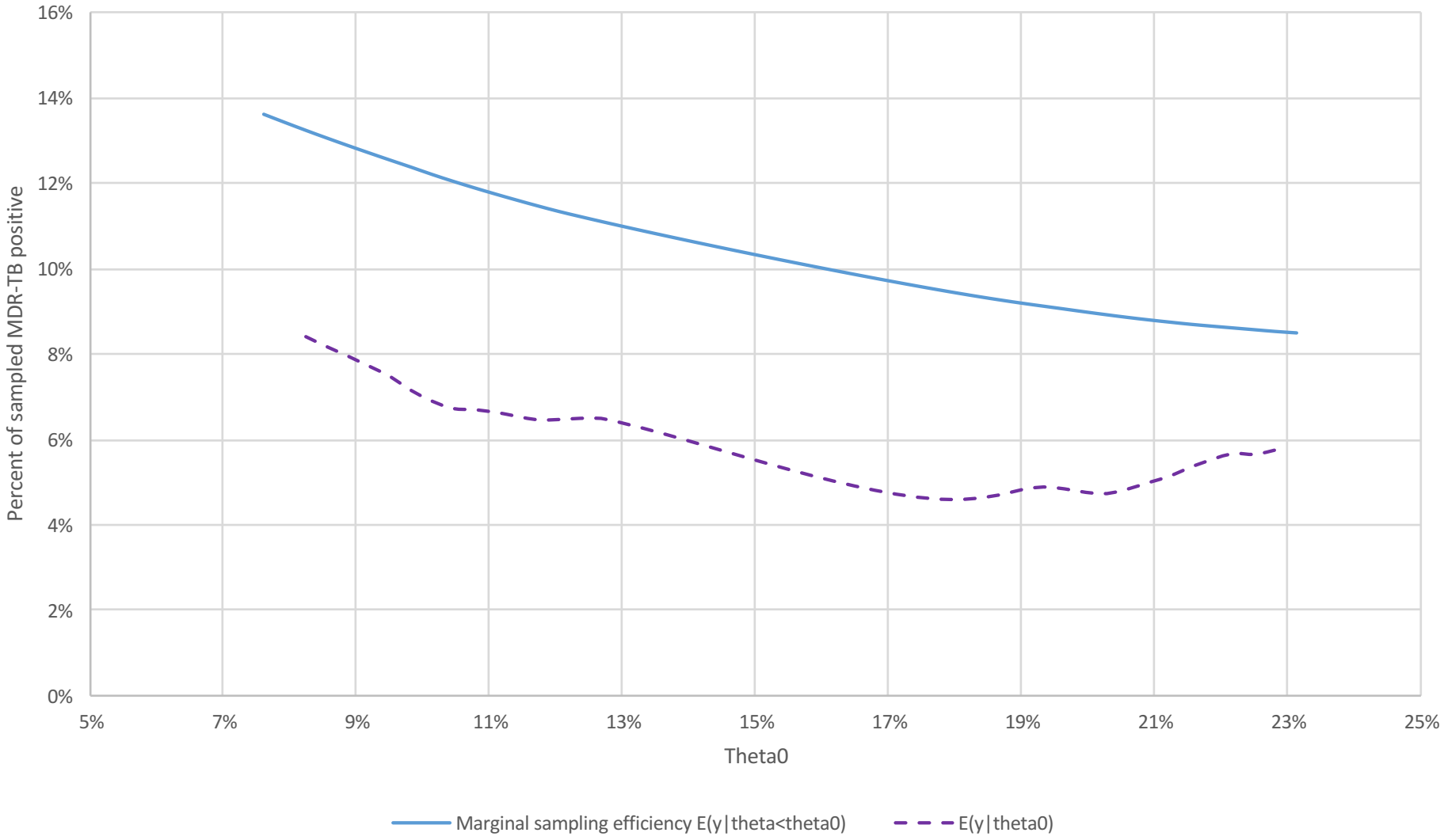
- Examine MDR-TB testing for those who are tested for TB, and have a TB+ result
- Minimal data requirements
- Extract number of patients:
  - TB+ result
  - Tested for MDR
  - MDR+ result

Results

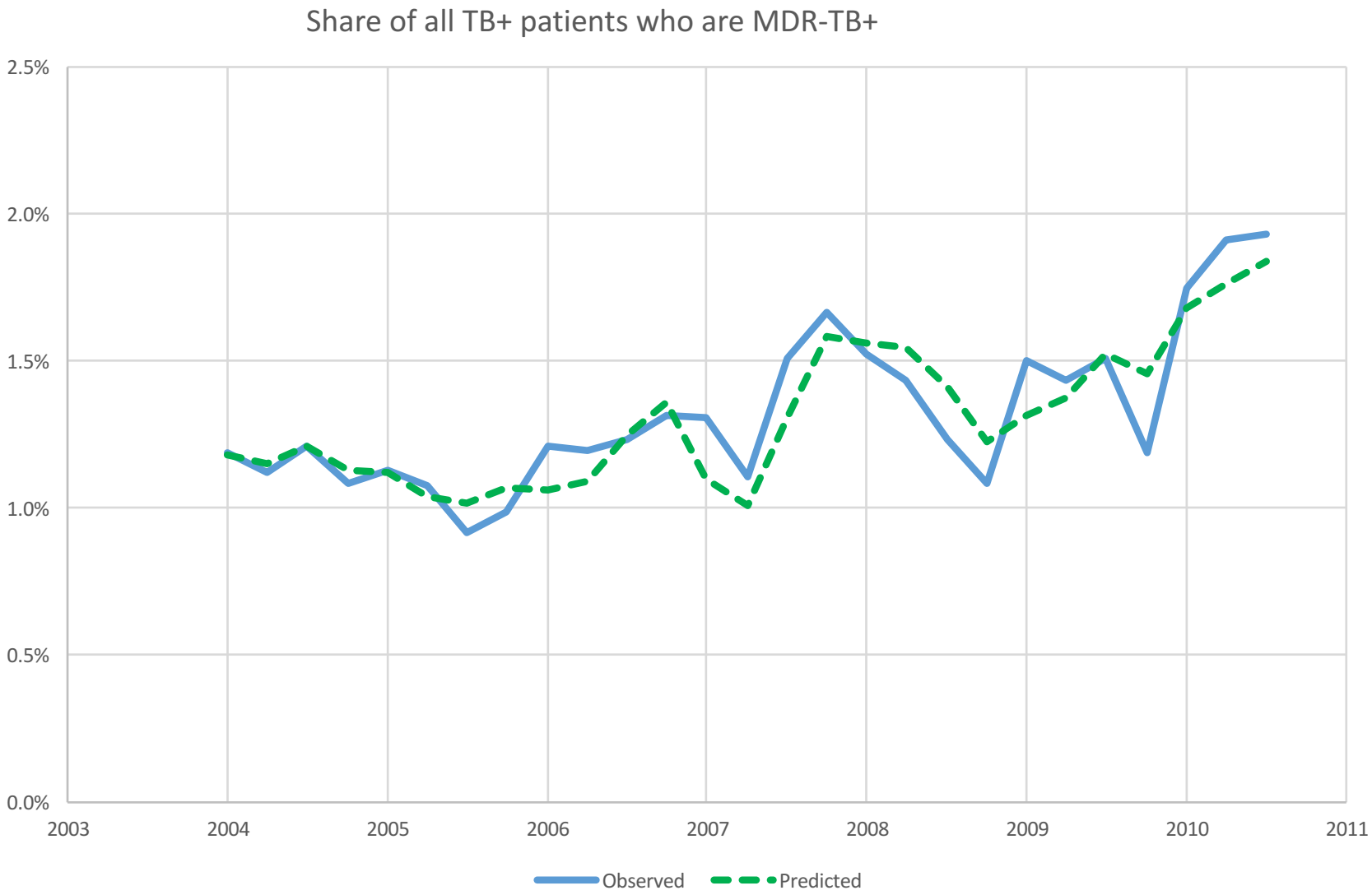
Fraction of TB+ tested for MDR



Sampling efficiency: percent of MDR tested who are MDR+



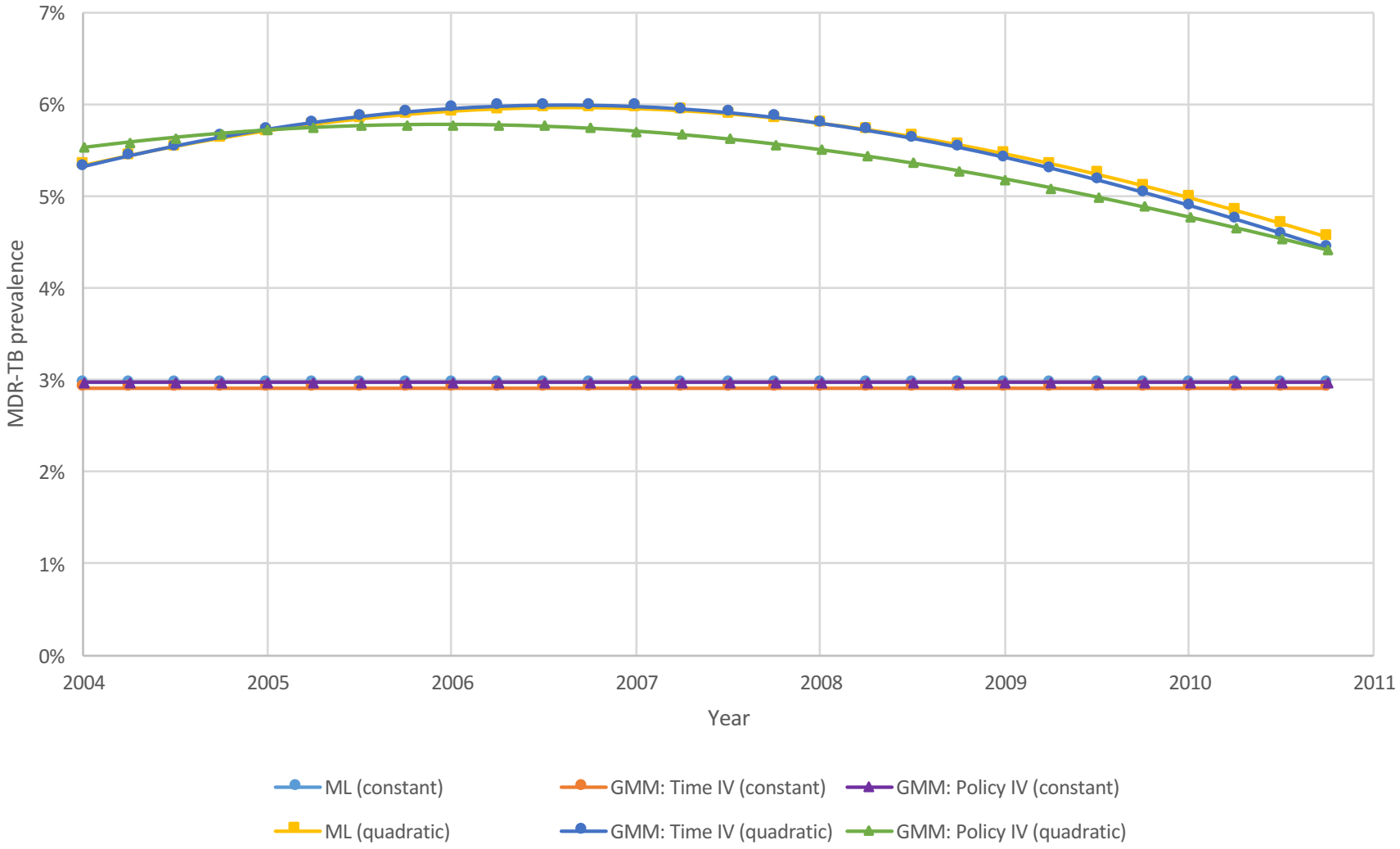
Predicted MDR+ matches observed MDR+ over time



Main results: robustness checks

Method:	ML	GMM	GMM	GMM	GMM
Instruments:		Time	Time	Policies	Policies
	(1)	(2)	(3)	(4)	(5)
Prevalence	0.0305***	0.0298***	0.0342***	0.0305***	0.0301***
	(0.0002)	(0.0013)	(0.0004)	(0.0016)	(0.0004)
Signal-to-noise	1.0749***	1.0891***	0.9336***	1.0716***	1.0846***
	(0.0003)	(0.0489)	(0.0117)	(0.057)	(0.0142)
Observations	262,845	262,845	262,853	262,850	262,842
Clustering	No	No	Yes	No	Yes
Pseudo R^2 #1	0.694	0.699	0.529	0.691	0.703
Pseudo R^2 #2	0.695	0.695	0.694	0.695	0.696
Log likelihood	-90646.845				
GMM criterion		0.00131	0.01626	0.00044	0.0078

Quadratic MDR-TB time trend estimates



Conclusions

- Evidence that “official” MDR rates are too low
  - 16-26% of MDR cases were undiagnosed
  - More resources are needed
- Routine data can provide real-time tracking
  - Widely available and under-used
  - Valuable where cannot test everyone or where compliance with guidelines < 100%
- Clear applications beyond TB

Thank you!