Tuberculosis Diagnosis and Treatment under Uncertainty

Rachel Cassidy\textsuperscript{a} and Charles F. Mansk\textsuperscript{b, c}

Draft: July 2019

\textsuperscript{a}Institute for Fiscal Studies, London, WC1E 7AE, UK; \textsuperscript{b}Department of Economics, Northwestern University, Evanston, IL 60208; and \textsuperscript{c}Institute for Policy Research, Northwestern University, Evanston, IL 60208.

Abstract

1.6 million people worldwide died from tuberculosis (TB) in 2017. A new TB diagnostic test – Xpert MTB/RIF (Cepheid, Sunnyvale, CA, USA) – was endorsed by the World Health Organization in 2010. Trials demonstrated that Xpert is faster and has greater sensitivity and specificity than smear microscopy – the most common sputum-based diagnostic test. However, subsequent trials found no impact of introducing Xpert on morbidity and mortality. We present a decision-theoretic model of how a clinician might decide whether to order Xpert or other tests for TB, and whether to treat a patient, with or without test results. Our first result characterizes the conditions under which it is optimal to perform empirical treatment: that is, treatment without diagnostic testing. We then examine the implications for decision-making of partial knowledge of TB prevalence or test accuracy. This partial knowledge generates ambiguity, also known as deep uncertainty, about the best testing and treatment policy. In the presence of such ambiguity, we show the usefulness of diversification of testing and treatment.

Significance Statement

Tuberculosis (TB) remains a serious global health problem. A new, more accurate test for diagnosis was endorsed by the World Health Organization in 2010. However, trials showed that using the test did not yield reductions in TB-related deaths. To help understand why, we model how a clinician might decide whether to order tests for TB and whether to treat a patient for TB, with or without test results. We highlight the role of uncertainty about the prevalence of TB and the accuracy of different tests, for patients with different characteristics. We show that, given such uncertainty, a reasonable policy may be to diversify testing and treatment, randomly assigning patients with certain characteristics to different combinations of testing and treatment.

1. Introduction

Addressing the continuing prevalence of tuberculosis (TB) in several regions of the world remains a major priority for global health policymakers and practitioners. 1.6 million people worldwide died from TB in 2017 (WHO, 2018), and the “End TB” strategy of the World Health Organization (WHO) is committed to reduce TB deaths by 95% during the period 2015-2035 (WHO, 2015). A key challenge in fighting TB is that of improving capacity for rapid and accurate diagnosis. A new TB diagnostic test – Xpert MTB/RIF (Cepheid, Sunnyvale, CA, USA) – was endorsed by the WHO in 2010 (WHO, 2010). Trials to establish Xpert’s diagnostic effectiveness demonstrated that Xpert is faster and has greater sensitivity and specificity than smear microscopy – the most common sputum-based diagnostic test under the status quo (Boehme \textit{et al}., 2010; Boehme \textit{et al}., 2011). Xpert is also much faster at diagnosing multi-drug-resistant (MDR) TB compared to existing culture tests. However, subsequent trials to establish Xpert’s impact on morbidity and mortality found no statistically significant effect across a range of settings (Yoon \textit{et al}., 2012; Hanrahan \textit{et al}., 2013; Theron \textit{et al}., 2014; Cox \textit{et al}., 2014; Mupfumi \textit{et al}., 2014; Calligaro \textit{et al}., 2015; Churchyard \textit{et al}., 2015), although they were not powered to detect modest effect sizes.
In light of this apparent paradox, we present a decision-theoretic model of how a clinician might decide: (1) whether to order one or more tests for TB; and (2) whether to treat a patient, with or without test results. The model is prescriptive: that is, it seeks to improve the performance of actual decision making. This is in contrast to a descriptive model, which would seek to understand and predict how actual decision makers behave. We begin by defining optimal decision-making, but then show that the clinician typically does not have the information required to assess optimality. We therefore ultimately focus on reasonable decision-making, as characterized in previous work (Manski, 2009, 2013, 2018).

The model highlights two key features of decision-making for TB diagnosis and treatment. First, evidence from the aforementioned trials suggests that an important driver of Xpert’s lack of impact on mortality is that clinicians frequently engage in empirical treatment under the status quo. That is, they treat patients for TB using observation alone, without obtaining a positive result from diagnostic tests (van Rie, 2015; Auld et al., 2016). Our model shows that there exist settings in which clinicians may find empirical treatment to be optimal for some groups of patients, even when one or more diagnostic tests are available. When a new, superior diagnostic is introduced, it may still be optimal to choose empirical treatment.

Second, important parameters relevant to decision-making may not be known by the clinician. Moreover, there may exist no credible basis for asserting a subjective probability distribution over the possibilities, as is supposed in the Bayesian paradigm for decision-making under uncertainty. Diagnosis and treatment is then a problem of decision-making under ambiguity (Ellsberg, 1961), also known as deep uncertainty.

There may, for example, be ambiguity regarding the prevalence of TB in specific sub-populations of patients. Reasons for this ambiguity may include underreporting, misdiagnosis, limited granularity in available data on patient characteristics, and absence of scientific consensus on the relationship between latent and active TB (Behr et al., 2018, and accompanying rapid responses). Prevalence surveys, expert opinion, and modelling may provide credible bounds on prevalence for certain sub-populations, but no basis for choosing a point estimate, or asserting a subjective probability distribution between the bounds.

Similarly, there may be ambiguity about the diagnostic effectiveness of Xpert for sub-populations that differ from the populations studied in trials. That is, the external validity of the available trials may be unclear. For example, trials performed on HIV-positive adults who are receiving antiretroviral therapy (ART) may not reveal much about diagnosis and treatment of children, or of ART-naïve HIV-positive adults with advanced levels of immunosuppression.

Considering this ambiguity, it is important to ask what the policy response to the availability of a new diagnostic test should be. We build on Manski (2009, 2013, 2018) to show that under ambiguity it is reasonable for clinicians to pursue diversification. That is, within groups of observationally identical patients, clinicians may want to randomly test and treat some fraction of patients, but not others. Diversification has the immediate benefit of eliminating gross errors. Over time, it also generates new evidence on the accuracy of new diagnostic tests and on treatment response, similar to the evidence produced by a randomized controlled trial.

The problem of tuberculosis diagnosis and treatment under ambiguity exemplifies a broad class of decision-making problems under ambiguity in global health. Randomized trials and observational studies of diagnostic tests and treatments may often yield credible bounds on test accuracy and treatment response in given settings, but not credible point estimates or subjective probability distributions between these bounds. Moreover, even when well-designed trials have high internal validity, they may lack external validity. That is, it may be difficult to extrapolate trial findings to different patient populations or different healthcare contexts.

2. A Model of Optimal Diagnosis and Treatment Decisions

We first abstract from ambiguity, and study decision-making when the clinician knows the population parameters that determine optimal diagnosis and treatment decisions. The idealized optimization model presented here applies to a broad spectrum of medical settings. We specify how it relates to TB.
2.1 Basic Concepts and Notation

To model diagnosis and treatment decisions, we first specify a decision maker and a set of feasible actions. We use the concepts and notation of Manski (2013), applying the abstract setup developed there to TB. As there, we consider a clinician who cares for the population of patients who present to her for examination. We consider this patient population to be predetermined, and we assume that patients always comply with the clinician's decisions.\(^1\)\(^2\) We also assume that treatment decisions for these patients do not affect disease transmission.\(^3\)

When a patient presents for examination, the clinician initially observes covariates that may include demographic attributes, medical history, indicators of health status, and patient statements of preferences regarding care and outcomes. The clinician can prescribe a treatment based on observing these covariates alone (empirical treatment) or order a test that may yield further evidence. In the latter case, the clinician prescribes a treatment after observation of the test result.

Let the notation \(x\) denote the initially observed covariates of a patient, and \(t\) denote a treatment. We suppose for simplicity that there are two feasible treatments, \(t = A\) and \(t = B\). In the context of TB, let treatment \(A\) indicate surveillance – a decision not to prescribe antibiotics. Let treatment \(B\) indicate aggressive treatment – prescription of antibiotics, perhaps complemented by nutritional supplements.\(^4\)

Let \(s\) indicate whether the clinician orders the diagnostic test, with \(s = 1\) if she orders the test and \(s = 0\) if she does not. Let \(r\) denote the test result and suppose that \(r\) can take one of two values: \(p\), positive, indicating the patient has the condition; or \(n\), negative, indicating the patient does not have the condition.\(^5\)

The actions that the clinician may choose, and the knowledge of patient covariates accompanying each action, may be expressed as a decision tree. The clinician chooses \(s = 0\) or \(s = 1\) with knowledge of \(x\). If she chooses \(s = 0\), she chooses \(t = A\) or \(t = B\) with knowledge of \(x\). If she chooses \(s = 1\), she chooses \(t = A\) or \(t = B\) with knowledge of \((x, r)\).

When the clinician makes testing decisions \(s\), patients with the same value of \(x\) are observationally identical, while those with distinct values are observationally distinct. Hence, the clinician can use \(x\) to profile, making systematically different testing decisions for groups of patients with different values of \(x\). The clinician cannot profile within the group of patients having the same value of \(x\). However, she can randomly differentiate within the group, ordering testing for some fraction and not testing for the remainder. We term this \textit{diversification} in testing.

To formalize this idea, let \(\delta_S(x)\) be the fraction of the patients with covariates \(x\) who are tested and \(1 - \delta_S(x)\) be the fraction not tested. The clinician can choose \(\delta_S(x)\) to be any fraction in the interval \([0, 1]\). This done, she tests a randomly drawn fraction \(\delta_S(x)\) of the patient group and does not test the remainder. Such

---

\(^1\) The impact of introducing a new, improved TB diagnostic test on patients’ decision to present for examination is likely to be marginal. This is because the generic nature of symptoms (persistent cough, fever) means that patients will likely present for examination suspecting a range of possible illnesses. Moreover, if treatment following new diagnostic tests substitutes for empirical treatment, patients may not perceive an increase in the overall probability of receiving treatment.

\(^2\) In the case of TB treatment, arguably the largest source of patient non-compliance arises from patients not completing the course of antibiotics. It is not clear what effect, if any, improved diagnostic tests would have on compliance.

\(^3\) This assumption is a simplification given that TB is an infectious disease. The assumption is least problematic when considering testing and treatment of isolated patients, most so when considering broad public-health efforts to reduce the prevalence of TB. We discuss the implications of relaxing this assumption in section 4.

\(^4\) For simplicity we consider just two treatments, antibiotics versus observation only, and two illness states, TB versus no TB. The model can be extended to include testing for and treating MDR-TB alongside regular TB, and testing for and treating HIV alongside TB. These extensions may be accomplished with further notation.

\(^5\) In the context of TB, the raw measurements from both microscopy and Xpert are continuous. However, the standard practice in the research literature and in clinical practice has been to set a threshold and binarize the outcome. That is, one views measurements above the threshold as a positive test result, and measurements below as negative.
randomization could be implemented in a similar way to random security screening at airports, or random
drug testing of athletes.

Applying similar reasoning, the clinician can profile treatment across groups of patients with different
observed covariates and randomly differentiate treatment among patients with the same observed
covariates. When considering treatment, we need to distinguish three types of patients. Patients who are not
tested have observed covariates $x$ when they are treated. Patients who are tested have observed covariates
$(x, r)$ when treated, where $r$ equals $n$ or $p$. Among patients who are not tested, let $\delta T_0(x)$ be the fraction of
the patients with covariates $x$ who receive treatment $B$ and $1 - \delta T_0(x)$ be the fraction who receive treatment
$A$. Among those who are tested, let $\delta T_1(x, r)$ be the fraction of the patients with covariates $(x, r)$ who receive
$B$ and $1 - \delta T_1(x, r)$ be the fraction who receive $A$.

2.2 Welfare function

We next specify a welfare function that embodies the objective of the decision maker. Rather than
consider each patient in isolation, we will suppose that the objective of the clinician is to optimize care on
average across the patients in her practice. Optimization in this sense does not require that the clinician be
certain what treatment is best for each patient. It only requires knowledge of mean treatment response within
groups of patients having the same observed covariates.

Discussions of health care often suppose that a clinician should optimize care for each patient in
isolation, without reference to care of other patients. However, it is feasible to optimize care for a single
patient only if the clinician knows enough about individual treatment response to be certain what treatment
is best for this patient. Clinicians typically lack this knowledge, particularly when considering whether to
order a diagnostic test. After all, the medical purpose of a diagnostic test is to provide evidence on health
status that may be useful in choosing a treatment. If the clinician were already to know what treatment is
best, there would be no medical reason to contemplate a test.

It remains to specify the welfare function. As in Manski (2013), we assume that the clinician aggregates
the benefits and harms of making a specific testing and treatment decision for a given patient into a scalar
welfare measure. Going beyond the abstract notation of Manski (2013), we make the dependence of welfare
on illness explicit. Let $z = 1$ if the patient is ill and $z = 0$ otherwise. The clinician does not know $z$ when
choosing $(s, t)$. Given this, testing and treatment decisions will depend on a patient’s risk of illness rather
than on realized illness outcomes.

Specifically, let $U(z, s, t)$ summarize the clinician's overall assessment of the benefits and harms that
would occur if she were to make testing decision $s$ and treatment decision $t$ for a patient whose illness state
was $z$. The welfare measure may express not only health outcomes but also patient preferences and financial
costs. In the case of TB, the welfare $U(z, s, A)$ from the decision not to treat a patient may also include the
value of possible future treatment, based on the probability that an untreated patient will come back to be
diagnosed and treated again at a later date. Patients may respond heterogeneously to testing and treatment,
so $U(z, s, t)$ may vary across patients.

Mean welfare across the population of patients is determined by the fraction of those in each covariate
group that the clinician assigns to each testing-treatment option. Suppose that $x$ lies in a finite set $X$ of
possible covariate values. For each $x \in X$, let $P(x)$ denote the fraction of patients with covariate value $x$. For
$r \in \{p, n\}$, let $f(r|x)$ denote the fraction of patients with covariates $x$ who would have test result $r$ if they
were to be tested.

For each possible value of $(s, t)$, let $E[U(z, s, t)|x]$ be the mean expected welfare that would result if all
patients with covariates $x$ were to receive $(s, t)$, and let $E[U(z, s, t)|x, r]$ be the mean expected welfare that
would result if all patients with covariates $x$ and test result $r$ were to receive $(s, t)$, where:

\[
E[U(z, s, t)|x] = P(z = 0|x)E[U(0, s, t)|x] + P(z = 1|x)E[U(1, s, t)|x], \tag{1a}
\]

\[
E[U(z, s, t)|x, r] = P(z = 0|x, r)E[U(0, s, t)|x, r] + P(z = 1|x, r)E[U(1, s, t)|x, r]. \tag{1b}
\]
Let \( \delta = [\delta_S(x), \delta_{T0}(x), \delta_{T1}(x, r), x \in X, r \in \{p, n\}] \) denote any specified testing-treatment allocation. Then the mean welfare \( W(\delta) \) that would result if the clinician were to choose allocation \( \delta \) is obtained by averaging the various mean welfare values \( E[U(z, s, t)|x] \) and \( E[U(z, s, t)|x, r] \) across the groups who receive them. Thus,

\[
W(\delta) = \sum_{x \in X} P(x)[1 - \delta_S(x)][1 - \delta_{T0}(x)]E[U(z, 0, A)|x] + [1 - \delta_S(x)]\delta_{T0}(x)E[U(z, 0, B)|x]
\]

\[\sum_{r \in \{p, n\}} f(r|x)\{\delta_S(x)[1 - \delta_{T1}(x, r)]E[U(z, 1, A)|x, r] + \delta_S(x)\delta_{T1}(x, r)E[U(z, 1, B)|x, r]\}].\]

2.3. Optimal Testing and Treatment

An optimal testing and treatment allocation is any \( \delta \) that maximizes \( W(\delta) \). Manski (2013) shows that an optimal allocation is

\[
\delta_S(x) = \begin{cases} 1 & \text{if } \sum_{r \in \{p, n\}} f(r|x)[\max\{E[U(z, 1, A)|x, r], E[U(z, 1, B)|x, r]\}] > \max\{E[U(z, 0, A)|x], E[U(z, 0, B)|x]\}, \\ 0 & \text{otherwise.} \end{cases}
\] [3a]

\[
\delta_{T0}(x) = \begin{cases} 1 & \text{if } E[U(z, 0, B)|x] > E[U(z, 0, A)|x], \\ 0 & \text{otherwise.} \end{cases}
\] [3b]

\[
\delta_{T1}(x, p) = \begin{cases} 1 & \text{if } E[U(z, 1, B)|x, p] > E[U(z, 1, A)|x, p], \\ 0 & \text{otherwise.} \end{cases}
\] [3c]

\[
\delta_{T1}(x, n) = \begin{cases} 1 & \text{if } E[U(z, 1, B)|x, n] > E[U(z, 1, A)|x, n], \\ 0 & \text{otherwise.} \end{cases}
\] [3d]

Each maximum is unique when the stated inequality is strict, while all allocations yield the same welfare when the values are equal.

The above derivation shows that empirical treatment (treatment with antibiotics without performing a diagnostic test) is optimal when the inequality in (3a) does not hold and the inequality in (3b) does hold. Empirical treatment is not optimal otherwise.

The analysis in Manski (2013) extends immediately to settings with more than two possible test results. To perform the extension, one simply sums over the feasible results in Eq. 3a and extends Eq. 3c-3d to consider each feasible result. For example, \( r \) could be an ordered measure of the magnitude of a test finding. Or one could undertake multiple tests, in which case \( r \) gives a combination of test results. If one undertakes multiple tests, the analysis assumes that they are ordered together, and their results are observed simultaneously. Sequential testing can in principle be accommodated, but it requires generalization of the framework.

2.4. Risk of Illness and Treatment Decisions

To simplify further computations, we henceforth use a more compact notation for \( E[U(z, s, t)|x] \) and \( E[U(z, s, t)|x, r] \), as follows:

\[
E[U(z, s, t)|x] = (1 - P_x)U_x(0, s, t) + P_xU_x(1, s, t),
\] [1a']
\[ E[U(z, s, t) | x, r] = (1 - P_{xr})U_{xr}(0, s, t) + P_{xr}U_{xr}(1, s, t). \]  \[1b'\]

We can simplify further if we assume that knowledge of a test result does not directly affect patient welfare in a given illness state: formally, \( U_{x}(0, s, t) = U_{x}(0, s, t) \) and \( U_{x}(1, s, t) = U_{x}(1, s, t) \). With this assumption, knowledge of the test result affects decision-making purely by changing risk assessment from \( P_{x} \) to \( P_{xr} \), not for any other reason. In the context of TB, the assumption means that, if one were to know that a patient is or is not ill with the disease, the result of a microscopy or Xpert test would not affect welfare. We think this assumption is realistic and maintain it below.

With the above notation and assumption, the treatment decision criteria in Eq. 3b-3d are as follows:

\begin{enumerate}
  \item \textbf{treatment with no test result:} [3b']
    \begin{align*}
      & \text{choose B if } (1 - P_{x})U_{x}(0, 0, B) + P_{x}U_{x}(1, 0, B) \geq (1 - P_{x})U_{x}(0, 0, A) + P_{x}U_{x}(1, 0, A), \\
      & \text{choose A otherwise.}
    \end{align*}
  \item \textbf{treatment with positive test result:} [3c']
    \begin{align*}
      & \text{choose B if } (1 - P_{xp})U_{x}(0, 1, B) + P_{xp}U_{x}(1, 1, B) \geq (1 - P_{xp})U_{xp}(0, 1, A) + P_{xp}U_{x}(1, 1, A), \\
      & \text{choose A otherwise.}
    \end{align*}
  \item \textbf{treatment with negative test result:} [3d']
    \begin{align*}
      & \text{choose B if } (1 - P_{xn})U_{x}(0, 1, B) + P_{xn}U_{x}(1, 1, B) \geq (1 - P_{xn})U_{x}(0, 1, A) + P_{xn}U_{x}(1, 1, A), \\
      & \text{choose A otherwise.}
    \end{align*}
\end{enumerate}

In the medical literature on diagnostic testing, \( P_{x} \) is called the \textit{base rate} or the \textit{prevalence} of the illness for patients with covariates \( x \). \( P_{xp} \) is called the \textit{positive predictive value} of a test and \( 1 - P_{xn} \) is called the \textit{negative predictive value}. In general, \( P_{xp} > P_{xn} \). An ideal test that perfectly predicts disease would have \( P_{xp} = 1 \) and \( P_{xn} = 0 \). In practice, tests are imperfect predictors, so \( 1 > P_{xp} > P_{xn} > 0 \).

A considerable part of the medical literature measures test accuracy in a different way, reporting the \textit{sensitivity} and \textit{specificity} of a test. Sensitivity is the probability that the test result is positive conditional on the patient being ill, \( P(r = p|x, z = 1) \). Specificity is the probability that the result is negative conditional on the patient being healthy, \( P(r = n|x, z = 0) \).

Sensitivity and specificity do not provide the information that a clinician would want to have to inform patient care. These measures of accuracy permit one to predict the test result conditional on patient health status, but the clinician’s problem is to predict health status conditional on the test result. Perceptive writers on diagnostic testing have long cautioned that sensitivity and specificity do not inform patient risk assessment. For example, Altman and Bland (1994) wrote (p. 102):

“The whole point of a diagnostic test is to use it to make a diagnosis, so we need to know the probability that the test will give the correct diagnosis. The sensitivity and specificity do not give us this information. Instead we must approach the data from the direction of the test results, using predictive values. Positive predictive value is the proportion of patients with positive test results who are correctly diagnosed. Negative predictive value is the proportion of patients with negative test results who are correctly diagnosed.”

Despite the cautions expressed by writers such as Altman and Bland, it has remained common to measure the accuracy of diagnostic tests by their sensitivity and specificity. We discuss the potential implications of this below in Section 4.

\textit{2.5. Threshold Risk Assessments for Choice between Surveillance and Aggressive Treatment}

It is often credible to make various assumptions about patient welfare when comparing surveillance and aggressive treatment. In particular,
(i) Health is better than illness: \( U_s(0, s, t) > U_s(1, s, t) \) for all \((s, t)\).

(ii) Testing is costly/harmful: \( U_s(z, 0, t) > U_s(z, 1, t) \) for all \((z, t)\).

(iii) Surveillance is better than aggressive treatment when healthy: \( U_s(0, s, A) > U_s(0, s, B) \) for all \(s\).

(iv) Aggressive treatment is better than surveillance when ill: \( U_s(1, s, B) > U_s(1, s, A) \) for all \(s\).

These assumptions are realistic in the TB context: (i) a patient is better off not having TB than having TB; (ii) performance of a microscopy or Xpert test does not harm patients, but does incur financial costs; (iii) when a patient is healthy, there is no benefit from prescription of antibiotics, but there are financial costs and possible harms to patients; (iv) when a patient is ill with TB, the health benefits of antibiotic treatment exceed the financial costs and harms.

Analysis in Manski (2018) shows that under assumptions (iii) and (iv), the treatment criteria in Eq. 3b'-3d' yield simple solutions. Aggressive treatment is the optimal decision if the risk of illness equals or exceeds a threshold that equalizes mean welfare under treatments A and B. Surveillance is better if risk is less than or equal to the threshold.

In the absence of testing, risk of illness is measured by \( P_x \) and the threshold yielded by criterion (3b') is

\[
P_{x0}^* = \frac{U_s(0, 0, A) - U_s(0, 0, B)}{[U_s(0, 0, A) - U_s(0, 0, B)] + [U_s(1, 0, B) - U_s(1, 0, A)]}.
\]  \[4a\]

With testing, risk of illness is measured by \( P_{xp} \) or \( P_{xn} \) respectively. The threshold yielded by both of the criteria in Eq. 2c' and 2d' is

\[
P_{x1}^* = \frac{U_s(0, 1, A) - U_s(0, 1, B)}{[U_s(0, 1, A) - U_s(0, 1, B)] + [U_s(1, 1, B) - U_s(1, 1, A)]}.
\]  \[4b\]

Under assumptions (iii) and (iv) both thresholds lie in the open interval \((0, 1)\).

It is important to keep in mind that some treatment errors occur with optimal decisions. Define a Type I error to be a choice of treatment B when A is optimal. Let a Type II error be a choice of A when B is optimal. Suppose that a clinician makes optimal decisions as derived above. Without testing, Type I errors do not occur when \( P_x < P_{x0}^* \) and Type II errors do not occur when \( P_x > P_{x0}^* \). Type II errors occur with probability \( P_x \) when \( P_x < P_{x0}^* \) and Type I errors occur with probability \((1 - P_x)\) when \( P_x > P_{x0}^* \). Analogous results hold with testing.

### 2.6. How Testing Affects Treatment

We observed in Section 2.3 that empirical treatment is optimal if the inequality in 3a does not hold and the inequality in 3b does hold. When choosing between surveillance and aggressive treatment, the inequality in 3b reduces to the condition \( P_x > P_{x0}^* \). We now ask how, if at all, testing affects treatment.

In general, the answer to this question is complex because the thresholds \( P_{x0}^* \) and \( P_{x1}^* \) in Eq. 4a and 4b may differ. Substantial simplification occurs if the thresholds are equal. A sufficient condition for equality is the assumption that testing imposes an additive treatment-invariant cost on welfare; that is, \( U_s(z, 0, t) - U_s(z, 1, t) = K > 0 \) for some positive \( K \), for all \( z \) and \( t \). This assumption is realistic in the case of TB, since treating a patient with antibiotics is generally no more or less costly depending on whether or not the patient has taken a diagnostic test. In contrast, the assumption may be violated for diseases where treatment is easier to perform after a test: for example, if a testing procedure is invasive, and treatment can be delivered at the same time.
Suppose that $P^{*}x_{0} = P^{*}x_{1}$ and let the common value be denoted $P^{*}x$. Then the implication of testing for treatment depends purely on the magnitudes relative to $P^{*}x$ of the pertinent probabilities of illness ($P_{x}$, $P_{xn}$, $P_{xp}$). It holds algebraically that $P_{x}$ lies between $P_{xn}$ and $P_{xp}$. Specifically,

$$P_{x} = P_{xn}(1 - f_{x}) + P_{xp}f_{x},$$

where $f_{x} \equiv f(r = p|x)$ is the probability of a positive test result. We assume that a positive result indicates a higher risk of illness than does a negative one, so $P_{xn} < P_{xp}$. This inequality and Eq. 5 yield the inequality $P_{xn} < P_{x} < P_{xp}$.

It follows that testing affects optimal treatment if and only if the inequality $P_{xn} < P^{*}x < P_{xp}$ holds. Given this inequality, a patient with a positive test result receives treatment B, and a patient with a negative test result receives A. In the absence of testing, the patient might receive either A or B, depending on whether the risk of illness is above or below the threshold characterized in Equation (4a).

Testing does not affect treatment otherwise. If $P_{xp} < P^{*}x$, treatment A is optimal with or without testing. If $P_{xn} > P^{*}x$, treatment B is optimal with or without testing.

### 3. Testing and Treatment under Ambiguity

#### 3.1 Sources of Ambiguity and Standard Decision Criteria

Within the model of Section 2, optimization of testing and treatment for patients with covariates $x$ is feasible if one knows the mean welfare function $U_{\gamma}(\cdot, \cdot, \cdot)$, the illness probabilities $(P_{x}$, $P_{xn}$, $P_{xp})$, and the probability $f_{x}$ of a positive test result. A clinician with incomplete knowledge may not be able to optimize and hence faces a problem of decision making under ambiguity. To formalize incomplete knowledge, let the state space, denoted $\Gamma$, list the vectors $(U_{\gamma x}$, $P_{\gamma x}$, $f_{\gamma x}$, $P_{\gamma xp}$, $P_{\gamma xn})$, $x \in X$, $\gamma \in \Gamma$ that satisfy Eq. 5 for each value of $x$ and that are deemed feasible based on available evidence and maintained assumptions.

To consider decision-making under ambiguity, Manski (2013) begins with the welfare function of Eq. 2, considered as a function over the state space. For each $\gamma \in \Gamma$,

$$W_{\gamma}(\delta) =$$

$$\sum_{x \in X} P(x)[(1 - \delta(x))[1 - \delta_{T0}(x)]E_{\gamma}[U(z, 0, A)|x] + [1 - \delta(x)]\delta_{T0}(x)E_{\gamma}[U(z, 0, B)|x]$$

$$+ \sum_{r \in \{p, n\}} f_{\gamma}(r|x)\{\delta_{T1}(x, r)\delta_{T1}(x, r)[1 - \delta_{T1}(x, r)]E_{\gamma}[U(z, 1, A)|x, r] + \delta_{T1}(x, r)\delta_{T1}(x, r)E_{\gamma}[U(z, 1, B)|x, r]}\},$$

where

$$E_{\gamma}[U(z, 0, t)|x] = (1 - P_{\gamma x})U_{\gamma x}(0, 0, t) + P_{\gamma x}U_{\gamma x}(1, 0, t),$$

$$E_{\gamma}[U(z, 1, t)|x, r] = (1 - P_{\gamma xr})U_{\gamma x}(0, 1, t) + P_{\gamma xr}U_{\gamma x}(1, 1, t).$$

With this structure, one may in principle study decision making using standard criteria, including maximization of subjective expected welfare, maximin, and minimax-regret. Maximization of subjective expected welfare is a standard dynamic programming problem and thus is tractable. However, it requires specification of a subjective distribution on the state space, which we find difficult to motivate. Study of the other criteria appears to require complex new analysis.
3.2. Piecemeal Minimax-Regret Decision Making

Rather than pursue any of the above approaches, we propose a piecemeal minimax-regret criterion. We consider each value of \( x \) and each of the four component decisions in isolation from one another. These components are the choices: (1) to test or not to test, (2) between A and B without testing, (3) between A and B with testing and a positive result, and (4) between A and B with testing and a negative result. Each choice is a decision between two options, making piecemeal decision making relatively simple to study. Piecemeal decision making may also be realistic in settings where each component decision may be performed by a different clinician, for example if a different clinician may be on duty for the follow-up consultation to make the treatment decision once the test results have been received. In such a setting, each clinician cannot control what the clinician who makes the subsequent decision will do, and may not even be able to communicate with him or her. Thus, a reasonable approach is to model each subsequent decision as separate.

We perform analysis that extends the study of minimax-regret decision making in Manski (2009). The extension is especially simple if we suppose that \( U_x(\cdot, \cdot, \cdot) \) is known; hence, the threshold risk assessment \( P^*_x \) is known. Considerable scope for ambiguity remains through incomplete knowledge of \( (P_x, P_{xn}, P_{xp}) \) and \( f_x \). We proceed abstractly here and characterize the ambiguity in the TB context in Section 4.

The minimax-regret analysis in Manski (2009), p. 1019 can be applied separately to each component decision. In each case, the result is a singleton allocation of patients in the absence of ambiguity and a fractional allocation with ambiguity. In non-technical language, a fractional allocation means diversification of treatment.

Consider decisions 3 and 4. The options are treatments A and B. For test result \( r \), let the smallest and largest feasible values of \( P_{xr} \) be denoted \( P_{xrL} \) and \( P_{xrH} \) respectively. Section 4 discusses how these lower and upper bounds may be generated in practice.

Ambiguity occurs when \( P_{xrL} < P^*_x < P_{xrH} \). Let \( M_{xr}(B) \) be the maximum value of the average treatment effect \( E_{x}[U(z, 1, B)|x, r] - E_{x}[U(z, 1, A)|x, r] \) across the state space. With ambiguity, the maximum is positive and occurs when \( P_{xr} = P_{xrH} \), i.e. at the maximum probability of being ill conditional on \( x \) and \( r \). Analogously, let \( M_{xr}(A) \) be the maximum value of the reverse average treatment effect \( E_{x}[U(z, 1, A)|x, r] - E_{x}[U(z, 1, B)|x, r] \), which occurs when \( P_{xr} = P_{xrL} \). Thus,

\[
M_{xr}(B) = (1 - P_{xrH})U_x(0, 1, B) + P_{xrH}U_x(1, 1, B) - (1 - P_{xrH})U_x(0, 1, A) - P_{xrH}U_x(1, 1, A), \tag{8a}
\]

\[
M_{xr}(A) = (1 - P_{xrL})U_x(0, 1, A) + P_{xrL}U_x(1, 1, A) - (1 - P_{xrL})U_x(0, 1, B) - P_{xrL}U_x(1, 1, B). \tag{8b}
\]

The analysis in Manski (2009) shows that, among patients with test result \( r \), the minimax-regret criterion yields

\[
\delta_{T1}(x, r) = 0 \quad \text{if} \quad P_{xrL} \leq P^*_x, \quad \text{if} \quad P_{xrL} < P^*_x < P_{xrH}, \quad \text{if} \quad P_{xrH} \leq P^*_x, \tag{9}
\]

Consider decision 2. The situation is the same except that the relevant probability of illness is \( P_x \) (and its bounds, \( P_{xL} \) and \( P_{xH} \)) rather than \( P_{xr} \). With this modification, the above result in Equations (8a)-(9) applies.

Consider decision 1. The options are to test and not to test. When making the testing decision, one should consider how decisions (2) through (4) will be made. Suppose that piecemeal minimax regret will be used to make decisions (2) through (4). It can be shown that the average effect of testing is then the average effect of treatment compared to surveillance, multiplied by the difference between the probabilities that the patient will be assigned to treatment with testing compared to without testing, minus the cost of testing. See Supporting Information S1 for the derivation.
Applying again the analysis in Manski (2009), the fraction of patients allocated to testing is similar to the result in Equation (9). The fraction is zero if the maximum “average treatment effect” of testing vs not testing (i.e. the maximum regret from not testing) is less than zero. It is one if the “average treatment effect” of not testing vs testing (i.e. the maximum regret from testing) is negative. It is a fractional allocation if both values are positive.

3.3. Adaptive Diversification

The above provides a full description of static piecemeal decision-making. Finally, as in Manski (2009), consider adaptive application of the piecemeal criteria across a sequence of cohorts. Suppose that the distributions of test results and treatment response among persons with covariates x remain stable over time. Suppose as well that observation of patients eventually reveals whether or not they are ill. Then complete learning eventually occurs if δS(x) > 0 for some cohort. Randomized testing of patients with covariates x reveals fx, and randomized treatment following testing reveals Pxp and Pxn. Px is revealed directly if δS(x) < 1 and indirectly by Eq. 5 if δS(x) = 1.

We caution that complete learning may not occur if observation of patients does not always reveal whether they are ill. For example, patients with the illness in question may self-cure without treatment. Or patients may respond to treatment even if they have a different illness – in the case of TB, antibiotics may also cure non-TB bacterial infections. Thus, one may never learn with certainty whether a patient was ill.

Learning occurs most quickly when δS(x) = 1, i.e. with universal testing. However, our model shows that universal testing may not be reasonable given the cost of testing. Complete learning does not occur if δS(x) = 0 for all cohorts. In this case, randomized treatment reveals Px. This yields only partial knowledge of (fx, Pxp, Pxn), which must satisfy Eq. 5. To avoid this outcome, which is generally undesirable from a multi-cohort perspective, one might set δS(x) > 0 for some cohort.

4. Implications for TB testing and treatment

4.1 Optimal TB testing and treatment

The model is useful for studying TB, firstly because it formalizes the conditions under which empirical treatment is optimal. This is important because the status quo diagnostic test in the absence of Xpert – microscopy – has a low positive predictive value: see, for example, Boehme et al. (2011) and Theron et al. (2014).

The model makes clear that a clinician should choose treatment B regardless of the test result if the inequality in (3a) does not hold and the inequality in (3b) does hold. In this case, the expected welfare following testing, taking into account the probabilities of negative and positive results, is no greater than the welfare of assigning the patient without testing to either no treatment or treatment. Without testing, the expected welfare of treatment is higher than that of no treatment.

The conditions for optimality of empirical treatment may hold if a patient’s probability of having TB is high even following a negative test result. For example, given the poor predictive value of microscopy among HIV-positive patients with advanced levels of immunosuppression, a clinician may find it optimal to treat such patients even following a negative test result. Empirical treatment may also be optimal if the probability of having TB after a negative test result is moderate and the welfare cost of untreated illness is high, as may occur with patients in intensive care units.

If a clinician chooses to treat a patient empirically, then a Type I error is more likely to occur than in treatment following testing. The costs of Type I errors may be substantial. For example, they may prevent correct diagnosis and treatment of another condition (Houben et al., 2018). Our model takes these costs into account when determining whether empirical treatment is optimal or not.

The model also allows us to formalize the possible effects of introducing Xpert on rates of empirical treatment. Xpert has a higher positive predictive value than microscopy, making condition (3a) more likely
to hold. For some patient populations, it may therefore be optimal to switch out of empirical treatment and into testing with Xpert. For other patient populations, condition (3a) may still not hold. Then empirical treatment may remain optimal.

Several of the trials examining Xpert’s impact on morbidity and mortality found only partial substitution away from empirical treatment when Xpert was introduced (Yoon et al., 2012; Theron et al., 2014; Calligaro et al., 2015; Churchyard et al., 2015). The studies did not find conclusive evidence of reduced morbidity and mortality, which one might expect if there was a reduction in the number of Type II errors in treatment. A possible reason is that the studies were generally only powered to detect relatively large effects. Another possible reason is if empirical treatment mainly leads to type I rather than type II errors. This may occur if clinicians err on the side of over-treating, in order to reduce the risk of not treating patients who truly have TB. If introduction of Xpert mainly leads to a reduction in type I errors, then this reduces unnecessary treatment for TB. Yet this may not translate into significant reductions in morbidity and mortality, unless patients incorrectly treated for TB have other serious conditions and are now more quickly treated under a correct diagnosis.

4.2 Model limitations

The model has limitations insofar as it relies on certain simplifying assumptions. One is the assumption that patient response is individualistic. TB is an infectious disease, implying that the decision to treat a given patient may have spillovers on the illness status of other future patients and, hence, on future testing and treatment decisions. A clinician should take this into account when making individual testing and treatment decisions, if the effect of treating a given patient on future TB transmission is non-negligible. Whether the spillover effect is non-negligible may depend on TB prevalence, among other factors. A policymaker may take spillovers into account in setting clinical guidelines, even if each individual treatment decision has a negligible effect on transmission. This is a key topic for future research.

We also made the simplifying assumption that the patient’s decision to present for examination is fixed conditional on x – which may include the patient’s symptoms, the distance from the patient’s home to the clinic, and so on – and is not affected by testing and treatment policy. Introduction of a new diagnostic could in principle influence the patient’s decision as to whether to incur the time and cost of presenting for examination at a clinic. As discussed earlier, we think that the magnitude of this effect is likely to be small. In the case of a different policy such as case-finding intervention, one would have to allow for the policy substantially increasing the probability of patient presentation within certain patient populations.

4.3 Ambiguity in the TB context

To optimize, the clinician must know a patient’s risk of illness P(z|x) before performance of a test, the risk P(z|x, r) after observation of a test result, and the probabilities f(r|x) of positive and negative test results. There are several reasons why these parameters are subject to ambiguity in the context of TB.

First, when epidemiological studies estimate prevalence, they typically report P(z|w), where w is a subset of the attributes x that a clinician observes. For example, the WHO reports prevalence by country, HIV status, age and sex, but not by factors such as socioeconomic status or other comorbidities (WHO 2018). This is convenient for reporting and monitoring purposes, but it means that a clinician will typically face ambiguity over P(z|x). Second, imperfect data quality implies there is often ambiguity in estimates of P(z|w). In the absence of prevalence surveys, estimates are based on notification rates. Underreporting is

6 Nevertheless, clinicians may fail to update their behavior at least in the short run. The extent to which clinicians’ behavior is characterized by biases is an important consideration for descriptive modelling, but it is outside the scope of our prescriptive model.

7 As outlined earlier, one might also expect a reduction in morbidity and mortality if introduction of Xpert leads to more rapid diagnosis and hence correct treatment of MDR-TB. However, most of the studies cited are conducted in sites where prevalence of MDR-TB is relatively low.
typically accounted for by expert opinion, or a constant adjustment factor, rather than data or modelling of
the patient presentation decision (WHO 2018 Technical Appendix - Glaziou, P., et al.).

Third, when trials of diagnostic tests report predictive values, the same issue emerges that they report conditional on a subset of attributes. Thus, these studies reveal \( f(r|w) \), rather than \( f(r|x) \). Moreover, these studies do not report \( P(z|r, x) \) or even \( P(z|r, w) \). Instead, they report sensitivity, \( P(r = p|w, z = 1) \), and specificity, \( P(r = n|w, z = 0) \), as discussed above. Thus, these studies do not provide the clinician with the predictive values that she needs for decision making.

Fourth, there may also be ambiguity in the welfare function. There may be incomplete knowledge of the effectiveness of antibiotic treatment in curing patients who have TB. Again, this can arise if clinical trials to determine the effectiveness of antibiotic treatments condition on \( w \) rather than \( x \). There may also be ambiguity about the cost of different errors in treatment. Regarding type I errors, there may be uncertainty as to what will happen to patients if they are treated for TB when they in fact have a different condition. Regarding type II errors, there may be uncertainty as to whether and when a patient will present for examination again, if they are not treated after a first visit.

4.4 Illustrative numerical exercises

We perform some illustrative numerical exercises to demonstrate the quantitative importance of some of these sources of ambiguity in TB diagnosis, and in understanding the impact (or lack thereof) when a superior diagnostic such as Xpert is introduced. Supporting Information S2 details these exercises in full.

Of the published efficacy and effectiveness trials for Xpert, Theron et al. (2014) provides particularly granular and extensive reporting of data. We draw on this study as an example, to highlight the ambiguity that remains even when data and results are reported in a relatively thorough and transparent manner.

First, in Supporting Information S2.1, we calculate the positive and negative predictive values for Xpert and microscopy using the data in Theron et al. This exercise underlines the potentially misleading nature of reporting sensitivity and specificity alone. Xpert offers a dramatically greater sensitivity compared to microscopy – 84.3% compared to 61.2%, in the Cape Town clinic which we use as a case study – and particularly so for patients who are HIV-positive – 78.3% compared to 41.0%, taken across all clinics since this is not reported by clinic. The gains in predictive value, while still sizeable, are of a much smaller magnitude. The largest difference across tests is in the negative predictive value, which is 96.4% for Xpert compared to 92.4% for microscopy across all patients in the Cape Town clinic, 91.9% for Xpert compared to 80.4% for microscopy for HIV-positive patients across all clinics, and 98.1% for Xpert compared to 92.6% for microscopy for HIV-negative patients across all clinics. These predictive values, not sensitivity and specificity, are the values clinicians should use when making testing and treatment decisions. These more modest differences may thus help explain the muted impact of Xpert on rates of empirical treatment, morbidity, and mortality, especially in clinics where most patients are HIV-negative.

Second (Supporting Information S2.1), we illustrate how substantial ambiguity can arise from a seemingly minor lack of granularity in data. A clinician making TB testing and treatment decisions should wish to know the positive and negative predictive value of test results, conditional on a patient’s HIV status, in the context of her clinic and patient population. The data in Theron et al. are reported by clinic and HIV status separately, but not by HIV status conditional on clinic. The positive and negative predictive values conditional on HIV status and clinic needed by the clinician can therefore only be bounded.

Given that the largest improvement offered by Xpert compared to microscopy appears to be in the negative predictive value for HIV-positive patients, we focus on bounding the probability that an HIV-positive patient (\( w = 1 \)) at the Cape Town (\( x = CT \) ) clinic has TB conditional on a negative result (\( r = n \)). We show that with weak credible assumptions, the bounds on this probability are \( P(z = 1|x = CT, w = 1, r = n) \in [0.036, 0.566] \) for Xpert and \( P(z = 1|x = CT, w = 1, r = n) \in [0.076, 0.566] \) for microscopy. Meanwhile \( P(z = 1|x = CT, w = 1) \in [0.181, 0.566] \) for an HIV-positive patient at the Cape Town clinic, in the absence of a test.

A negative test result therefore substantially reduces the lower bound on the probability that an HIV-positive patient at the Cape Town clinic has TB, and this effect is larger for an Xpert test compared to a
microscopy test. However, without further assumptions or more granular data, the probability that such a patient has TB conditional on a negative test result still encompasses large values. A clinician may therefore reasonably treat such a patient even conditional on a negative test result, and hence reasonably perform empirical treatment in anticipation of this; that is, not order a test. Moreover, the fact that trials observe only a partial substitution away from empirical treatment when moving from microscopy to Xpert may be reasonable using the data available to clinicians.

5. Conclusion

The model we have presented provides an idealized yet helpful characterization of optimal clinical decision-making when testing for and treating TB. The model also highlights the role of ambiguity in such decision-making. Ambiguity may arise from imperfect data quality and lack of granularity in data reporting, reporting of sensitivity and specificity rather than predictive values, and incomplete knowledge of the welfare function.

The model and numerical exercises may help shed light on the apparent paradox that the recent introduction of a superior TB diagnostic – Xpert – has had little impact on morbidity and mortality. In particular, the model shows how empirical treatment (treatment without testing) may be optimal under full information and may be reasonable under ambiguity.

Under ambiguity, we showed that a reasonable policy is to diversify treatment and testing; that is, randomly assign observationally identical patients to different treatment and testing regimes, in proportions that can be calculated from available data. As well as having reasonable decision-theoretic properties, an additional benefit of diversification is that it produces learning. Diversification mimics a trial with multiple arms, one for each possible testing and treatment decision. Thus, over time it yields information on the distribution of test results and the risks of illness, with and without testing. Adaptive diversification would update the proportion of patients assigned to each treatment and testing regime as this information became available.

Implementation of diversification may pose practical and ethical challenges. Similar to a randomized trial, diversification generates equal treatment of patients ex ante, but not ex post. Procedures for obtaining patients’ informed consent to participate in a diversification scheme at the level of a clinic could be based on similar procedures for randomised controlled trials. If diversification were to be implemented on a larger scale, for example at the level of a region or country, the ethical considerations would be similar to those faced by large-scale policy experiments.

Adaptive diversification aside, the ambiguity currently faced by clinicians could be reduced by making relatively straightforward changes to the ways in which data from trials and prevalence studies are reported. Trials of diagnostic tests could report positive and negative predicted values, rather than focus on sensitivity and specificity. Studies could report more granular data; that is, data conditional on richer covariates. This would allow clinicians to condition their decision-making on a richer set of the patient characteristics that they observe.

Acknowledgements: Dr Cassidy’s work on this project was funded by the ESRC Centre for the Microeconomic Analysis of Public Policy (CPP) at the Institute for Fiscal Studies. We thank Kalipso Chalkidou, Michael Gneimer, Rein Houben, and seminar audiences at the Centre for Microdata Methods and Practice at University College London and the Institute for Policy Research at Northwestern Univesrity for valuable comments.
References (in order of citation)


World Health Organisation, “The End TB Strategy”,

World Health Organisation, “WHO endorses new rapid tuberculosis test”,


Rapid responses, BMJ 2018;362:k2738


Supporting Information

S1. Application of the Minimax-Regret Criterion to the Piecemeal Testing Decision

Suppose that piecemeal minimax regret will be used to make decisions (2) through (4). Then the average testing effect is

\[
\sum_{r \in \{p, n\}} f_{\gamma}(r|x) \left\{ [1 - \delta_{T1}(x, r)]E_{\gamma}[U(z, 1, A)|x, r] + \delta_{T1}(x, r)E_{\gamma}[U(z, 1, B)|x, r] \right\}
\]

\[
- \left\{ [1 - \delta_{T0}(x)]E_{\gamma}[U(z, 0, A)|x] + \delta_{T0}(x)E_{\gamma}[U(z, 0, B)|x] \right\}, \quad [S1]
\]

where \(\delta_{T1}(x, r)\) and \(\delta_{T0}(x)\) are the fractions allocated to B with and without testing, as defined above. Recall the identities \(f_{\gamma}(r = p|x) = 1 - f_{\gamma}(r = n|x)\) and \(P_{\gamma} = f_{\gamma}P_{\gamma p} + (1 - f_{\gamma})P_{\gamma n}\). These imply that the average testing effect is a function of the three unknowns \((f_{\gamma}, P_{\gamma p}, P_{\gamma n})\).

The expression for the average testing effect simplifies when the welfare cost of testing is constant across treatments and illness outcomes; that is, if \(U(z, 1, t) = U(z, 0, t) - K\) for some \(K > 0\) that does not vary with \((z, t)\). This additional assumption, that the cost of testing is illness-invariant as well as treatment-invariant, simplifies the analysis below, but it can be relaxed without changing the main intuition of the subsequent results. The assumption would not be realistic when considering diagnostic tests where the physical demands of taking the test or the time required to wait for results are more costly for a patient who is ill than for one who is not. These possibilities do not seem to be major concerns for TB tests. Producing sputum for microscopy or Xpert may be more difficult for sick patients, but it does not have adverse severe health consequences. The results of both microscopy and Xpert are returned relatively quickly.

Combining this assumption with the identity \(E_{\gamma}[U(z, 0, t)|x] = \sum_{r \in \{p, n\}} f_{\gamma}(r|x) E_{\gamma}[U(z, 0, t)|x, r]\) implies that the average testing effect reduces to

\[
\sum_{r \in \{p, n\}} f_{\gamma}(r|x) \left[ \delta_{T1}(x, r) - \delta_{T0}(x) \right] \left\{ E_{\gamma}[U(z, 0, B)|x, r] - E_{\gamma}[U(z, 0, A)|x, r] \right\} - K. \quad [S2]
\]

That is, the average testing effect is the average effect of treatment compared to surveillance, multiplied by the difference between the probabilities that the patient will be assigned to treatment with testing compared to without testing, minus the cost of testing.

To obtain the fraction of patients who are tested, define maximum regret when one chooses not to test or to test (or equivalently, maximum “average treatment effect” of testing vs of not testing). These respectively are

\[
M_{\gamma}(1) = \text{Max} \sum_{r \in \{p, n\}} f_{\gamma}(r|x) \left\{ [1 - \delta_{T1}(x, r)]E_{\gamma}[U(z, 1, A)|x, r] + \delta_{T1}(x, r)E_{\gamma}[U(z, 1, B)|x, r] \right\} - \left\{ [1 - \delta_{T0}(x)]E_{\gamma}[U(z, 0, A)|x] + \delta_{T0}(x)E_{\gamma}[U(z, 0, B)|x] \right\}, \quad [S3]
\]

\[
M_{\gamma}(0) = -\text{Min} \sum_{r \in \{p, n\}} f_{\gamma}(r|x) \left\{ [1 - \delta_{T1}(x, r)]E_{\gamma}[U(z, 1, A)|x, r] + \delta_{T1}(x, r)E_{\gamma}[U(z, 1, B)|x, r] \right\} - \left\{ [1 - \delta_{T0}(x)]E_{\gamma}[U(z, 0, A)|x] + \delta_{T0}(x)E_{\gamma}[U(z, 0, B)|x] \right\}. \quad [S4]
\]

Applying again the analysis in Manski (2009), the fraction of patients allocated to testing is

\[
\delta_{S}(x) = 0 \text{ if } M_{\gamma}(1) \leq 0,
\]

\[
= M_{\gamma}(1)/[M_{\gamma}(1) + M_{\gamma}(0)] \text{ if } M_{\gamma}(1) \geq 0 \text{ and } M_{\gamma}(0) \geq 0,
\]

\[
= 1 \text{ if } M_{\gamma}(0) \leq 0. \quad [S5]
\]

16
Computation of $M_1(x)$ and $M_2(x)$ requires maximization and minimization of the average testing effect over the feasible values of the three unknowns ($f_{\gamma x}$, $P_{\gamma x}$, $P_{\gamma x}$). This may be accomplished numerically in general. It has an analytical solution when the state space is simple. For example, suppose that ($P_{\gamma x}$, $P_{\gamma x}$) are known. Then the average testing effect is linear in $f_{\gamma x}$. Hence, the extrema occur at boundary values of $f_{\gamma x}$. Or suppose that $f_{\gamma x}$ is known. Then the average testing effect is linear in $P_{\gamma x}$ and $P_{\gamma x}$. If the state space for ($P_{\gamma x}$, $P_{\gamma x}$) is rectangular, then the extrema occur at boundary values.

S2. Numerical exercises

In this section we draw on Theron et al. (2014). We highlight the ambiguity arising from the fact that, although the data on testing outcomes is reported by clinic and by HIV status, it is not reported by HIV status at the clinic level.

S2.1 Positive and negative predictive values

Define

- $x =$ clinic.
- $w = 1$ if patient has HIV, $= 0$ if not.
- $r = p$ with positive TB test result, $= n$ with negative TB test result.
- $z = 1$ if culture shows that patient has TB, $= 0$ if not.

We refer to $z = 1$ as “illness,” taking culture as the reference standard. To apply the kind of decision-making formalized in the model, a clinician needs to know $P(z|x, w)$ and $P(z|x, r, w)$.

We use $x =$ Cape Town (CT) as an example. Table 1 of Theron et al. (2014) shows that HIV status is known at the time of TB testing for all but eight patients, placing $P(w = 1|x = CT)$ in the range [0.32, 0.34]. For the bounding analysis below, we take 0.32 as the lower bound, which yields the same identification region as if we knew 0.32 was the proportion (Manski, 2018).

Table 2 of Theron et al. contains the following raw data used to compute sensitivity and specificity at the Cape Town clinic, from which we can derive $P(r|x = CT)$, $P(z|x = CT)$, and $P(z|x = CT, r)$.

Table S1: Raw data extracted from Theron et al. (2014), Table 2.

<table>
<thead>
<tr>
<th></th>
<th>Gugulethu Clinic, Cape Town, South Africa</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Smear microscopy</strong></td>
<td></td>
</tr>
<tr>
<td>Sensitivity n/N</td>
<td>22/36</td>
</tr>
<tr>
<td>Specificity n/N</td>
<td>171/172</td>
</tr>
<tr>
<td><strong>Point-of-care Xpert</strong></td>
<td></td>
</tr>
<tr>
<td>Sensitivity n/N</td>
<td>32/38</td>
</tr>
<tr>
<td>Specificity n/N</td>
<td>161/163</td>
</tr>
</tbody>
</table>

The computations below refer to patients for whom complete data are available.

- The number of patients in both treatment groups, conditional on $x = CT$, is $36 + 172 + 38 + 163 = 409$.\(^8\)

\(^8\)The total number of patients in Cape Town appears to be 416 (208 tested with microscopy, 208 with Xpert). However, implicitly seven of the Xpert patients are not considered in the sensitivity/specificity calculations. The authors state that four patients have a failed Xpert result before repeat, and one a failed Xpert after repeat. This leaves two Xpert patients unaccounted for, perhaps due to a failed culture.
At recruitment, the number with culture-confirmed TB is $36 + 38 = 74$.
Hence, $P(z = 1| x = \text{CT}) = \frac{74}{409} = 0.181$.

The number of patients in the Xpert group is $38 + 163 = 201$.
The number with $r = p$ is $32 + 2 = 34$. Hence, $P(r = p| x = \text{CT}, \text{Xpert}) = \frac{34}{201} = 0.169$ and $P(r = n| x = \text{CT}, \text{Xpert}) = 0.831$.
The number of patients in the Xpert group with $z = 1$ is $38$. Hence, $P(z = 1| x = \text{CT}, \text{Xpert}) = \frac{38}{201} = 0.189$.
The number of patients in the Xpert group with $z = 1$ and $r = p$ is $32$. Hence, $P(z = 1| x = \text{CT}, \text{Xpert}, r = p) = \frac{32}{34} = 0.941$. This is the positive predictive value of Xpert at the Cape Town clinic.
The number of patients in the Xpert group with $z = 1$ and $r = n$ is $6$. Hence, $P(z = 1| x = \text{CT}, \text{Xpert}, r = n) = \frac{6}{167} = 0.036$ and $P(z = 0| x = \text{CT}, \text{Xpert}, r = n) = 0.964$. This is the negative predictive value of Xpert at the Cape Town clinic.

The number of patients in the microscopy group is $36 + 172 = 208$.
The number with $r = p$ is $22 + 1 = 23$. Hence, $P(r = p| x = \text{CT}, \text{microscopy}) = \frac{23}{208} = 0.111$ and $P(r = n| x = \text{CT}, \text{microscopy}) = 0.889$.
The number of patients in the microscopy group with $z = 1$ is $36$. Hence, $P(z = 1| x = \text{CT}, \text{microscopy}) = \frac{36}{208} = 0.173$.
The number of patients in the microscopy group with $z = 1$ and $r = p$ is $22$. Hence, $P(z = 1| x = \text{CT}, \text{microscopy}, r = p) = \frac{22}{23} = 0.957$. This is the positive predictive value of microscopy at the Cape Town clinic.
The number of patients in the microscopy group with $z = 1$ and $r = n$ is $14$. Hence, $P(z = 1| x = \text{CT}, \text{microscopy}, r = n) = \frac{14}{185} = 0.076$ and $P(z = 0| x = \text{CT}, \text{microscopy}, r = n) = 0.924$. This is the negative predictive value of microscopy at the Cape Town clinic.

Table S2 presents these positive and negative predictive values at the Cape Town clinic, and contrasts these figures with the sensitivity and specificity at the Cape Town clinic as reported by Theron et al. Table S3 in the Supplementary Appendix of Theron et al. also reports sensitivity and specificity, as well as positive and negative predictive values, by HIV status, aggregated across all clinics. Table S2 below also reports these values.

Table S2: Positive and negative predictive values compared to sensitivity and specificity, at recruitment, Theron et al. (2014)

<table>
<thead>
<tr>
<th></th>
<th>Point-of-care Xpert</th>
<th>Microscopy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cape Town, All HIV Statuses (Theron et al. Table 1)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity</td>
<td>84.3% (32/38)</td>
<td>61.2% (22/36)</td>
</tr>
<tr>
<td>Specificity</td>
<td>98.8% (161/163)</td>
<td>99.5% (171/172)</td>
</tr>
<tr>
<td>Positive predictive value</td>
<td>94.1% (32/34)</td>
<td>95.7% (22/23)</td>
</tr>
<tr>
<td>Negative predictive value</td>
<td>96.4% (161/167)</td>
<td>92.4% (171/185)</td>
</tr>
<tr>
<td><strong>All Clinics, by HIV status (Theron et al. Table S3)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity, HIV+</td>
<td>78.3% (97/124)</td>
<td>41.0% (50/122)</td>
</tr>
<tr>
<td>Specificity, HIV+</td>
<td>93.6% (304/325)</td>
<td>95.2% (294/309)</td>
</tr>
<tr>
<td>Sensitivity, HIV-</td>
<td>93.4% (56/60)</td>
<td>68.4% (41/60)</td>
</tr>
<tr>
<td>Specificity, HIV-</td>
<td>97.2% (206/212)</td>
<td>98% (236/241)</td>
</tr>
<tr>
<td>Positive predictive value, HIV+</td>
<td>82.3% (97/118)</td>
<td>77% (50/65)</td>
</tr>
</tbody>
</table>
Negative predictive value, HIV+ 91.9% (304/331) 80.4% (294/366)
Positive predictive value, HIV- 90.4% (56/62) 89.2% (41/46)
Negative predictive value, HIV- 98.1% (206/210) 92.6% (236/255)

S2.2 Bounding the probability of illness conditional on clinic, HIV status and test result

Manski (2018) shows how to use the data on P(z|x) and P(w|x) to produce worst-case bounds for P(z = 1|x, w). Applying Equation 2 of Manski (2018), the identification region for P(z = 1|x = CT, w = j) is the interval:

\[ [0,1] \cap \left[ \frac{P(z = 1|x = CT)}{P(w = j|x = CT)} \cdot \frac{P(z = 1|x = CT)}{P(w = j|x = CT)} \right]. \]  \[ \text{[S6]} \]

As shown in Section S2.1, the data from Theron et al. suggest that the largest improvement offered by Xpert compared to microscopy is in negative predictive value for HIV-positive patients. We therefore focus on bounding the predictive values of Xpert and microscopy in the case where w = 1 (the patient is HIV-positive) and r = n.

Applying Equation (S6) gives us the following identification region for P(z = 1|x = CT, w = 1):

\[ P(z = 1|x = CT, w = 1) \in [0,1] \cap [(0.181-0.68)/0.32, 0.181/0.32] = [0, 0.566]. \]  \[ \text{[S7]} \]

Here and later, we compute P(z = 1|x = CT) by the overall Cape Town illness rate 0.181 rather than by the slightly different rates realized in the Xpert and microscopy groups, which were 0.189 and 0.173 respectively.

We can use a credible assumption to tighten these bounds: namely that conditional on a given clinic, the probability of TB is higher conditional on a patient being HIV-positive than conditional on a patient being HIV-negative. That is,

\[ P(z = 1|x, w = 1) > P(z = 1|x, w = 0). \]  \[ \text{[S8]} \]

This implies that a lower bound on P(z=1|x=CT, w=1) is P(z=1|x=CT) =0.181. Hence we have the following bounds on the probability that an HIV-positive patient at the Cape Town clinic has TB in the absence of a test result:

\[ P(z = 1|x = CT, w = 1) \in [0.181, 0.566] \]  \[ \text{[S7']} \]

To obtain bounds for P(z|x, w = 1, r = n) requires conditioning on x and on the joint event that w = 1 and r = n. Again, due to the way that data is reported in Theron et al. (2014), we do not know P(w = 1, r = n|x = CT). We can apply Equation 2 of Manski (2018) to use P(r|x = CT) and P(w|x = CT) to bound P(r|x = CT, w) for each of Xpert and microscopy. This gives the following identification regions for P(r = n|x = CT, w = 1):

Xpert: \[ P(r = n|x = CT, w = 1) \in [0.1] \cap [(0.831-0.68)/0.32, 0.831/0.32] = [0.472, 1] \]  \[ \text{[S9]} \]

Microscopy: \[ P(r = n|x = CT, w = 1) \in [0.1] \cap [(0.889-0.68)/0.32, 0.889/0.32] = [0.653, 1] \]  \[ \text{[S10]} \]

---

9 Table S3 of Theron et al. shows that this is true in the full sample, i.e. unconditional on a given clinic.
The joint probability \( P(w = 1, r = n|x = CT) \) equals \( P(r = n|x = CT, w = 1) \) multiplied by \( P(w = 1|x = CT) = 0.32 \). Hence we have that:

**Xpert:** \( P(w = 1, r = n|x = CT) \in [0.151, 0.32] \). \[S11\]

**Microscopy:** \( P(w = 1, r = n|x = CT) \in [0.209, 0.32] \). \[S12\]

These bounds can be used to obtain bounds for \( P(z = 1|x = CT, w = 1, r = n) \) by replacing \( P(w = j) \) in Equation (S6) with \( P(w = 1, r = n|x = CT) \) and \( P(w \neq j) \) with \( (1 - P(w = 1, r = n|x = CT) \). The bounds in (S11) and (S12) can then be inserted into this modified version of Equation (S6) to obtain the following bounds:

**Xpert:** \( P(z = 1|x = CT, w = 1, r = n) \in [0, 1] \cap \left[ \frac{0.181 - 0.849}{0.32}, \frac{0.181}{0.151} \right] = [0, 1] \) \[S13\]

**Microscopy:** \( P(z = 1|x = CT, w = 1, r = n) \in [0, 1] \cap \left[ \frac{0.181 - 0.791}{0.32}, \frac{0.181}{0.209} \right] = [0, 0.866] \) \[S14\]

A reasonable way to tighten the upper bound is to make the credible assumption that the probability of illness, conditional on clinic and HIV status, is lower conditional on a negative test result than conditional on a positive test result; that is, a test is informative conditional on clinic and HIV status:

\[ P(z = 1|x, w = 1, r = p) > P(z = 1|x, w = 1, r = n). \] \[S15\]

Hence, \( P(z = 1|x, w = 1, r = n) < P(z = 1|x, w = 1) \). Taking the upper bound of (S7'), this yields an upper bound on \( P(z = 1|x = CT, w = 1, r = n) \) of 0.566 for both Xpert and microscopy.

To tighten the lower bound, we can make the credible assumption that, conditional on clinic and a negative test result, the probability of illness is higher for patients who are HIV-positive than for those who are HIV-negative:

\[ P(z = 1|x, w = 1, r = n) > P(z = 1|x, w = 0, r = n). \] \[S16\]

This is reasonable because not only are HIV-positive patients more likely to have TB in general, but they are also more likely to have TB conditional on a negative result, given the lower negative predictive value of both Xpert and microscopy for HIV-positive compared to HIV-negative patients. Assumption (S16) applies that the lower bound of \( P(z = 1|x = CT, w = 1, r = n) \) is \( P(z = 1|x = CT, r = n) \) which from the above data is 0.036 for Xpert and 0.076 for microscopy. This leads us to the following bounds:

**Xpert:** \( P(z = 1|x = CT, w = 1, r = n) \in [0.036, 0.566] \). \[S13'\]

**Microscopy:** \( P(z = 1|x = CT, w = 1, r = n) \in [0.076, 0.566] \) \[S14'\]

These bounds could be tightened further by imposing additional assumptions. For example, one might impose bounded variation assumptions; see Manski (2018).

---

10 We continue to set \( P(z = 1|x = CT) = 0.181 \), i.e. the probability of illness in the full Cape Town sample. This is a more precise estimate of the population probability than the rate of illness observed in the Xpert and the microscopy sub-samples separately.