Tuberculosis diagnosis and treatment under uncertainty

Rachel Cassidy* and Charles F. Manski**

*Institute for Fiscal Studies, London, WC1E 7AE, United Kingdom; **Department of Economics, Northwestern University, Evanston, IL 60208; and ***Institute for Policy Research, Northwestern University, Evanston, IL 60208

Contributed by Charles F. Manski, October 2, 2019 (sent for review July 15, 2019; reviewed by Toru Kitagawa and Thomas Trikalinos)

In 2017, 1.6 million people worldwide died from tuberculosis (TB). A new TB diagnostic test—Xpert MTB/RIF from Cepheid—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.

Addressing the continuing prevalence of tuberculosis (TB) in several regions of the world remains a major priority for global health policymakers and practitioners. In 2017, 1.6 million people worldwide died from TB (1), and the “End TB” strategy of the World Health Organization (WHO) is committed to reduce TB deaths by 95% during the period 2015 to 2035 (2). A key challenge in fighting TB is that of improving capacity for rapid and accurate diagnosis. A new TB diagnostic test—Xpert MTB/RIF from Cepheid—was endorsed by the WHO in 2010 (3). 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 (4, 5). Xpert is also much faster at diagnosing multidrug-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 (6–12), although they were not powered to detect modest effect sizes.

Considering 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 contrasts with 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 (13–15).

The model highlights 2 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 (16, 17). 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 are then a problem of decision making under ambiguity (18), also known as deep uncertainty.

There may, for example, be ambiguity regarding the prevalence of TB in specific subpopulations 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 (see ref. 19 and accompanying rapid responses). Prevalence surveys, expert opinion, and modeling may provide credible bounds on prevalence for certain subpopulations 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 subpopulations 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 refs. 13–15 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, like the evidence produced by a randomized controlled trial.

The problem of TB 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 those 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.

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.

Basic Concepts and Notation. To model diagnosis and treatment decisions, we first specify a decision maker and the feasible actions. We use the concepts and notation of ref. 14, applying the abstract setup developed there to TB. As there, we consider a clinician who cares for the patients who present for examination. We consider these patients to be predetermined, and we assume that they comply with the clinical decisions.\(^1\) We also assume that treatment decisions for these patients do not affect disease transmission.\(^2\)

The clinician initially observes patient covariates that may include medical history, demographic attributes, measures of health status, and patient statements expressing their preferences for care and outcomes. The clinician can choose a treatment based on observing these covariates alone (empirical treatment) or order a test providing further evidence. In the latter case, the clinician chooses a treatment after observing the test result.

We denote observationally identical covariates and let \(t\) denote a treatment. We suppose that \(x\) is discrete and there are 2 feasible treatments, denoted A and B. These conditions can be weakened in principle. 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.\(^3\)

Let \(s\) indicate whether the clinician orders the test, with \(s = 1\) if she orders it and \(s = 0\) otherwise. Let \(r\) denote the test result. Suppose that \(r\) can take 2 values: \(p\), positive, indicating the patient has the condition, or \(n\), negative, indicating the patient does not have the condition.\(^4\)

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

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

To formalize this idea, let \(\delta_r(x)\) be the fraction of patients with covariates \(x\) who are tested and let \(1 - \delta_r(x)\) be the fraction who are not tested. The clinician can choose \(\delta_r(x)\) to be any fraction in the interval \([0, 1]\). This done, she tests a randomly drawn fraction \(\delta_r(x)\) of the patient group and does not test the remainder. Such 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 distinguish 3 groups of patients. Those who are not tested have observed covariates \(x\) when treated. Those who are tested have observed covariates \((x, r)\) when treated, with \(r\) equaling \(n\) or \(p\). Among patients who are not tested, let \(\delta_T(\theta)\) be the fraction with covariates \(x\) who receive treatment \(B\) and let \(1 - \delta_T(\theta)\) be the fraction who receive \(A\). Among those who are tested, let \(\delta_T(x, r)\) be the fraction with covariates \((x, r)\) receiving \(B\) and let \(1 - \delta_T(x, r)\) be the fraction receiving \(A\).

Welfare Function. We next specify a welfare function embodying the objective of the clinician. Rather than consider a patient in isolation, we will suppose that the objective is to optimize care on average across the patients in her practice. This sense of optimization does not require certainty about what treatment is best for each patient. It only requires knowing mean treatment response for patients with the same observed covariates.

Discussions of patient care often suppose optimization for each patient separately, without reference to care of other patients. However, a clinician can optimize care for a single patient only if she knows enough to be certain what treatment is best for this patient. Clinicians typically lack this knowledge, particularly when deciding whether to order a test. After all, the medical purpose of testing is to yield evidence on health status that may be helpful when choosing a treatment. If the clinician were already to know the best treatment, there would be no medical reason to perform a test.

It remains to specify the welfare function. As in ref. 14, we assume that the clinician aggregates the benefits and harms of making a specific testing and treatment decision into a scalar measure of welfare. Going beyond the abstract notation of ref. 14, 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 \(U(\zeta, s, t)\) summarize the clinician’s 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,

---

\(^{\text{a}}\) 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 and 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.

\(^{\text{b}}\) In the case of TB treatment, arguably the largest source of patient noncompliance arises from patients not completing the course of antibiotics. It is not clear what effect, if any, improved diagnostic tests will have on compliance.

\(^{\text{c}}\) 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 later.

\(^{\text{d}}\) For simplicity we consider just 2 treatments, antibiotics versus observation only, and 2 illness states, TB versus no TB. The model can be extended to include testing for and treating MDR TB alongside regular TB, as well as testing for and treating HIV alongside TB. These extensions may be accomplished with further notation.
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. Given that testing for TB is noninvasive and does not affect treatment outcomes directly, welfare may have the additive form $U(z, s, t) = u(z, t) - K$, where $u(z, t)$ measures life-cycle expected utility for a patient with illness state $z$ and treatment $t$ (including direct utility from health, financial benefits from being fit to work, and so on), while $K$ measures the financial cost of testing.

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

For each value of $(s, t)$, let $E[U(z, s, t)|x]$ be the mean expected welfare that results if all patients with $x$ receive $(s, t)$. Let $E[U(z, s, t)|x, r]$ be the mean expected welfare that results if all those with covariate value $x$ and test result $r$ receive $(s, t)$. We assume that, for each value of $(s, t)$, the potential welfare function $U(z, s, t)$ varies across patients in a way that is mean independent of their illness outcomes $z$, conditional on $x$ and on $(s, t)$; that is, $E[U(z, s, t)|x, r] = E(U(0, s, t)|x)$. It then follows from the law of iterated expectations that

$$E[U(z, s, t)|x] = P(z = 0|x)E[U(0, s, t)|x] + P(z = 1|x)E[U(1, s, t)|x],$$

and $E[U(z, s, t)|x, r] = P(z = 0|x, r)E[U(0, s, t)|x, r] + P(z = 1|x, r)E[U(1, s, t)|x, r]$. Let $\delta = [\delta_0(x), \delta_1(x), \delta_2(x, t), r = \{p, n\}]$ denote a specified testing–treatment allocation. The mean welfare $W(\delta)$ that results if the clinician chooses $\delta$ 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. Hence,

$$W(\delta) = \sum_{x \in X} P(x)[1-\delta_0(x)]E[U(z, 0, A)|x] + [1-\delta_0(x)]\delta_0(x)E[U(z, 0, B)|x] + \sum_{r \in \{p, n\}} f(r|x)\delta_2(x, |1-\delta_1(t, x, r)|E[U(z, 1, A)|x, r] + \delta_2(x, \delta_1(x, r)E[U(z, 1, B)|x, r]|).$$

**Optimal Testing and Treatment.** An optimal testing and treatment allocation maximizes $W$; ref. 14, shows that an optimal allocation is $\delta_0(x) = 1$ if $\sum_{r \in \{p, n\}} f(r|x)\max\{E[U(z, 1, A)|x, r], E[U(z, 1, B)|x, r]| \geq \max\{E[U(z, 0, A)|x], E[U(z, 0, B)|x]\}, = 0$ otherwise. $\delta_1(x, p) = 1$ if $E[U(z, 1, B)|x, p] \geq E[U(z, 1, A)|x, p]$, = 0 otherwise.

$\delta_1(x, n) = 1$ if $\sum_{r \in \{p, n\}} f(r|x)\max\{E[U(z, 1, A)|x, r], E[U(z, 0, A)|x]\] \geq E[U(z, 0, B)|x], = 0$ otherwise. $\delta_1(x, p) = 1$ if $E[U(z, 1, B)|x, p] \geq E[U(z, 1, A)|x, p]$, = 0 otherwise.

Each maximum is unique when the 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 ref. 14 extends immediately to settings with more than 2 possible test results. To perform the extension, one simply sums over the feasible results in $[3a]$ and extends $[3c]$ and $[3d]$ 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.

**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:

$$E[U(z, s, t)|x] = (1 - P_{x1})U_0(0, s, t) + P_{x1}U_1(1, s, t),$$

$$E[U(z, s, t)|x, r] = (1 - P_{xr})U_{x0}(0, s, t) + P_{xr}U_{x1}(1, s, t).$$

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, $U_0(0, s, t) = U_0(0, s, t)$ and $U_1(1, s, t) = U_1(1, s, t)$. With this assumption, knowledge of the test result affects decision making purely by changing risk assessment from $P_{x1}$ to $P_{x1}$, 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 $[3b]$ through $[3d]$ are as follows:

**treatment with no test result:**
choose $B$ if $(1 - P_{p1})U_0(0, 0, B) + P_{p1}U_1(1, 0, B) \geq (1 - P_{n1})U_0(0, 0, A) + P_{n1}U_1(1, 0, A),$
choose $A$ otherwise.

**treatment with positive test result:**
choose $B$ if $(1 - P_{p1})U_0(0, 1, B) + P_{p1}U_1(1, 1, B) \geq (1 - P_{n1})U_0(0, 1, A) + P_{n1}U_1(1, 1, A),$
choose $A$ otherwise.

**treatment with negative test result:**
choose $B$ if $(1 - P_{p1})U_0(1, 0, B) + P_{p1}U_1(1, 1, B) \geq (1 - P_{n1})U_0(1, 0, A) + P_{n1}U_1(1, 1, A),$
choose $A$ otherwise.

In the medical literature on diagnostic testing, $P_{x}$ is called the base rate or the prevalence of the illness for patients with covariates $x$. $P_{xp}$ is called the positive predictive value of a test and $1 - P_{xn}$ is
called the 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 sensitivity and specificity of a test. Sensitivity is the probability that the test result is positive conditional on the patient’s being ill, \( P(t = p|x, z = 1) \). Specificity is the probability that the result is negative conditional on the patient being healthy, \( P(t = n|x, z = 0) \).

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 (ref. 20, p. 102) wrote:

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.

Despite the cautions expressed in articles such as ref. 20, it has remained common to measure the accuracy of diagnostic tests by their sensitivity and specificity, without providing the value of the prevalence required to derive predictive values. We discuss the potential implications of this later.

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,

1) Health is better than illness: \( U_x(0, s, t) > U_x(1, s, t) \) for all \((s, t)\).
2) Testing is costly/harmful: \( U_x(z, 0, t) > U_x(z, 1, t) \) for all \((z, t)\).
3) Surveillance is better than aggressive treatment when healthy: \( U_x(0, s, A) > U_x(0, s, B) \) for all \(s\).
4) Aggressive treatment is better than surveillance when ill: \( U_x(1, s, B) > U_x(1, s, A) \) for all \(s\).

These assumptions are realistic in the TB context: 1) a patient is better off not having TB than having TB; 2) performance of a microscopy or Xpert test does not harm patients but does incur financial costs; 3) when a patient is healthy, there is no benefit from prescription of antibiotics, but there are financial costs and possible harms to patients; and 4) when a patient is ill with TB, the health benefits of antibiotic treatment exceed the financial costs and harms.

Analysis in ref. 15 shows that under assumptions 3 and 4 the treatment criteria in \([3b']\) through \([3d']\) yield simple solutions. Aggressive treatment is the optimal decision if the risk of illness equals or exceeds a threshold that equals 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

\[
P_{x,0} \equiv \frac{U_x(0,0,A) - U_x(0,0,B)}{[U_x(0,0,A) - U_x(0,0,B)] + [U_x(1,0,B) - U_x(1,0,A)]}
\]

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

\[
P_{x,1} \equiv \frac{U_x(0,1,A) - U_x(0,1,B)}{[U_x(0,1,A) - U_x(0,1,B)] + [U_x(1,1,B) - U_x(1,1,A)]}
\]

Under assumptions 3 and 4 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 \( P_x < P_{x,0} \) and type II errors do not occur when \( P_x > P_{x,0} \). Type II errors occur with probability \( P_x \) when \( P_x < P_{x,0} \) and type I errors occur with probability \( 1 - P_x \) when \( P_x > P_{x,0} \). Analogous results hold with testing.

**How Testing Affects Treatment.** We observed earlier 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 the condition \( P_x > P_{x,0} \). We now ask how, if at all, testing affects treatment.

In general, the answer to this question is complex because the thresholds \( P_{x,0} \) and \( P_{x,1} \) in \([4a]\) 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, \( U_x(z, 0, t) - U_x(z, 1, t) = K > 0 \) 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 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,

\[
P_x = P_{x,0}(1 - f_t) + P_{x,1}f_t,
\]

where \( f_t \equiv f(t = p|x) \) 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 \([5]\) 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 \([4a]\).

Testing does not affect treatment otherwise. If \( P_{xp} < P_{x,0} \), treatment A is optimal with or without testing. If \( P_{xn} > P_{x,0} \), treatment B is optimal with or without testing.

**Testing and Treatment under Ambiguity**

**Sources of Ambiguity and Standard Decision Criteria.** Within the model of optimal diagnosis and treatment, optimization for patients with covariates \( x \) is feasible if one knows the mean welfare function \( U_x(\cdot, \cdot) \), the illness probabilities (\( P_x, P_{xn}, P_{xp} \)), and the probability \( f_t \) 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 $(U_{tx}, P_{tx}, f_{tx}, P_{txp}, P_{txh})$, $x \in X$, $\gamma \in \Gamma$ that satisfy [5] for each value of $x$ and that are deemed feasible based on available evidence and maintained assumptions.

To consider decision making under ambiguity (14), begins with the welfare function of [2], considered as a function over the state space. For each $\gamma \in \Gamma$,

$$W_\gamma(\delta) = \sum_{x \in X} \mathbb{P}(x) \left( [1 - \delta_S(x)][1 - \delta_{T_0}(x)]E_{1}(U(x, 0, A)) | x \right)
+ [1 - \delta_S(x)][1 - \delta_{T_0}(x)]E_{2}(U(x, 0, B)) | x \right)
+ \sum_{r \in \{0, 1\}} f_r(\delta(x)) \left[ [1 - \delta_{T_1}(x, r)]E_{1}(U(x, 1, A)) | x, r \right]
+ \delta_S(x)\delta_{T_1}(x, r)E_{2}(U(x, 1, B)) | x, r \right),$$

where

$$E_{1}(U(x, 0, A)) | x = (1 - P_{tx})U_{tx}(0, 0, t) + P_{tx}U_{tx}(1, 0, t), \quad [7a]$$
$$E_{2}(U(x, 1, t)) | x = (1 - P_{tx})U_{tx}(0, 1, t) + P_{tx}U_{tx}(1, 1, t). \quad [7b]$$

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.

**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 4 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 2 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 other clinician 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 ref. 13. The extension is especially simple if we suppose that $U(x, \cdot, \cdot)$ is known; hence, the threshold risk assessment $P^*_x$ is known. Considerable scope for ambiguity remains through incomplete knowledge of $(P_{tx}, P_{txp}, P_{txh})$. We proceed abstractly here and characterize the ambiguity in the TB context later.

The minimax-regret analysis in ref. 13 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 nontechnical 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 $P_{tx}$ be denoted $P_{txl}$ and $P_{txh}$, respectively. We later discuss how these lower and upper bounds may be generated in practice.

Ambiguity occurs when $P_{txl} < P^*_x < P_{txh}$. Let $M_{tx}(B)$ be the maximum value of the average treatment effect $E_{1}(U(z, 1, B) | x, r) - E_{1}(U(z, 1, A) | x, r)$ across the state space. With ambiguity, the maximum is positive and occurs when $P_{tx} = P_{txh}$. That is, at the maximum probability of being ill conditional on $x$ and $r$. Analogously, let $M_{tx}(A)$ be the maximum value of the reverse average treatment effect $E_{2}(U(z, 1, A) | x) - E_{2}(U(z, 1, B) | x) | r$, which occurs when $P_{tx} = P_{txl}$. Thus,

$$M_{tx}(B) = (1 - P_{txh})U_{tx}(0, 1, B) + P_{txh}U_{tx}(1, 1, B), \quad [8a]$$
$$M_{tx}(A) = (1 - P_{txh})U_{tx}(0, 1, A) + P_{txh}U_{tx}(1, 1, A). \quad [8b]$$

The analysis in ref. 13 shows that, among patients with test result $r$, the minimax-regret criterion yields

$$\delta_{T_1}(x, r) = 0 \text{ if } P_{txh} \leq P^*_x,$n$$
$$= M_{tx}(B) / [M_{tx}(B) + M_{tx}(A)] \text{ if } P_{txl} < P^*_x < P_{txh}, \quad [9]$$
$$= 1 \text{ if } P_{txl} \geq P^*_x.$$n

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 $P_{tx}$. With this modification, the above result in [8a] through [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 SI Appendix for the derivation.

Applying again the analysis in ref. 13, the fraction of patients allocated to testing is similar to the result in [9]. The fraction is 0 if the maximum “average treatment effect” of testing vs. not testing (i.e., the maximum regret from not testing) is less than 0. It is 1 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.

**Adaptive Diversification.** The above provides a full description of static piecemeal decision making. Finally, as in ref. 13, 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 they are ill. Then, complete learning eventually occurs if $\delta_S(x) > 0$ for some cohort. Randomized testing of patients with covariates $x$ reveals $f_x$, and randomized treatment following testing reveals $P_{tx}$ and $P_{txh}$. $P_x$ is revealed directly if $\delta_S(x) < 1$ and indirectly by [5] if $\delta_S(x) = 1$.

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 $\delta_S(x) = 1$, that is, 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 \( d_\delta(x) = 0 \) for all cohorts. In this case, randomized treatment reveals \( P_\delta \). This yields only partial knowledge of \((f, P_\delta, P_n)\), which must satisfy [5]. To avoid this outcome, which is generally undesirable from a multicohort perspective, one might set \( d_\delta(x) > 0 \) for some cohort.

**Implications for TB Testing and Treatment**

**Optimal TB Testing and Treatment.** The model is useful for studying TB, first 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 (4, 8).

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, considering 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 (21). 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.4 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 (6, 8, 11, 12). 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 overtreating 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. However, 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.

**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 spillover 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 nonnegligible.

Whether the spillover effect is nonnegligible may depend, among other things, on the prevalence of TB and of risk factors such as HIV/AIDS, the infectivity of the specific strain of TB, and the nature of social networks. A policymaker may take spillovers into account in setting clinical guidelines, even if each individual treatment decision has a negligible effect on transmission. We cannot calculate on the magnitude of the effect, but we think it reasonable to expect that concern for spillovers would increase the public health incentive to treat TB actively. This is a key topic for future research.

When performing future research on spillovers, researchers will have to confront forms of ambiguity beyond those considered in this paper. There exists a large public health literature on optimal treatment of infectious diseases under the assumption that the mechanism of disease transmission is completely known. However, this assumption often is not credible. A vexing problem impeding progress has been the infeasibility of using large-scale randomized trials to learn how transmission varies with treatment policies. Hence, epidemiologists have had to rely on fitting mathematical models to available observational data. In the case of TB, there is an ongoing debate over the extent to which new cases of active TB are caused by recent infection or by progression of latent TB (see ref. 19 and accompanying rapid responses).

We are aware of only 2 studies that address formation of public health policy for treatment of infectious disease with partial knowledge of disease transmission. These studies (22, 23) consider reasonable choice of vaccination policy by a social planner, with specific attention to maximin and minimax-regret policies. Vaccination is somewhat distant from TB testing and treatment. Nevertheless, the general methodology used in these studies—specification of a social welfare function and a set of policy options, characterization of the available knowledge, and derivation of policies that have desirable properties—should provide some general guidance 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’s substantially increasing the probability of patient presentation within certain patient populations.

Yet another simplification of the model is its assumption that available evidence yields known bounds on the various probabilities needed to optimize decision making. In practice, the sources of ambiguity that have been discussed above are further exacerbated by the ordinary sampling imprecision that occurs when one uses finite-sample data to draw inferences. In principle, statistical decision theory as envisioned by ref. 24 provides a coherent framework for public health planning with sampling imprecision. This theory has been applied to study minimax-regret treatment choice with sample data from randomized trials.
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 $P(z|w)$ 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 (1). 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 typically accounted for by expert opinion, or a constant adjustment factor, rather than data or modeling of the patient presentation decision (1, 27).

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,61.2% publish only report $P(z|w)$ or even $P(z|w, r)$, rather than $P(r|z)$ 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.

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. SI Appendix details these exercises in full.

Of the published efficacy and effectiveness trials for Xpert, ref. 8 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 SI Appendix we calculate the positive and negative predictive values for Xpert and microscopy using the data in ref. 8. 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 larger 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, in SI Appendix 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 ref. 8 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, $P(z = 1|x = CT, w = 1) = [0.181, 0.566]$ 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.

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. The piecewise minimax-regret procedure studied in the paper offers a specific practical way to implement diversification. A public health agency may want to consider this procedure, or another procedure with desirable properties.

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. As in 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 randomized 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 like 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.

Data Availability. The research performed in this article involved no collection of new data, nor did we perform any secondary analysis of previously collected data. The article only performs illustrative numerical exercises that directly use empirical findings reported in ref. 8, which is a published article. The details of the numerical exercises are provided in SI Appendix.

ACKNOWLEDGMENTS. R.C.’s work on this project was performed while she was a Postdoctoral Fellow at the Institute for Fiscal Studies (IFS) and was funded by the Economic and Social Research Council Centre for the Microeconomic Analysis of Public Policy at the IFS. We thank Kalipso Chalkidou, Michael Gmeiner, 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 University for valuable comments.