Machine predictions and human decisions with variation in payoffs and skill: the case of antibiotic prescribing*

Michael Allan Ribers† Hannes Ullrich‡

October 2023

Abstract

We analyze how machine learning predictions may improve antibiotic prescribing in the context of the global health policy challenge of increasing antibiotic resistance. Estimating a binary antibiotic treatment choice model, we find variation in the skill to diagnose bacterial urinary tract infections and in how general practitioners trade off the expected cost of resistance against antibiotic curative benefits. In counterfactual analyses we find that providing machine learning predictions of bacterial infections to physicians increases prescribing efficiency. However, to achieve the policy objective of reducing antibiotic prescribing, physicians must also be incentivized. Our results highlight the potential misalignment of social and heterogeneous individual objectives in utilizing machine learning for prediction policy problems.

*We benefited from helpful comments by Jérôme Adda, David Chan, Tomaso Duso, Daniel Ershov, Mogens Fosgerau, Matthew Gentzkow, Qing Gong, Shan Huang, Paul Heidhues, Günter Hitsch, Yufeng Huang, Ulrich Kaiser, Jonathan Kolstad, Chuck Manski, Jeanine Miklós-Thal, Ziad Obermeyer, Yeşim Orhun, Imke Reimers, Bertel Schjerning, Stephan Seiler, Michelle Sovinsky, Jann Spiess, Florian Szücs, Christoph Wolf, and participants at the 13th TSE Digital Economics Conference, IC^{2}S^2 2020, ESWC 2020, ASHEcon 2020, the German IO Committee Meeting, the Economics of Antibiotics Workshop at TSE, the EHEC seminar, the European Quant Marketing Seminar, the Economics+AI workshop at ETH Zurich, the 1st CEPR Health Economics conference at TSE, the American-European Health Economics workshop at Harvard Kennedy School, as well as in seminars in Berlin, Copenhagen, Essen, Helsinki, Konstanz, Mannheim, and Paris. We are indebted to Lars Bjerrum for providing expertise on UTI treatment in Denmark and to Jenny Dahl Knudsen, Sidsel Kyst, and Rolf Magnus Arpi at Herlev and Hvidovre hospitals for enabling the sharing of the laboratory data. We thank Adam Lederer for proofreading. Financial support from the European Research Council (ERC) under the EU Horizon 2020 research and innovation programme (grant agreement no. 802450) is gratefully acknowledged.

†Department of Economics, University of Copenhagen. michael.ribers@econ.ku.dk

‡Department Firms and Markets, DIW Berlin; Department of Economics, University of Copenhagen; and Berlin School of Economics. hullrich@diw.de.

1 Introduction

Antibiotic resistant infections are among the leading causes of death worldwide. Every year, more people die due to antibiotic resistance than due to either breast cancer, HIV, malaria, or opioid overdose.[1] In 2019, an estimated 4.95 million deaths have been associated with antibiotic resistance, with an estimated 1.27 million directly attributable deaths (Laxminarayan 2022; Murray et al. 2022).[2]

Antibiotics are vital pharmaceuticals for treating bacterial infections but their use is also considered the main driver of antibiotic resistance (Costelloe et al. 2010; WHO 2014; Adda 2020). Hence, treatment decisions must solve a trade off between patients' expected sickness cost under diagnostic uncertainty and the external cost of increased antibiotic resistance.[3] Policies have focused mostly on affecting how physicians weigh the external cost, for example by making their prescribing intensities salient (Hallsworth et al. 2016). When physicians make decisions under uncertainty, diagnostic skill becomes an important determinant of treatment outcomes and policies focusing only on incentives may lead to large inefficiencies (Mullainathan and Obermeyer 2022). To design policy interventions, it is crucial to quantify to what extent antibiotic misuse is driven by diagnostic uncertainty as opposed to physicians' socially inefficient trade-offs.

In this paper, we view the challenge of reducing antibiotic use through the lens of a prediction policy problem, as in Kleinberg et al. (2015, 2018), where large-scale data and machine learning may help reduce patient-level diagnostic uncertainty and improve treatment outcomes. In high-risk situations, like medical decision-making, evaluating the potential effects of interventions *ex ante* is important. Yet, such an evaluation is difficult when human agents' information and preferences are unknown and heterogeneous. Prior evaluations have relied on quasi-experimental designs with crucial monotonicity assumptions implying homogeneous skill or preferences (Currie and MacLeod 2017; Kleinberg et al. 2018). We propose a structural model of physician prescribing decisions that accommodates flexible heterogeneity in physicians' diagnostic information and payoffs, thus providing a framework for counterfactual policy evaluation using machine learning predictions.

Specifically, we study antibiotic treatment decisions for 36,972 initial urinary tract infection (UTI) consultations in 175 primary care clinics in Denmark. UTI are one of the most common

[1] See Murray et al. (2022) for estimates on the toll of antibiotic resistance and WHO Malaria Fact Sheet (https://www.who.int/news-room/fact-sheets/detail/malaria), WHO Breast Cancer Fact Sheet (https://www.who.int/news-room/fact-sheets/detail/breast-cancer), and WHO Opioid Overdose Fact Sheet (https://www.who.int/news-room/fact-sheets/detail/opioid-overdose), accessed on 30 March 2022.

[2] In the US, 2.8 million antibiotic-resistant infections result in 35,000 deaths, $20 billion in health care costs, and $35 billion in lost productivity per year (CDC 2013, 2019; Kwon and Powderly 2021).

[3] Improved diagnostics are a key factor in improving antibiotic use but investment in diagnostic technologies is lacking, see WHO Antimicrobial Resistance Fact Sheet (accessed on 30 March 2022).

classes of bacterial infections, and primary care accounts for 75 percent of prescriptions in Denmark (Danish Ministry of Health 2017).[4] Due to the acute nature of UTI, its symptomatic burden and risk of complications, immediate antibiotic therapy is recommended.[5] Therefore, physicians make treatment decisions prior to observing definitive test results, which become available after several days.[6]

This delay poses a challenge for physicians but provides us with a microscope for measuring heterogeneous physician skill and payoffs to perform counterfactual evaluations.[7] Ribers and Ullrich (2023) have analyzed the potential of machine learning in this setting following simple threshold-based policies without uncovering the mechanisms of decision improvements. We make use of this unique setting to quantify physician error conditional on testing, its variation across clinics, and how a non-targeted policy using machine learning predictions affects prescription outcomes. Combining diagnostic outcomes from microbiological laboratories with administrative data on individual patients, we see physicians prescribe an antibiotic in 39 percent of initial consultations, which corresponds closely to the mean rate of bacterial test results. However, bacterial test results and antibiotic prescribing decisions do not match for a large share of patients, with significant variation across physicians.

To analyze physician decisions, we first gather observable information at the time of initial consultation to predict the binary test outcome, whether significant bacteria are present or not, using machine learning following Ribers and Ullrich (2023). We then incorporate patient-specific machine learning predictions in a binary choice model proposed by Chan et al. (2022), which follows two steps physicians take in treating patients. First, assessing the risk of a bacterial cause of infection depends on diagnostic skill. Physicians observe patient characteristics and medical histories, amenable to machine learning methods, which they can relate to the prevalence of bacterial UTI. Physicians also receive a signal from clinical assessment including patients' symptom descriptions and point-of-care tests which we do not observe. Physicians may use

[4]Foxman (2002) reports 50 percent of women contract a UTI at least once in their lifetime. In the US, yearly UTI-related health care costs, including workplace absences, are estimated at $3.5 billion (Flores-Mireles et al. 2015). In Denmark, 10 percent of all women have received antibiotic treatment for UTI (Bjerrum and Lindbæk 2015). In Europe, primary care accounts for 90 percent of all antibiotic prescriptions (Llor and Bjerrum 2014).

[5]For UTI, patients often seek medical attention when symptoms are already advanced, increasing the urgency to treat. The estimated short term cost of delaying treatment are six symptomatic days, including 2.4 days of restricted activity (Foxman 2002). In 76 percent of community-acquired UTI patients, symptoms persist without treatment (Ferry et al. 2004). Without treatment, natural progression of an infection can often lead to hospitalization. In an evidence review, Grigoryan et al. (2014) conclude that immediate antimicrobial therapy is recommended for bacterial UTI.

[6]In many common situations in health care, diagnostic results are delayed or unavailable but delaying treatment decisions carries important costs (Cassidy and Manski 2019; Manski 2021); for example, in biopsies for malignant tumors, testing for tuberculosis, or testing for SARS-CoV-2 virus. In these situations, machine learning may provide opportunities for earlier access to valuable diagnostic information.

[7]This feature has been used in the medical literature and related work in Yelin et al. (2019), Kanjilal et al. (2020), Huang et al. (2022), and Ribers and Ullrich (2023).

both sources of diagnostic information to form beliefs about a patient's sickness state. Second, given their assessment, physicians decide whether to prescribe an antibiotic by weighing patients' expected sickness cost while waiting for diagnostic certainty against the cost of increased antibiotic resistance.

Estimating the model, we find significant heterogeneity in skill and payoffs across clinics. Notably, we find clinical diagnostic skill is negatively correlated with physician age and positively associated with the extent of point-of-care testing. In counterfactual policy evaluations, we compare outcomes of a simple threshold-based decision rule with outcomes generated by the model of payoff-maximizing physicians. Improvements induced by the threshold-based decision rule, a reduction of 8.9 percent in prescribing and 22.7 percent in overprescribing, are not only due to diagnostic information generated using machine learning. They are also driven by imposing payoff weights which differ from estimated physician preferences. To achieve reductions in antibiotic prescribing, physicians need to be incentivized in addition to receiving improved diagnostic information.

The best policy for a social planner depends on her weight on the antibiotic resistance externality. If the social planner's weight on the externality exceeds the mean physicians' weight, incentivizing physicians to reduce prescribing is necessary to maximize social welfare increases. Yet, incentivizing physicians without improving diagnostic information leads to undertreatment and potential welfare losses. We conclude that the improvement of diagnostic information and incentives to reduce prescribing should be seen as complementary policy tools.

The remainder of the paper is organized as follows. Section 2 places our contribution in the existing literature. Section 3 presents the institutional background and data. Section 4 develops the model of physician prescription choice and Section 5 describes the empirical analysis. Section 6 presents the estimation results, Section 7 describes counterfactual policy evaluations, and Section 8 shows robustness checks. Section 9 concludes.

2 Prior literature

We bridge two literatures by focusing on physician heterogeneity in assessing the potential of machine learning predictions to help reduce antibiotic prescribing. One investigates *ex ante* the potential of machine learning to help achieve varying policy objectives. The other identifies heterogeneity in skill and preferences as sources of observed variation in health care provision.

A growing literature has evaluated prediction policy problems by replacing observed human decisions with threshold rules based on prediction-based rankings, assuming decision makers follow these perfectly (Kang et al. 2013; Bayati et al. 2014; Kleinberg et al. 2015; Chalfin et al.

2016; Kleinberg et al. 2018; Andini et al. 2018; Yelin et al. 2019; Hastings et al. 2020; Dobbie et al. 2021; Mullainathan and Obermeyer 2022). The literature identifying supplier-driven variation in health care provision as well as potential drivers of this variation is also large (Chandra and Staiger 2007; Epstein and Nicholson 2009; Chandra et al. 2011; Skinner 2011; Finkelstein et al. 2016; Currie et al. 2016; Abaluck et al. 2016; Currie and MacLeod 2017; Abaluck et al. 2020). In both these literatures, empirical designs often rely on assuming homogeneity in either skill or preferences. Chan et al. (2022) discuss the use of restrictive homogeneity assumptions and propose a framework to identify both skill and preferences as drivers of practice variation.

Machine learning predictions may improve decisions by augmenting or replacing skill, while facilitating decision rules that may not be aligned with individual decision makers' objectives (Agrawal et al. 2018; Cowgill and Stevenson 2020). Hence, an evaluation of prediction-based policies needs to take skill and preferences as determinants of decisions into account. Huang and Ullrich (2023) and Ribers and Ullrich (2023) find indicative, model-free evidence that variation in antibiotic treatment quality may be related to variation in diagnostic information. In high risk settings such as health care, experimental work on the interplay of machine learning predictions and human skill is difficult and rare, for example Agarwal et al. (2023). To evaluate the potential improvements of prediction-based policies over human decisions in the treatment of UTI *ex ante* using counterfactual policy evaluation, we estimate a structural model of treatment decisions to measure variation in payoffs and two-dimensional skill, separated into observable and unobservable information, across physicians.[8]

Finally, we contribute to the literature exploring demand side policies aimed at curbing antibiotic resistance. This literature includes Laxminarayan et al. (2013) on prescription surveillance and stewardship programs, Bennett et al. (2015) on general practitioner competition in Taiwan, Currie et al. (2014) and Das et al. (2016) on financial incentives for physicians in China and India, Kwon and Jun (2015) on peer effects in Korea, Hallsworth et al. (2016) on communication of social norms in the UK, McAdams et al. (2019) and McAdams (2021) on optimal targeting of antibiotics, and Dubois and Gokkoca (2023) on the relevance of antibiotic resistance information for the choice of antibiotic.

[8]Rambachan (2022) proposes a non-parametric framework to identify human prediction quality from observational data. Our parametric model provides a tractable way to separate the two dimensions of skill at the clinic-level, which is essential for counterfactual evaluations.

3 Health care context and data

3.1 General practice in Denmark

Denmark has several regulations that impact decision making in primary care. General practitioners act as the primary gatekeepers in a universal and tax financed single payer health care system. Every person living in Denmark is assigned to a general practitioner, who needs to be consulted for any primary care needs, by a list-system within a fixed geographic radius around the home address. Hence, short-term 'shopping' for antibiotic prescriptions is generally not possible. General practitioners work as privately owned businesses but all service fees are collectively negotiated and fixed between the national union of general practitioners and the public health authority. The majority of clinics in general practice are single-physician establishments.

Physicians do not generate earnings by prescribing drugs to patients who have to purchase their prescriptions from local pharmacies. General practitioners are responsible for prescribing approximately 75 percent of the human consumed systemic antibiotics in Denmark (Danish Ministry of Health 2017). Pharmacies earn a fixed fee per processed prescription regardless of price or other drug attributes, for example branded versus generic drugs. Prescription drugs are subsidized but patients co-pay a fraction of the list price. The Danish market for prescription drugs is highly regulated resulting in low and uniform prices for antibiotics nationwide, about 100 Danish Kroner (15 US Dollars) per complete treatment.

3.2 Diagnosis and treatment of UTI

UTI are one of the most common types of bacterial infections and, hence, a leading cause for antibiotic use in primary care (Grigoryan et al. 2014; Gupta et al. 2017). UTIs occur when bacteria infect the urinary tract, the bladder, or kidneys. Without treatment, they can lead to debilitating symptoms, at the extreme including sepsis and death. In the US alone, The healthcare system bears an annual burden of approximately $1.6-3.5 billion due to community-acquired UTI. (Foxman 2002; Flores-Mireles et al. 2015). Once a diagnosis is established, clinical guidelines recommend the use of antibiotics.[9]

Foxman (2002) documents that almost half of all women contract at least one UTI during their lifetime. Numerous other groups in the population face an elevated UTI risk, including children, the elderly, and those with medical conditions like diabetes, weakened immune systems, or underlying urological abnormalities (Foxman 2002). Many of these groups can be identified

[9]See *Medicinrådets behandlingsvejledning vedrørende urinvejsinfektioner* (https://medicinraadet.dk/anbefalinger-og-vejledninger/behandlingsvejledninger/urinvejsinfektioner-uvi) or *Urinary Tract Infections* (https://www.mayoclinic.org/diseases-conditions/urinary-tract-infection/symptoms-causes/syc-20353447) by the Cleveland Clinic, accessed 11/2/2022.

in observable data by personal attributes such as age and gender, or by observing past diagnoses and health care utilization.

Medical attention is required for UTI symptoms, which encompass discomfort, pain, urinary frequency, urgency, and new-onset incontinence. Signs of a systemic infection, including fever, shivering, or overall illness, may also occur. Attributing these symptoms to a UTI can be challenging, as they can also be linked to other conditions such as sexually transmitted urethritis or vaginitis, early pyelonephritis, noninfectious urethritis, overactive bladder, bladder or kidney stones, benign prostatic hyperplasia, or a bladder tumor (Wilson and Gaido 2004; Gupta et al. 2017; Nik-Ahd et al. 2018; Holm et al. 2021). Sometimes, fungi or viruses can also cause UTI. Importantly, encoding these symptoms systematically is difficult. For instance, assessing "pain" necessitates contextual elicitation and judgement regarding its nature, severity, chronology, and location. Beyond symptom assessment, by speaking to patients physicians may obtain contextual information, including behavioral factors.

At a consultation, diagnostic information can be obtained by point-of-care testing, first and foremost urinary dipstick and microscopic analysis. These diagnostics may exhibit very low specificity, the true negative rate, and sensitivity, the true positive rate, well below 0.5 (Devillé et al. 2004; Wilson and Gaido 2004; Chu and Lowder 2018). To obtain a reliable assessment of the true infection state, urine samples can be analyzed at a specialized hospital laboratory. These laboratories offer the gold standard diagnosis for UTI with high accuracy and minimal reliance on human judgment but require about three days to obtain results (Schmiemann et al. 2010). This test can be used to confirm treatment decisions *ex post*, ensure full information is available to administer optimal treatment later, or provide antibiotic resistance information.

3.3 Microbiological laboratory test data

Individual-level clinical microbiological laboratory test results comprise the central data set of our analysis. We acquired clinical microbiological laboratory test results from Herlev hospital and Hvidovre hospital, two major hospitals in Denmark's capital region covering a catchment area of roughly 1.7 million inhabitants, nearly one third of the Danish population, for the period of January 2010 to December 2012. The laboratory data provide information on whether significant bacteria are found, the bacterial species, and an antibiotic resistance profile when bacteria are detected in a patient sample. In addition, patient and clinic identifiers as well as information on the microbiological sample type, the test acquisition date, sample arrival date at the laboratory, and test response date are provided. A total of 2,579,617 microbiological samples are observed including samples sent in from general practitioners and hospitals.

3.4 Administrative data

The administrative data provided by Statistics Denmark cover the entire population of Denmark between January 1, 2002, and December 31, 2012. For each person, we observe a comprehensive set of socioeconomic and demographic variables, the complete prescription history of systemic antibiotics (*Lægemiddeldatabasen*), hospitalizations (*Landspatientregisteret*), and general practitioner insurance claims (*Sygesikringsregisteret*).[10] All population-level administrative data can be linked using patients' personal identifiers and general practitioners' clinic license numbers. Household member identifiers allow us to also link administrative and laboratory data of patients' family members.

The demographic data include gender, age, education, occupation, income, marriage and family status, home municipality, immigration status, and place of origin. The data on systemic antibiotic prescriptions contain the date of purchase, patient and prescribing clinic identifiers, anatomical therapeutic chemical drug classification, drug name, price, indication of use, purchased package size, and defined daily dose.[11] The hospitalization data comprise all patient contacts with hospitals including admission and discharge dates, procedures performed, type of hospitalization, primary and secondary diagnoses, and the number of total bed days. The insurance claims data cover all general practitioner clinic services provided to the Danish population of patients, including clinic and patient identifiers, the week of consultation, and services used.

4 Model of physicians' antibiotic treatment decision

We propose a framework that combines machine learning predictions with a model of general practitioner treatment choice that allows for flexible heterogeneity in physicians' payoff functions and skill levels. The model follows Chan et al. (2022) by separating an individual physician's treatment choice problem from the preceding step of forming diagnostic predictions. We depart from their model by introducing heterogeneous patient types, making explicit that sampled patients may vary in their likelihood of being sick conditional on observable characteristics. Physicians can in principle observe such characteristics prior to patients' sickness realisations and use these to inform their prescription decisions, but the extent to which they do is a priori unknown.

Physician skill manifests in two dimensions. Diagnosis based on observable patient characteristics, i.e. information on a patient's type, and diagnosis based on clinical examination,

[10] See Statistics Denmark (2012b,d,e,a,f,g,c,h,i) and The Danish Health Data Authority (2012a,b).

[11] While observing a purchase is not equivalent to observing a prescription, Koulayev et al. (2017) document that prescription medication adherence is high in Denmark.

i.e. information based on a patient's sickness state. This distinction of the two types of diagnostic skill provides a way to analyze the effects of counterfactual policies that improve either diagnostic skill independently. This is crucial for the *ex ante* evaluation of policies informed by machine learning risk predictions, e.g. providing information on patient types, because in many situations experts such as physicians may hold valuable private information that needs to be accommodated in the design of effective policies.

Sickness

We define patient i's binary sickness realization, y_i, as determined by a latent index, ν_i, such that the patient has a bacterial infection according to

$$y_i = \mathbf{1}[\nu_i > 0]. \tag{1}$$

The latent index ν_i is normally distributed with mean τ_i, the patient's type, such that $\nu_i \sim \mathcal{N}(\tau_i, 1)$.[12] Hence, a patient's probability of acquiring a urinary tract infection as a function of type τ_i is given by

$$P(y_i = 1|\tau_i) = \Phi(\tau_i), \tag{2}$$

where $\Phi(\cdot)$ is the standard normal CDF. For physician j, patient types are distributed

$$\tau_i \sim \mathcal{N}(\tau_j, \sigma^2_{\tau_j}), \tag{3}$$

where the physician-specific distribution of patient types accommodates variation in patients assigned to general practice clinics, for example based on geographic location, physician characteristics, or systematic differences in physicians' laboratory testing decisions. We make no distributional assumptions on τ_j or σ_{τ_j} across physicians.

Prediction

When a patient consults a physician in clinical practice, the physician gathers information about the patient's true sickness state from two sources. Physician j's signal on patient i's type based on observable characteristics is given by

$$\xi_{ij} \sim \mathcal{N}(\tau_i, \sigma^2_{\xi_j}). \tag{4}$$

[12] We normalize the sickness threshold to 0 and the variance of ν_i to 1.

The parameter σ_{ξ_j} represents the physician's diagnostic skill where low signal variance reflects high skill and high signal variance reflects low skill.

Clinical examination of patient i provides a direct signal on a patient's sickness state given by

$$\eta_{ij} \sim \mathcal{N}(y_i, \sigma^2_{\eta_j}). \tag{5}$$

It can include, for example, information gathered based on symptom assessment, patient-reported behavioral factors, and diagnostic tests at the point-of-care.[13] The parameter σ_{η_j} represents the physician's clinical diagnostic skill. Again, low signal variance reflects high skill, high signal variance reflects low skill.

We assume the signals ξ_{ij} and η_{ij} are independent.[14] That is, information related to the observable patient-type specific disease prevalence is assumed to be independent of information acquired via clinical assessment at an in-person consultation. For example, this assumption implies that information coming from the assessment of a dipstick or microscopy rapid diagnostic test, which signals the presence of bacteria in the urine (St John et al. 2006), is independent of the knowledge of different disease risk between an older and a younger woman.

Assuming a physician knows her own skill levels and her patient type distribution, her posterior probability of the patient's sickness state conditional on type and diagnostic signals is given by $P(y_i = 1 \mid \xi_{ij}, \eta_{ij}; \tau_j, \sigma_{\tau_j}, \sigma_{\eta_j}, \sigma_{\xi_j})$, which is derived formally in Appendix A.

Treatment choice

A physician's payoff function at an initial consultation reflects the trade-off between a patient suffering the sickness cost from delaying prescribing until a test result is available and the social cost of prescribing associated with a potential increase in antibiotic resistance due to antibiotic use. While the social cost is incurred for every antibiotic prescribed, the sickness cost of waiting is only incurred by untreated patients suffering from a bacterial infection. Likewise, antibiotic treatment is only curative and alleviates sickness if a patient suffers from a bacterial infection. We abstract from the choice of antibiotic molecule and focus on the extensive margin, the decision

[13] Physicians typically perform either one or both of the rapid diagnostic technologies available today: urine dipstick and microscopic analysis (Davenport et al. 2017). Dipstick analysis is a standard procedure but microscopic analysis requires additional equipment and training. Errors in interpreting dipstick results and performing microscopic analysis introduce variation in diagnostic skill in this setting (Holm et al. 2017), an observation documented in medical decision-making more generally (Hoffrage et al. 2000; Pallin et al. 2014).

[14] The signal structure is given by the bivariate normal

$$\begin{pmatrix} \xi_{ij} \\ \eta_{ij} \end{pmatrix} \sim N\left(\begin{pmatrix} \tau_i \\ y_i \end{pmatrix}, \begin{bmatrix} \sigma^2_{\xi_j} & 0 \\ 0 & \sigma^2_{\eta_j} \end{bmatrix} \right). \tag{6}$$

whether to prescribe an antibiotic versus delaying or avoiding antibiotic treatment altogether.[15] Thus, the general payoff function at a patient's initial consultation can be written as

$$\pi(d, y; \beta) = -y(1-d) - \beta d, \tag{7}$$

where d is an indicator for the decision to prescribe an antibiotic prior to observing the patient test result, y is the sickness state, and β is the social cost of prescribing.[16] Note that we normalize the weight on the sickness cost, the first term in equation (7), to one because the social cost of prescribing can only be identified relative to the sickness cost.[17]

A physician who maximizes expected payoff conditional on signals ξ_{ij} and η_{ij} proceeds to prescribe an antibiotic if

$$\begin{aligned} d_{ij} = 1 \quad &\Leftrightarrow \quad \mathrm{E}\{\pi(1, y_i; \beta_j) \mid \xi_{ij}, \eta_{ij}\} > \mathrm{E}\{\pi(0, y_j; \beta_j) \mid \xi_{ij}, \eta_{ij}\} \\ &\Leftrightarrow \quad P(y_i = 1 \mid \xi_{ij}, \eta_{ij}, \tau_j, \sigma_{\tau_j}, \sigma_{\eta_j}, \sigma_{\xi_j}) > \beta_j, \end{aligned} \tag{8}$$

that is, if the expected sickness cost for the patient while awaiting the test result is larger than the social cost of prescribing.

5 Empirical Analysis

5.1 Analysis sample

We apply several restrictions to define the sample for our main analysis. Urine samples in the hospital laboratory data constitute 477,609 samples out of which 156,694 are submitted by general practitioners. Firstly, we exclude 39,592 test observations from pregnant women in our analysis as both the test decision, including mandatory screening tests, and the prescription

[15] In our data, two molecules, Pivmecillinam and Sulfamethizole, account for 82 percent of all UTI-indicated prescriptions. Conditional on observing a positive test result, Yelin et al. (2019) evaluate how prediction of resistance probabilities can improve the choice of molecule using electronic health records from Israel. Kanjilal et al. (2020) study molecule choice for an emergency department in the US.

[16] An alternative payoff function that would include the social cost of follow-up prescriptions to sick patients who did not receive an initial prescription has the following form:

$$\begin{aligned} \pi(d, y; \beta_j) &= -y(1-d) - \beta_j d - \beta_j(1-\rho)y(1-d) \\ &= -(1 + \beta_j(1-\rho))y(1-d) - \beta_j d \\ &\propto -y(1-d) - \tilde{\beta}_j d, \end{aligned}$$

where $\rho \in (0, 1)$ is the spontaneous recovery rate while awaiting test results and $\tilde{\beta}_j = \beta_j/(1 + \beta_j(1-\rho))$. The term $\tilde{\beta}_j(1-d)y$ is the social cost accrued from patients that did not get an initial antibiotic prescription but did test positive for bacteria and were given a follow-up prescription if they did not spontaneously recover. The counterfactual predictions using this payoff function are identical to our main specification and only the interpretation of the weight on the externality changes.

[17] We refrain from using a monetary measure of these cost because existing research is lacking reliable estimates, see Jit et al. (2020) for a recent survey. Hence, the parameter β is a composite measure of a physician's subjective assessment of the social cost of antibiotic resistance and her preference weight on this cost.

decision do not represent typical cases of suspected UTI.

To focus on consultations that constitute a first contact with a physician within the patient's treatment spell, we further exclude 51,183 test observations where the patient received a systemic antibiotic prescription or was tested within 28 days prior to the sample acquisition date. The full set of test results used for machine learning comprises 65,919 urine samples taken during initial consultations with men or non-pregnant women in 2010, 2011, and 2012. For estimating the model and performing counterfactual policy analysis, we focus on the years 2011 and 2012 and require that clinics have at least 100 observations. The final analysis sample comprises 175 clinics and 36,972 observations.

By focusing on consultations during which physicians collected a urine sample for microbiological laboratory testing, we ensure that definitive test outcomes are observed regardless of the physicians' prescription decisions. Hence, our results may not generalize to prescription occasions that did not include microbiological testing. Even so, Ribers and Ullrich (2023) document that the predicted risk distribution in the sample of tested patients is not distinctly different from that of the general population of UTI cases and that selectivity of testing does not affect the policy results of a threshold-based decision rule.

For all patients with UTI-indicated prescriptions between 2011 and 2012, the mean age is 56, the share of female patients is 86 percent, the share of patients with migration background, reflecting a group of patients potentially less well known to Danish physicians, is 11 percent, and the share of patients living in a single household is 57 percent. In our analysis sample of tested patients, those who received a prescription have a mean age of 49, are women in 85 percent of cases, 15 percent have a migration background, and 51 percent live in single households. Individuals with a positive bacterial test outcome have a mean age of 52, the share of females is 87 percent, 13 percent have a migration background, and 54 percent live in single households.

5.2 Sample descriptives

Table 1 shows descriptive statistics for the 175 clinics included in the estimation sample. The top panel presents summary statistics for the sample of tested patients used for estimation. The bottom panel summarizes the number of all patients visiting a sampled clinic in the sample period.

The mean sample size per clinic is 211 initial consultation observations with a laboratory test, comprising 169 unique patients. On average, the number of patients per clinic for whom a clinic is their primary general practitioner is 3,491. In the two-year period of our estimation sample, on average 486 unique patients received a urine dipstick diagnostic in the sample period, which is

typically performed at urinary tract-related consultations in general practice. Laboratory testing for bacterial UTI is indicated in clinical practice when point-of-care diagnostics are inconclusive (Davenport et al. 2017). The 169 unique patients who received an initial laboratory diagnostic correspond to 35 percent of patients with at least one urinary tract consultation.[18]

Table 1 Clinic-level summary statistics

	Mean	St.dev.
Microbiological laboratory data, tested patients		
Initial consultations with laboratory test, per clinic	211	104
Unique patients with laboratory test, per clinic	169	68
Laboratory test result delay in days	3.1	0.3
Initial antibiotic prescribing rate	0.39	0.12
Bacterial rate	0.38	0.09
Initial prescribing rate, bacterial infections	0.60	0.14
Initial prescribing rate, no bacterial infections	0.26	0.10
Claims data, all patients		
Unique patients, per clinic	3,491	1,486
Unique patients with dipstick claim, per clinic	486	269
Unique patients with microscopy claim, per clinic	91	203
Clinics	175	
Initial consultation observations	36,972	

Notes: This table reports the means and standard deviations across clinics for the estimation sample in years 2011 and 2012.

Laboratory test procedures take two or more days during which general practitioners must decide under uncertainty. In our sample, the mean waiting time was 3.1 days with a standard deviation of 0.3. Since we know the precise timing of urine sample acquisitions and the test response date, we can determine whether physicians prescribe antibiotics with or without knowledge of the test result. Before knowing the test result, physicians prescribe an antibiotic in 39 percent of cases, on average. This rate corresponds nearly to the true bacterial rate of 38 percent. However, on average only 60 percent of patients with bacterial infections receive a prescription at the initial consultation while 26 percent of patients without a bacterial infection receive one, pointing to a substantial mismatch between initial prescriptions and bacterial infections.

5.3 Physician heterogeneity in prescription decisions

To inspect heterogeneity in observed treatment decisions, we view the physician's problem through the lens of binary classification where the prescription decision at an initial consul-

[18] We observe the date of a laboratory test but not of a dipstick test or consultation claimed by a clinic. Hence, we cannot compute the laboratory test rates based on a perfect patient-consultation match.

tation is the predicted condition and the laboratory test outcome indicating a bacterial infection is the true condition. The true positive rate (TPR) is the share of patients prescribed an antibiotic out of all patients with bacterial infections. The false positive rate (FPR) is the share of patients prescribed an antibiotic out of all patients without bacterial infections. Hence, the FPR is a measure of overprescribing.

Figure 1 shows a heat map of all individual clinics' locations in the TPR-FPR space. No clinics are below the diagonal which would indicate a higher rate of prescribing to negative test results than to patients with positive test results. On the diagonal physicians would prescribe equally to both groups indicating that they cannot distinguish the two and prescribe at random. Moving away from the diagonal towards the top-left is indicative of higher diagnostic skill in that physicians prescribe more to patients with infections than to those without. At the top-left, a physician with perfect skill would prescribe to 100 percent of positive test results and to 0 percent of non-bacterial test results.

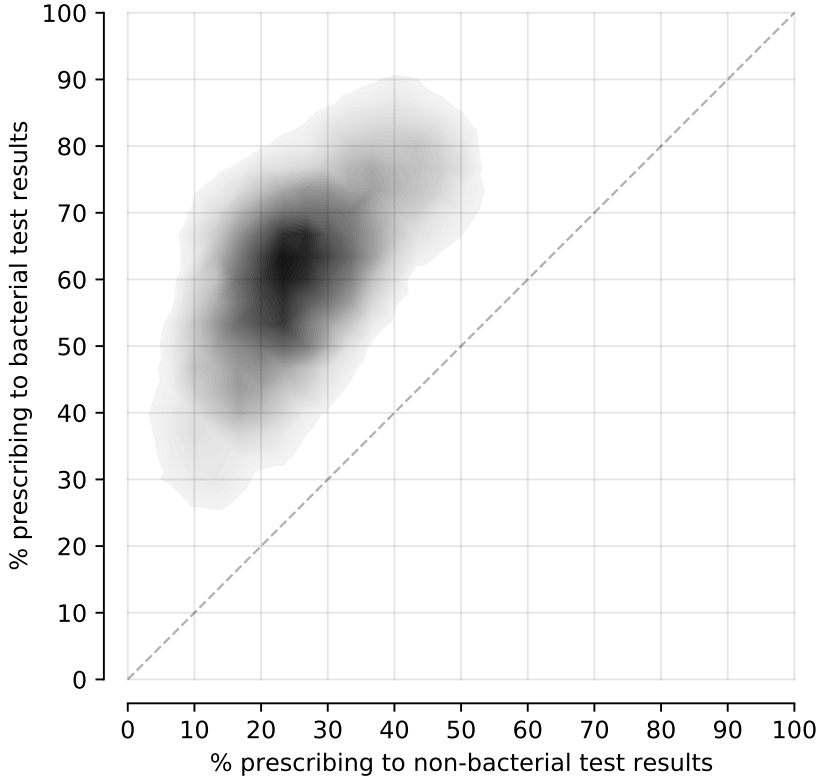


Figure 1: Heat map of clinics' prescribing rates conditional on test result

Notes: To ensure anonymity, the figure shows a heat map of an underlying scatter plot, with a minimum of five clinics used for local means. Darker areas represent higher clinic density.

Being located close to the origin indicates a large weight on the antibiotic resistance externality relative to individual patient sickness cost, reflected in low levels of overprescribing but also

low levels of appropriate prescribing. Being located towards the top right corresponds to more intense prescribing to patients who have both bacterial and non-bacterial test results, reflecting a low weight on the antibiotic resistance externality relative to individual sickness cost.

Overall, general practitioners appear to do well in avoiding prescribing to non-bacterial cases while prescribing to a high share of bacterial infections. Yet, the significant variation both perpendicular as well as parallel to the diagonal line indicates variation in skill and payoffs across clinics.

However, because the observed variation does not only reflect skill and payoffs but also variation in clinics' distributions of patients conditional on their testing decisions, a model incorporating the patient type distribution is needed to identify skill and payoffs.

5.4 Machine learning predictions

Machine learning tools can provide a personalized estimate of a patient's risk of UTI based on observables. We build on results from Huang et al. (2022) and Ribers and Ullrich (2023) who compute $m(x_i) = E[y_i|x_i]$, that is the expected microbiological test outcome y_i that determines if patient i with observable covariates x_i, available at the time of consultation, suffers from a bacterial UTI. We use the extreme gradient boosting algorithm (XGBoost), a fast and flexible ensemble method for structured, tabular data (Friedman et al. 2000; Friedman 2001; Chen and Guestrin 2016). The sample period 2010 serves as data for tuning and training the machine learning algorithm. We compute out-of-sample predictions for the analysis sample composed of all consultations from January 1st, 2011 to December 31st, 2012.

The area under the receiver operating curve (AUC) for the predictions in the analysis sample is 0.725. Figure 4 in Appendix B shows mean bacterial rates over bins of 100 consecutive observations after sorting all observations based on machine learning predicted risk. The close alignment along the diagonal illustrates the quality of predictions throughout the range of predicted risk.

The top predictors for bacterial outcomes reported in Ribers and Ullrich (2023) include patient age, gender, consulted clinic identifier, recent antibiotic prescriptions, recent antibiotic resistance results, clinic-specific resistance levels, regional prescription intensity, and recent hospital stays. These predictors can be plausibly related to the prevalence of bacterial infections such as UTI but no causal interpretation can be given to predictors selected by current machine learning methods. Details on the setup and implementation of the prediction algorithm, including replications using LASSO-based predictions, are described in detail in Ribers and Ullrich (2023). We take these predictions as given and focus on the policy evaluation problem.

5.5 Patient type distributions

The distributions of true patient types are *a priori* not observable to us and, importantly, determined by clinics' propensities to test patients. We make use of the machine learning predictions to estimate the patient type distribution for each clinic. To map machine learning predictions $m(x_i)$ into the model, we obtain patient types $\tau_i = \Phi^{-1}(m(x_i))$ by inversion, where $\Phi(\cdot)$ is the standard normal CDF.

However, machine learning predictions contain error. For inferring the patient type distribution, we avoid making the assumption that machine learning predictions represent the true patient type. Instead, similar to Mullainathan and Obermeyer (2022), we require the weaker assumption that the ordering of machine learning-predicted patient types reflects their true ordering. We run a binary logit regression for each clinic,

$$y_i = \frac{exp(\lambda_i)}{1 + exp(\lambda_i)}, \tag{9}$$

where $\lambda_i = \beta_0 + \beta_1 m(x_i) + \epsilon_i$.[19] The predicted outcomes of this regression provide our adjusted estimates of predicted risk, $\tilde{m}(x_i)$. The patient type estimates, which we will use in the estimation of the structural model, follow by inversion, $\tilde{\tau}_i = \Phi^{-1}(\tilde{m}(x_i))$.

To infer clinic-specific type distributions, we define the structural parameters $\tau_j = \mathrm{E}[\tilde{\tau}_{i(j)}]$ and $\sigma_{\tau_j} = \sqrt{\mathrm{E}[(\tilde{\tau}_{i(j)} - \tau_j)^2]}$ for the set of patients $\mathcal{I}_j$ consulting clinic j. We estimate $\hat{\tau}_j = \frac{1}{N_j} \sum_{i \in \mathcal{I}_j} \tilde{\tau}_i$ and $\hat{\sigma}_{\tau_j} = \sqrt{\frac{1}{N_j - 1} \sum_{i \in \mathcal{I}_j} (\tilde{\tau}_i - \hat{\tau}_j)^2}$, where N_j is the number of patients consulting clinic j.

5.6 Identification of physician skill and payoff parameters

Identification relies on a key feature reflected in many medical treatment decision contexts: the initial treatment decision must be made before test results from diagnostic procedures become available. In our data, on average, the waiting period for laboratory test results is 3.1 days, after which the true sickness state of the patient is revealed independent of the initial treatment decision. Consequently, we do not experience selection on labels and observe the joint distribution of prescription decisions and sickness realizations. Hence, we can directly observe true positives, true negatives, false positives, and false negatives.

Chan et al. (2022) show that a single skill parameter and a preference parameter can be identified from observing a decision maker's TPR-FPR location. Their model generates unique non-overlapping receiver operating characteristic (ROC) curves as a function of the skill parameter. An ROC curve represents all possible trade-offs between the TPR and FPR attainable for

[19] More flexible functional forms such as mean and rolling mean outcomes over bins of predicted risk do not change the results but require researcher decisions on bin sizes.

a given skill level. Once the ROC curve is determined, the position along the curve reflects the physician's preference weight, that is, the exact choice of trade off between TPR and FPR. This approach requires that patient type distributions are the same across clinics. Several papers have employed random assignment designs to plausibly assume this condition holds (Dobbie et al. 2021; Chan et al. 2022; Marquardt 2022). When disease prevalence and the distribution of patient types vary across clinics, observing a physician's location in the TPR-FPR space is not sufficient for identification. The clinic's patient type distribution is also needed. We make the identifying assumption that physicians know the type distribution for patients they test, with mean type τ_j and variance σ_{τ_j}.

In our model, the shape of physician j's ROC curve is jointly determined by the two skill parameters σ_{ξ_j} and σ_{η_j}. To see this, we show simulated ROC curves and their associated AUC values in Figure 5 in Appendix C. Higher levels in either skill dimension result in ROC curves shifted towards the top-left, implying better prediction ability, higher AUC values, and more favorable TPR-FPR trade-offs for the physician. This observation implies that multiple skill level combinations can result in different ROC curves that cross the same TPR-FPR point, as a decrease in one skill can be compensated by an increase in the other.

The two skill parameters on patient types and sickness realizations are separately identified by how physician prescribing differs as a function of continuous patient types and binary sickness realisations. At the one extreme, a physician relying only on the patient type signal, that is $\sigma_{\xi_j} < \infty$ and $\sigma_{\eta_j} = \infty$, will not be able to diagnose patients on their sickness realisations as no clinical information is used. Such a physician must have $E(d_{ij} \mid \tau_i, y_i) = E(d_{ij} \mid \tau_i)$ so that prescription decisions will be an increasing function of patient types only. At the other extreme, a physician relying only on the clinical examination signal, that is $\sigma_{\xi_j} = \infty$ and $\sigma_{\eta_j} < \infty$, will have $E(d_{ij} \mid \tau_i, y_i) = E(d_{ij} \mid y_i)$ so that prescription decisions will only be an increasing function of sickness realizations. Figure 6 in Appendix D shows expectations over simulated decisions as a function of continuous patient types and binary sickness realisations for multiple parameter values. Larger σ_{ξ_j} is reflected in a steeper slope of the function with respect to patient type, while larger σ_{η_j} separates the function into two curves conditional on y.[20]

Given skill levels, the preference parameter β_j is identified by the physician's observed TPR-FPR location along the ROC curve ranging from never prescribing for $\beta = 1$, with $(\text{TPR}, \text{FPR}) = (0, 0)$, to always prescribing for $\beta = 0$, with $(\text{TPR}, \text{FPR}) = (1, 1)$.

[20]We provide simulations here because the comparative statics with respect to the two skill parameters are difficult to derive with no closed form solution.

Interpretation of σ_{η_j} with varying patient-level diagnostic difficulty We model the signals obtained from clinical examination, η_{ij}, as centered on the sickness state $y_i \in \{0, 1\}$. Hence, the skill parameter σ_{η_j} is interpreted relative to the normalization of the distance between the binary sickness states to the unit interval. If patients differ in diagnostic difficulty, physicians with an easier patient group might have a distance between sickness states larger than one, making it easier to distinguish sickness states for a given σ_{η_j}. A patient group with harder diagnostic difficulty would be reflected in a distance between sickness states smaller than one. Hence, potential unobserved variation in patient difficulty limits the comparison of clinical diagnostic skill across physicians as it is not clear if differing σ_{η_j} are due to differences in skill or patient difficulty. However, the normalization does not affect counterfactual outcomes where we hold σ_{η_j} and the patient sample fixed.

5.7 Variation in testing

Physicians likely have valuable information that allows them to test with high yield, that is with a high bacterial rate. While the observable data contain rich patient- and clinic-level information that also capture persistent practice styles in testing, for example by including clinic identifiers as predictors, we cannot exclude that remaining systematic variation in unobservables exist.[21] Even though the structural parameters required for the evaluation of counterfactuals are identified, to inspect further external validity and to have an indication of potential variation in diagnostic difficulty of tested patients across clinics, we inspect whether variation in testing is driven by unobservables which may plausibly be correlated with test outcomes.

In Danish general practice, evidence suggests that decisions to send urine cultures to a laboratory lack systematic patterns. Córdoba et al. (2018) find that cases defined as suspected complicated UTI or the use and results of of rapid dipstick and microscopy tests, unobservable to us, do not predict the use of laboratory tests. Holm et al. (2021) analyze clinical management of UTI in Denmark and document that symptom assessment in general practice is highly noisy. Yet, rapid point-of-care diagnostics are known to help identify certain bacterial strains which may inform the initial decision to test.[22] Furthermore, virulence and severity of disease can vary

[21] Variation in diagnostic difficulty may in principle also be introduced by patients' selection of which physician to consult. In Denmark, general practitioners are assigned by an individual's residential address. Switching away from these default assignments is possible but uncommon. One reason for the lack of switching is the small choice set patients have in practice due to capacity constraints in Danish general practice (Kristiansen and Sheng 2022). Therefore, physicians treating UTI are almost completely determined by location of residence. The data we use for prediction contain information about patients' location of residence in addition to the described socioeconomic and health data, allowing for the prediction algorithm to use this information. Using data from Denmark, Huang and Ullrich (2023) provide evidence that patients do not sort into general practice clinics based on antibiotic prescribing style.

[22] The ability to detect bacterial strains varies across in-clinic diagnostics. Nitrite dipstick diagnostics can help detect so-called gram negative organisms of which E.coli is the most common for UTI. However, nitrite dipstick

between bacterial strains which may lead to variation in testing (Flores-Mireles et al. 2015).

To see if information acquired at initial consultations may lead to variation in testing in our sample, we first inspect the balance of the types of bacteria found in tests conditional on clinic-specific test yield, $E_j[y]$. If physicians systematically use information on the presence of bacteria from point-of-care tests for deciding whether to obtain a laboratory diagnostic, we should see differing rates of bacterial species across clinics with different test yields. We split clinics into two groups, above and below the median bacterial rate. The top panel in Table 4 in Appendix F reports the observation-weighted shares of bacterial strains for these two groups. The differences across bacteria are small. After controlling for mean predicted risk, which maps into the patient type distribution in the structural model, by comparing clinics below and above the median of deviations in mean test yield and predicted risk, $E_j[y] - E_j[m(x)]$, the differences become markedly smaller and not statistically significantly different from zero.

Physicians may also vary in their knowledge about the prevalence of antibiotic resistance for their patients. Such knowledge may influence the decision to use laboratory diagnostics. For the five molecules commonly used to treat UTI, the bottom panel in Table 4 shows variation in resistance rates between clinics with low and high test yield. Controlling for predicted risk, these differences become much smaller and not statistically significantly different from zero, providing further indication that unobserved, consultation-specific diagnostic information at the point-of-care is an unlikely driver of laboratory testing.

The choice of antibiotic at the initial consultation may also be informative of physicians' expectations about potential bacterial species or resistances which may influence the decision to test. The top panel of Table 5 in Appendix F shows that differences in the shares prescribed between clinics with high and low test yield are small and mostly not statistically significantly different from zero. Finally, the bottom panel of Table 5 shows differences in clinics' use of diagnostics. The number of laboratory test observations and usage of point-of-care tests such as urine dipstick and microscopic analysis do not differ significantly between clinics with low versus high bacterial rates in tested patients.

These findings fail to support notable selection on unobservables for the tested patient pool, in particular after conditioning on predicted risk. Hence, noise in the decision to test may be large, for example influenced by variation in logistical and organizational constraints in general practice.

diagnostics can not detect bacterial strains such as Enterococci, Staphylococci, and other organisms that do not convert nitrate to nitrite.

5.8 Model parameter estimation

We estimate the model parameters in two steps. In the first step, we recover individual patient types, $\tilde{\tau}_i$, and clinic-specific type distribution parameters $\hat{\tau}_j$ and $\hat{\sigma}_{\tau_j}$ as described in Section 5.5. We plug these into the treatment choice model, assuming that physicians' priors are equal to their clinics' patient type distribution.

In the second step, we estimate the model for each clinic by maximum likelihood using observed data on prescription decisions, d_{ij}, sickness realizations, y_i, and estimated individual patient types, $\tilde{\tau}_i$. The simulated likelihood contribution from a single observation is

$$\ell_{ij}(\sigma_{\xi_j}, \sigma_{\eta_j}, \beta_j \mid d_{ij}, y_i, \tilde{\tau}_i, \hat{\tau}_j, \hat{\sigma}_{\tau_j}) = P(d_{ij} \mid y_i, \tilde{\tau}_i, \hat{\tau}_j, \hat{\sigma}_{\tau_j}, \sigma_{\xi_j}, \sigma_{\eta_j}, \beta_j) \tag{10}$$

where the probability of prescribing is computed in equation 23 in Appendix E. The joint log-likelihood for the sample of clinic j is given by

$$\mathcal{L}_j(\sigma_{\xi_j}, \sigma_{\eta_j}, \beta_j \mid \boldsymbol{d}_j, \boldsymbol{y}_j, \tilde{\boldsymbol{\tau}}_j, \hat{\tau}_j, \hat{\sigma}_{\tau_j}) = \sum_{i \in \mathcal{I}_j} \log\big[\ell_{ij}(\sigma_{\xi_j}, \sigma_{\eta_j}, \beta_j \mid d_{ij}, y_i, \tilde{\tau}_i, \hat{\tau}_j, \hat{\sigma}_{\tau_j})\big], \tag{11}$$

with prescription decisions $\boldsymbol{d}_j = \{d_{ij}\}_{i \in \mathcal{I}_j}$, sickness realizations $\boldsymbol{y}_j = \{y_{ij}\}_{i \in \mathcal{I}_j}$, and patient types $\tilde{\boldsymbol{\tau}}_j = \{\tilde{\tau}_i\}_{i \in \mathcal{I}_j}$ for all patients of clinic j. Physician skill and payoff parameters are recovered as

$$\left(\hat{\sigma}_{\xi_j}, \hat{\sigma}_{\eta_j}, \hat{\beta}_j\right) = \underset{\sigma_{\xi_j}, \sigma_{\eta_j}, \beta_j}{\arg\min}\, \mathcal{L}_j(\sigma_{\xi_j}, \sigma_{\eta_j}, \beta_j \mid \boldsymbol{d}_j, \boldsymbol{y}_j, \tilde{\boldsymbol{\tau}}_j, \hat{\tau}_j, \hat{\sigma}_{\tau_j}). \tag{12}$$

6 Estimation results

Table 2 reports the means and standard deviations of the distribution of the estimated parameters $\{\hat{\sigma}_{\xi_j}, \hat{\sigma}_{\eta_j}, \hat{\beta}_j\}$ across clinics. The mean of $\hat{\sigma}_{\xi_j}$ is markedly larger than $\hat{\sigma}_{\eta_j}$, implying that physicians on average rely more on clinical diagnostic information than on information obtained from observing patient types. This result suggests that providing patient type information in the form of machine learning predicted risk should improve physicians' ability to predict the bacterial cause of infections. The extent to which patient type and clinical diagnostic information is used in decisions varies significantly between clinics, as reflected in the standard deviations of the estimates of $\hat{\sigma}_{\xi_j}$ and $\hat{\sigma}_{\eta_j}$. The mean value of 0.43 of the preference parameter estimates suggests conservative physicians on average. The mean physician weighs the social cost of increasing antibiotic resistance due to one antibiotic prescription slightly below one half the health benefit of instantly giving an effective treatment to one patient. The standard deviation of 0.10 reflects existing heterogeneity in how physicians solve this trade-off.

Table 2 Distribution of parameter estimates

	Mean	(St.dev.)
Patient type distribution, $\hat{\tau}_j$	-0.33	(0.27)
Patient type distribution, $\hat{\sigma}_{\tau_j}$	0.50	(0.14)
Type signal noise, $\hat{\sigma}_{\xi_j}$	3.01	(3.64)
Diagnostic signal noise, $\hat{\sigma}_{\eta_j}$	1.28	(0.86)
Payoff function parameter, $\hat{\beta}_j$	0.43	(0.10)

Notes: This table reports the means and standard deviations of the nonparametric distribution of parameter estimates over 175 clinics. The model is estimated separately for each clinic.

Figures 7 to 9 in Appendix G show the distributions of parameter estimates. Due to anonymization requirements, we show heat maps of scatter plots where no values in areas containing fewer than five clinics are reported. The distributions of both skill parameters σ_{ξ_j} and σ_{η_j} are concentrated in the area between 0 and 3. Correlation between both parameters appears to be small across clinics. Yet, we find a relevant number of physicians with very large σ_{ξ_j} estimates. In Figure 7, physicians with estimated $\sigma_{\xi_j} > 5$ account for 21% of all physicians. This group does not appear to make particular use patient type information encoded in observable data. Therefore, combining systematic information in predictions $m(x_i)$ with valuable clinical diagnostic information used by these physicians may substantially improve decisions. Figures 8 and 9 do not show a systematic relationship between the estimated payoff weights and both noise parameters. Figure 10 in Appendix I shows the distributions of the observed mean, over-, and underprescribing rates as well as their simulated in-sample counterparts based on the estimated model. The simulated distributions closely resemble the observed data.

Observed heterogeneity To investigate potential sources of heterogeneity across clinics, we correlate parameter estimates with observable clinic characteristics. For clinics with more than one physician, we aggregate individual physician characteristics to the clinic level because prescriptions are observed for clinics. Because the complete registry linking clinics with individual physician identifiers could not be obtained, we can merge characteristics for a subset of 107 of the total of 175 clinics. Linear regression results of the parameter estimates on clinic characteristics in Tables 6 to 8 in Appendix H show several interesting patterns. In Table 6, the correlation of $\hat{\sigma}_{\xi_j}$ with physician age, the use of diagnostics and the number of patients per physician is positive, reflecting lower use of patient type information, but none of the coefficients are statistically significant. In Table 7, for the estimates of the skill parameter for clinical diagnostic information, $\hat{\sigma}_{\eta_j}$, higher noise is associated with higher physician age. Older physicians might rely more on

their clinical experience and personal knowledge of patients than on point-of-care tests. Lower noise is associated with higher propensity to perform point-of-care dipstick analyses of urine samples, which can provide clinical information instantly. In Table 8, for $\hat{\beta}_j$, the coefficients are close to zero except for a small negative coefficient on physician age.

7 Counterfactual policy evaluation

We consider four counterfactual policies to assess how measuring preferences and information is essential for understanding the potential of machine learning predictions to help achieve socially desirable outcomes. For example, if β_j differs from socially desirable weights on the antibiotic resistance externality, physicians' adherence to prediction-based decision rules must be enforced or incentivized. Without accounting for heterogeneity and diverging incentives, *ex ante* expected improvements based on prediction-based decision rules may be mistakenly ascribed to improved prediction technologies instead of to the imposition of socially desirable objectives. Policy approaches relying on improved predictive information while allowing full discretion to expert decision makers may then fail to achieve their intended outcomes.

7.1 Decision outcomes

We evaluate policies by comparing counterfactual decisions d_{ij}^{CF} with observed physician decisions d_{ij}. For the full set of patients $\mathcal{I}$ across all clinics, changes in initial prescribing are defined by

$$\Delta d = \sum_{i \in \mathcal{I}} (d_i^{CF} - d_i).$$

Changes in overprescribing, i.e. antibiotic prescriptions given to patients without a bacterial infection, are defined

$$\Delta d(1-y) = \sum_{i \in \mathcal{I}} (d_i^{CF} - d_i)(1-y_i)$$

and changes in the number of initially treated bacterial UTI patients are defined

$$\Delta dy = \sum_{i \in \mathcal{I}} (d_i^{CF} - d_i) y_i.$$

Table 3 shows policy outcomes for the four counterfactual interventions. For every counterfactual outcome, we estimate 95 percent confidence intervals using bootstrapping based on 1000 samples and including all estimation steps.[23]

[23]We use the Bayesian bootstrap by re-weighting bootstrap observations as proposed in Rubin (1981), which offers higher speed and stability with nonlinear models. Standard bootstrapping using discrete weights by re-

The first counterfactual (CF1) serves as a benchmark, in which we reproduce an algorithmic prescription rule as in Huang et al. (2022). Here we do not make use of our choice model but instead include the physician decisions as a predictor in the machine learning algorithm.[24] In this policy, prescriptions for patients with low predicted risk are delayed until test results are available. All patients with high predicted risk receive prescriptions before test results arrive. Hence, counterfactual decisions are determined by a threshold k and the rule $d^{CF} = \mathbf{1}[m(x,d) > k]$. This type of policy resembles the approach taken by the prior literature evaluating machine learning predictions (Bayati et al. 2014; Chalfin et al. 2016; Kleinberg et al. 2018; Yelin et al. 2019; Hastings et al. 2020; Huang et al. 2022; Ribers and Ullrich 2023). Similar to this literature we focus on a solution for k that guarantees a welfare increase to a social planner for all potential payoff weights $\beta^S \in [0,1]$. We set k to maximize reductions in antibiotic use without changing the number of prescriptions to bacterial infections, that is keeping $\Delta dy = 0$. This policy relies on the assumption that human discretion can be overruled or that decision makers adhere perfectly to prediction-based prescription rules.

The first counterfactual policy reduces overall prescribing by 8.9 percent and overprescribing by 22.7 percent while, by construction, the change in the number of prescriptions to bacterial infections is zero. Without a model, we remain agnostic to the mechanism leading to these improved outcomes.

The policy in the second counterfactual (CF2) provides physicians with the machine learning prediction of type τ_i for every patient and assumes that physicians use it without noise by setting $\sigma_{\xi_j} = 0$. In this counterfactual, the clinical diagnostic skill and payoff function parameter are held fixed. We find that overall prescribing increases by 3.5 percent and overprescribing decreases by 3.9 percent. The number of treated bacterial infections increases by 8.3 percent. Hence, the improved information on patient type provided to the physicians leads to more efficient prescribing but fails to achieve the aim of reducing antibiotic prescribing overall. Comparing these results with the redistribution policy in CF1, we conclude that the reductions in overall prescribing and in overprescribing documented in the first counterfactual cannot be driven only by a potential superiority of machine learning predictions over physicians' diagnostic information.

In the third counterfactual (CF3), we again provide physicians with the machine learning prediction of type τ_i for every patient and hold their clinical diagnostic information fixed. In

sampling observations results in very similar confidence intervals. For computational reasons, we do not retrain the machine learning algorithm for every bootstrap sample and, hence, keep $m(x_i)$ fixed.

[24] Figure 11 in Appendix J shows the quality of predictions, with an AUC of 0.79, is markedly higher than when physician decisions are not used as a predictor because all predictive information encoded in physician decisions can now be used. Basing an algorithmic rule with a single threshold on these predictions, which we do here for convenience, is nearly equivalent to excluding physician decisions as a predictor and using a two-threshold rule that delegates some decisions to physicians as in Ribers and Ullrich (2023).

Table 3 Counterfactual policy outcomes

	ML redistribution	Provide ML-based τ_i		Incentives only
	$\Delta dy = 0$ $d^{CF} = \mathbf{1}[m(x,d) > k]$	$\sigma_{\xi_j} = 0$	$\sigma_{\xi_j} = 0$ $\beta_j = \hat{\beta}_j + \kappa_1$	$\beta_j = \hat{\beta}_j + \kappa_2$
Overall prescribing, Δd, in percent of $N_{\boldsymbol{d}} = 14,359$	-8.9 [-9.5, -8.4]	3.5 [3.1, 4.0]	-8.1 [-8.5, -7.7]	-8.1 [-8.5, -7.7]
Treated bacterial cases, Δdy, in percent of $N_{dy} = 8,704$	0	8.3 [7.9, 8.7]	0	-5.7 [-6.0, -5.4]
Overprescribing, $\Delta d(1-y)$, in percent of $N_{d(1-y)} = 5,655$	-22.7 [-23.9, -21.4]	-3.9 [-4.6, -3.0]	-20.6 [-21.5, -19.7]	-11.8 [-12.3, -11.3]
Mean change in payoffs, $\frac{1}{J}\sum_{j=1}^{J} W_j(\boldsymbol{d}_j)$	0.074 [0.067, 0.079]	0.119 [0.115, 0.123]	0.114 [0.111, 0.118]	-0.001 [-0.002, -0.001]

Notes: This table reports changes to the status quo in percent across 175 clinics and 36,972 patients. The left column shows further relevant absolute totals. The risk threshold for prescribing in counterfactual one is $k = 0.510$ $[0.508, 0.513]$. We set $\kappa_1 = 0.038$ $[0.037, 0.040]$ to obtain $\Delta dy = 0$ in counterfactual three and set $\kappa_2 = 0.021$ $[0.020, 0.021]$ to obtain $\Delta d = -8.1$ in counterfactual four. Bootstrapped 95 percent confidence intervals based on 1000 bootstrap samples in brackets.

addition, we increase the payoff parameter β_j by a constant κ_1 to maximize the reduction in overall prescribing while holding the number of prescriptions to bacterial infections fixed. We define the counterfactual payoff parameter as $\beta_j = \hat{\beta}_j + \kappa_1$, where setting $\kappa_1 = 0.038$ $[0.037, 0.040]$ attains an overall reduction in prescribing by 8.1 percent and in overprescribing by 20.6 percent.[25] This intervention could be implemented using nudges or an antibiotic tax that shifts the relative weight on the social cost of antibiotic resistance and a patients' sickness cost of delayed antibiotic treatment. Compared with the redistribution policy in counterfactual one, the overall reductions in (over)prescribing are now similar. However, clinic-level outcomes differ markedly between CF1 and CF3 which we will discuss in further detail below.

Finally, in the fourth counterfactual (CF4) we leave diagnostic information unchanged in both dimensions but adjust clinics' payoffs. We increase β_j by a constant $\kappa_2 = 0.021$ $[0.020, 0.021]$ such that the same reduction in overall antibiotic use is attained as in the third counterfactual. We motivate this counterfactual by policies that aim to raise awareness and increase the perceived social cost of antibiotic use due to antibiotic resistance (Hallsworth et al. 2016). The reduction in antibiotic use is 8.1 percent by construction. However, overprescribing is only reduced by 11.8 percent and the intervention comes at the cost of prescribing antibiotics to 5.7 percent fewer patients with a bacterial infection. Thus, policies focusing only on preferences may do harm when decisions are made under uncertainty. Physicians use antibiotics inefficiently because they mis-predict risk and not only because they do not care about the problem of antibiotic resistance. Hence, the improvement of diagnostic quality appears to play an important role for ensuring

[25] Recalling that the mean estimated β_j is 0.43, setting $\kappa_1 = 0.038$ implies an increase in β_j by 8.8 percent on average.

sustainable reductions in antibiotic use.

These results illustrate the value of separating the prediction and decision step in the structural model. The effects of interventions attempting to incentivize behavior according to social objectives can be considered independently from interventions aimed at purely improving diagnostic information. This is in contrast to situations studied by Cowgill and Stevenson (2020) in which algorithm outputs may be manipulated to communicate not only predictions but also social objectives. They argue that such manipulations can lead to refusal by human experts to use predictions. Our model allows for interventions in which the two channels, providing machine learning predictions to experts and incentivizing social behavior, can be evaluated as complements.

7.2 Clinic-level payoffs

Table 3 also reports the mean change of physician payoffs in Equation (7) for each policy, defined as

$$W_j(\boldsymbol{d}_j^{CF}) = \frac{\Pi_j(\boldsymbol{d}_j^{CF}) - \Pi_j(\boldsymbol{d}_j)}{\bar{\Pi}_j - \Pi_j(\boldsymbol{d}_j)} \tag{13}$$

where $\bar{\Pi}_j = -\hat{\beta}_j \sum_{i \in \mathcal{I}j} y_i$ is the first best outcome realized if the physician only gives prescriptions to patients with a bacterial infection and $\Pi_j(\boldsymbol{d}_j) = \sum_{i \in \mathcal{I}j} \pi(d_{ij}, y_i \,;\, \hat{\beta}_j)$ is the physician's payoff for the set of observed decisions $\boldsymbol{d}_j$. Payoff gains are largest for the counterfactual policy which provides patient type information and smallest for the policy in which only incentives are used. The policy increasing physicians' weights on the externality and providing predictions generates gains which are nearly as large as the policy providing patient type information. Redistributing prescriptions in the first counterfactual results in positive but lower gains than policies that provide prediction information.

Figure 2 shows the distribution of clinic-level changes in payoffs for all counterfactuals. For the first, using physician decisions as prediction input but allowing for no physician discretion, a sizable share of clinics obtain decreased payoffs. All clinics benefit from information provision and full discretion in the second counterfactual. When incentivized and given information, all payoffs still increase but the distribution is shifted to the left, illustrating the misalignment of objectives but also the offsettting value of new information. When only incentivized, the distribution of payoff changes is symmetric around zero.

These results are intuitive given that information provision strictly improves efficiency while changing physicians' weights, or decisions altogether, conflicts with their preferences. This conflict can hinder the use of a redistribution policy and require a more sophisticated policy design. In particular, these results further demonstrate that in policy contexts where diagnostic infor-

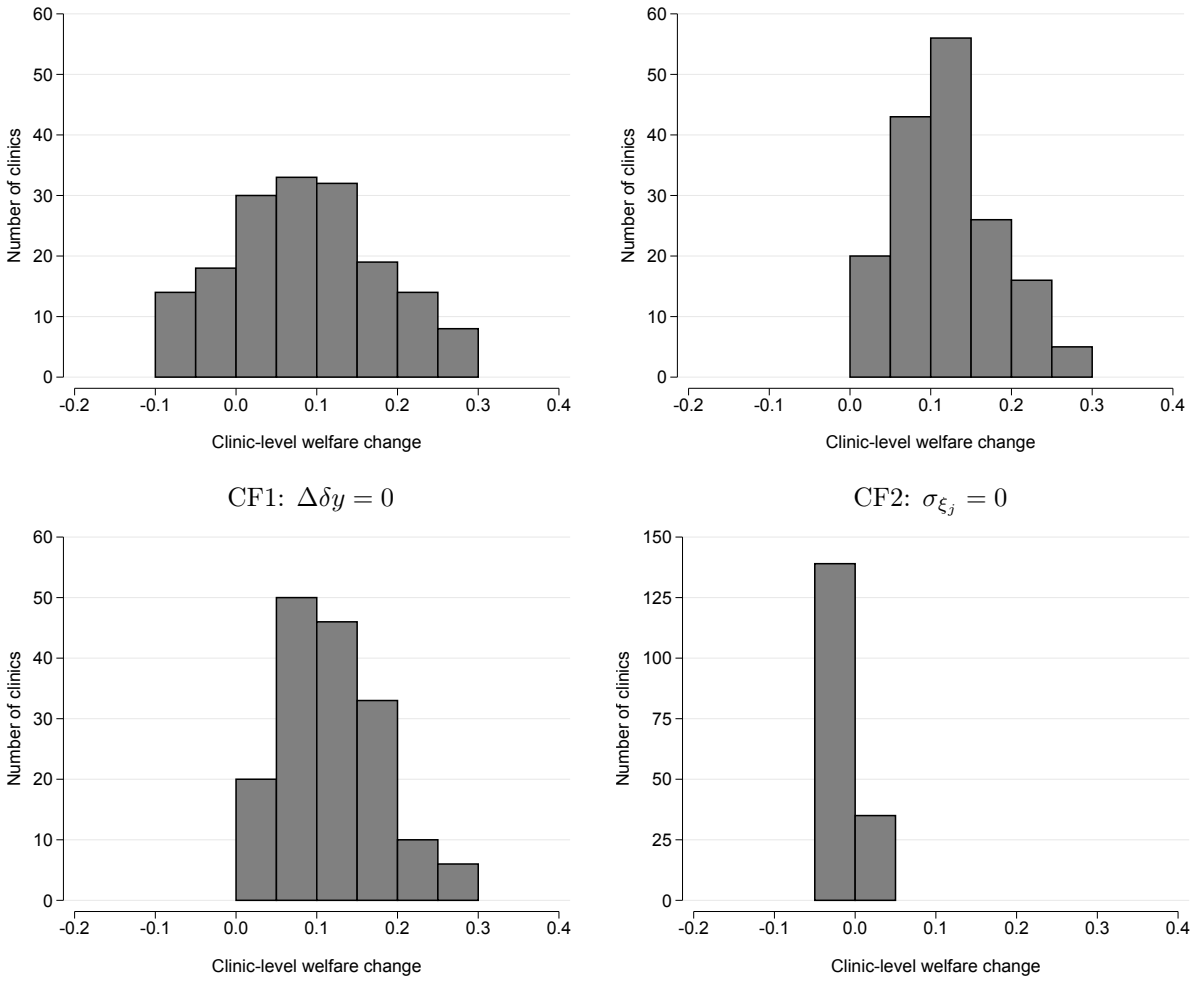


Figure 2: Distribution of counterfactual payoff changes, W_j

mation is limited, manipulating incentives without improving information can lead to adverse outcomes.

7.3 Clinic heterogeneity in counterfactual changes of decision outcomes

To investigate heterogeneity in policy outcomes further, we plot changes in overall antibiotic prescribing and in prescriptions to bacterial infections for all four counterfactual policies in Figure 12 in Appendix K. In Figure 12a the first counterfactual leads to large heterogeneity in outcomes, including large increases in antibiotic prescribing as well as in foregone treatment. Instead, providing predictions results in significantly more concentrated counterfactual outcomes, shifted upwards to increase the number of treated bacterial cases in the second counterfactual in Figure 12b and shifted left and downwards when incentivized to prescribe less in Figure 12c. As depicted in Figure 12d, changing only the preference parameter without providing new information is effective in reducing antibiotic use for all clinics but at the cost of fewer treated

bacterial infections for all clinics.

7.4 Social planner

Finally, taking the perspective of a social planner, we provide a ranking of counterfactual policies based not only on reported counts of outcomes but on gains in payoffs given social preferences. We calculate counterfactual welfare effects over the continuum of potential social planner preference parameter values $\beta^S \in [0, 1]$ as

$$W(\boldsymbol{d}^{CF}, \beta^S) = \frac{\Pi(\boldsymbol{d}^{CF}, \beta^S) - \Pi(\boldsymbol{d}, \beta^S)}{\bar{\Pi} - \Pi(\boldsymbol{d}, \beta^S)} \tag{14}$$

where $\bar{\Pi} = -\beta^S \sum_{i \in \mathcal{I}} y_i$ is the first best aggregate outcome over the full set of patients, $\mathcal{I}$, realized if and only if prescriptions are given to patients with a bacterial infection. $\Pi(\boldsymbol{d}, \beta^S) = \sum_{i \in \mathcal{I}} \pi(d_{ij(i)}, y_i\,;\, \beta^S)$ is the aggregated payoff function for the set of decisions $\boldsymbol{d}$.

Figure 3 shows $W(\boldsymbol{d}^{CF}, \beta^S)$ over the full support of β^S, revealing that the best policy depends on the social planner's weight on the antibiotic resistance externality. If the social planner's weight on the externality is small, below approximately the average estimated physician β_j of 0.43, the counterfactual policy CF2, providing physicians with machine learning predictions and leaving them full discretion, maximizes the social planner's payoff gains. This is intuitive because it is only welfare-increasing to reduce prescriptions below the observed levels of physician prescribing if society places a sufficiently large weight on the resistance externality. Otherwise, treating more bacterial infections becomes the more important objective.

If the social planner's weight on the externality is larger than the average estimated physicians' weight, then welfare gains are larger for policies CF1 and CF3 which implement the social planner's objective function by replacing physician decisions or by manipulating physicians' weights' on the externality to reduce prescribing. If a policy only incentivizes physicians, as in CF4, the change in social welfare is negative for β^S below the mean estimated physician β_j due to a reduction in treated bacterial infections. If the social planner places a larger weight on the externality, policy CF4 leads to positive welfare gains although below those of counterfactuals CF1 and CF3. Hence, counterfactual CF4 appears to be suboptimal throughout the social planner preference range.

Policies targeting preferences may however be easier to implement in practice than policies improving diagnostic information. Depending on social preferences, targeting preferences can be worthwhile in the context we consider here. We find this can have adverse effects when diagnostic information is limited. Our results suggest that the largest welfare gains can be achieved when improved diagnostics and targeting physician preferences are seen as complementary policy tools.

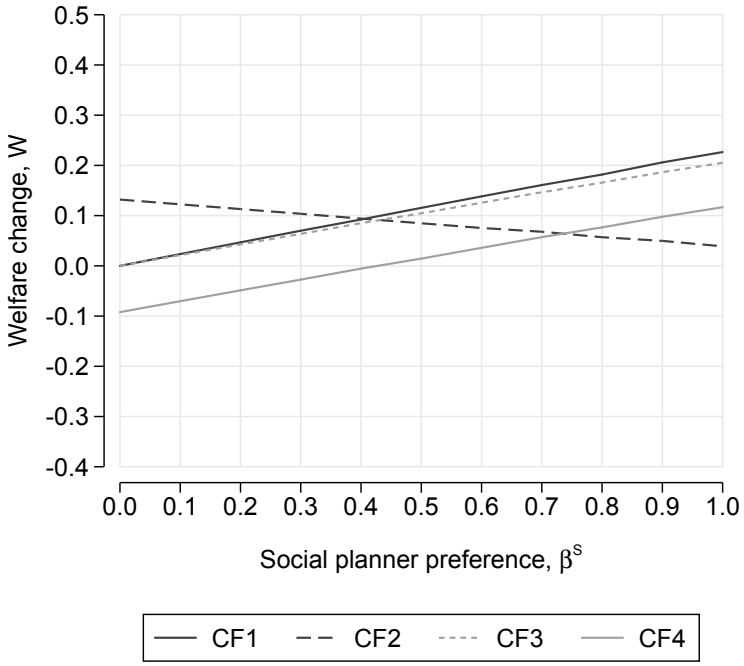


Figure 3: Counterfactual welfare for the social planner as a function of β^S

8 Robustness checks

To assess the robustness of our results and qualitative conclusions, we estimate the model and perform counterfactual policy evaluations for two alternative samples. The results are presented in Appendix L.

First, we increase the period prior to an observed consultation based on which we define an "initial" consultation. In the main analysis, we require four weeks without antibiotic treatment or laboratory testing. If this period is too short, we might include some patients who are still currently in treatment. If so, physicians may hold private information about the current treatment spell for a patient, which would affect the decision to use a laboratory test as well as to prescribe an antibiotic. Extending this period to 12 weeks of no prior antibiotic treatment or laboratory testing, we obtain counterfactual policy changes in Table 9 that are slightly smaller than our main result but lead to the same qualitative conclusions. Second, including clinics with as low as 100 observations may lead to too small samples at the clinic-level. To check if our results are sensitive to this potential issue, we restrict our sample to include only clinics with at least 200 test observations. For this smaller sample of 68 clinics, we obtain counterfactual policy results in Table 10 which closely align with the main results.

9 Conclusion

Antibiotic resistance is driving an increasingly pressing global health crisis, making efficient antibiotic use a prime policy concern. Through the lens of this policy problem, we consider how policies enabled by machine learning predictions can be evaluated when humans hold heterogeneous, private information. It is typically difficult to determine whether private information is complementary to or can be substituted by machine learning predictions. If such information, or the skill required to obtain it, is difficult to measure and varying across decision makers, assessing the added value of machine learning predictions *ex ante* is challenging. Field trials may be designed to provide reliable assessments but are often difficult to implement for ethical, legal, or practical reasons, particularly so in health care (Stern et al. 2022). Therefore, it is important to develop model-based tools to evaluate potential implementations *ex ante*.

The specific setting we consider, antibiotic prescribing for suspected urinary tract infections, resembles many situations in primary care provision and more generally expert decision problems under initial uncertainty. Exploiting that we observe the true health outcome *ex post*, we provide evidence that information generated by machine learning predictions and information held by human experts can be complements. We further show that improved information alone may not be sufficient to achieve socially desirable policy goals, requiring policies that combine information-provision and incentives. Ludwig and Mullainathan (2021) discuss the difficulty of designing algorithms to provide recommendations for decision-support systems that efficiently incorporate preferences and information. Hence, to curb antibiotic use, policy initiatives should promote diagnostic innovations such as new rapid point-of-care tests and solutions using large-scale patient data combined with incentives that internalize social cost, for example via a tax on antibiotic prescriptions.

Several important avenues for further research specific to the context of antibiotic prescribing remain. It would be worthwhile to encode further clinical information, for example, from electronic health records, such as reported symptoms and results from in-clinic diagnostics to further improve machine learning predictions. This would require combining electronic health records with administrative data. Further research is needed to better understand physicians' potential behavioral reactions to the introduction of prediction tools. An interesting question in this regard is how to design information provision to physicians to achieve the policy objective of reduced prescribing as discussed in Cowgill and Stevenson (2020). Results from such studies may provide insights on how to optimally communicate machine learning predictions, to what extent to explain prediction outcomes, and potential effects on decision makers' incentives to acquire and use information and expertise.

More generally, as data availability and the quality of prediction algorithms are improving at a rapid pace, the rate at which such technologies will be more broadly adopted and productively exploited will depend on the kind and quality of human expertise it can complement. Investment in prediction technologies as well as in human capital, while emphasizing policy objectives, is key to induce welfare-improving technological progress.

References

Abaluck, J., Agha, L., Chan, D., Singer, D., and Zhu, D. (2020). Fixing Misallocation with Guidelines: Awareness vs. Adherence. NBER Working Paper No. 27467.

Abaluck, J., Agha, L., Kabrhel, C., Raja, A., and Venkatesh, A. (2016). The determinants of productivity in medical testing: Intensity and allocation of care. *American Economic Review*, 106(12):3730–64.

Adda, J. (2020). Preventing the spread of antibiotic resistance. *AEA Papers and Proceedings*, 110:255–259.

Agarwal, N., Moehring, A., Rajpurkar, P., and Salz, T. (2023). Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology. NBER Working Paper No. 31422.

Agrawal, A., Gans, J., and Goldfarb, A. (2018). *Prediction Machines: The Simple Economics of Artificial Intelligence*. Harvard Business Press.

Andini, M., Ciania, E., de Blasio, G., D'Ignazio, A., and Salvestrini, V. (2018). Targeting with machine learning: An application to a tax rebate program in Italy. *Journal of Economic Behavior and Organization*, 156:86–102.

Bayati, M., Braverman, M., Gillam, M., Mack, K. M., Ruiz, G., Smith, M. S., and Horvitz, E. (2014). Data-driven decisions for reducing readmissions for heart failure: General methodology and case study. *PLoS ONE*, 9(10):e109264.

Bennett, D., Hung, C.-L., and Lauderdale, T.-L. (2015). Health care competition and antibiotic use in Taiwan. *Journal of Industrial Economics*, 63(2):371–393.

Bjerrum, L. and Lindbæk, M. (2015). Which treatment strategy for women with symptoms of urinary tract infection? *BMJ*, 351:h6888.

Cassidy, R. and Manski, C. F. (2019). Tuberculosis diagnosis and treatment under uncertainty. *Proceedings of the National Academy of Sciences*, 116(46):22990–22997.

CDC (2013). Antibiotic resistance threats in the United States. Technical report.

CDC (2019). Antibiotic resistance threats in the United States. Technical report.

Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J., and Mullainathan, S. (2016). Productivity and selection of human capital with machine learning. *American Economic Review*, 106(5):124–127.

Chan, D. C., Gentzkow, M., and Yu, C. (2022). Selection with variation in diagnostic skill: Evidence from radiologists. *Quarterly Journal of Economics*, 137(2):729–783.

Chandra, A., Cutler, D., and Song, Z. (2011). Chapter Six - Who Ordered That? The Economics of Treatment Choices in Medical Care. In Pauly, M. V., Mcguire, T. G., and Barros, P. P., editors, *Handbook of Health Economics*, volume 2, pages 397–432.

Chandra, A. and Staiger, D. O. (2007). Productivity spillovers in health care: Evidence from the treatment of heart attacks. *Journal of Political Economy*, 115(1):103–140.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, San Francisco, California, USA.

Chu, C. M. and Lowder, J. L. (2018). Diagnosis and treatment of urinary tract infections across age groups. *American Journal of Obstetrics and Gynecology*, 219(1):40–51.

Córdoba, G., Holm, A., Sørensen, T. M., Siersma, V., Sandholdt, H., Makela, M., Frimodt-Møller, N., and Bjerrum, L. (2018). Use of diagnostic tests and the appropriateness of the treatment decision in patients with suspected urinary tract infection in primary care in Denmark–observational study. *BMC family practice*, 19(1):65.

Costelloe, C., Metcalfe, C., Lovering, A., Mant, D., and Hay, A. D. (2010). Effect of antibiotic prescribing in primary care on antimicrobial resistance in individual patients: Systematic review and meta-analysis. *BMJ*, 340:c2096.

Cowgill, B. and Stevenson, M. T. (2020). Algorithmic social engineering. *AEA Papers and Proceedings*, 110:96–100.

Currie, J., Lin, W., and Meng, J. (2014). Addressing antibiotic abuse in China: An experimental audit study. *Journal of Development Economics*, 110:39–51.

Currie, J. and MacLeod, W. B. (2017). Diagnosing expertise: Human capital, decision making, and performance among physicians. *Journal of Labor Economics*, 35(1):1–43.

Currie, J., MacLeod, W. B., and Van Parys, J. (2016). Provider practice style and patient health outcomes: The case of heart attacks. *Journal of Health Economics*, 47:64–80.

Danish Ministry of Health (2017). National action plan on antibiotics in human healthcare. Three measurable goals for a reduction of antibiotic consumption towards 2020.

Das, J., Holla, A., Mohpal, A., and Muralidharan, K. (2016). Quality and accountability in health care delivery: Audit-study evidence from primary care in India. *American Economic Review*, 106(12):3765–3799.

Davenport, M., Mach, K. E., Shortliffe, L. M. D., Banaei, N., Wang, T.-H., and Liao, J. C. (2017). New and developing diagnostic technologies for urinary tract infections. *Nature Reviews Urology*, 14(5):296.

Devillé, W. L., Yzermans, J. C., van Duijn, N. P., Bezemer, P. D., van der Windt, D. A., and Bouter, L. M. (2004). The urine dipstick test useful to rule out infections. A meta-analysis of the accuracy. *BMC Urology*, 4(1):4.

Dobbie, W., Liberman, A., Paravisini, D., and Pathania, V. (2021). Measuring bias in consumer lending. *Review of Economic Studies*, 88(6):2799–2832.

Dubois, P. and Gokkoca, G. (2023). Antibiotic Demand in the Presence of Antimicrobial Resistance. TSE Working Paper Nr. 23-1457.

Epstein, A. J. and Nicholson, S. (2009). The formation and evolution of physician treatment styles: An application to cesarean sections. *Journal of Health Economics*, 28(6):1126–1140.

Ferry, S. A., Holm, S. E., Stenlund, H., Lundholm, R., and Monsen, T. J. (2004). The natural course of uncomplicated lower urinary tract infection in women illustrated by a randomized placebo controlled study. *Scandinavian Journal of Infectious Diseases*, 36(4):296–301.

Finkelstein, A., Gentzkow, M., and Williams, H. (2016). Sources of geographic variation in health care: Evidence from patient migration. *Quarterly Journal of Economics*, 131(4):1681–1726.

Flores-Mireles, A. L., Walker, J. N., Caparon, M., and Hultgren, S. J. (2015). Urinary tract infections: Epidemiology, mechanisms of infection and treatment options. *Nature Reviews Microbiology*, 13:269–284.

Foxman, B. (2002). Epidemiology of urinary tract infections: Incidence, morbidity, and economic costs. *American Journal of Medicine*, 113(1):5–13.

Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, 28(2):337–407.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5):1189–1232.

Grigoryan, L., Trautner, B. W., and Gupta, K. (2014). Diagnosis and management of urinary tract infections in the outpatient setting: A review. *JAMA*, 312(16):1677–1684.

Gupta, K., Grigoryan, L., and Trautner, B. (2017). Urinary Tract Infection. *Annals of Internal Medicine*, 167(7):ITC49–ITC64.

Hallsworth, M., Chadborn, T., Sallis, A., Sanders, M., Berry, D., Greaves, F., Clements, L., and Davies, S. C. (2016). Provision of social norm feedback to high prescribers of antibiotics in general practice: A pragmatic national randomised controlled trial. *The Lancet*, 387(10029):1743–1752.

Hastings, J. S., Howison, M., and Inman, S. E. (2020). Predicting high-risk opioid prescriptions before they are given. *Proceedings of the National Academy of Sciences*, 117(4):1917–1923.

Hoffrage, U., Lindsey, S., Hertwig, R., and Gigerenzer, G. (2000). Communicating statistical information. *Science*, 290(5500):2261–2.

Holm, A., Cordoba, G., Sørensen, T. M., Jessen, L. R., Frimodt-Møller, N., Siersma, V., and Bjerrum, L. (2017). Clinical accuracy of point-of-care urine culture in general practice. *Scandinavian Journal of Primary Health Care*, 35(2):170–177.

Holm, A., Siersma, V., and Cordoba, G. C. (2021). Diagnosis of urinary tract infection based on symptoms: How are likelihood ratios affected by age? a diagnostic accuracy study. *BMJ Open*, 11(1):e039871.

Huang, S., Ribers, M. A., and Ullrich, H. (2022). Assessing the value of data for prediction policies: The case of antibiotic prescribing. *Economics Letters*, page 110360.

Huang, S. and Ullrich, H. (2023). Provider effects in antibiotic prescribing: Evidence from physician exits. Berlin School of Economics Discussion Papers Nr. 18.

Jit, M., Ng, D. H. L., Luangasanatip, N., Sandmann, F., Atkins, K. E., Robotham, J. V., and Pouwels, K. B. (2020). Quantifying the economic cost of antibiotic resistance and the

impact of related interventions: Rapid methodological review, conceptual framework and recommendations for future studies. *BMC Medicine*, 18(1):38.

Kang, J. S., Kuznetsova, P., Luca, M., and Choi, Y. (2013). Where not to eat? Improving public policy by predicting hygiene inspections using online reviews. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1443–1448, Seattle, Washington, USA. Association for Computational Linguistics.

Kanjilal, S., Oberst, M., Boominathan, S., Zhou, H., Hooper, D. C., and Sontag, D. (2020). A decision algorithm to promote outpatient antimicrobial stewardship for uncomplicated urinary tract infection. *Science Translational Medicine*, 12(568).

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2018). Human decisions and machine predictions. *Quarterly Journal of Economics*, 133(1):237–293.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, 105(5):491–495.

Koulayev, S., Simeonova, E., and Skipper, N. (2017). Can physicians affect patient adherence with medication? *Health Economics*, 26(6):779–794.

Kristiansen, I. L. and Sheng, S. Y. (2022). Doctor Who? The Effect of Physician-Patient Match on The SES-Health Gradient. CEBI Working Paper No. 05/22.

Kwon, I. and Jun, D. (2015). Information disclosure and peer effects in the use of antibiotics. *Journal of Health Economics*, 42:1–16.

Kwon, J. H. and Powderly, W. G. (2021). The post-antibiotic era is here. *Science*, 373(6554):471–471.

Laxminarayan, R. (2022). The overlooked pandemic of antimicrobial resistance. *The Lancet*, 399(10325):P606–607.

Laxminarayan, R., Duse, A., Wattal, C., Zaidi, A. K. M., Wertheim, H. F. L., Sumpradit, N., Vlieghe, E., Hara, G. L., Gould, I. M., Goossens, H., Greko, C., So, A. D., Bigdeli, M., Tomson, G., Woodhouse, W., Ombaka, E., Peralta, A. Q., Qamar, F. N., Mir, F., Kariuki, S., Bhutta, Z. A., Coates, A., Bergstrom, R., Wright, G. D., Brown, E. D., and Cars, O. (2013). Antibiotic resistance – the need for global solutions. *The Lancet Infectious Diseases Commission*, 13(12):1057–1098.

Llor, C. and Bjerrum, L. (2014). Antimicrobial resistance: Risk associated with antibiotic overuse and initiatives to reduce the problem. *Therapeutic Advances in Drug Safety*, 5(6):229–241.

Ludwig, J. and Mullainathan, S. (2021). Fragile Algorithms and Fallible Decision-Makers: Lessons from the Justice System. *Journal of Economic Perspectives*, 35(4):71–96.

Manski, C. F. (2021). Probabilistic Prediction for Binary Treatment Choice: With Focus on Personalized Medicine. NBER Working Paper No. 29358.

Marquardt, K. (2022). Mis(sed) diagnosis: Physician decision-making and ADHD. FRB of Chicago Working Paper No. 2022-23.

McAdams, D. (2021). The Blossoming of Economic Epidemiology. *Annual Review of Economics*, 13:539–570.

McAdams, D., Waldetoft, K. W., Tedijanto, C., Lipsitch, M., and Brown, S. P. (2019). Resistance diagnostics as a public health tool to combat antibiotic resistance: A model-based evaluation. *PLOS Biology*, 17(5):e3000250.

Mullainathan, S. and Obermeyer, Z. (2022). Diagnosing physician error: A machine learning approach to low-value health care. *Quarterly Journal of Economics*, 137(2):679–727.

Murray, C. J. et al. (2022). Global burden of bacterial antimicrobial resistance in 2019: A systematic analysis. *The Lancet*, 399(10325):629–655.

Nik-Ahd, F., Lenore Ackerman, A., and Anger, J. (2018). Recurrent Urinary Tract Infections in Females and the Overlap with Overactive Bladder. *Current Urology Reports*, 19(11):94.

Pallin, D. J., Ronan, C., Montazeri, K., Wai, K., Gold, A., Parmar, S., and Schuur, J. D. (2014). Urinalysis in acute care of adults: Pitfalls in testing and interpreting results. *Open Forum Infectious Diseases*, 1(1):ofu019.

Rambachan, A. (2022). Identifying prediction mistakes in observational data. Working paper.

Ribers, M. A. and Ullrich, H. (2023). Machine learning and physician prescribing: A path to reduced antibiotic use. Berlin School of Economics Discussion Paper Nr. 19.

Rubin, D. B. (1981). The Bayesian Bootstrap. *The Annals of Statistics*, 9(1):130–134.

Schmiemann, G., Kniehl, E., Gebhardt, K., Matejczyk, M. M., and Hummers-Pradier, E. (2010). The diagnosis of urinary tract infection: A systematic review. *Deutsches Ärzteblatt International*, 107(21):361.

Skinner, J. (2011). Chapter Two - Causes and Consequences of Regional Variations in Health Care. In Pauly, M. V., Mcguire, T. G., and Barros, P. P., editors, *Handbook of Health Economics*, volume 2, pages 45–93.

St John, A., Boyd, J. C., Lowes, A. J., and Price, C. P. (2006). The use of urinary dipstick tests to exclude urinary tract infection: A systematic review of the literature. *American Journal of Clinical Pathology*, 126(3):428–436.

Statistics Denmark (2012a). Arbejdstilknytning (IDA, The Integrated Database for Labour Market Research), 2005-2012.

Statistics Denmark (2012b). Befolkningen (BEF, Population Demographics), 2005-2012.

Statistics Denmark (2012c). Døde (DOD, Deaths), 2005-2012.

Statistics Denmark (2012d). Familie (FAM, Families), 2005-2012.

Statistics Denmark (2012e). Husstandsforhold (HUST, Households), 2005-2012.

Statistics Denmark (2012f). Indvandring (IEPE, Migration), 2005-2012.

Statistics Denmark (2012g). Landspatientregistret (LPR, The National Patient Registry), 2005-2012.

Statistics Denmark (2012h). Sygesikringsydelser (SSSY, The National Health Insurance Service), 2005-2012.

Statistics Denmark (2012i). Uddannelsesoplysninger (UDDA, Education), 2005-2012.

Stern, A. D., Goldfarb, A., Minssen, T., and Price, I. W. N. (2022). AI Insurance: How Liability Insurance Can Drive the Responsible Adoption of Artificial Intelligence in Health Care. *NEJM Catalyst*, 3(4):CAT.21.0242.

The Danish Health Data Authority (2012a). Lægemiddeldatabasen (LMDB, The Danish National Prescription Registry), 2005-2012.

The Danish Health Data Authority (2012b). Yderregister (YDER, General Practice Providers), 2005-2012.

WHO (2014). Antimicrobial resistance: 2014 global report on surveillance. Technical report, World Health Organization.

Wilson, M. L. and Gaido, L. (2004). Laboratory Diagnosis of Urinary Tract Infections in Adult Patients. *Medical Microbiology*, 38:1150–1158.

Yelin, I., Snitser, O., Novich, G., Katz, R., Tal, O., Parizade, M., Chodick, G., Koren, G., Shalev, V., and Kishony, R. (2019). Personal clinical history predicts antibiotic resistance of urinary tract infections. *Nature Medicine*, 25(7):1143–1152.

A Derivation of the posterior sickness probability conditional on diagnostic signals ξ_i and η_i

Physician j has a normal prior on patient types, $N(\tau_j, \sigma^2_{\tau_j})$, and receives a patient type signal, $\xi_{ij} \sim N(\tau_j, \sigma^2_{\xi_j})$, as given in equation (4). She hence forms the posterior belief about patient i's type given by

$$\tau_i \mid \xi_{ij}, \tau_j, \sigma_{\tau_j}, \sigma_{\xi_j} \sim N(\mu_{ij}, \sigma^2_{\mu_j}) \tag{15}$$

where

$$\mu_{ij} = \frac{\tau_j \sigma^2_{\xi_j} + \xi_{ij} \sigma^2_{\tau_j}}{\sigma^2_{\xi_j} + \sigma^2_{\tau_j}} \qquad \text{and} \qquad \sigma^2_{\mu_j} = \frac{\sigma^2_{\xi_j} \sigma^2_{\tau_j}}{\sigma^2_{\xi_j} + \sigma^2_{\tau_j}}. \tag{16}$$

The physician's latent sickness expectation conditional on the patient's type signal is thus given by

$$\nu_i \mid \xi_{ij}, \tau_j, \sigma_{\tau_j}, \sigma_{\xi_j} \sim N(\mu_{ij}, \sigma^2_{\mu_j} + 1), \tag{17}$$

which implies that the physician believes the patient is sick with probability

$$P(y_i = 1 \mid \xi_{ij}, \tau_j, \sigma_{\tau_j}, \sigma_{\xi_j}) = P(\nu_i > 0 \mid \xi_{ij}, \tau_j, \sigma_{\tau_j}, \sigma_{\xi_j}) = \Phi\left(\frac{\mu_{ij}}{\sqrt{\sigma^2_{\mu_j} + 1}} \right), \tag{18}$$

where $\Phi(\cdot)$ is the standard normal CDF.

Following a normal distributed signal on the patient sickness state from clinical assessment, η_{ij}, the physician's posterior on the patient's sickness realization becomes

$$\begin{aligned}
&P(y_i = 1 \mid \xi_{ij}, \eta_{ij}, \tau_j, \sigma_{\tau_j}, \sigma_{\xi_j}, \sigma_{\eta_j}) \\
&= \frac{P(\eta_{ij} \mid y_i = 1, \sigma_{\eta_j}) P(y_i = 1 \mid \xi_{ij}, \tau_j, \sigma_{\tau_j}, \sigma_{\xi_j})}{\sum_{y_i = 0,1} P(\eta_{ij} \mid y_i, \sigma_{\eta_j}) P(y_i \mid \xi_{ij}, \tau_j, \sigma_{\tau_j}, \sigma_{\xi_j})} \\
&= \frac{\phi(\frac{\eta_{ij} - 1}{\sigma_{\eta_j}}) \Phi\left(\frac{\mu_{ij}}{\sqrt{\sigma^2_{\mu_j} + 1}} \right)}{\phi(\frac{\eta_{ij} - 1}{\sigma_{\eta_j}}) \Phi\left(\frac{\mu_{ij}}{\sqrt{\sigma^2_{\mu_j} + 1}} \right) + \phi(\frac{\eta_{ij} - 0}{\sigma_{\eta_j}}) \Phi\left(\frac{-\mu_{ij}}{\sqrt{\sigma^2_{\mu_j} + 1}} \right)} \\
&= \frac{1}{1 + e^{\frac{1 - 2\eta_{ij}}{2\sigma^2_{\eta_j}}} \frac{\Phi\left(\frac{-\mu_{ij}}{\sqrt{\sigma^2_{\mu_j} + 1}} \right)}{\Phi\left(\frac{\mu_{ij}}{\sqrt{\sigma^2_{\mu_j} + 1}} \right)}},
\end{aligned} \tag{19}$$

where $\phi(\cdot)$ is the standard normal density function and we use that

$$\frac{\phi\left(\frac{\eta_{ij}-0}{\sigma_{\eta_j}}\right)}{\phi\left(\frac{\eta_{ij}-1}{\sigma_{\eta_j}}\right)} = \frac{\frac{1}{\sigma_{\eta_j}\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(\eta_{ij}-0)^2}{\sigma^2_{\eta_j}}}}{\frac{1}{\sigma_{\eta_j}\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(\eta_{ij}-1)^2}{\sigma^2_{\eta_j}}}} = e^{\frac{1}{2}\frac{(\eta_{ij}-1)^2}{\sigma^2_{\eta_j}} - \frac{1}{2}\frac{(\eta_{ij}-0)^2}{\sigma^2_{\eta_j}}} = e^{\frac{1-2\eta_{ij}}{2\sigma^2_{\eta_j}}}. \tag{20}$$

B Machine learning predictions

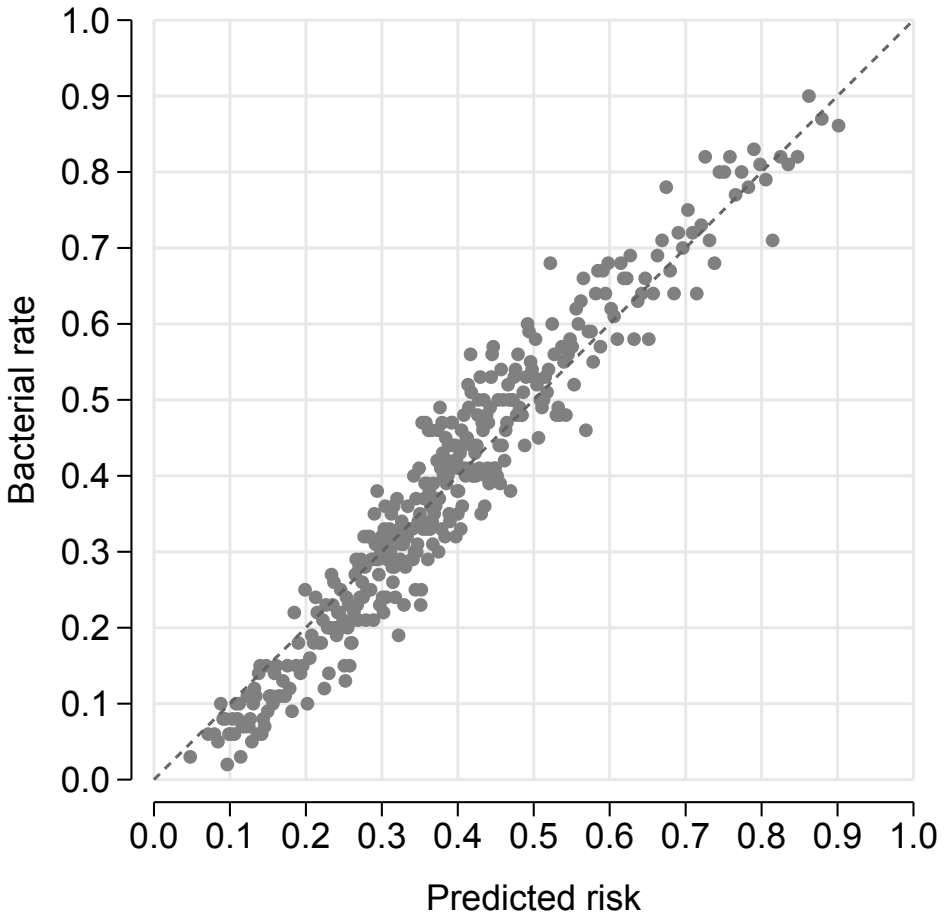
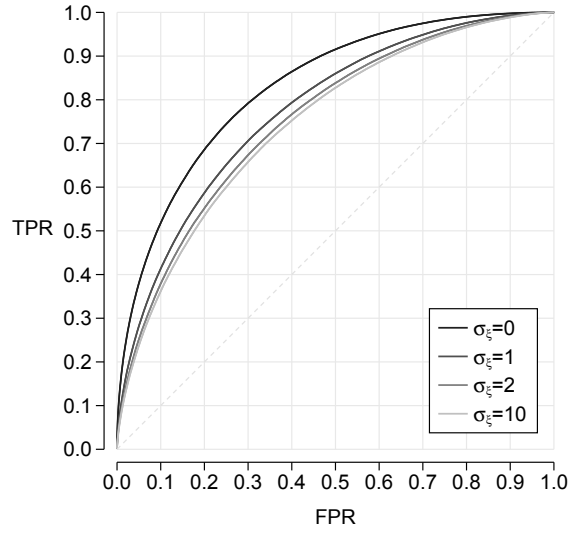


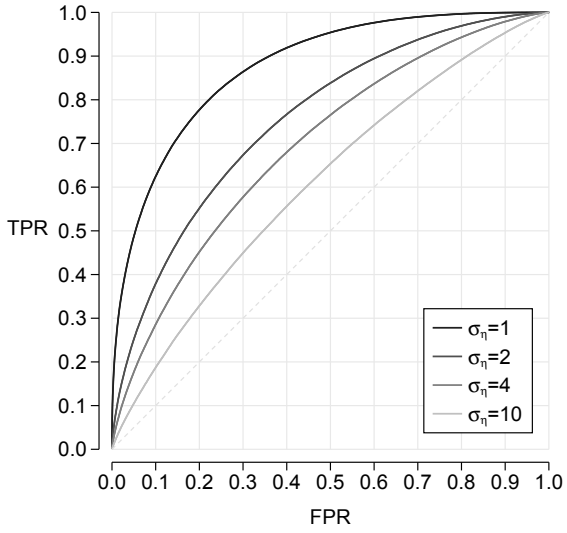
Figure 4: Bacterial rate $E[y_i]$ and machine learning predictions $m(x_i)$

Notes: Mean bacterial test outcomes relative to predicted risk of bacterial UTI. Spheres and triangles represent bins of 100 patients sorted by predicted risk.

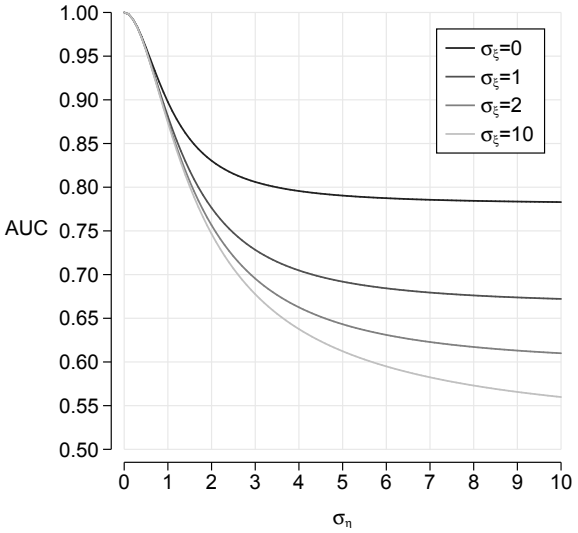
C Simulated ROCs and AUCs



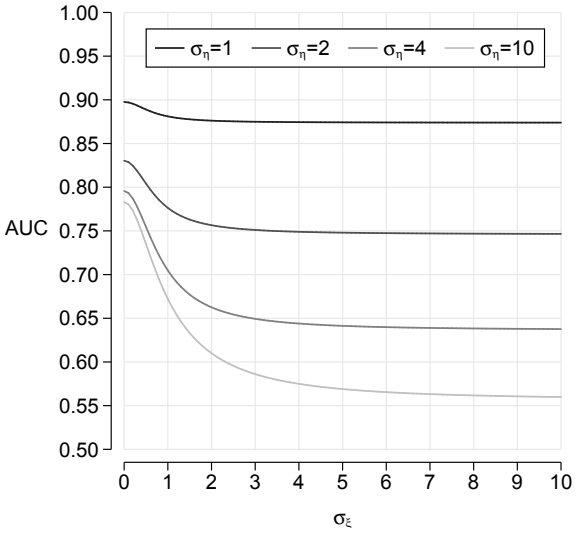
(a) Simulated ROCs for $\sigma_\eta = 2$



(b) Simulated ROCs for $\sigma_\xi = 2$



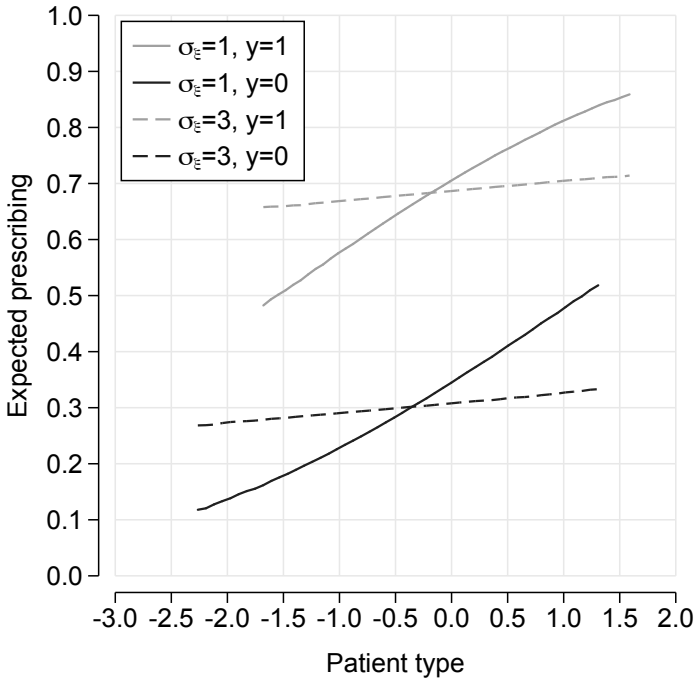
(c) Simulated AUCs as a function of σ_η



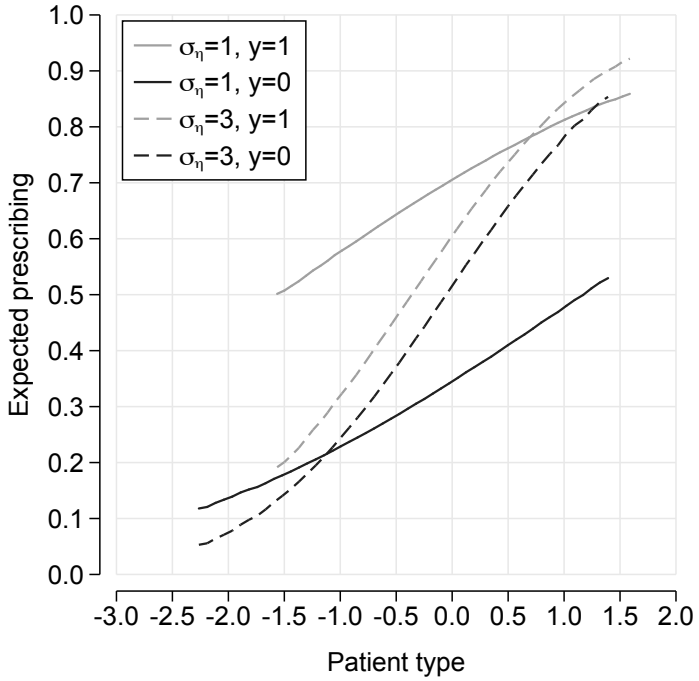
(d) Simulated AUCs as a function of σ_ξ

Figure 5: Simulated AUCs and ROCs as a function of σ_ξ and σ_η

D Simulated and observed decisions by y and $\tilde{m}(x_i)$



(a) Simulated $E[y]$ for $\sigma_{\eta_j} = 1$



(b) Simulated $E[y]$ for $\sigma_{\xi_j} = 1$

Figure 6: Simulated expected decisions as function of patient type and parameters

E Derivation of prescription probability for unobserved diagnostic signals ξ_i and η_i

Given ξ_{ij}, equations (8) and (19) allow the computation of an $\eta^*_{ij}(\xi_{ij})$ above which realizations of η_{ij} will be equivalent to a physician prescribing an antibiotic:

$$\begin{aligned} d_{ij} = 1 \quad &\Leftrightarrow \quad P(y_i = 1 \mid \xi_{ij}, \eta_{ij}, \tau_j, \sigma_{\tau_j}, \sigma_{\xi_j}, \sigma_{\eta_j}) > \beta_j \\ &\Leftrightarrow \quad \eta_{ij} > \frac{1}{2} - \sigma^2_{\eta_j} \log \left(\left(\frac{1}{\beta_j} - 1 \right) \frac{\Phi\left(\frac{\mu_{ij}}{\sqrt{\sigma^2_{\mu_j} + 1}} \right)}{\Phi\left(\frac{-\mu_{ij}}{\sqrt{\sigma^2_{\mu_j} + 1}} \right)} \right) = \eta^*_{ij}(\xi_{ij}), \end{aligned} \tag{21}$$

where, for notational ease, we suppress that η^*_{ij} is also conditional on a physician's prior and skill parameters. Hence, to arrive at the probability that physician j prescribes an antibiotic to patient i, we must integrate over realizations of ξ_{ij}:

$$P(d_{ij} = 1 \mid y_i, \tau_i, \tau_j, \sigma_{\tau_j}, \sigma_{\xi_j}, \sigma_{\eta_j}, \beta_j) = \int_{-\infty}^{\infty} \Phi\left(\frac{y_i - \eta^*_{ij}(\xi_{ij})}{\sigma_{\eta_j}} \right) \frac{1}{\sigma_{\xi_j}} \phi\left(\frac{\xi_{ij} - \tau_i}{\sigma_{\xi_j}} \right) d\xi_{ij}. \tag{22}$$

We approximate the integral using Gauss-Hermite quadrature:

$$\begin{aligned} P(d_{ij} = 1 \mid y_i, \tau_i, \tau_j, \sigma_{\tau_j}, \sigma_{\xi_j}, \sigma_{\eta_j}, \beta_j) &= \int_{-\infty}^{\infty} \Phi\left(\frac{y_i - \eta^*_{ij}(\xi_{ij})}{\sigma_{\eta_j}} \right) \frac{1}{\sigma_{\xi_j}} \phi\left(\frac{\xi_{ij} - \tau_i}{\sigma_{\xi_j}} \right) d\xi_{ij} \\ &= \int_{-\infty}^{\infty} \Phi\left(\frac{y_i - \eta^*_{ij}(\xi_{ij}, \tau_j, \sigma_{\tau_j})}{\sigma_{\eta_j}} \right) \frac{1}{\sigma_{\xi_j}\sqrt{2\pi}} e^{\left(\frac{\xi_{ij} - \tau_i}{2\sigma_{\xi_j}} \right)^2} d\xi_{ij} \\ &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \Phi\left(\frac{y_i - \eta^*_{ij}(u_{ij}\sigma^2_{\xi_j}\sqrt{2} + \tau_i)}{\sigma_{\eta_j}} \right) e^{u^2_{ij}} du_{ij} \\ &\approx \frac{1}{\sqrt{\pi}} \sum_k w_k \Phi\left(\frac{y_i - \eta^*_{ij}(x_k\sigma^2_{\xi_j}\sqrt{2} + \tau_i)}{\sigma_{\eta_j}} \right), \end{aligned} \tag{23}$$

where x_k and w_k are Gauss-Hermite nodes and weights, respectively.

F Balance tables

Table 4 Balance of types of bacterial infection causes

	$E_j[y]$			$E_j[y] - E_j[m(x)]$		
	$\leq med$	$> med$	Δ	≤ 0	> 0	Δ
Bacterial species isolated						
E. coli	0.70	0.73	*0.026*	0.70	0.72	0.019
	(0.07)	(0.06)	(0.010)	(0.07)	(0.06)	(0.010)
E. faecalis	0.07	0.05	*-0.016*	0.06	0.06	-0.007
	(0.04)	(0.03)	(0.005)	(0.04)	(0.03)	(0.005)
K. pneumoniae	0.04	0.04	0.005	0.04	0.04	-0.007
	(0.03)	(0.03)	(0.004)	(0.03)	(0.02)	(0.004)
S. agalactiae	0.04	0.04	-0.007	0.04	0.04	0.0002
	(0.04)	(0.02)	(0.005)	(0.04)	(0.03)	(0.005)
Others	0.15	0.14	-0.008	0.15	0.14	-0.004
	(0.05)	(0.04)	(0.007)	(0.05)	(0.05)	(0.007)
Molecule-specific resistance						
Mecillinam (J01CA11)	0.23	0.19	*-0.037*	0.22	0.21	-0.012
	(0.06)	(0.04)	(0.008)	(0.06)	(0.06)	(0.009)
Trimethoprim (J01EA01)	0.22	0.21	-0.016	0.22	0.21	-0.008
	(0.07)	(0.05)	(0.009)	(0.07)	(0.04)	(0.009)
Sulfamethizole (J01EB02)	0.36	0.33	*-0.025*	0.35	0.34	-0.004
	(0.08)	(0.06)	(0.011)	(0.08)	(0.06)	(0.011)
Ciprofloxacin (J01MA02)	0.14	0.12	*-0.022*	0.13	0.13	-0.004
	(0.06)	(0.04)	(0.008)	(0.06)	(0.04)	(0.008)
Nitrofurantoin (J01XE01)	0.06	0.06	-0.005	0.06	0.05	-0.008
	(0.04)	(0.03)	(0.005)	(0.03)	(0.03)	(0.005)
Number of clinics	88	87		96	79	

Notes: This table reports mean bacterial species and resistance rates for clinics above and below the median (med) of mean bacterial rates $E_j[y]$ and mean deviations $E_j[y] - E_j[m(x)]$. Physician-level means and standard deviations are weighted by physician-level numbers of observations. The listed molecules are antibiotics mainly prescribed for urinary tract infections in Denmark, accounting for 92.5 percent of prescriptions in our sample, Pivmecillinam and Sulfamethizole for 82 percent. For differences in italic, the null hypothesis of $\Delta = 0$ is rejected at the five percent level.

Table 5 Balance of molecules initially prescribed and use of diagnostics

	$E_j[y]$			$E_j[y] - E_j[m(x)]$		
	$\leq med$	$> med$	Δ	≤ 0	> 0	Δ
Molecule initially prescribed:						
Pivmecillinam (J01CA08)	0.52	0.59	*0.072*	0.54	0.58	0.037
	(0.20)	(0.19)	(0.029)	(0.20)	(0.20)	(0.030)
Trimethoprim (J01EA01)	0.02	0.02	-0.002	0.03	0.02	-0.008
	(0.04)	(0.03)	(0.006)	(0.05)	(0.03)	(0.006)
Sulfamethizole (J01EB02)	0.28	0.27	-0.009	0.26	0.28	0.020
	(0.19)	(0.19)	(0.028)	(0.18)	(0.19)	(0.029)
Ciprofloxacin (J01MA02)	0.05	0.02	*-0.020*	0.04	0.02	*-0.021*
	(0.05)	(0.02)	(0.006)	(0.05)	(0.03)	(0.006)
Nitrofurantoin (J01XE01)	0.03	0.03	-0.003	0.03	0.03	-0.002
	(0.04)	(0.04)	(0.006)	(0.04)	(0.05)	(0.006)
Use of diagnostics:						
Test observations	254	271	17.0	265	259	-6.5
	(117.3)	(139.5)	(19.49)	(131.9)	(124.9)	(19.14)
Urine dipsticks per patient	0.25	0.24	-0.012	0.26	0.24	-0.018
	(0.12)	(0.10)	(0.017)	(0.11)	(0.11)	(0.017)
Microscopy per patient	0.03	0.04	0.017	0.04	0.03	-0.001
	(0.07)	(0.08)	(0.011)	(0.08)	(0.07)	(0.011)
Number of clinics	88	87		96	79	

Notes: This table reports mean prescribed molecules and clinics' usage intensity of diagnostics for clinics above and below the median (med) of mean bacterial rates $E_j[y]$ and mean deviations $E_j[y] - E_j[m(x)]$. Physician-level means and standard deviations are weighted by physician-level numbers of observations. The listed molecules are antibiotics mainly prescribed for urinary tract infections in Denmark, accounting for 92.5 percent of prescriptions in our sample, Pivmecillinam and Sulfamethizole for 82 percent. For differences in italic, we reject the null hypothesis of $\Delta = 0$ at the five percent level.

G Model parameter estimates

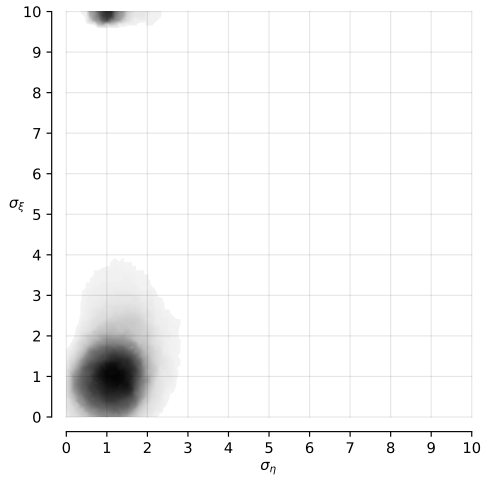


Figure 7: Heat map of physician-level estimates for σ_η and σ_ξ

Notes: To ensure anonymity, the figure shows a heat map of an underlying scatter plot, with a minimum of five clinics used for local means. Darker areas represent higher clinic density.

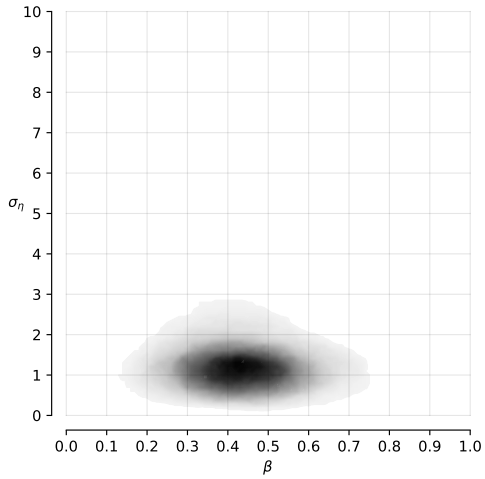


Figure 8: Heat map of physician-level estimates for β and σ_η

Notes: To ensure anonymity, the figure shows a heat map of an underlying scatter plot, with a minimum of five clinics used for local means. Darker areas represent higher clinic density.

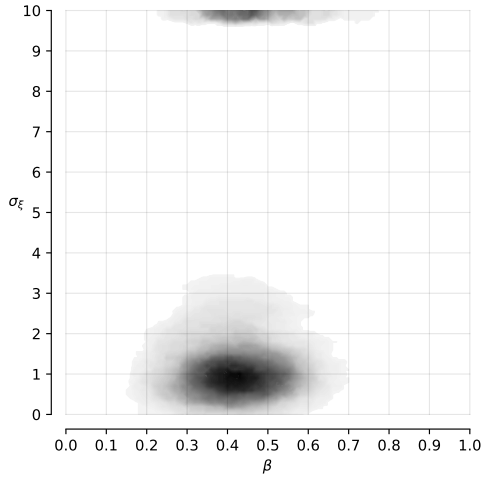


Figure 9: Heat map of physician-level estimates for β and σ_ξ

Notes: To ensure anonymity, the figure shows a heat map of an underlying scatter plot, with a minimum of five clinics used for local means. Darker areas represent higher clinic density.

H Observed heterogeneity: correlates of $\hat{\sigma}_{\eta_j}$, $\hat{\sigma}_{\xi_j}$, and $\hat{\beta}_j$

Table 6 Correlation of diagnostic skill estimate $\hat{\sigma}_{\xi_j}$ with clinic characteristics

	Linear regression for clinical signal noise $\hat{\sigma}_{\xi_j}$				
$N = 107$	(1)	(2)	(3)	(4)	(5)
Mean number of physicians	0.01				-0.22
	[-1.50, 1.52]				[-2.14, 1.69]
Mean age of physicians	2.16				2.14
	[-3.70, 8.01]				[-3.88, 8.160]
Share of female physicians	-0.35				-0.43
	[-1.39, 0.69]				[-1.53, 0.66]
Dipstick tests per consultation		0.57		0.49	0.83
		[-2.10, 3.24]		[-2.25, 3.23]	[-2.07, 3.74]
Microscopy analyses per consultation			0.44	0.35	0.39
			[-2.89, 3.78]	[-3.08, 3.77]	[-3.29, 4.08]
Patients per physician		0.35	0.32	0.36	-0.54
		[-1.93, 2.64]	[-1.93, 2.56]	[-1.92, 2.64]	[-3.49, 2.41]
Constant	1.65	2.70	3.09	2.70	1.86
	[-5.10, 8.40]	[-0.58, 5.98]	[0.67, 5.52]	[-0.58, 5.99]	[-6.38, 10.10]
R^2	0.02	0.002	0.001	0.002	0.02

Notes: This table presents coefficients for linear regressions where the outcome is the physician-level estimate of the patient type signal noise parameter summarized in Table 2. The correlates are clinic-level physician characteristics scaled at the mean. Mean values are used for multi-physician clinics. Heteroskedasticity-robust 95% confidence intervals are reported in brackets.

Table 7 Correlation of clinical skill estimate $\hat{\sigma}_{\eta_j}$ with clinic and physician characteristics

	Linear regression for clinical signal noise $\hat{\sigma}_{\eta_j}$				
$N = 107$	(1)	(2)	(3)	(4)	(5)
Mean number of physicians	-0.07				-0.05
	[-0.22, 0.08]				[-0.22, 0.11]
Mean age of physicians	0.82*				0.93*
	[-0.07, 1.71]				[-0.07, 1.94]
Share of female physicians	-0.12				-0.07
	[-0.35, 0.11]				[-0.24, 0.10]
Dipstick tests per consultation		-1.07		-1.07	-1.00
		[-3.12, 0.98]		[-3.15, 1.00]	[-2.98, 0.98]
Microscopy analyses per consultation			-0.20	0.00	-0.01
			[-0.69, 0.29]	[-0.34, 0.34]	[-0.41, 0.38]
Patients per physician		0.11	0.20	0.11	-0.16
		[-0.23, 0.45]	[-0.14, 0.55]	[-0.24, 0.45]	[-0.65, 0.34]
Constant	0.65*	1.96**	1.10***	1.96**	1.36*
	[-0.10, 1.39]	[0.25, 3.66	[0.82, 1.39]	[0.24, 3.67]	[-0.23, 2.96]
R^2	0.04	0.07	0.01	0.07	0.10

Notes: This table presents coefficients for linear regressions where the outcome is the physician-level estimate of the clinical signal noise parameter summarized in Table 2. The correlates are clinic-level physician characteristics scaled at the mean. Mean values are used for multi-physician clinics. Heteroskedasticity-robust 95% confidence intervals are reported in brackets.

Table 8 Correlation of payoff parameter estimate $\hat{\beta}_j$ with clinic characteristics

	Linear regression for payoff parameter $\hat{\beta}_j$				
$N = 107$	(1)	(2)	(3)	(4)	(5)
Mean number of physicians	-0.02				-0.02
	[-0.05, 0.002]				[-0.06, 0.01]
Mean age of physicians	-0.10				-0.09
	[-0.23, 0.04]				[-0.23, 0.04]
Share of female physicians	-0.01				-0.01
	[-0.04, 0.02]				[-0.04, 0.01]
Dipstick tests per consultation		0.04		0.05	0.06
		[-0.04, 0.12]		[-0.03, 0.13]	[-0.03, 0.14]
Microscopy analyses per consultation			-0.03	-0.04	-0.03
			[-0.11, 0.04]	[-0.12, 0.03]	[-0.10, 0.05]
Patients per physician		0.02	0.02	0.02	0.01
		[-0.04, 0.08]	[-0.04, 0.07]	[-0.04, 0.08]	[-0.07, 0.08]
Constant	0.57***	0.39***	0.43***	0.39***	0.52***
	[0.42, 0.72]	[0.32, 0.47]	[0.37, 0.49]	[0.31, 0.47]	[0.35, 0.69]
R^2	0.03	0.01	0.01	0.02	0.05

Notes: This table presents coefficients for linear regressions where the outcome is the physician-level estimate of the payoff parameter summarized in Table 2. The correlates are clinic-level physician characteristics scaled at the mean. Mean values are used for multi-physician clinics. Heteroskedasticity-robust 95% confidence intervals are reported in brackets.

I Model fit

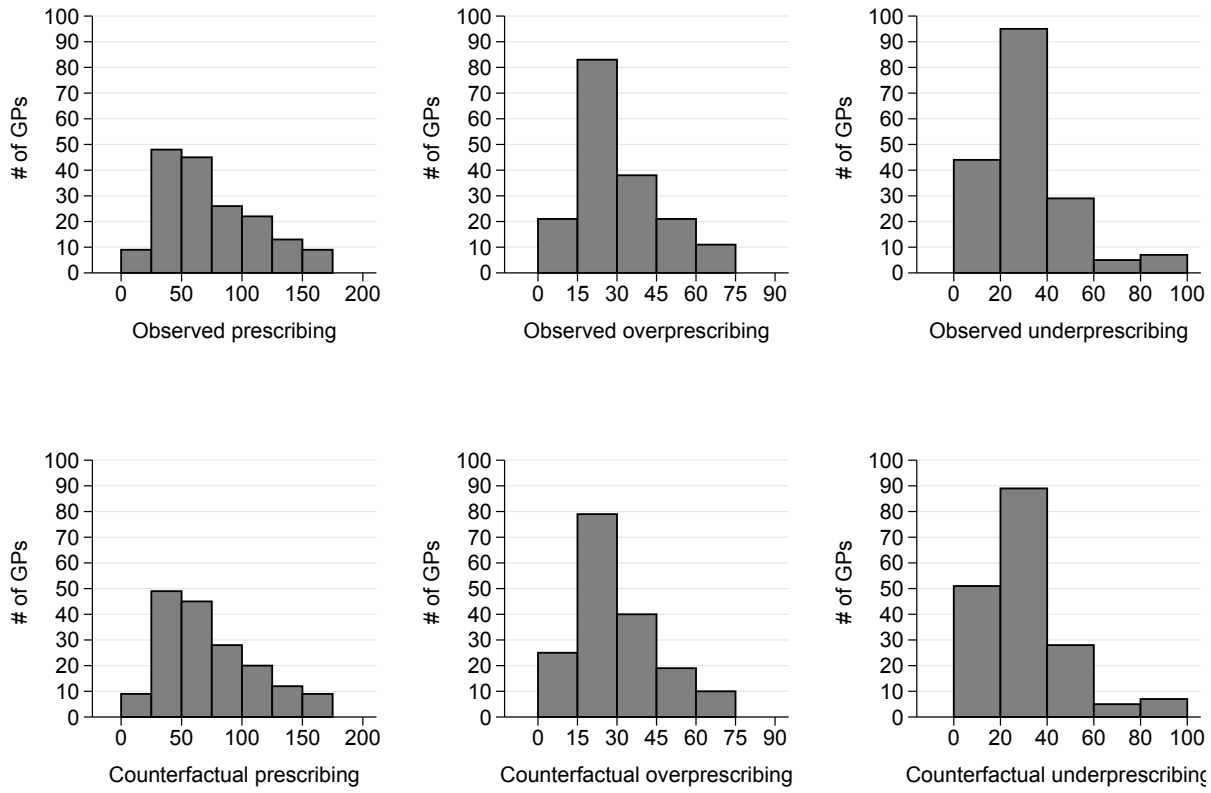


Figure 10: Observed and simulated in-sample moments

J Machine learning predictions with physician decisions

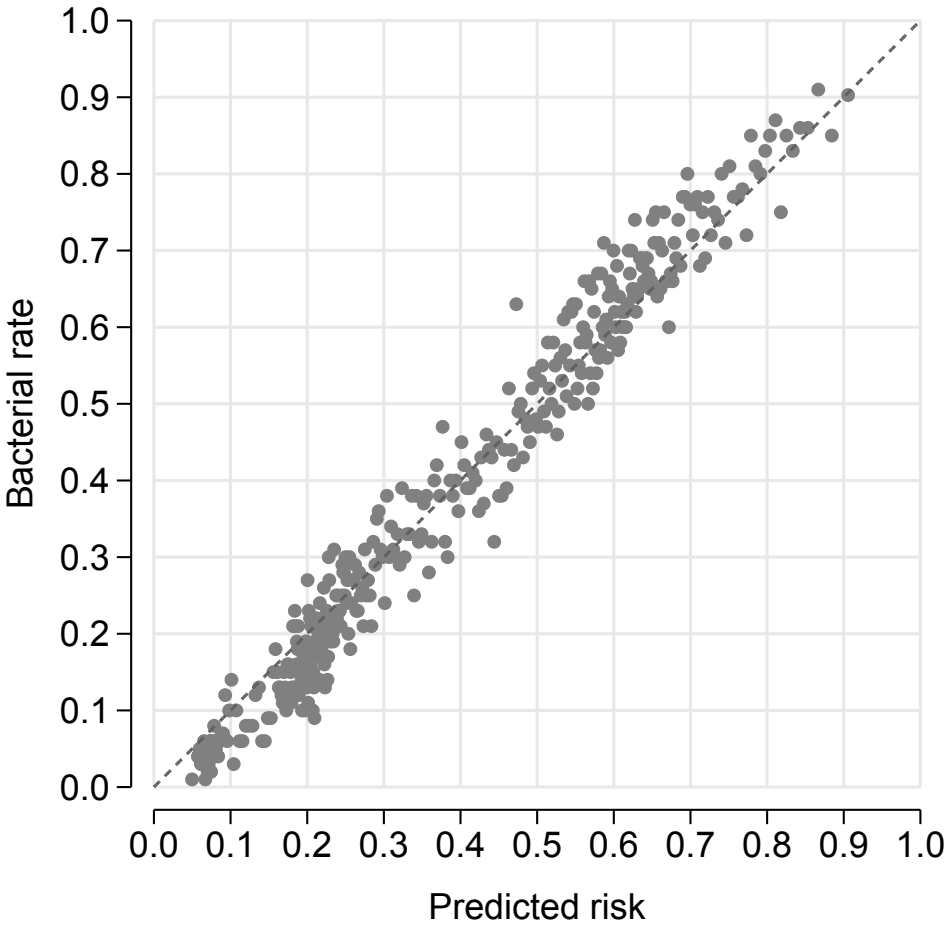
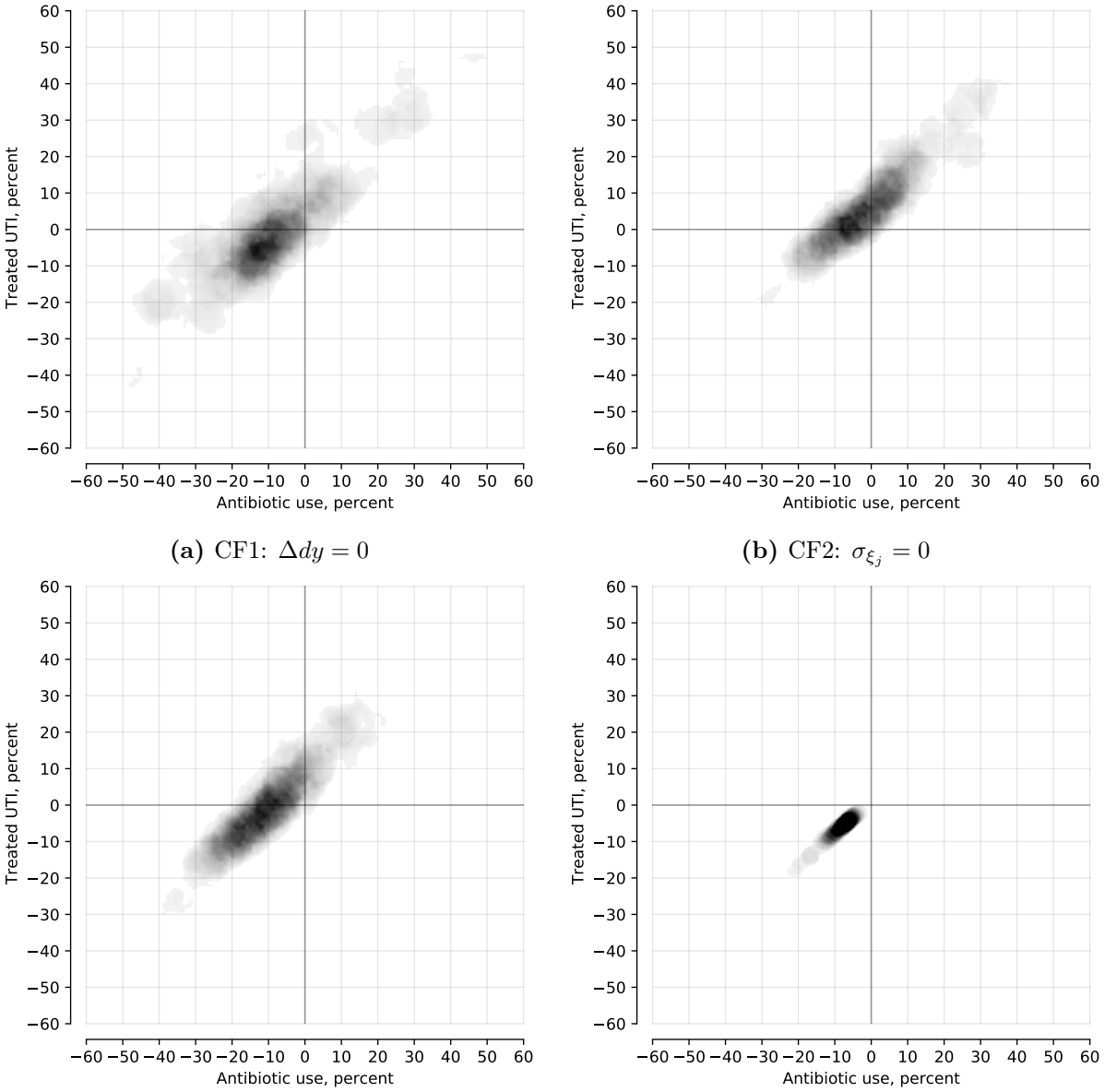


Figure 11: Bacterial rate $E[y_i]$ and machine learning predictions $m(x_i)$

Notes: Mean bacterial test outcomes relative to predicted risk of bacterial UTI. Spheres and triangles represent bins of 100 patients sorted by predicted risk.

K Clinic-level counterfactual outcomes



(a) CF1: $\Delta dy = 0$

(b) CF2: $\sigma_{\xi_j} = 0$

(c) CF3: $\sigma_{\xi_j} = 0$ and $\beta_j = \hat{\beta}_j + \kappa_1$

(d) CF4: $\beta_j = \hat{\beta}_j + \kappa_2$

Figure 12: Heat maps of clinic-level counterfactual outcomes

Notes: To ensure anonymity, the figure shows a heat map of an underlying scatter plot, with a minimum of five clinics used for local means. Darker areas represent higher clinic density.

L Robustness

L.1 Patients without treatment or tests for minimum 12 weeks

Table 9 Counterfactual policy outcomes

	ML redistribution	Provide ML-based τ_i		Incentives only
	$\Delta dy = 0$ $d^{CF} = \mathbf{1}[m(x,d) > k]$	$\sigma_{\xi_j} = 0$	$\sigma_{\xi_j} = 0$ $\beta_j = \hat{\beta}_j + \kappa_1$	$\beta_j = \hat{\beta}_j + \kappa_2$
Overall prescribing, Δd, in percent of $N_{\boldsymbol{d}} = 10,215$	-7.9 [-8.4, -7.4]	4.1 [3.6, 4.7]	-6.3 [-6.7, -5.9]	-6.3 [-6.7, -5.9]
Treated bacterial cases, Δdy, in percent of $N_{dy} = 6,062$	0	7.0 [6.6 ,7.6]	0	-4.3 [-4.5, -4.0]
Overprescribing, $\Delta d(1-y)$, in percent of $N_{d(1-y)} = 4,153$	-19.5 [-20.5 ,-18.2]	-0.2 [-1.0 ,-1.0]	-15.5 [-16.5, -14.5]	-9.3 [-9.9, -8.7]
Mean change in payoffs, $\frac{1}{J}\sum_{j=1}^{J} W_j(\boldsymbol{d}_j)$	0.066 [0.058, 0.070]	0.099 [0.094, 0.103]	0.096 [0.092, 0.100]	-0.001 [-0.002, 0.000]

Notes: This table reports changes to the status quo in percent across 175 clinics and 27,319 patients. The left column shows further relevant absolute totals. The risk threshold for prescribing in counterfactual one is $k = 0.478$ $[0.474, 0.480]$. We set $\kappa_1 = 0.033$ $[0.31, 0.36]$ to obtain $\Delta dy = 0$ in counterfactual three and set $\kappa_2 = 0.016$ $[0.15, 0.17]$ to obtain $\Delta d = -6.3$ in counterfactual four. Bootstrapped 95 percent confidence intervals in brackets.

L.2 Clinics with minimum 200 consultations

Table 10 Counterfactual policy outcomes

	ML redistribution	Provide ML-based τ_i		Incentives only
	$\Delta dy = 0$ $d^{CF} = \mathbf{1}[m(x,d) > k]$	$\sigma_{\xi_j} = 0$	$\sigma_{\xi_j} = 0$ $\beta_j = \hat{\beta}_j + \kappa_1$	$\beta_j = \hat{\beta}_j + \kappa_2$
Overall prescribing, Δd, in percent of $N_d = 8,274$	-7.9 [-8.4, -7.4]	3.3 [2.8 ,3.9]	-7.0 [-7.5 ,-6.6]	-7.0 [-7.5 ,-6.6]
Treated bacterial cases, Δdy, in percent of $N_{dy} = 5,094$	0	7.2 [6.8 ,7.7]	0	-4.9 [-5.2 ,–4.5]
Overprescribing, $\Delta d(1-y)$, in percent of $N_{d(1-y)} = 3,180$	-20.6 [-21.9, -19.2]	-2.9 [-3.9, -1.9]	-18.3 [-19.4, -17.3]	-10.5 [-11.2, -10.0]
Mean change in payoffs, $\frac{1}{J}\sum_{j=1}^{J} W_j(\boldsymbol{d}_j)$	0.063 [0.056, 0.069]	0.111 [0.106, 0.116]	0.107 [0.102, 0.111]	-0.001 [-0.002, 0.000]

Notes: Counterfactual changes relative to the status quo in percent across 68 clinics and 21,258 patients. The left column shows further relevant absolute totals. The risk threshold for prescribing in counterfactual one is $k = 0.50$ $[0.50, 0.51]$. We set $\kappa_1 = 0.036$ $[0.033, 0.038]$ to obtain $\Delta dy = 0$ in counterfactual three and set $\kappa_2 = 0.019$ $[0.018, 0.020]$ to obtain $\Delta d = -7.0$ in counterfactual four. Bootstrapped 95 percent confidence intervals in brackets.