Machine Predictions and Human Decisions with Variation in Payoffs and Skill

Michael Allan Ribers* Hannes Ullrich†

September 2021

Abstract

Human decision-making varies due to differences in information and incentives. This constitutes a substantial challenge for evaluating how machine learning predictions can improve decision outcomes. We tackle this challenge in the context of the global health policy problem of improving efficiency in antibiotic prescribing, the leading cause of antibiotic resistance. We incorporate machine learning predictions on large-scale administrative data into a treatment choice model featuring heterogeneity in patients' disease risk, physician payoffs, and diagnostic skill. We find substantial variation in the skill to diagnose bacterial urinary tract infections and in how physicians trade off the antibiotic resistance externality against curative benefits. Counterfactual policy evaluation shows that providing predictions to physicians increases efficiency but does not reduce antibiotic use. To reduce prescribing, by close to 10 percent, physicians must be incentivized. Our results highlight the importance of potential misalignment of decision-makers' and social planners' objectives in considering prediction policies.

*Postdoctoral researcher, Department Firms and Markets, DIW Berlin; Department of Economics, University of Copenhagen; and Berlin Center for Consumer Policies (BCCP). 1353 Copenhagen, Denmark. michael.ribers@econ.ku.dk

†Research Associate and Associate Professor, Department Firms and Markets, DIW Berlin; Department of Economics, University of Copenhagen; Berlin School of Economics; Berlin Center for Consumer Policies (BCCP); and CESifo. Mail address: DIW Berlin, Mohrenstraße 58, 10117 Berlin, Germany. Phone: +49 30 897 89 521, hullrich@diw.de.

1 Introduction

Machine learning methods and the increasing availability of large-scale data provide new opportunities to design welfare improving policies for a broad set of problems with prediction at their core (Kleinberg et al. 2015, 2018; Agrawal, Gans, and Goldfarb 2018; Athey 2018, Kahnemann, Sibony, and Sunstein 2021). Prominent examples include bail decisions in criminal justice, hiring, detecting social service fraud, health care provision, and labor market assistance programs. Yet, evaluating the potential improvements of prediction-based decision rules over human decisions is challenging because human decisions are outcomes of individual incentives and prediction skills. Importantly, variation in observed decisions can be a result of variation in both (Chan, Gentzkow, and Yu 2021). Without quantifying human skill and incentives, it is difficult to evaluate *ex ante* whether the potential outcomes of policies using prediction-based decision rules are due to rich data and the superior prediction quality of machine learning methods or the imposition of an altered objective function. Hence, the separation of prediction quality and decision maker objectives is crucial for designing solutions to prediction policy problems (Cowgill and Stevenson 2020).

In this paper, we separately measure human skill and payoff functions to evaluate prediction policies in the context of the global health crisis caused by increasing antibiotic resistance. This is a policy challenge where prediction is key to reduce diagnostic uncertainty; human decisions are made by highly trained experts; and decisions involve a socially relevant trade-off. Antibiotics are vital pharmaceuticals for treating bacterial infections but their use is also considered the main driver of antibiotic resistance (WHO 2014, Adda 2020).[1] We use machine learning predictions and a structural model to identify diagnostic skill and payoffs underlying antibiotic prescribing decisions. Decomposing diagnostic skill into the use of information encoded in observable data, amenable to machine learning, and the use of information not encoded in data, we evaluate counterfactual policies combining physician skill and machine learning.

We study antibiotic treatment decisions for urinary tract infections (UTI) in general

[1] Antibiotic effectiveness is decreasing due to antibiotic resistant bacteria, threatening to render simple infections, like pneumonia or infections in wounds, fatal. Annually in the US, 2.8 million antibiotic-resistant infections result in an estimated 35,000 deaths, $20 billion in health care costs, and $35 billion in lost productivity (CDC 2013/2019, Kwon and Powderly 2021).

practice in Denmark.[2] Our analysis uses a sample of 40,955 initial consultations across 189 clinics, where we combine administrative data on individual patients with their diagnostic outcomes from gold standard microbiological laboratory tests. Due to the acute nature of UTI, with its significant symptomatic burden and risk of complications, immediate antibiotic therapy is recommended for bacterial UTI.[3] Therefore, physicians make treatment decisions prior to observing test results, which only become available after several days. In our data, physicians prescribe an antibiotic in approximately 40 percent of initial consultations, which corresponds to the mean rate of bacterial infections. Observing *ex post* test outcomes enables us to evaluate prescription decisions and machine learning predictions based on the *realized* sickness state. 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.

We propose a binary treatment choice model that incorporates machine learning predictions of individual patients' binary sickness state of having a bacterial urinary tract infection. The model follows two main steps physicians take when treating patients. First, the underlying cause of illness must be assessed. Risk assessment of a bacterial cause of infection depends on a physician's diagnostic skill. Importantly, patients may vary in their prior risk of bacterial infections, which is determined by their patient type. Physicians observe patient's personal characteristics and medical histories, amenable to machine learning methods, which they can relate to the prevalence of bacterial UTI and form a signal on a patient's type. Physicians also receive a signal from clinical assessment

[2]UTI are one of the most common classes of bacterial infections. Foxman (2002) reports almost 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), with 10 percent of all women receiving antibiotic treatment for UTI (Bjerrum and Lindæk 2015). General practice accounts for 90 percent of prescriptions across Europe and for 75 percent of prescriptions in Denmark (Llor and Bjerrum 2014; Danish Ministry of Health 2017).

[3]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 lead to hospitalization with high costs. In an evidence review, Grigoryan et al. (2014) conclude that immediate antimicrobial therapy is recommended for bacterial UTI.

including patients' symptom descriptions and point-of-care tests. This information is not observed, so it cannot be used for machine learning predictions. Physicians use signals from both sources of diagnostic information to form posterior beliefs about a patient's sickness state. In the second step, given their assessment, physicians solve a trade off to decide whether to prescribe an antibiotic weighing patients' expected sickness cost while waiting for diagnostic certainty against the cost of increased antibiotic resistance.

Identification of diagnostic skill and payoffs relies on observing the joint distribution of treatment decisions and test outcomes for each clinic. Variation in decisions across clinics can come from differences in skill, preferences, or the types of patients tested. Chan, Gentzkow, and Yu (2021) show that characterizing a physician by her true positive rate (TPR) and false positive rate (FPR) identifies skill and payoffs, relying on a random assignment design to ensure the distribution of patients is plausibly the same across physicians. Without random assignment, clinics face heterogeneous patient distributions.[4] Exploiting that we observe (delayed) test outcomes, we use machine learning predictions to estimate clinic-specific patient type distributions. We provide evidence suggesting that the machine learning algorithm is consistent at the clinic-level and that patients are unlikely to be systematically tested based on unobservables relevant for the initial treatment decision. Given the patient type distribution, the two skill parameters are identified by differential sorting on risk implied by physicians' treatment decisions, machine learning predictions, and sickness realizations.

We find significant heterogeneity across clinics. The mean and standard deviation of the estimated patient type signal noise are larger than for the clinical signal noise, suggesting that physicians rely more on clinical assessment than on assessing patient type-specific disease risk. The mean estimated weight on the antibiotic resistance externality relative to patients' sickness cost is 0.43 with a standard deviation of 0.08. Hence, the mean physician weighs the social cost of increasing antibiotic resistance due to one antibiotic prescription slightly below one half the health cost for a patient with a bacterial infection while waiting for diagnostic certainty. To further document physician heterogeneity, we correlate the parameter estimates with clinic characteristics. The noise parameter on

[4]In many settings, heterogeneity in prior sickness probabilities is important. Older patients may be more likely to have pneumonia than younger patients, or women may be more likely to receive clinical assessment for mental health problems resulting in varying prior risk (Marquardt 2021). For UTI, older patients may have higher prior risk than young patients.

clinical diagnostic information is positively correlated with physician age and negatively associated with the intensity of point-of-care testing, suggesting that physicians with higher skill are younger on average and rely more on high-quality diagnostic technologies.

In counterfactual policy evaluations, we compare a threshold-based decision rule replacing physician decisions using a predicted-risk ranking with decisions based on our model of payoff-maximizing physicians. We find that improvements, a reduction of nearly 10 percent in prescribing and 25 percent in overprescribing achieved by replacing physician decisions, are not only due to improved diagnostic information via machine learning. They are also driven by imposing the policy maker's payoff weights, which differ from physician preferences. Hence, to achieve reductions in antibiotic prescribing, physicians need to be incentivized in addition to receiving improved diagnostic information. Computing gains in payoffs, we find 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 is larger than, approximately, the mean estimated physicians' weight, incentivizing physicians to reduce prescribing is necessary to maximize social welfare increases.

We contribute to several literatures. Prior work considers prediction policy problems by replacing human decisions using threshold rules on prediction-based rankings. We allow for physician discretion in a model that separates human skill and preferences, which drive decisions machine learning predictions aim to improve. Thus, we can evaluate complementarities between human skill and machine learning, a crucial step in assessing the potential of machine learning (Brynjolfsson, Wang, and McElheran 2021). Kang et al. (2013) assess how online reviews can help predict restaurants' sanitary conditions for hygiene inspections, Chalfin et al. (2016) evaluate prediction-based worker rankings for hiring and promotion decisions, and Andini et al. (2018) assess household consumption response predictions for the targeting of a tax rebate program. Kleinberg et al. (2018) evaluate potential improvements of bail decisions using decision rules based on risk predictions, assuming that risk prediction skill is homogeneous across judges. Dobbie et al. (2021) evaluate profit gains in lending decisions when replacing loan examiners by a decision rule based on machine learning predictions.

Similarly, considering health care provision, Bayati et al. (2014) use a prediction-based decision rule to reduce hospital readmissions for heart failure. Hastings, Howison, and Inman (2019) predict the riskiness of opioid prescriptions and impose constraints on

decisions based on predictions. Yelin et al. (2019) predict molecule-specific antibiotic resistance probabilities, conditional on knowing which bacteria are present, and improve prescription efficiency by redistributing molecules while holding the distribution of prescribed molecules fixed. Currie and MacLeod (2017) allow for heterogeneity in skill but assume homogeneous preferences to evaluate the counterfactual of reassigning C-sections from low- to high-risk pregnancies. Mullainathan and Obermeyer (2021) analyze the role of average physician skill in testing for heart attacks, partly driven by physician information that is unobserved and, hence, not amenable for use in machine learning. They evaluate counterfactual rules that replace physician decisions by a prediction-based ranking. We focus on the information physicians' hold and the objective function they solve to shed light on how socially desirable decision outcomes can be achieved.

Our work also relates to the literature identifying drivers of variation in health care provision. Chandra and Staiger (2007) study heterogeneity in heart attack treatment driven by specialization due to learning and local knowledge spillovers, concluding that standardization could lead to welfare losses. Chandra and Staiger (2020) find that inefficiencies in treatment for heart attacks in hospitals are driven by hospitals' expectations of their comparative advantage. They conclude that standardization and information-provision to hospitals could increase efficiency. Chan, Gentzkow, and Yu (2021) find large heterogeneity in radiologists' skill and preferences. Abaluck et al. (2021) find that physicians fail to follow decision guidelines for prescribing anticoagulants but not because they have superior diagnostic information. They propose that individualized predictions can help improve decision-making. Standardization, for example, by improving adherence to uniform decision guidelines, may be ineffective when skill is heterogeneous. Marquardt (2021) explores causes for diagnostic disparities in mental health and finds disparities are largely driven by differences in underlying disease prevalence and physician decision-making. We measure heterogeneity in physician skill and payoffs when patients are heterogeneous in disease risk within and across clinics to evaluate the potential of enhancing physicians' diagnostic information using individualized risk predictions.

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, Hung, and Lauderdale (2015) on general practitioner competition in Taiwan, Currie, Lin, and Meng (2014) and Das et al. (2016)

on financial incentives for physicians in China and India, Kwon and Jun (2015) on peer effects in Korea, and Hallsworth et al. (2016) on communication of social norms in the UK. Huang and Ullrich (2021) use exogenous variation of physician-patient assignment in Denmark and find that antibiotic prescribing style varies importantly across general practice clinics. We analyze diagnostic information and preferences as sources of decision noise, which is pivotal for designing and evaluating efficient policy measures.

The remainder of the paper is organized as follows. Section 2 presents the institutional background and data. Section 3 shows the machine learning prediction results and Section 4 inspects observed heterogeneity in prescription decisions. Section 5 develops the structural model of physician prescription choice when skill and preferences vary, discusses identification and estimation, and estimation results. Section 6 presents counterfactual policy evaluations and Section 7 concludes.

2 Danish administrative data and laboratory tests

We use Danish administrative registry data that cover a vast array of information including patient and patient household members' detailed socioeconomic data as well as antibiotic prescription histories, general practice insurance claims, and hospitalization records. Notably, the coherent use of unique personal identifiers enables us to merge registries as well as connect individuals to laboratory test results acquired from two major Danish hospitals.

2.1 The Danish health care system

Denmark has a universal and tax financed single payer health care system with general practitioners as the primary gatekeepers. Every Danish resident is allocated to a general practitioner by a list-system within a fixed geographic radius around their home address. Patients can switch physicians from their initial assignment at a small cost but most remain with their assigned general practitioner. Although general practice clinics operate as privately owned businesses, all fees for services are collectively negotiated between the national union of general practitioners and the public health insurer. Physicians do not generate earnings by prescribing drugs; these are dispensed to patients at local pharmacies. In 2012, Denmark had 2,200 general practice clinics with a median size

of just above one general practitioner per clinic (Møller Pederson, Sahl Andersen, and Søndergaard 2012). General practitioners are responsible for prescribing approximately 75 percent of the human consumed systemic antibiotics in Denmark (Danish Ministry of Health 2017).

2.2 Danish national registries

The administrative data provided by Statistics Denmark covers the entire population of Denmark between January 1, 2002, and December 31, 2012. The registries can be linked for all individuals. 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*).

The demographic data include gender, age, education, occupation, income, marriage and family status, home municipality, immigration status and place of origin, and, lastly, includes household member identifiers that allows us to identify the patients' family members and add their demographic data as well as the laboratory data and the data from the following registries. The data on systemic antibiotic prescriptions contain slightly more than 35 million purchased prescriptions. We observe the date of purchase, patient and prescribing general practice clinic identifiers, anatomical therapeutic chemical drug classification, drug name, price, indication of use, purchased package size, and defined daily dose.[5] The hospitalization data comprise all patient contacts with hospitals. The data contain observations on hospitalizations of more than 2 million unique individuals per year and include information on hospitalization admission and discharge dates, procedures performed, type of hospitalization (ambulatory, emergency, etc.), primary and secondary diagnoses, and the number of total bed days. Lastly, the insurance claims data cover all general practitioner clinic services provided to the Danish population of patients. The claims data are comprised of approximately 100 million claims per year and include physician and patient identifiers, the week of consultation, and services used. For example, claims allow us to identify pregnant women from mandatory pregnancy-associated examinations.

[5]While observing a purchase is not equivalent to observing a prescription, Koulayev, Simeonova, and Skipper (2017) document that prescription medication adherence is high in Denmark.

The administrative registers yield a vector x_{it} of 1,215 predictor variables for patient i at time t. The predictor variables can be grouped into categories including patient characteristics and test timing, patient past prescriptions, patient past laboratory test results, patient past hospitalizations, patient past general practice insurance claims, household members' past prescriptions, household members' past laboratory test results, household members' past hospitalizations, household members' past hospitalizations, and household members' past general practice insurance claims. These historic data could, in principle, be observed by the physician at the time of testing.

2.3 Microbiological laboratory test results

Individual-level clinical microbiological laboratory test results comprise the central data set of our analysis. We aim to predict a binary outcome indicating if bacteria were isolated when a urine sample was acquired from a patient consulting a general practice physician. 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 provides the bacterial species and relevant antibiotic resistances 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 is provided. A total of 2,579,617 microbiological samples are observed in the time period with submissions from both general practitioner clinics and hospitals.

2.4 Analysis sample

We apply several restrictions to define the sample for our main analysis. Urine samples constitute 477,609 samples out of which 156,694 are submitted by general practitioners, treating community acquired health conditions, the focus of our application. Bacteria were isolated in approximately one out of three urine samples, both overall and among the general practitioner submitted samples. We further restrict the number of observations in order to focus on consultations that constitute a first contact with a physician within the patient's treatment spell. We exclude test observations where the patient received a systemic antibiotic prescription or was previously tested within 28 days prior to the

sample acquisition date. Lastly, we also exclude pregnant women from our analysis as both the test decision, including mandatory tests during pregnancy, and the prescription decision cannot be compared to the typical non-emergency patient. The set of test observations used for machine learning comprises 74,511 test results for urine samples taken during initial consultations with men or non-pregnant women. We use all data from 2010 exclusively for tuning and training the machine learning algorithm. We evaluate predictions and estimate the structural model using data from 2011 and 2012, keeping all clinics with more than 100 observations.

Table 1 shows descriptives for the 189 clinics included in the estimation sample, based on the laboratory data used for estimation and the claims data comprising the population of physician claims. The sample size for each clinic has 216.7 initial consultation observations, comprising 172.1 unique patients. The number of patients per clinic, for whom a clinic is their primary general practitioner, is 3563.4, on average. Out of these, 475.9 unique patients received a urine dipstick diagnostic in the sample period, which is essentially performed at every urinary tract-related consultation in general practice. Hence, over one third of unique patients who consulted for a urinary tract-related ailment, received a laboratory test result. Laboratory test procedures last 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 0.28 across clinics. 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 to the true bacterial rate of 0.38. However, we will see that the match between initial prescriptions and bacteria infections is significantly lower than suggested by these averages.

By focusing on consultations during which physicians collected a urine sample for microbiological laboratory testing, we ensure that test outcomes are observed for all patients regardless of the physicians' prescription decisions. By conditioning on laboratory testing our results may not easily generalize to prescription occasions that did not include patient microbiological testing. However, laboratory testing for bacterial UTI is common and indicated in clinical practice when point-of-care diagnostics are inconclusive (Davenport et al. 2017). In Danish general practice clinics, evidence suggests that the decision

Table 1 Clinic descriptives for the estimation sample

	Mean	St.dev.
Microbiological laboratory data, tested patients		
Initial consultations with laboratory test, per clinic	216.7	108.8
Unique patients with laboratory test, per clinic	172.1	80.4
Initial antibiotic prescribing rate	0.39	0.13
Bacterial rate	0.38	0.09
Laboratory test result delay in days	3.1	0.28
Claims data, all patients		
Unique patients, per clinic	3563.4	1531.3
Unique patients with dipstick claim, per clinic	475.9	174.6
Unique patients with microscopy claim, per clinic	90.4	197.9
Clinics	189	

Notes: This table reports unweighted means and standard deviations over 189 clinics. The sum of observations for these clinics is 40,955.

to send cultures to the laboratory lack systematic patterns. Córdoba et al. (2018) find that neither cases typically defined as suspected complicated UTI nor rapid dipstick or microscopy test results predict the use of laboratory tests.

To see who is tested, we compare basic demographic information of all patients who received a UTI-indicated antibiotic prescription in the clinics in our sample with the subset of patients who received a laboratory test and an antibiotic prescription. For all patients with UTI-indicated prescriptions between 2010 and 2012, the mean age is 54.5, the share of female patients is 84.8 percent, the share of patients with migration background, reflecting a group of patients potentially less well known to Danish physicians, is 15 percent, and the share of patients living in a single household is 53.1 percent. In our sample of tested patients, those who received a prescription have a mean age of 49.3, are women in 84.9 percent of cases, 17.7 percent have a migration background, and 51.9 percent live in single households. Individuals with a positive bacterial test outcome have a mean age of 52.66, the share of females is 86.9 percent, 15.5 percent have a migration background, and 54.7 percent live in single households.

3 Machine learning predictions

We use prediction results from Ribers and Ullrich (2021) who train an extreme gradient boosting algorithm, proposed by Friedman, Hastie, and Tibshirani (2000) and Friedman (2001), to predict if patients suffer from bacterial UTI based on information contained in laboratory tests and a rich set of individual patient data. The sample period in 2010 serves as training data and for tuning the machine learning algorithm to optimize out-of-sample prediction quality. For prediction, we create 24 monthly out-of-sample evaluation partitions from January 2011 to December 2012 and use all data prior to the respective test partition as training data. Because we use machine learning purely for prediction, we treat it as a black box.[6]

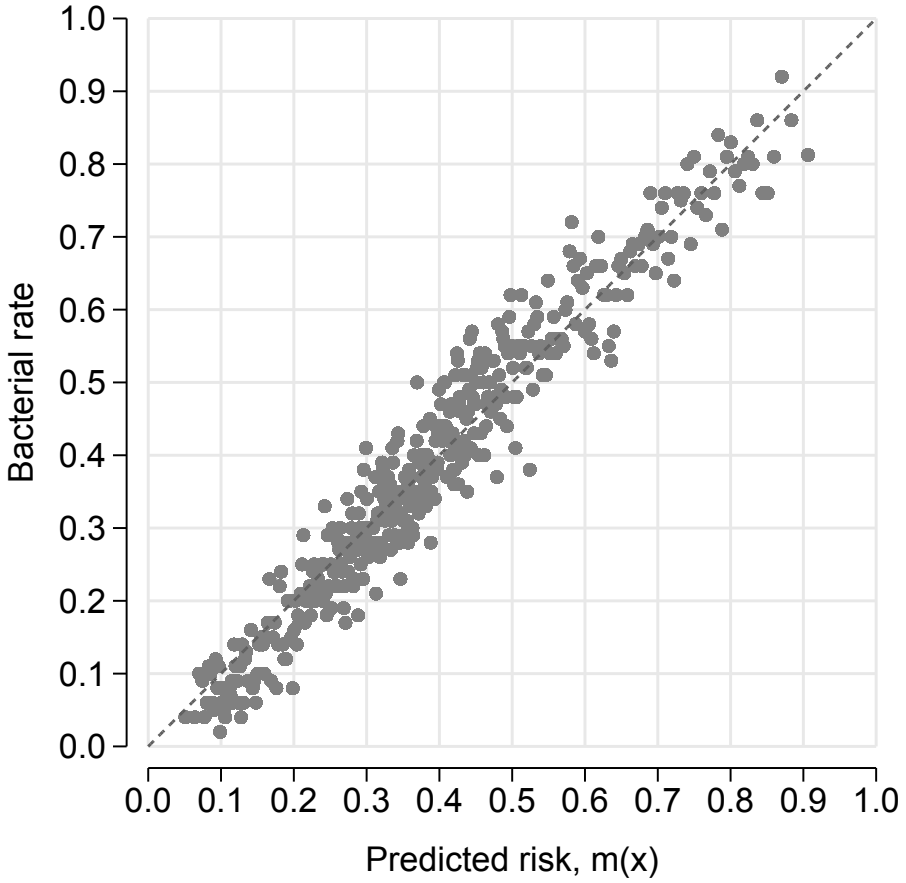


Figure 1: Laboratory test results relative to machine predictions of bacterial test outcomes. Spheres represent averages over 100 tested patients sorted by predicted risk. Based on the estimation sample holding 40,955 observations.

We illustrate the quality of the machine learning predictions $m(x)$ of the binary out-

[6]Details on the implementation using XGBoost in R are described in Ribers and Ullrich (2021).

come y, indicating the result of a microbiological test result, for our main estimation sample in Figure 1, which plots the average test results against the average out of sample predicted risk. Every sphere represents a bin containing 100 patients, with patients assigned to bins based on their predicted risk. Outcomes are close to the 45 degree line throughout the risk distribution, showing that, on average, the algorithm correctly predicts bacterial risk. A common measure of prediction quality for binary outcomes is the area under the receiver operating curve (AUC) for out-of-sample observations. Our prediction function for positive bacterial test outcomes has an AUC equal to 0.728.

4 Heterogeneity in prescription decisions

To describe heterogeneity in physicians' treatment decisions, we view the diagnostic problem through the lens of a classification problem. The receiver operating characteristic (ROC) curve is a common tool for summarizing a binary classification problem. It represents the set of all trade-offs between FPR and TPR that a given classification technology allows. For antibiotic prescribing, a false positive is considered an overprescription, that is a prescription to a person who did not suffer from a bacterial infection. A true positive is a prescription of an antibiotic to a person with a bacterial infection. At one extreme of this set every patient with a bacterial infection can be given an antibiotic, at the cost of complete overprescribing. Conversely, overprescribing can be completely avoided at the cost of giving no antibiotics to any patients.

The achievable trade-offs between these extremes depend on a physicians' skill to diagnose whether an illness is caused by a bacterial infection or not. Given this skill, the position on the associated ROC curve reflects the physician's choice of trade off between false and true positive rates. We can directly calculate physicians' false and true positives rates and plot their location in the ROC space because we observe the disease state for every tested patient irrespective of prescription decisions.

Figure 2 shows a heat map of prescription rates relative to negative and positive bacterial test outcomes. Physician prescribing relative to bacterial outcomes varies widely. Physicians' location close to the origin place is suggestive of a large weight on the antibiotic resistance externality relative to individual sickness cost, hence low levels of overprescribing but also low levels of appropriate prescribing. Physicians to the top right are

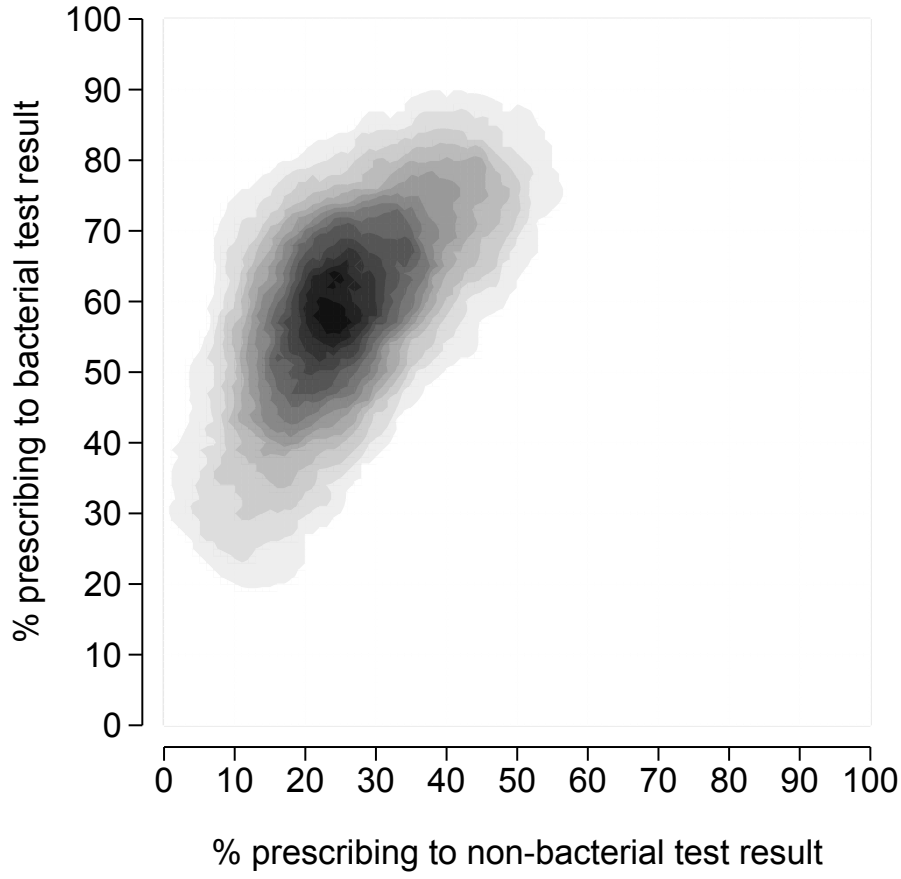


Figure 2: Heat map of physicians' true and false positive rates.

Notes: To ensure required anonymization, this figure shows a heatmap of the underlying scatter plot, with a minimum of five clinics used for local means. Darker grey represents a higher density of clinics.

more intense prescribers, which suggests a low weight on the antibiotic resistance externality relative to individual sickness cost. The plot suggests that general practitioners in Denmark do remarkably well in avoiding prescribing to non-bacterial cases while at the same time prescribing to a high share of bacterial infections. Yet, the significant variation both away from as well as parallel to the diagonal line suggest that policies may be able to improve decision outcomes by enhancing diagnostic prediction as well as incentivizing physicians to choose different trade offs.

5 Treatment choice model

We propose a formal framework to combine machine learning predictions and general practitioners' treatment choice in a model that allows for heterogeneous payoff functions and skill levels. The model follows Chan, Gentzkow, and Yu (2021) by separating the in-

dividual physicians' treatment choice problem from the preceding step of forming predictions. We depart from their model by introducing heterogeneous patient types, allowing for patients to vary in their likelihood of being sick conditional on observable characteristics. Such characteristics may be observable to physicians and can also be used to predict individual patient types using machine learning. Physician skill manifests in two dimensions: diagnosis based on observable background information and diagnosis based on unobservable clinical information available only to the physician. The distinction of these two types of diagnostic skill and the payoff function in a model of physician prescription choice provides a systematic framework to analyze the effects of counterfactual policies that improve diagnostic skill or manipulate physicians' payoff functions.

Sickness

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

$$y_i = \mathbb{1}[\nu_i > \bar{\nu}], \tag{1}$$

where $\bar{\nu}$ is a common threshold across all patients. The latent patient index, ν_i, is normally distributed with mean τ_i, the patient's type, such that

$$\nu_i \sim \mathcal{N}(\tau_i, \sigma_\nu^2). \tag{2}$$

where σ_ν^2 is common across all patients. The distribution of types vary across physicians according

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

where we make no assumptions on the distribution of τ_j or $\sigma_{\tau_j}^2$ across physicians. As patient types are only identified relative to their distance from the sickness threshold $\bar{v}$ in terms of units of σ_ν, we normalize $\bar{v} = 0$ and $\sigma_\nu = 1$.

Prediction

In clinical practice when a patient consults a physician, the physician gathers information about the patient's true sickness state from two sources. First, from patient observable

background information, that is, the patient type; and second, from instant clinical examination.[7] Physician j's signal on patient i's type is given by

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

where the variance $\sigma^2_{\xi_j}$ represents physician diagnostic skill using observable patient characteristics where low signal variance reflects high skill. Clinical examination of patient i provides a noisy signal on the patient sickness state:

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

where the variance $\sigma^2_{\eta_j}$ represents physician clinical diagnostic skill where low signal variance again reflects high skill. We assume the signals ξ_{ij} and η_{ij} are independent.[8] That is, we assume that information related to the observable patient-type specific disease prevalence is independent of information acquired via clinical assessment, a dipstick or microscopy analysis, at an in-person consultation. For example, knowledge of the difference in disease risk between an older and a younger woman is independent of the assessment of a dipstick test, which signals the presence of bacteria in the urine (John et al. 2006).

Assuming a physician knows her own skill levels and her patient type distribution, the physician posterior on sickness realization conditional on type and diagnostic signals is

$$\nu_{ij} \mid \xi_{ij}, \eta_{ij}, \tau_j, \sigma^2_{\tau_j} \sim \mathcal{N}(\mu_{ij}, \sigma^2_j), \tag{6}$$

where the posterior mean μ_{ij} and variance σ^2_j are derived in Appendix A.

[7] Physicians acquire instant clinical diagnostic information by performing either one or both of the rapid diagnostic technologies available today: urine dipstick and microscopic analysis (Davenport et al. 2017). The dipstick analysis is standard procedure but microscopic analysis requires additional equipment and specific training. Errors in interpreting dipstick results and performing microscopic analysis introduce variation in diagnostic skill in this setting, an observation documented in medical decision making more generally (Hoffrage et al. 2000, Pallin et al. 2014).

[8] The full covariance structure conditional on patient type is given by

$$\begin{pmatrix} \nu_i \\ \eta_{ij} \\ \xi_{ij} \end{pmatrix} \mid \tau_i \quad \sim \quad N\left(\begin{pmatrix} \tau_i \\ \tau_i \\ \tau_i \end{pmatrix}, \begin{bmatrix} 1 & \sigma_{\eta_j} & 0 \\ \sigma_{\eta_j} & 1+\sigma_{\eta_j} & 0 \\ 0 & 0 & \sigma_{\xi_j} \end{bmatrix} \right)$$

Treatment choice

A physician's payoff function reflects the trade off between a patient suffering a sickness cost from delaying prescribing until a test result is available and a 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 in treatment decisions, that is the decision whether to prescribe an antibiotic versus delaying or avoiding antibiotic treatment altogether.[9] 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 test results.[10] Note that we normalize the weight on the sickness cost, the first term in equation (7), to one because β can only be identified relative to sickness cost.[11] Given her posterior sickness belief for patient i, physician j maximizes expected payoff and proceeds

[9] 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. Kanilaij et al. (2020) study a similar problem for an emergency department in the US.

[10] An alternative payoff function would include the social cost of follow-up prescriptions to sick patients that 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.

[11] The parameter β reflects physicians subjective assessment of the social cost of antibiotic resistance and their weight on this cost. 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.

to prescribe an antibiotic if

$$\mathrm{E}\{\pi(1, y_i; \beta_j) \mid \mu_{ij}, \sigma_j\} > \mathrm{E}\{\pi_j(0, y_j; \beta_j) \mid \mu_{ij}, \sigma_j\} \quad \Leftrightarrow \quad \Phi\left(\frac{\mu_{ij}}{\sigma_j}\right) > \beta_j, \tag{8}$$

that is, if the expected sickness cost while waiting for the test result is larger than the social cost of prescribing. We obtain the final prescription rule:

$$d_{ij} \mid \mu_{ij}, \sigma_j = \mathbb{1}[\, \mu_{ij} > v_j^* \,], \tag{9}$$

where $v_j^* \equiv \sigma_j \Phi^{-1}(\beta_j)$. The comparative statics with respect to β_j are straight-forward. The larger a physician's weight on the antibiotic resistance externality relative to individual patients' sickness cost, the less likely she is to prescribe an antibiotic. The effect of the two skill parameters σ_{ξ_j} and σ_{η_j} is ambiguous and can both increase or decrease the probability of prescriptions depending on patient type and sickness realization.

5.1 Identification

Identification of skill and payoffs relies on one key feature in our setting, reflecting many medical treatment decision contexts. The initial treatment decision must be made before test results from diagnostic procedures become available. In our data, the waiting period for laboratory test results is typically two to four days, averaging 3.1 days. Consequently, we observe the joint distribution of prescription decisions and *ex post* sickness realizations, that is, true positives ($d = 1, y = 1$), true negatives ($d = 0, y = 0$), false positives ($d = 1, y = 0$), and false negatives ($d = 0, y = 1$) for each patient consultation.[12]

In the absence of heterogeneity in underlying patient risk, measured by population disease prevalence, Chan et al. (2021) show that skill and preferences can be identified from observing a decision maker's location in the ROC space, the TPR and FPR. In the presence of heterogeneity in patient risk, for example due to variation in disease prevalence across groups of patients, observing a physician's location in the ROC space is not sufficient to identify skill and preferences. The reason is two-fold. First, with a non-degenerate patient type distribution, the mapping of patient risk into latent type is biased due to the non-linearity of the inverse normal cumulative distribution function. Second, different patient type distributions lead to different distributions of latent sickness

[12]Physicians prescribe antibiotics prior to observing test outcomes in 38.3 percent of initial consultations, which corresponds closely to the mean bacterial rate of 37.8 percent.

realizations. As the model-implied physician ROC curve is a function of the latent sickness distribution, the ROC curve must also vary as a function of patient types.

To see how skill can be identified in the presence of heterogeneity in patient risk, assume for the moment the distribution of tested patient types at each clinic is known. Recall that, at each consultation, the physician receives a noisy signal from two sources of information: patient observables on the patient's type and clinical examination of the patient's sickness state. Having received both signals, the physician forms a risk assessment for the patient. Lower signal noise for either signal improves the physician's risk assessment since both signals are correlated with the true sickness state. But multiple skill levels can result in different ROC curves intersecting one physician's observed TPR-FPR. If we assume a physician's skill-levels are such that her model-implied ROC curve intersects the observed TPR-FPR, we can lower one skill level, shifting the ROC curve down-right, and increase the other skill level, shifting the ROC curve top-left, exactly such that a new different ROC curve still intersects the observed TPR-FPR. Varying the two skill levels in this way recovers all combinations of skill-levels that map into ROC curves intersecting the observed TPR-FPR. Combinations of skill parameters are unique in the sense that any skill level in one dimension in this set corresponds to a unique skill level in the other dimension.

To identify the physician's unique pair of skill levels, we use that signals from patient types and signals from clinical examination map onto different sorting of patients by posterior risk. At one extreme, a physician relying exclusively on type information will sort patients according to patient types only. Such a physician would have $E(d \mid \tau_i, y_i = 1) = E(d \mid \tau_i, y_i = 0)$ as no clinical information is used to distinguish patients once types are observed. A physician making use of clinical examination signals will have $E(d \mid \tau_i, y_i = 1) > E(d \mid \tau_i, y_i = 0)$. The difference $E(d \mid \tau_i, y_i = 1) - E(d \mid \tau_i, y_i = 0)$ is increasing in patient type signal noise and decreasing in clinical examination signal noise. Ultimately, patients are split in two groups conditional on sickness realization and type information is used only to sort patients within each group.

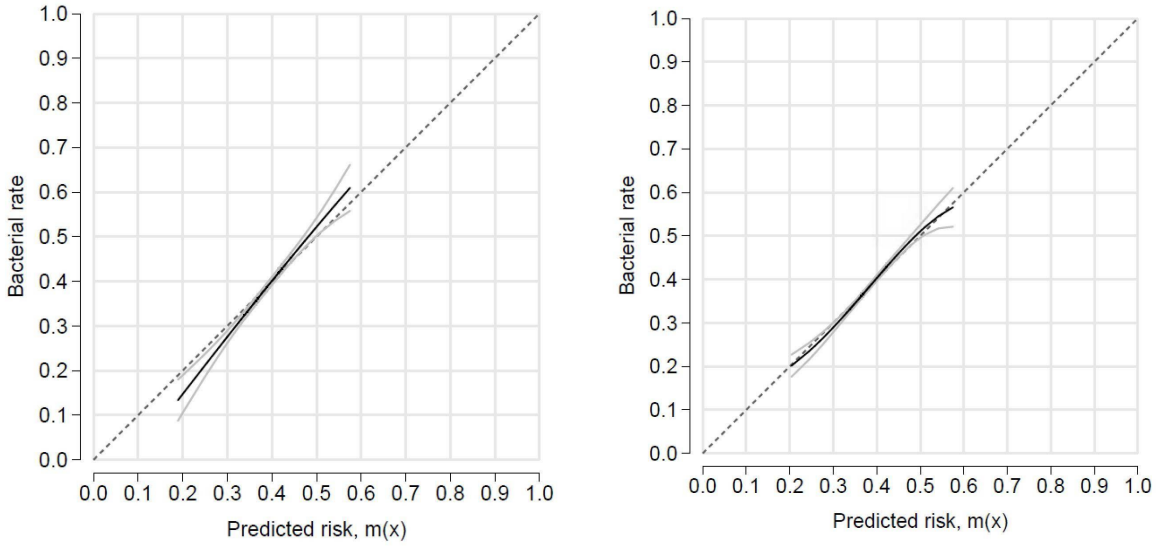
Given physician skill levels, the preference parameter β_j is identified by the physician's observed location along the ROC curve ranging from never prescribing $(\text{TPR}, \text{FPR}) = (0, 0)$ for $\beta = 1$ to always prescribing $(\text{TPR}, \text{FPR}) = (1, 1)$ for $\beta = 0$.

Clinic-specific patient type distribution

As the distribution of patient types is *a priori* unknown and determined by physicians' propensity to test patients, we use machine learning predictions for each clinic's set of tested patients to infer their distribution of types. Therefore, it is crucial for identification that predictions are not systematically biased. Bias would be introduced if physicians systematically relied on unobservables to select whom to test.[13] We cannot test directly whether the machine learning predictions consistently recover patient types because we only observe a single sickness realization for each patient. However, if machine learning predicted patient types are consistent, the physician-level sum of sickness realizations follows the Poisson-Binomial distribution, which is the distribution of the sum of successes for non-identical independent Bernoulli trials. We test whether the physician-level sum of the sickness indicator y concurs with the corresponding set of patient type predictions. Using a two-sided test of the observed number of patients with a bacterial infection, we cannot reject machine learning patient type predictions for 154 out of 189 physicians, that is 82 percent, at the five percent level. Our estimation results are robust to leaving out physicians for whom this test rejects that patient type predictions are consistent.

To inspect consistency of machine learning predictions graphically, Figure 3 shows the fit of clinic-level mean bacterial rates and mean predicted risk using fractional polynomials. Figure 3a is based on all 189 clinics. Figure 3b is based on the sample of 154 clinics for which we cannot reject machine learning patient type predictions at the five percent level. Both lines are centered at 0.4 and follow the diagonal, suggesting that risk predictions do not exhibit bias at the clinic-level. For the restricted sample of clinics in

[13]For example, the test decision may depend on unobservables correlated with the sickness state. Holm et al. (2021) analyze clinical management of UTI in Denmark and document that symptom assessment of UTI in general practice is highly noisy in general. Bias 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 important capacity constraints in Danish general practice (Kristiansen and Sheng 2020). 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 (2021) provide evidence that patients do not sort into general practice clinics based on antibiotic prescribing style.



(a) Full sample

(b) Restricted sample

Figure 3: Clinic-specific bacterial rates and machine learning predicted risk

Notes: Fractional polynomial fit of the observed clinic-specific bacterial rates and mean machine learning predicted risk. Grey lines show 95 confidence intervals. Panel 3b is based on the sample of 154 clinics for which we cannot reject machine learning patient type predictions at the five percent level for the two-sided test of the Poisson-binomial distributed sum of sick patients.

Figure 3b the fitted line almost perfectly overlaps with the diagonal.

Selective testing based on unobservables

We investigate several plausible channels for selection on unobservables in test decisions. Rapid point-of-care diagnostics can help identify some bacterial strains but not all.[14] Abaluck et al. (2016) show that variation in practice style drives test yield, the bacterial rate of tested. We inspect the balance of the types of bacteria found in tests conditional on clinic-specific test yield. If skill varies with the signal physicians receive from point of care tests and this simultaneously affects the decision to obtain a laboratory diagnostic, we should see differing rates of bacterial strains across clinics with different test yield. For

[14]The ability to detect different bacterial strains varies across in-clinic diagnostics such as dipstick and microscopic analysis. Nitrite dipstick diagnostics can help detect so-called gram negative organisms of which E.coli is the most common for UTI. However, nitrite dipstick diagnostics can not detect bacterial strains such as Enterococci, Staphylococci, and other organisms that do not convert nitrate to nitrite. Finally, virulence and severity of disease can vary between bacterial strains (Flores-Mireles et al. 2015).

example, if there is variation in the ability to spot E.coli bacteria in point of care tests, then we would see variation in the share of E.coli prevalence across clinics with varying test yield. We split clinics into two groups, above and below the median bacterial rate. The top panel in Table 2 reports the observation-weighted shares of bacterial strains for these two groups. The small differences across bacteria suggest little systematic selection into testing based on informative signals from point of care tests. We repeat the analysis by splitting the sample into physicians above and below the mean of their residual test yield conditional on predicted risk. We find similarly small differences across bacteria.

Physicians may also vary in their knowledge about the prevalence of antibiotic resistance for specific patients or in the population. Such knowledge may influence the decision use laboratory diagnostics. For the five molecules commonly used to treat UTI, the bottom panel in Table 2 shows little variation in resistance rates between clinics with low and high test yield, $E_j[y]$. Comparing clinics below and above the median of deviations in mean test yield and predicted risk, $E_j[y] - E_j[m(x)]$, these differences are even smaller, except for Nitrofurantoin, providing further evidence that information at the point of care, unobservable to us, is unlikely a driver of selection into laboratory testing.

As an alternative to resistance test results, revealed after the initial consultation, the choice of molecule at the initial consultation may also be informative of physicians' expectation about potential bacterial species or resistances. The top panel of Table 3 shows that differences in the shares of molecules prescribed between clinics with high and low test yield, as well as between clinics above and below the median of deviations in means, are small and mostly not statistically significantly different from zero. The bottom panel of Table 3 shows differences in clinics' use of diagnostics. The number of laboratory test observations and usage intensity of point of care tests such as urine dipsticks and microscopic analysis do not differ between clinics with low versus high bacterial rates in tested patients.

In Tables 7 and 8 in the Appendix B, we find similar patterns with smaller differences when restricting the sample to the 154 clinics for which we cannot reject that the patient type distribution estimated using machine learning predictions generated the observed sickness realizations. Overall, while we find small differences between clinics with differing test yield, these findings do not seem to support a strong selection mechanism for the tested patient pool, in particular after conditioning on predicted risk $m(x)$. Hence, noisi-

Table 2 Balance of types of bacterial infection causes

	$E_j[y]$			$E_j[y] - E_j[m(x)]$		
	Low	High	Δ	Low	High	Δ
Bacterial species isolated						
E.coli	0.70	0.72	*0.026*	0.70	0.72	*0.022*
	(0.07)	(0.06)	(0.009)	(0.07)	(0.06)	(0.010)
E.faecalis	0.07	0.06	*-0.016*	0.07	0.06	*-0.011*
	(0.04)	(0.03)	(0.005)	(0.04)	(0.03)	(0.005)
K. pneumoniae	0.04	0.04	0.006	0.04	0.04	*-0.007*
	(0.03)	(0.03)	(0.004)	(0.03)	(0.02)	(0.004)
S. agalactiae	0.05	0.04	*-0.010*	0.04	0.04	-0.007
	(0.04)	(0.02)	(0.005)	(0.04)	(0.03)	(0.005)
Others	0.15	0.15	-0.006	0.15	0.15	0.002
	(0.05)	(0.04)	(0.007)	(0.05)	(0.05)	(0.007)
Molecule-specific resistance						
Mecillinam (J01CA11)	0.24	0.20	*-0.038*	0.23	0.21	*-0.019*
	(0.06)	(0.04)	(0.008)	(0.06)	(0.06)	(0.008)
Trimethoprim (J01EA01)	0.23	0.21	*-0.018*	0.23	0.21	-0.014
	(0.07)	(0.05)	(0.009)	(0.07)	(0.05)	(0.009)
Sulfamethizole (J01EB02)	0.37	0.34	*-0.027*	0.36	0.35	-0.009
	(0.08)	(0.06)	(0.010)	(0.08)	(0.06)	(0.010)
Ciprofloxacin (J01MA02)	0.14	0.12	*-0.018*	0.14	0.13	-0.007
	(0.06)	(0.04)	(0.008)	(0.06)	(0.05)	(0.008)
Nitrofurantoin (J01XE01)	0.06	0.06	-0.003	0.07	0.05	*-0.012*
	(0.04)	(0.03)	(0.005)	(0.04)	(0.03)	(0.005)
Number of cases	20072	20883		21877	19078	
Number of clinics	95	94		101	88	

Notes: This table reports mean bacterial species and resistance rates for clinics above and below the median 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. 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.

ness in the decision to test may be large, for example influenced by variation in logistical and organizational constraints in the management of general practitioner clinics.

Table 3 Balance of molecules initially prescribed and use of diagnostics

	$E_j[y]$			$E_j[y] - E_j[m(x)]$		
	Low	High	Δ	Low	High	Δ
Molecule initially prescribed						
Pivmecillinam (J01CA08)	0.54	0.58	0.044	0.55	0.58	0.028
	(0.18)	(0.19)	(0.027)	(0.18)	(0.19)	(0.027)
Trimethoprim (J01EA01)	0.03	0.02	-0.003	0.03	0.02	-0.005
	(0.04)	(0.03)	(0.005)	(0.04)	(0.03)	(0.005)
Sulfamethizole (J01EB02)	0.25	0.27	0.014	0.25	0.27	0.020
	(0.14)	(0.19)	(0.025)	(0.16)	(0.18)	(0.025)
Ciprofloxacin (J01MA02)	0.05	0.03	*-0.018*	0.05	0.03	*-0.020*
	(0.05)	(0.03)	(0.006)	(0.05)	(0.03)	(0.006)
Nitrofurantoin (J01XE01)	0.04	0.04	0.0003	0.04	0.04	-0.003
	(0.04)	(0.05)	(0.007)	(0.05)	(0.05)	(0.007)
Number of prescriptions	6467	9228		7476	8219	
Use of diagnostics						
Test observations	261.8	280	18.1	273.6	268.0	-5.6
	(124.3)	(145.2)	(19.67)	(140.3)	(130.1)	(19.68)
Urine dipsticks per patient	0.24	0.24	0.000	0.25	0.23	-0.013
	(0.13)	(0.10)	(0.017)	(0.12)	(0.11)	(0.017)
Microscopy per patient	0.03	0.04	0.017	0.03	0.04	0.005
	(0.06)	(0.08)	(0.011)	(0.08)	(0.07)	(0.011)
Number of clinics	95	94		101	88	

Notes: This table reports mean prescribed molecules and clinics' usage intensity of diagnostics for clinics above and below the median 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. 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.

5.2 Estimation

We estimate the model parameters in two steps, separately for each clinic. In the first step, we estimate the clinic-specific patient type distribution using machine learning predicted risk conditional on observables, $m(x_i)$. We recover predicted individual patient types by inversion as $\tau_i = \Phi^{-1}(m(x_i))$, where $\Phi(\cdot)$ is the standard normal CDF. Let $\mathcal{I}_j$ be the set of patients consulting in clinic j. Using the distribution of individual patient types, we estimate the structural parameters $\hat{\tau}_j = \mathrm{E}_{i\in\mathcal{I}_j}\{\tau_i\}$ and $\hat{\sigma}^2_{\tau_j} = \mathrm{Var}_{i\in\mathcal{I}_j}\{\tau_i\}$.

In the second step, we estimate the model by simulated maximum likelihood using observed data on prescription decisions, d_{ij}, sickness realizations, y_i, and individual patient types, τ_i, recovered from machine learning predictions, $m(x_i)$. The simulated likelihood contribution from a single observation is

$$\mathcal{L}_{ij}\Big(d_{ij} \mid \Theta_j, y_i, \tau_i, \hat{\tau}_j, \hat{\sigma}^2_{\tau_j}\Big) = \begin{cases} \Pr\Big\{\mu_{ij} \;>\; v^*_j \mid \Theta_j, y_i, \tau_i, \hat{\tau}_j, \hat{\sigma}^2_{\tau_j}\Big\} & \text{if } d_{ij} = 1 \\ \Pr\Big\{\mu_{ij} \;\leq\; v^*_j, \mid \Theta_j, y_i, \tau_i, \hat{\tau}_j, \hat{\sigma}^2_{\tau_j}\Big\} & \text{if } d_{ij} = 0, \end{cases} \tag{10}$$

where $\Theta = \{\beta_j, \sigma_{\xi_j}, \sigma_{\eta_j}\}$ and μ_{ij} is computed based on simulated ξ_{ij} and η_{ij} conditional on physician priors, predicted patient type, and observed y_i using the distributional assumptions in equations (2), (4), and (5). Appendix C explains the procedure to simulate the probabilities in equation (10).

The joint log-likelihood for the two-step procedure is given by

$$\ell_j(\boldsymbol{d}_j \mid \Theta_j, \boldsymbol{y}_j, \boldsymbol{\tau}_j) = \sum_{i\in\mathcal{I}_j} \log\Big(\mathcal{L}_{ij}\Big(d_{ij} \mid \Theta_j, y_i, \tau_i, \hat{\tau}_j, \hat{\sigma}^2_{\tau_j}\Big)\Big), \tag{11}$$

where $\boldsymbol{d}_j$, $\boldsymbol{y}_j$, and $\boldsymbol{\tau}_j$ are vectors over prescription decisions, sickness realizations, and patient types for all patients of clinic j. Physician skill and preferences are recovered for clinic j from

$$\hat{\Theta}_j = \arg\min_{\Theta_j} \ell_j(\boldsymbol{d}_j \mid \Theta_j, \boldsymbol{y}_j, \boldsymbol{\tau}_j) \tag{12}$$

Estimating $\hat{\Theta}_j$ independently for every physician, we recover the distribution of physician skill and payoff parameters.[15]

[15] We estimate each set of parameters by maximizing the simulated likelihood function using 1,000 Modified Latin Hypercube Sampling draws proposed by Hess, Train, and Polak (2006) and a quasi-Newton method. The Nelder-Mead method and simulated annealing, a global optimization algorithm, yield nearly identical results.

5.3 Estimation results

Table 4 reports the means and standard deviations of $\hat{\Theta}_j$. The means of the noise parameters for patient type (σ_{ξ_j}) and clinical diagnostic information (σ_{η_j}) are large, 5.54 and 2.14. The mean of σ_{ξ_j} is markedly larger than σ_{η_j}, implying that physicians 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 σ_{ξ_j} and σ_{η_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 0.08 reflects relevant heterogeneity in how physicians solve this trade-off.

Table 4 Distribution of parameter estimates

	Mean	(St.dev.)
Type signal noise, σ_{ξ_j}	5.54	(4.19)
Diagnostic signal noise, σ_{η_j}	2.14	(1.26)
Payoff function parameter, β_j	0.43	(0.08)

Notes: This table reports the means and standard deviations of the distribution of parameter estimates over 189 clinics. The model is estimated separately for each clinic.

Figures 6 to 8 in Appendix D show the distributions of parameter estimates. For anonymization we show heat maps and do not report values in areas containing fewer than five clinics. The distribution of the clinical diagnostic skill parameter σ_{η_j} is concentrated in the area between 1 and 3. The noise parameter σ_{ξ_j} measuring the extent to which physicians make use of patient type information is more dispersed between 1 and 7. The more pronounced concentration of σ_{η_j} estimates suggests that the majority of physicians makes use of clinical diagnostic information even if significant heterogeneity remains. The large estimate of σ_{ξ_j} on average suggests that providing machine learning predictions can

improve physician information. In particular, we find a relevant number of physicians with very large σ_{ξ_j} estimates. In Figure 6, physicians with estimated $\sigma_{\xi_j} > 5$ account for 40% of all physicians. This group does not appear to 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 7 and 8 do not show a systematic relationship between the estimated payoff weights and both noise parameters. Figures 9 to 11 in Appendix E show projections of the physician-level parameter estimates, sorted by mean parameter values, and their 95% confidence intervals computed by bootstrapping at the physician-level. The variance of skill parameter estimates increases in the size of the estimates. The estimates of σ_{η_j} have tight confidence intervals at lower values and throughout from below, and wider upper confidence bounds.

We also investigate how well the model fits the data. Figure 12 in Appendix G shows the distributions of the observed mean, over-, and underprescribing rates as well as their simulated in-sample counterparts based on the parameter estimates. The simulated distributions closely resemble the observed data.

5.4 Observed heterogeneity

To investigate potential sources of heterogeneity across clinics, we correlate parameter estimates with observable clinic characteristics. 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 113 out of the total of 189 clinics.

Linear regression results of the parameter estimates on clinic characteristics in Table 5 show several interesting patterns. For the estimates of the noise parameter for clinical diagnostic information, $\hat{\sigma}_{\eta_j}$, higher noise is associated with age of physicians working at the clinic. Noise is negatively associated with the share of female physicians. Higher skill is also associated with higher propensity to perform point-of-care microscopic analyses of urine samples, which reflect more extensive training for diagnosing bacterial urinary tract infections. The number of physicians per clinic does not appear to be correlated with the clinical signal noise estimates. Several narratives would be consistent with these correlations. For example, older physicians might rely more on their clinical experience

and personal knowledge of patients than on extensive diagnostic tests. Alternatively, they may be less likely to purchase new diagnostic equipment and rely on existing tools they are accustomed to.

Table 5 Correlation of clinical skill estimate with clinic and physician characteristics

	Linear regression for clinical signal noise $\hat{\sigma}_{\eta_j}$				
$N = 113$	(1)	(2)	(3)	(4)	(5)
Mean number of physicians	0.04				0.07
	[-0.26,0.34]				[-0.31,0.44]
Mean age of physicians	**1.41**				**1.65**
	[0.09,2.72]				[0.24,3.05]
Share of female physicians	-0.14				-0.12
	[-0.37,0.08]				[-0.36,0.12]
Dipstick tests per physician		0.16		0.27	0.25
		[-0.16,0.49]		[-0.02, 0.55]	[-0.05,0.55]
Microscopy analyses per physician			-0.05	**-0.07**	**-0.09**
			[-0.12,0.01]	[-0.15,-0.003]	[-0.18,-0.003]
Patients per physician		0.03	0.19	0.01	-0.22
		[-0.51,0.56]	[-0.31,0.70]	[-0.53, 0.55]	[-0.79,0.35]
R^2	0.07	0.01	0.01	0.03	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 4. 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. Coefficients in bold are statistically significantly different from zero at the five percent level.

We report results of the same regressions for the patient type signal noise parameter $\hat{\sigma}_{\xi_j}$ and the payoff parameter $\hat{\beta}_j$ in Tables 9 and 10 in Appendix F. None of the coefficients are significantly different from zero for these two parameters. For $\hat{\beta}_j$, the coefficient are close to zero except for the number of physicians in a clinic which shows a negative correlation. Larger clinics may place a lower weight on the antibiotic resistance externality relative to the benefit of antibiotic treatment. The correlation of $\hat{\sigma}_{\xi_j}$ with physician age is positive, similar to the result for $\hat{\sigma}_{\eta_j}$. The number of dipstick and microscopy analyses performed at a clinic, reflecting the number of UTI patients as well as the diagnostic effort exerted, are negatively correlated with patient type signal noise.

6 Counterfactual policy evaluation

We consider three counterfactual policies to illustrate the importance of incentives and information for prediction policy design. In particular, we show how measuring incentives and information is essential for understanding how machine learning predictions can be used to achieve socially desirable policy outcomes. For example, heterogeneous payoff weights β_j may differ from socially desirable β^S. If so, physicians' adherence to prediction-based decision rules must be enforced or incentivized. Without acknowledging heterogeneity and diverging incentives, improvements seen due to prediction-based decision rules may be mistakenly ascribed to improved prediction technology instead of imposition of socially desirable objectives. This is important because policy approaches relying on improved predictive information while allowing full discretion to expert decision makers may not be sufficient to achieve specific policy objectives, requiring alternative policy approaches.

We evaluate policies by comparing the set of physicians' counterfactual decision outcomes $\boldsymbol{d}^{CF}$ with the observed physician prescriptions $\boldsymbol{d}$. 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

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

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 6 shows policy outcomes for three counterfactual interventions.[16] The first counterfactual serves as a benchmark, in which we reproduce a prescription rule as in Huang, Ribers, and Ullrich (2021). This type of policy follows the prior literature evaluating machine learning predictions (Bayati et al. 2014; Chalfin et al. 2016; Kleinberg et al.

[16] We report nearly identical results in Table 12 in Appendix H for the subset of clinics for which we cannot reject that the machine learning predictions are consistent.

2018; Ribers and Ullrich 2019; Yelin et al. 2019; Hastings, Howison, and Inman 2020). To make use of physician information without resorting to our choice model, in the first counterfactual we include the physician decision as a predictor in the machine learning algorithm. The decision rule relies on the assumption that human discretion can be overruled or that decision makers adhere perfectly to prediction-based prescription rules, where the risk threshold leading to an antibiotic prescription is defined by the policy maker. 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. Here, β_j is typically unknown. Therefore, we follow the literature and focus on a solution that guarantees a welfare increase to the social planner for all $\beta^S \in [0, 1]$ by maximizing reductions in antibiotic use without reducing the number of prescriptions to bacterial infections, $\Delta \boldsymbol{dy} = 0$. This counterfactual policy reduces overall prescribing by 9.30 percent (1,460 prescriptions) and overprescribing by 23.32 percent (1,462 prescriptions) while, by construction, the change in the number of prescriptions to bacterial infections is zero. In this counterfactual, the mechanism to achieve these improved outcomes, improved information or manipulating payoffs, is unknown.

Table 6 Counterfactual policy outcomes

	ML redistribution	Provide ML-based τ_i	
	$\Delta dy \stackrel{!}{=} 0$	$\sigma_{\xi_j} = 0$	$\sigma_{\xi_j} = 0$ $\beta_j = \hat{\beta}_j + \kappa$
Overall prescribing, Δd, in percent of $N_{\boldsymbol{d}} = 15,695$	-9.30 [-10.56, -8.58]	1.61 [1.16, 2.24]	-9.47 [-10.24, -8.54]
Treated bacterial cases, Δdy, in percent of $N_{dy} = 9,435$	0	8.16 [7.48, 8.84]	0
Overprescribing, $\Delta d\neg y$, in percent of $N_{d\neg y} = 6,267$	-23.32 [-26.45, -21.52]	-8.26 [-9.18, -6.85]	-23.7 [-25.42, -21.67]
Mean change in payoffs, $\frac{1}{J}\sum_{j=1}^{J} W_j(\boldsymbol{d}_j)$	0.077	0.107	0.102

Notes: This table reports changes to the status quo in percent across 189 clinics. The left column shows absolute totals. In counterfactual three, we set $\kappa = 0.035$ $[0.031, 0.036]$ to obtain $\Delta dy = 0$. Bootstrapped 95 percent confidence intervals in brackets.

The second counterfactual policy 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, clinical diagnostic skill and the payoff function parameter are held fixed. We find that overall prescribing slightly increases by 1.61 percent (253 prescriptions) and overprescribing decreases by 8.26 percent (518 prescriptions). The number of treated bacterial infections increases by 8.16 percent (770 prescriptions). Hence, the improved and more precise information on patient type leads to more efficient prescribing but fails to achieve the aim of reducing antibiotic prescribing overall. Comparing these results with the redistribution policy, 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, we again provide physicians with the machine learning prediction of type τ_i for every patient and hold their clinical diagnostic information fixed. However, in addition, we increase the payoff parameter β_j by a constant κ to maximize the reduction in overall prescribing while holding the number of prescriptions to bacterial infections fixed. We define the counterfactual payoff parameter as $\beta_j = \hat{\beta}_j + \kappa$, where setting $\kappa = 0.035\,[0.031, 0.036]$ attains an overall reduction in prescribing by 9.47 percent (1,486 prescriptions) and in overprescribing by 23.7 percent (1,485 prescriptions). Such an intervention can be interpreted as a nudge or an antibiotic tax that shifts the relative weights on the social cost of increasing antibiotic resistance and an individual patients' sickness cost of delayed antibiotic treatment. Compared with the redistribution policy in counterfactual one, the reductions in (over)prescribing are now the same. Recalling that the mean estimated β_j is 0.43, the results achieved by the redistribution policy with no physician discretion, hence, would be achievable by providing physicians the machine learning predictions and reducing β_j by, on average, 8.1 percent.

These results illustrate the usefulness 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 are 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. The framework we consider allows for interventions in which the two channels, providing machine learning predictions to experts and incentivizing social behavior, can be evaluated as complements.

Table 6 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 decisions $\boldsymbol{d}_j$.

Payoff gains are largest for the counterfactual policy which provides patient type information and smallest for the policy leaving no discretion to physicians. The policy increasing physicians' weights on the externality and providing predictions lies in between. Figure 4 shows the distribution of clinic-level changes in payoffs for the three counterfactuals.

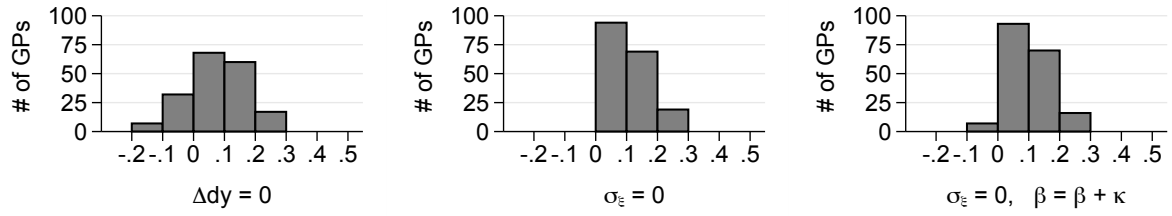


Figure 4: Distribution of counterfactual changes in payoffs

For the first redistribution policy, allowing for no physician discretion, a sizable share of clinics has negative changes in payoffs. All clinics benefit from information provision in the second counterfactual. When incentivized, for example by imposing a tax or fee on antibiotic prescriptions, some clinics' payoffs decrease even when given additional diagnostic information. These results are intuitive given that information provision strictly improves efficiency while changing physicians weights, or decisions altogether, conflicts with their revealed preferences. This conflict may make the implementation of a redistribution policy difficult and require a more sophisticated policy design.

Finally, taking the perspective of a social planner, we evaluate welfare effects to provide a ranking of counterfactual policies based not only on reported counts of outcomes but on gains in payoffs given social preferences. We calculate the welfare effects for the

three counterfactual policies 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 set of decisions $\boldsymbol{d}$.

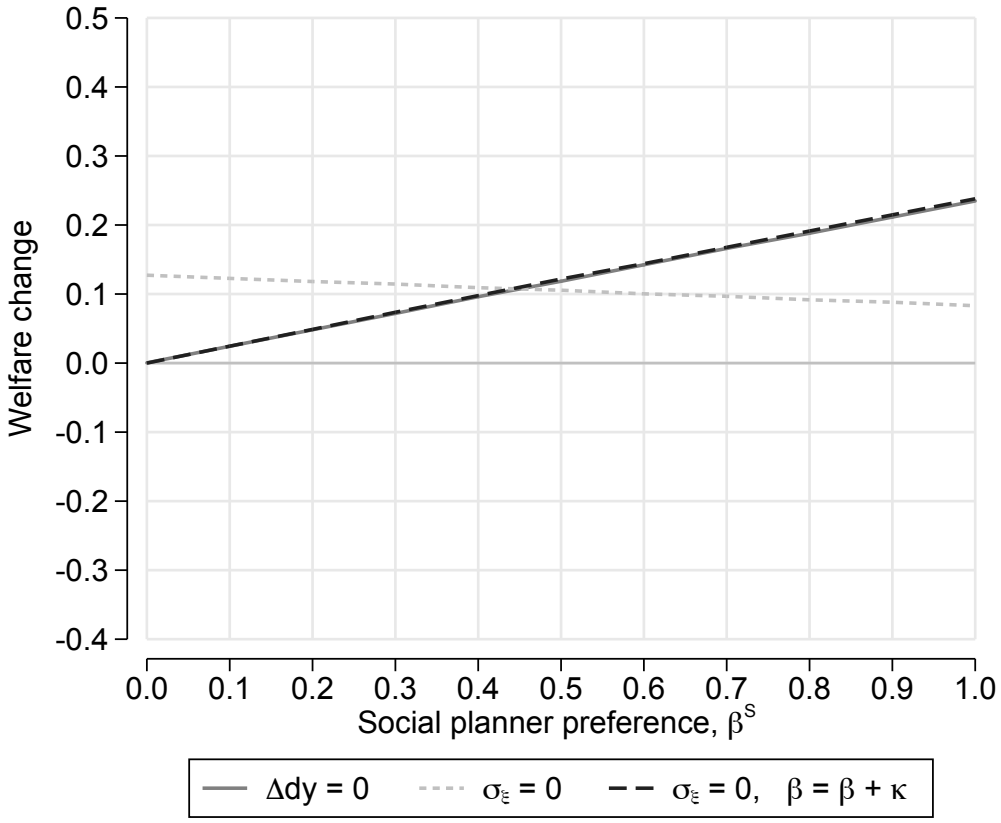


Figure 5: Counterfactual welfare effects for social planner $\beta^S \in [0, 1]$.

Figure 5 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. All counterfactuals generate positive aggregate payoff gains. However, if the social planner's weight on the externality is small, below approximately the average estimated physician β_j of 0.44, the policy providing physicians with machine learning predictions and leaving them full discretion maximizes the social planner's payoff gains. This is intuitive because, if society places a sufficiently large weight on the externality, then it is even worthwhile to reduce prescriptions to some bacterial infections in exchange for a reduction of over-prescribing. If the social planner's weight on the externality is larger than the average

estimated physicians' weight, then physicians must be incentivized to reduce prescribing by increasing their weight on the externality, even if they are given the machine learning predictions as additional diagnostic information.

7 Robustness checks

To assess the robustness of our results and qualitative conclusions, we estimate the model and perform counterfactual policy evaluations for three alternative estimation samples. All results are presented in Tables 11 to 16 in Appendix H.

We first reduce the sample from 189 to 154 clinics, for which we reject, at the five percent level, that our machine learning predictions generated the observed sickness realizations based on the Poisson-binomial test described in Section 5.1. For these clinics, the type distribution estimated based on machine learning predictions are expected to have smallest potential clinic-level bias. Table 11 shows the distribution of parameter estimates is very close to the results for the full sample. The means of both signal noise parameters are slightly smaller but the mean of the payoff parameter β_j is the same at 0.43. The counterfactual policy results in Table 12 are also very close to the main results, with confidence intervals largely overlapping.

In the second robustness check, 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 of laboratory testing. If this period is too short, we might include patients who are 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 slightly lower signal noise estimates and nearly the same mean for the payoff parameter β_j at 0.41. The counterfactual policy changes in Table 14 are slightly smaller but lead to the same qualitative conclusion that physicians must be incentivized to reduce overall antibiotic prescribing.

Finally, we extend our sample to include physicians with at least 50 test observations, compared to the minimum of 100 test observations in the full sample. For a significantly larger number of 280 clinics, we obtain similar parameter estimates, reported in Table

15, with a moderately larger mean estimate of the clinical signal noise parameter σ_{η_j} of 2.45 and a slightly larger mean estimate for β_j of 0.44. The counterfactual policy results in Table 16 are nearly the same as for the main sample.

8 Conclusion

We show how policies enabled by machine learning predictions can be evaluated when humans hold decision-relevant information. It is typically difficult to determine whether such 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. Therefore, it is important to develop model-based tools to evaluate potential implementations *ex ante*.

The setting we consider for this analysis, antibiotic prescribing for suspected urinary tract infections, resembles many situations in primary care provision and expert decision problems more generally. While we consider and exploit some idiosyncrasies of the setting in general practice, our analysis provides a more generally applicable framework for the evaluation of machine learning in data-rich environments. Whether and how our analysis can be helpful in alternative settings depends on measurement of the target outcome and on the availability of data to allow consistent predictions at the level of decision makers. However, we document the general result that comparing machine learning predictions to human decisions is insufficient to learn about how to design solutions to prediction policy problems. We provide evidence that information generated by machine learning predictions and held by human experts can be complements. We further show that information alone may not be sufficient to achieve socially desirable policy goals, requiring policies that combine information-provision and incentivization.

Several important avenues for further research specific to the context of antibiotic prescribing remain. It would also be worthwhile to attempt encoding 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. Further research is needed to better understand experts' potential behavioral reactions

to the introduction of prediction tools. An interesting avenue for further research in this regard would be to consider designing the information provided 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 information and expertise.

While we consider one specific type of medical diagnostic and treatment problem, our results indicate the potential of using machine learning predictions in many relevant health care situations. Data availability and the quality of prediction algorithms are improving at a rapid pace in and beyond health care. The rate at which such technologies will be more broadly adopted and productively exploited, however, will depend on the kind and quality of human expertise it can complement. If our findings are suggestive more generally, human experts are far from being replaced. Instead, investment in human capital is key to induce welfare-improving technological progress.

Acknowledgements

We benefited from helpful comments by Jason Abaluck, David Chan, Tomaso Duso, Daniel Ershov, Mogens Fosgerau, Matthew Gentzkow, Qing Gong, 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, Jann Spiess, Florian Szücs, Christoph Wolf, and participants at the 13th Digital Economics Conference in Toulouse, the 6th International Conference on Computational Social Science, the Econometric Society World Congress 2020, the Electronic Health Economics Colloquium, the IO Committee Meeting of the German Economic Association, the European Quant Marketing Seminar, the 10th Annual Conference of the American Society of Health Economists, as well as in seminars at DIW Berlin, University of Copenhagen, University of Konstanz, and University Paris-Sud. We are deeply indebted to Lars Bjerrum and Gloria Cristina Cordoba Currea for providing their expertise on diagnostics and antibiotic prescribing in Danish general practice as well as to Jenny Dahl Knudsen, Sidsel Kyst, and Rolf Magnus Arpi for enabling us to work with the microbiological laboratory data. We thank Adam Lederer for proofreading.

Funding

Financial support from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 802450) is gratefully acknowledged.

Conflict of interest

Michael Allan Ribers and Hannes Ullrich have no engagements or affiliations to disclose in relation to the context of this research.

References

[1] Abaluck, Jason, Leila Agha, Chris Kabrhel, Ali Raja, and Arjun Venkatesh (2016), "The determinants of productivity in medical testing: intensity and allocation of care," *American Economic Review*, 106 (12), 3730-3764.

[2] Abaluck, Jason, Leila Agha, David Chan, Daniel Singer, and Diana Zhu (2021), Fixing misallocation with guidelines: awareness vs. adherence, NBER Working Paper No. w27467.

[3] Adda, Jérôme (2020), "Preventing the spread of antibiotic resistance," *AEA Papers and Proceedings*, 110, 255-259.

[4] Agrawal, Ajay, Joshua Gans, and Avi Goldfarb (2018), *Prediction Machines: The Simple Economics of Artificial Intelligence*, Harvard Business Press.

[5] Andini, Monica, Emanuele Ciania, Guido de Blasio, Alessio D'Ignazio, and Viola Salvestrini (2018), "Targeting with machine learning: an application to a tax rebate program in Italy," *Journal of Economic Behavior and Organization*, 156, 86-102.

[6] Athey, Susan (2018), "The impact of machine learning on economics," in *The Economics of Artificial Intelligence: An Agenda* ed. Ajay K. Agrawal, Joshua Gans, and Avi Goldfarb, University of Chicago Press.

[7] Bayati, Mohsen, Mark Braverman, Michael Gillam, Karen M. Mack, George Ruiz, Mark S. Smith, and Eric Horvitz (2014), "Data-driven decisions for reducing read-

missions for heart failure: general methodology and case study," *PLoS ONE*, 9 (10), e109264.

[8] Bennett, Daniel, Che-Lun Hung, and Tsai-Ling Lauderdale (2015), "Health care competition and antibiotic use in Taiwan," *The Journal of Industrial Economics*, 63 (2), 371-393.

[9] Brynjolfsson, Erik, Wang Jin, and Kristina McElheran (2021), The power of prediction: predictive analytics, workplace complements, and business performance, working paper.

[10] Cassidy, Rachel, and Charles F. Manski (2019), "Tuberculosis diagnosis and treatment under uncertainty," *Proceedings of the National Academy of Sciences*, 116 (46), 22990-22997.

[11] CDC (2013), Antibiotic resistance threats in the United States, `https://www.cdc.gov/drugresistance/threat-report-2013/pdf/ar-threats-2013-508.pdf`, accessed 4/2/2019.

[12] CDC (2019), Antibiotic resistance threats in the United States, `http://dx.doi.org/10.15620/cdc:82532`, accessed 8/5/2021.

[13] Chalfin, Aaron, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan (2016), "Productivity and selection of human capital with machine learning," *American Economic Review*, 106 (5), 124-127.

[14] Chan, David C., Matthew Gentzkow, and Chuan Yu (2021), "Selection with variation in diagnostic skill: evidence from radiologists," *Quarterly Journal of Economics*, forthcoming.

[15] Chandra, Amitabh and Douglas O. Staiger (2007), "Productivity spillovers in health care: evidence from the treatment of heart attacks," *Journal of Political Economy*, 115(1), 103-140.

[16] Chandra, Amitabh and Douglas O. Staiger (2020), "Identifying sources of inefficiency in healthcare," *Quarterly Journal of Economics*, 135(2), 785-843.

[17] Córdoba, Gloria, Anne Holm, Tina Møller Sørensen, Volkert Siersma, Håkon Sandholdt, Marjukka Makela, Niels Frimodt-Møller, and Lars Bjerrum (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), 1-7.

[18] Cowgill, Bo, and Megan T. Stevenson (2020), "Algorithmic social engineering," *AEA Papers & Proceedings*, 110.

[19] Currie, Janet, Wanchuan Lin, and Juanjuan Meng (2014), "Addressing antibiotic abuse in China: an experimental audit study," *Journal of Development Economics*, 110, 39-51.

[20] Currie, Janet and W. Bentley MacLeod (2017), "Diagnosing expertise: human capital, decision making, and performance among physicians," *Journal of Labor Economics*, 35 (1), 1-43.

[21] Davenport, Michael, Kathleen E. Mach, Linda M. Dairiki Shortliffe, Niaz Banaei, Tza-Huei Wang, and Joseph C. Liao (2017), "New and developing diagnostic technologies for urinary tract infections," *Nature Reviews Urology*, 14 (5), 296.

[22] Danish Health and Medicines Authority (2013), Guidelines on prescribing antibiotics for physicians and others in Denmark, November 2013, Copenhagen.

[23] Danish Ministry of Health (2017), National handlingsplan for antibiotika til mennesker. Tre målbare mål for en reduktion af antibiotikaforbruget frem mod 2020.

[24] Das, Jishnu, Alaka Holla, Aakash Mohpal, and Karthik Muralidharan (2016), "Quality and accountability in health care delivery: audit-study evidence from primary care in India," *American Economic Review*, 106 (12), 3765-3799.

[25] Dobbie, Will, Andres Liberman, Daniel Paravisini, and Vikram Pathania (2021), "Measuring bias in consumer lending," *Review of Economic Studies*, forthcoming.

[26] Ferry, Sven A., Stig E. Holm, Hans Stenlund, Rolf Lundholm, and Tor J. Monsen (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.

[27] Flores-Mireles, Ana L., Jennifer N. Walker, Michael Caparon, and Scott J. Hultgren (2015), "Urinary tract infections: epidemiology, mechanisms of infection and treatment options," *Nature Reviews Microbiology*, 13, 269-284.

[28] Foxman, Betsy (2002), "Epidemiology of urinary tract infections: incidence, morbidity, and economic costs," *The American Journal of Medicine*, 113 (1), Suppl. 1, 5-13.

[29] Friedman, Jerome H., Trevor Hastie, and Robert Tibshirani (2000), "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)," *Annals of Statistics*, 28 (2), 337-407.

[30] Friedman, Jerome H. (2001), "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, 29 (5), 1189-1232.

[31] Grigoryan, Larissa, Trautner, Barbara W., and Kalpana Gupta (2014), "Diagnosis and management of urinary tract infections in the outpatient setting: a review," *JAMA*, Vol. 312, No. 16, 1677-1684.

[32] Hallsworth, Michael, Tim Chadborn, Anna Sallis, Michael Sanders, Daniel Berry, Felix Greaves, Lara Clements, and Sally C. Davies (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.

[33] Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009), *The Elements of statistical learning: data mining, inference, and prediction*, 2nd Edition, New York: Springer.

[34] Hastings, Justine S., Mark Howison, and Sarah E. Inman (2020), "Predicting high-risk opioid prescriptions before they are given," *Proceedings of the National Academy of Sciences*, 117(4), 1917-23.

[35] Hess, Stephane, Kenneth E. Train, and John W. Polak (2006), "On the use of a Modified Latin Hypercube Sampling (MLHS) method in the estimation of a mixed logit model for vehicle choice," *Transportation Research Part B: Methodological*, 40 (2), 147-163.

[36] Hoffrage, Ulrich, Samuel Lindsey, Ralph Hertwig, and Gerd Gigerenzer (2000), "Communicating statistical information," *Science*, 290 (5500), 2261-2262.

[37] Holm, Anne, Volkert Dirk Siersma, and Gloria Cristina Cordoba Currea (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.

[38] Huang, Shan and Hannes Ullrich (2021), Physician effects in antibiotic prescribing: evidence from physician exits, DIW Discussion Paper Nr. 1958.

[39] Huang, Shan, Michael A. Ribers, and Hannes Ullrich (2021), The value of data for prediction policy problems: evidence from antibiotic prescribing, DIW Discussion Paper Nr. 1939.

[40] Jit, Mark, Dorothy Hui Lin Ng, Nantasit Luangasanatip, Frank Sandmann, Katherine E. Atkins, Julie V. Robotham, and Koen B. Pouwels (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), 1-14.

[41] Kahneman, Daniel, Olivier Sibony, Cass R. Sunstein (2021), *Noise: a flaw in human judgment*, Little, Brown.

[42] Kang, Jun Seok, Polina Kuznetsova, Michael Luca, and Yejin Choi (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*, 1443-1448.

[43] Kanjilal, Sanjat, Michael Oberst, Sooraj Boominathan, Helen Zhou, David C. Hooper, and David Sontag (2020), "A decision algorithm to promote outpatient antimicrobial stewardship for uncomplicated urinary tract infection," *Science Translational Medicine*, 12 (568).

[44] Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer (2015), "Prediction policy problems," *American Economic Review*, 105 (5), 491-495.

[45] Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan (2018), "Human decisions and machine predictions," *Quarterly Journal of Economics*, 133 (1), 237-293.

[46] Koulayev, Sergei, Emilia Simeonova, and Niels Skipper (2017), "Can physicians affect patient adherence with medication?," *Health Economics*, 26, 779–794.

[47] Kristiansen, Ida Lykke and Yanying (Sophie) Sheng (2020), "Doctor Who - Can the Doctor-Patient Match Reduce the Socioeconomic Gradient in Health?," working paper.

[48] Kwon, Illoong and Daesung Jun (2015), "Information disclosure and peer effects in the use of antibiotics," *Journal of Health Economics*, 42, 1-16.

[49] Kwon, Jennie H. and William G. Powderly (2021), "The post-antibiotic era is here," *Science*, 373 (6554), 471-471.

[50] Laxminarayan, Ramanan, Adriano Duse, Chand Wattal, Anita K.M. Zaidi, Heiman F.L. Wertheim, Nithima Sumpradit, Erika Vlieghe, Gabriel Levy Hara, Ian M. Gould, Herman Goossens, Christina Greko, Anthony D. So, Maryam Bigdeli, Göran Tomson, Will Woodhouse, Eva Ombaka, Arturo Quizhpe Peralta, Farah Naz Qamar, Fatima Mir, Sam Kariuki, Zulfiqar A. Bhutta, Anthony Coates, Richard Bergstrom, Gerard D. Wright, Eric D. Brown, and Otto Cars (2013), "Antibiotic resistance – the need for global solutions," *The Lancet Infectious Diseases Commission*, 1-42.

[51] Llor, Carl and Lars Bjerrum (2014), "Antimicrobial resistance: risk associated with antibiotic overuse and initiatives to reduce the problem," *Therapeutic Advances in Drug Safety*, 5 (6), 229-241.

[52] Marquardt, Kelli (2021), Mis(sed) diagnosis: physician decision making and ADHD, working paper.

[53] Manski, Charles F. (2021), Probabilistic prediction for binary treatment choice: with focus on personalized medicine, working paper.

[54] Mullainathan, Sendhil and Ziad Obermeyer (2021), Diagnosing physician error: a machine learning approach to low-value health care, working paper.

[55] Pallin, Daniel J., Clare Ronan, Kamaneh Montazeri, Katherine Wai, Allen Gold, Siddharth Parmar, and Jeremiah D. Schuur (2014), "Urinalysis in acute care of adults: pitfalls in testing and interpreting results," *Open Forum Infectious Diseases*, 1 (1), ofu019.

[56] Møller Pedersen, Kjeld, John Sahl Andersen, and Jens Søndergaard (2012), "General practice and primary health care in Denmark," *Journal of the American Board of Family Medicine*, 25 (Suppl 1), S34-S38.

[57] Ribers, Michael A. and Hannes Ullrich (2019), "Battling antibiotic resistance: can machine learning improve prescribing?," DIW Discussion Paper Nr. 1803.

[58] Ribers, Michael A. and Hannes Ullrich (2021), "Battling antibiotic resistance: can machine learning improve prescribing?," working paper.

[59] St John, Andrew, James C. Boyd, Andrew J. Lowes, and Christopher P. Price (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.

[60] World Health Organization (2014), Antimicrobial Resistance: Global Report on Surveillance, Geneva, Switzerland.

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

Appendices

Appendix A Derivation of posterior distribution of ν_{ij}

Physician j has a normal prior on patient types, $N(\hat{\tau}_j, \hat{\sigma}_{\tau_j})$, where we assume rational expectations such that

$$\hat{\tau}_j = \frac{1}{N_j} \sum_{i \in \mathcal{I}_j} \tau_i \tag{15}$$

and

$$\hat{\sigma}^2_{\tau_j} = \frac{1}{N_j - 1} \sum_{i \in \mathcal{I}_j} (\tau_i - \hat{\tau}_j)^2 \tag{16}$$

where N_j is the number of patients at clinic j and types $\tau_i = \Phi^{-1}(m(x_i))$ follow from machine learning predictions where $\Phi(\cdot)$ is the standard normal CDF. Given the normal prior and a normal likelihood of type signals, ξ_{ij}, as stated in equation (4), physicians form posterior beliefs about patient i's type

$$\tau_{ij} \mid \xi_{ij}, \hat{\tau}_j, \hat{\sigma}^2_{\tau_j} \sim N(\tilde{\tau}_{ij}, \tilde{\sigma}^2_j) \tag{17}$$

where

$$\tilde{\tau}_{ij} = \frac{\hat{\tau}_j \sigma^2_{\xi_j} + \xi_{ij} \hat{\sigma}^2_{\tau_j}}{\sigma^2_{\xi_j} + \hat{\sigma}^2_{\tau_j}} \qquad \text{and} \qquad \tilde{\sigma}^2_j = \frac{\sigma^2_{\xi} \hat{\sigma}^2_{\tau_j}}{\sigma^2_{\xi_j} + \hat{\sigma}^2_{\tau_j}}. \tag{18}$$

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

$$\nu_{ij} \mid \xi_{ij}, \hat{\tau}_j, \hat{\sigma}^2_{\tau_j} \sim \int_{-\infty}^{\infty} \underbrace{\phi(t \mid \tilde{\tau}_j, \tilde{\sigma}^2_j) \phi(\nu_i \mid t, \sigma^2_v)}_{P(\text{type } t|\xi) P(v|\text{type } t)} \mathrm{d}t = N(\tilde{\tau}_j, \tilde{\sigma}^2_j + \sigma^2_v), \tag{19}$$

which is also the physician's prior before observing the patient's signal η_{ij} from clinical assessment. With a normal likelihood of signals in equation (5), the physician's posterior on the patient's sickness realization is

$$\nu_{ij} \mid \xi_{ij}, \eta_{ij}, \hat{\tau}_j, \hat{\sigma}^2_{\tau_j} \sim N(\mu_{ij}, \sigma^2_j) \tag{20}$$

where

$$\mu_{ij} = \frac{\tilde{\tau}_j \sigma^2_{\eta_j} + \eta_{ij} (\tilde{\sigma}^2_j + \sigma^2_v)}{\sigma^2_{\eta_j} + (\tilde{\sigma}^2_j + \sigma^2_v)} \qquad \text{and} \qquad \sigma^2_j = \frac{\sigma^2_{\eta_j} (\tilde{\sigma}^2_j + \sigma^2_v)}{\sigma^2_{\eta_j} + (\tilde{\sigma}^2_j + \sigma^2_v)} \ . \tag{21}$$

Appendix B Balance table for subsample

Table 7 Balance of types of bacterial infection causes

	$E_j[y]$			$E_j[y] - E_j[m(x)]$		
	Low	High	Δ	Low	High	Δ
Bacterial species isolated						
E.coli	0.70	0.72	*0.023*	0.70	0.71	0.017
	(0.07)	(0.06)	(0.010)	(0.06)	(0.07)	(0.010)
E.faecalis	0.07	0.06	-0.009	0.07	0.06	-0.004
	(0.04)	(0.03)	(0.005)	(0.04)	(0.03)	(0.005)
K. pneumoniae	0.04	0.05	*0.009*	0.05	0.04	-0.007
	(0.03)	(0.03)	(0.004)	(0.03)	(0.02)	(0.004)
S. agalactiae	0.04	0.03	-0.010	0.04	0.04	-0.005
	(0.05)	(0.02)	(0.006)	(0.04)	(0.03)	(0.006)
Others	0.16	0.14	-0.013	0.15	0.15	-0.001
	(0.05)	(0.04)	(0.008)	(0.05)	(0.05)	(0.008)
Molecule-specific resistance						
Mecillinam (J01CA11)	0.23	0.20	*-0.038*	0.22	0.21	-0.013
	(0.06)	(0.05)	(0.008)	(0.06)	(0.06)	(0.009)
Trimethoprim (J01EA01)	0.23	0.21	*-0.018*	0.23	0.21	*-0.019*
	(0.06)	(0.05)	(0.009)	(0.06)	(0.05)	(0.009)
Sulfamethizole (J01EB02)	0.37	0.34	*-0.027*	0.36	0.35	-0.009
	(0.08)	(0.06)	(0.010)	(0.08)	(0.06)	(0.011)
Ciprofloxacin (J01MA02)	0.15	0.12	*-0.018*	0.14	0.13	-0.012
	(0.06)	(0.05)	(0.008)	(0.06)	(0.05)	(0.009)
Nitrofurantoin (J01XE01)	0.06	0.06	-0.003	0.07	0.06	-0.010
	(0.04)	(0.03)	(0.005)	(0.04)	(0.03)	(0.006)
Number of cases	15668	17502		17124	16046	
Number of clinics	76	79		81	74	

Notes: This table reports mean bacterial species and resistance rates for clinics above and below the median of mean bacterial rates $E_j[y]$ and mean residuals $E_j[y - m(x)]$. Sample includes the 154 clinics for which we cannot reject the machine learning patient type distribution at the five percent level. Physician-level means and standard deviations are weighted by physician-level numbers of observations. 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 8 Balance of molecules initially prescribed and use of diagnostics

	$E_j[y]$			$E_j[y] - E_j[m(x)]$		
	Low	High	Δ	Low	High	Δ
Molecule initially prescribed						
Pivmecillinam (J01CA08)	0.55	0.60	0.051	0.56	0.60	0.037
	(0.18)	(0.19)	(0.030)	(0.18)	(0.19)	(0.030)
Trimethoprim (J01EA01)	0.03	0.02	-0.002	0.03	0.02	-0.005
	(0.03)	(0.03)	(0.005)	(0.04)	(0.03)	(0.005)
Sulfamethizole (J01EB02)	0.26	0.25	-0.010	0.26	0.25	-0.006
	(0.15)	(0.18)	(0.027)	(0.17)	(0.17)	(0.027)
Ciprofloxacin (J01MA02)	0.05	0.03	*-0.016*	0.05	0.03	*-0.018*
	(0.05)	(0.03)	(0.007)	(0.05)	(0.03)	(0.006)
Nitrofurantoin (J01XE01)	0.03	0.04	0.005	0.04	0.04	-0.0001
	(0.03)	(0.04)	(0.005)	(0.04)	(0.03)	(0.006)
Number of prescriptions	5354	7612		6292	6674	
Use of diagnostics						
Test observations	259.6	284.6	25.0	273.1	272.5	-0.7
	(129.2)	(154.2)	(22.81)	(149.0)	(130.1)	(23.00)
Urine dipsticks per patient	0.24	0.24	-0.009	0.25	0.23	-0.023
	(0.12)	(0.09)	(0.017)	(0.11)	(0.10)	(0.017)
Microscopy per patient	0.03	0.05	0.017	0.04	0.04	0.002
	(0.07)	(0.09)	(0.012)	(0.08)	(0.07)	(0.013)
Number of clinics	76	79		81	74	

Notes: This table reports mean prescribed molecules and clinics' usage intensity of diagnostics for clinics above and below the median of mean bacterial rates $E_j[y]$ and mean deviations $E_j[y] - E_j[m(x)]$. Sample includes the 154 clinics for which we cannot reject the machine learning patient type distribution at the five percent level. Physician-level means and standard deviations are weighted by physician-level numbers of observations. 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.

Appendix C Simulated choice probabilities

To estimate parameters Θ_j using maximum likelihood estimation, we simulate choice probabilities for patient i and physician j as follows:

1. Compute $\hat{\tau}_j$ and $\hat{\sigma}^2_{\tau_j}$ following Appendix A.

2. Given $m(x)$ and y, draw simulated sickness realizations

$$\nu_i^r \sim \mathcal{N}(\tau_i, 1 \mid y_i = \mathbb{1}[\nu_i^r > 0]).$$

   and signals

$$\xi_{ij}^r \sim \mathcal{N}(\tau_i, \sigma^2_{\xi_{ij}})$$
$$\eta_{ij}^r \sim \mathcal{N}(\nu_i^r, \sigma^2_{\eta})$$

   for physician parameters σ_{ξ_j} and σ_{η_j} and compute posterior μ_{ij}^r and σ_j^r following Appendix A.

3. Compute the expected payoffs:

$$EU_{ij}^r(d) = \mathrm{E}\{\pi(d, y_i, \beta_j) \mid \mu_{ij}^r, \sigma_j^r\} = \begin{cases} -\Phi\left(\frac{\mu_{ij}^r}{\sigma_j^r}\right) & \text{if } d = 0 \\ -\beta_j & \text{if } d = 1. \end{cases}$$

4. We compute choice probabilities using the Logit-smoothed Accept-Reject simulator with smoothing parameter $\lambda = 0.01$.[17] Inserting the utilities into the logit formula yields:

$$S_{ij0}^r = \frac{e^{EU_{ij}^r(0)/\lambda}}{e^{EU_{ij}^r(0)/\lambda} + e^{EU_{ij}^r(1)/\lambda}} \qquad \text{and} \qquad S_{ij1}^r = \frac{e^{EU_{ij}^r(1)/\lambda}}{e^{EU_{ij}^r(0)/\lambda} + e^{EU_{ij}^r(1)/\lambda}}.$$

5. Repeat steps two to four R times letting r takes the values 1 through R.

6. The simulated choice probabilities are obtained by averaging over simulations:

$$\hat{P}_{ij0} = \frac{1}{R}\sum_{r=1}^{R} S_{ij0}^r \qquad \text{and} \qquad \hat{P}_{ij1} = 1 - \hat{P}_{ij0} = \frac{1}{R}\sum_{r=1}^{R}\left(1 - S_{ij0}^r\right) = \frac{1}{R}\sum_{r=1}^{R} S_{ij1}^r$$

[17] Computing the simulated choice probabilities straightforward by

$$\hat{P}_{ij0} = \frac{1}{R}\sum_{r=1}^{R} \mathbb{1}[EU_{ij}^r(0) > EU_{ij}^r(1)] \qquad \text{and} \qquad \hat{P}_{ij1} = \frac{1}{R}\sum_{r=1}^{R} \mathbb{1}[EU_{ij}^r(0) \leq EU_{ij}^r(1)]$$

yields step-functions of σ_{ξ_j}, σ_{η_j} and β_j, which are problematic to minimize.

Appendix D Model parameter estimates

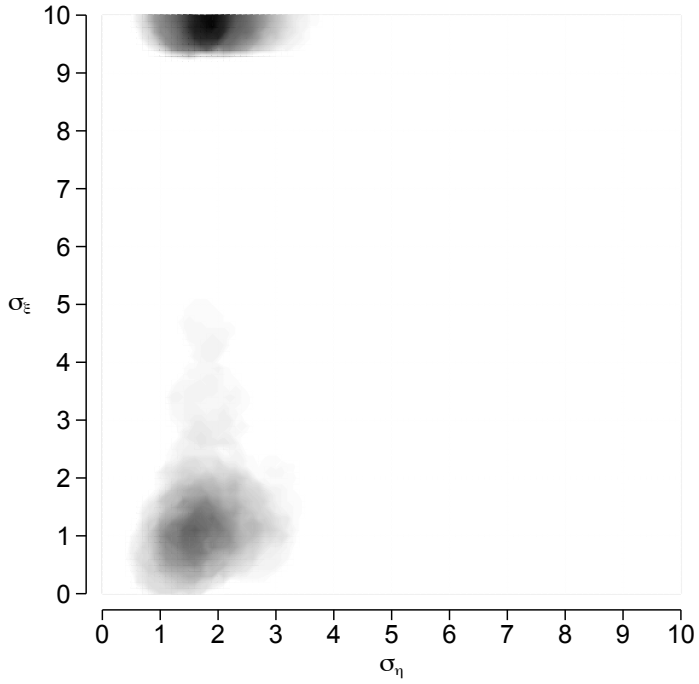


Figure 6: Heat map of physician-level estimates for σ_ξ and σ_η

Notes: To ensure anonymity, the heat map covers only areas of five physicians or more.

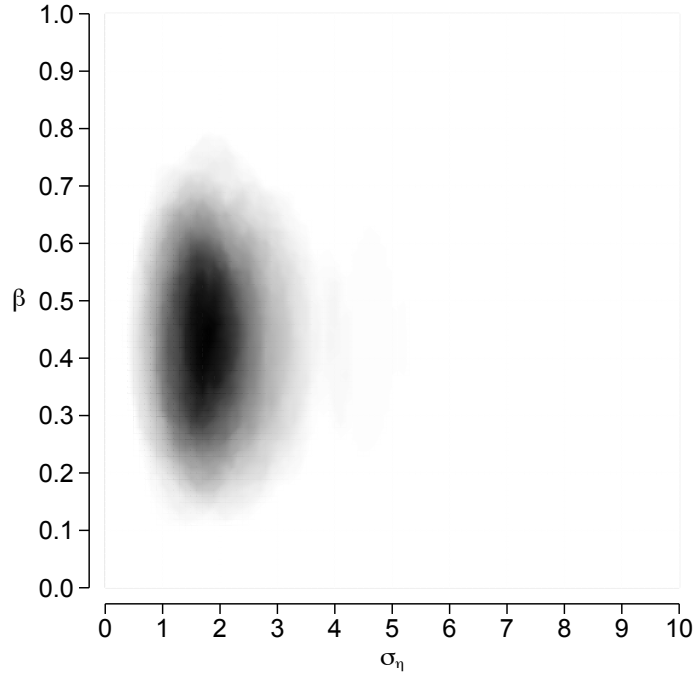


Figure 7: Heat map of physician-level estimates for β and σ_η

Notes: To ensure anonymity, the heat map covers only areas of five physicians or more.

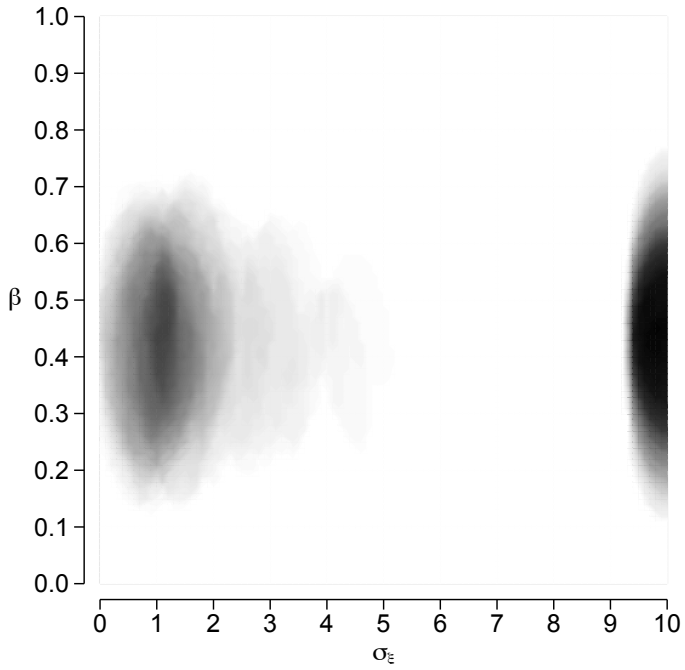


Figure 8: Heat map of physician-level estimates for β and σ_ξ

Notes: To ensure anonymity, the heat map covers only areas of five physicians or more.

Appendix E Projections of model parameter estimates

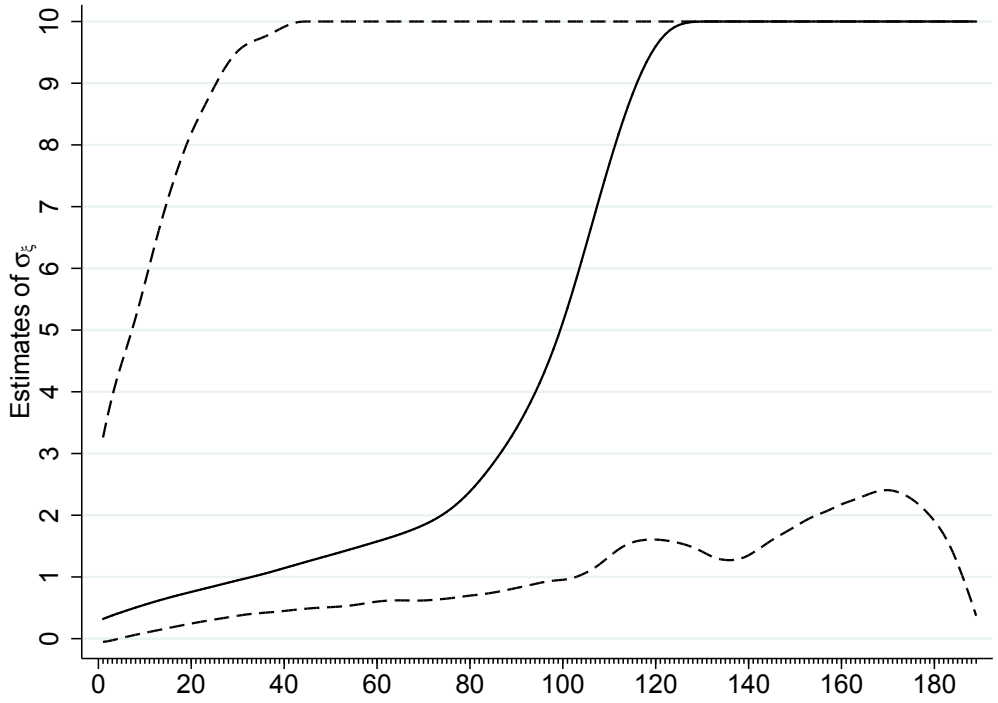


Figure 9: Estimated patient type signal noise parameters σ_ξ

Notes: Physician-level parameter estimates and bootstrapped 95% confidence intervals, sorted by size of the parameter estimate. To ensure anonymity, the figure is LOWESS-smoothed with bandwidth 0.2.

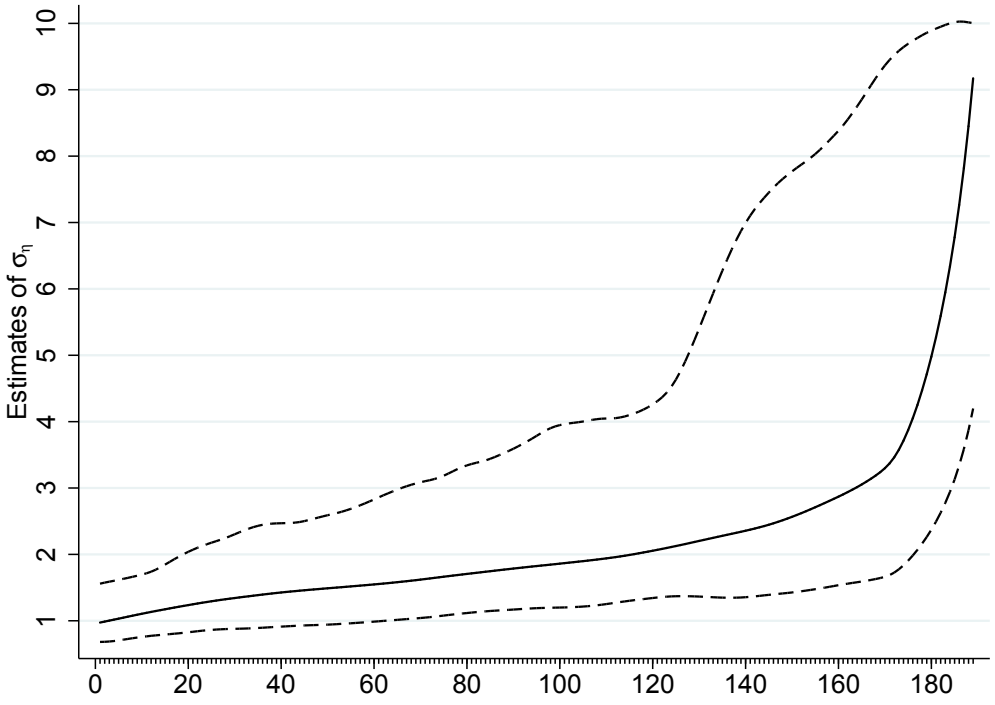


Figure 10: Estimated clinical diagnostic signal noise parameters σ_η

Notes: Physician-level estimates and bootstrapped 95% confidence intervals, sorted by size of the parameter estimate. To ensure anonymity, the figure is LOWESS-smoothed with bandwidth 0.2.

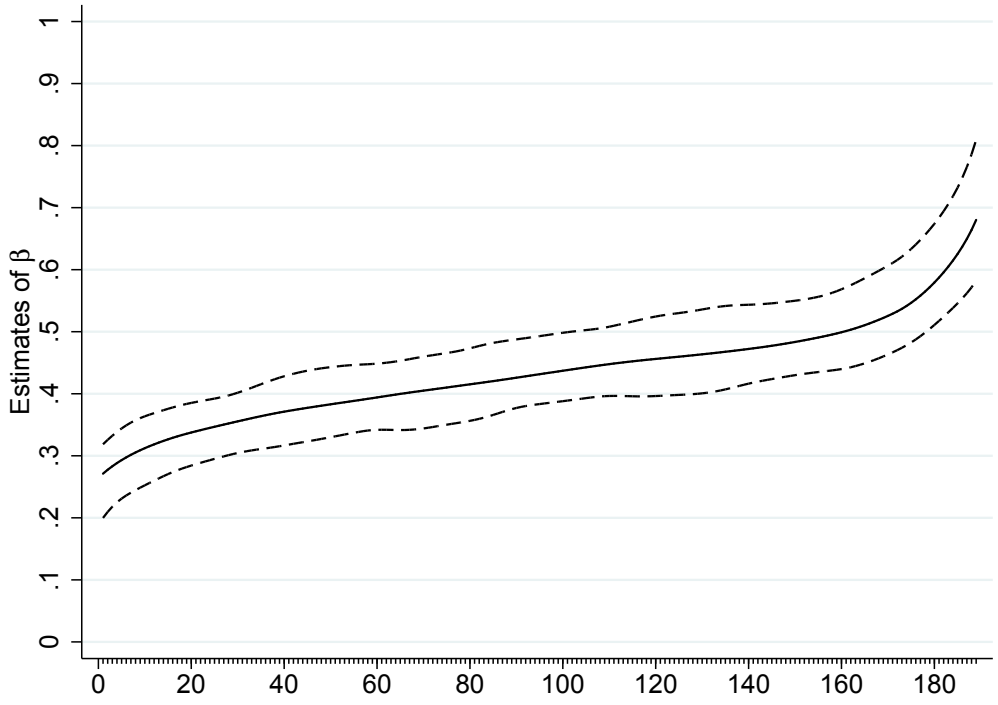


Figure 11: Estimated payoff parameters β

Notes: Physician-level estimates and bootstrapped 95% confidence intervals, sorted by size of the parameter estimate. To ensure anonymity, the figure is LOWESS-smoothed with bandwidth 0.2.

Appendix F Observed heterogeneity: $\hat{\sigma}_{\xi_j}$ and $\hat{\beta}_j$

Table 9 Correlation of clinical skill estimate with clinic and physician characteristics

	Linear regression for clinical signal noise $\hat{\sigma}_{\xi_j}$				
$N = 113$	(1)	(2)	(3)	(4)	(5)
Mean number of physicians	-0.27				-0.06
	[-1.77,1.23]				[-1.85,1.74]
Mean age of physicians	2.08				2.30
	[-3.34,7.49]				[-3.34,7.93]
Share of female physicians	0.08				0.11
	[-0.94,1.09]				[-0.97,1.19]
Dipstick tests per physician		-0.49		-0.37	-0.36
		[-2.12,1.15]		[-2.07, 1.33]	[-2.16,1.44]
Microscopy analyses per physician			-0.12	-0.09	-0.11
			[-0.50,0.26]	[-0.49, 0.32]	[-0.53,0.31]
Patients per physician		1.26	0.99	1.24	0.94
		[-1.25,3.77]	[-1.39,3.37]	[-1.28, 3.76]	[-2.17,4.05]
R^2	0.01	0.01	0.01	0.01	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 4. 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 10 Correlation of clinical skill estimate with clinic and physician characteristics

	Linear regression for payoff parameter $\hat{\beta}_j$				
$N = 113$	(1)	(2)	(3)	(4)	(5)
Mean number of physicians	-0.02 [-0.04,0.004]				-0.012 [-0.04,0.01]
Mean age of physicians	-0.002 [-0.10,0.10]				0.003 [-0.10,0.11]
Share of female physicians	-0.008 [-0.03,0.01]				-0.007 [-0.03,0.02]
Dipstick tests per physician		0.004 [-0.03,0.04]		0.008 [-0.03, 0.05]	0.006 [-0.03,0.04]
Microscopy analyses per physician			-0.002 [-0.007,0.002]	-0.003 [-0.01, 0.003]	-0.003 [-0.008,0.003]
Patients per physician		0.02 [-0.02,0.07]	0.03 [-0.01,0.07]	0.02 [-0.02, 0.07]	0.006 [-0.05,0.06]
R^2	0.03	0.01	0.02	0.02	0.03

Notes: This table presents coefficients for linear regressions where the outcome is the physician-level estimate of the payoff parameter summarized in Table 4. 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.

Appendix G Model fit

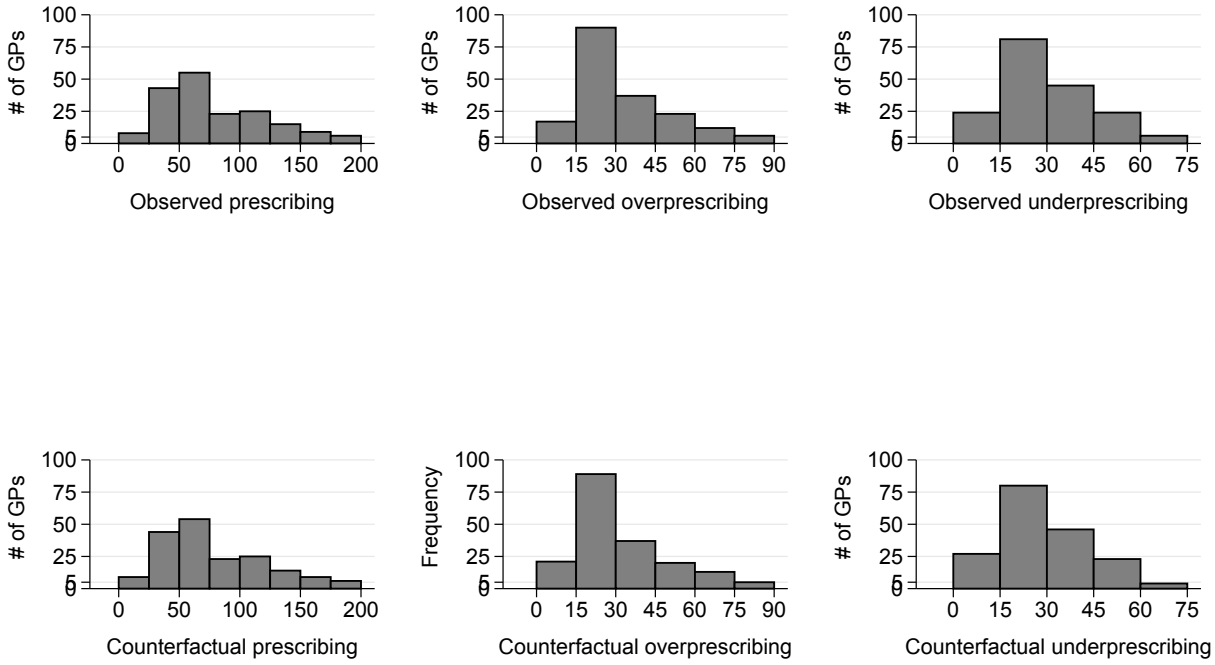


Figure 12: Observed and simulated in-sample moments

Appendix H Robustness

Estimation sample: clinics not rejected by Poisson-binomial test

Table 11 Parameter estimates

Type signal noise, σ_{ξ_j}	5.41 (4.14)
Clinical signal noise, σ_{η_j}	2.01 (0.96)
Payoff function parameter, β_j	0.43 (0.07)

This table reports the means and standard deviations of the distribution of parameter estimates for the reduced sample of 154 clinics. The model is estimated separately for each clinic.

Table 12 Counterfactual policy outcomes

	ML redistribution	Provide ML-based τ_i	
	$\Delta dy \stackrel{!}{=} 0$	$\sigma_{\xi_j} = 0$	$\sigma_{\xi_j} = 0$ $\beta_j = \hat{\beta}_j + \kappa$
Overall prescribing, Δd, in percent of $N_{\boldsymbol{d}} = 13,484$	-8.86 [-9.82, -7.87]	0.84 [0.05, 1.46]	-9.54 [-10.17, -8.75]
Treated bacterial cases, Δdy, in percent of $N_{dy} = 8,199$	0	7.72 [6.96, 8.55]	0
Overprescribing, $\Delta d\neg y$, in percent of $N_{d\neg y} = 5,285$	-22.48 [-24.92, -20.19]	-9.72 [-10.28, -8.40]	-24.19 [-25.50, -22.19]
Mean change in payoffs, $\frac{1}{J}\sum_{j=1}^{J} W_j(\boldsymbol{d}_j)$	0.081	0.110	0.104

Counterfactual changes relative to the status quo in percent across 154 clinics. The observed absolute numbers are reported in the left column. In counterfactual three, we set $\kappa = 0.033$ $[0.031, 0.036]$ to obtain $\Delta dy = 0$ as in counterfactual one. Bootstrapped 95 percent confidence intervals are reported in brackets.

Estimation sample: 12 weeks period without treatment or test

Table 13 Parameter estimates

Type signal noise, σ_{ξ_j}	4.73 (4.29)
Clinical signal noise, σ_{η_j}	1.99 (1.12)
Payoff function parameter, β_j	0.41 (0.07)

This table reports the means and standard deviations of the distribution of parameter estimates over 189 clinics. The model is estimated separately for each clinic.

Table 14 Counterfactual policy outcomes

	ML redistribution	Provide ML-based τ_i	
	$\Delta dy \stackrel{!}{=} 0$	$\sigma_{\xi_j} = 0$	$\sigma_{\xi_j} = 0$ $\beta_j = \hat{\beta}_j + \kappa$
Overall prescribing, Δd, in percent of $N_{\boldsymbol{d}} = 13,484$	-9.30 [-10.56, -8.58]	1.50 [1.31, 2.53]	-8.74 [-9.35, -7.92]
Treated bacterial cases, Δdy, in percent of $N_{dy} = 8,199$	0	7.27 [6.47, 8.18]	0
Overprescribing, $\Delta d\neg y$, in percent of $N_{d\neg y} = 5,285$	-23.32 [-26.45, -21.52]	-6.79 [-6.90, -4.80]	–21.29 [-22.80, -19.19]
Mean change in payoffs, $\frac{1}{J}\sum_{j=1}^{J} W_j(\boldsymbol{d}_j)$	0.073	0.095	0.091

Counterfactual changes relative to the status quo in percent across 189 clinics. The observed absolute numbers are reported in the left column. In counterfactual three, we set $\kappa = 0.031$ $[0.028, 0.034]$ to obtain $\Delta dy = 0$ as in counterfactual one. Bootstrapped 95 percent confidence intervals are reported in brackets.

Estimation sample: clinics with 50 observations and more

Table 15 Parameter estimates

Type signal noise, σ_{ξ_j}	5.25 (4.13)
Clinical signal noise, σ_{η_j}	2.45 (1.86)
Payoff function parameter, β_j	0.44 (0.08)

This table reports the means and standard deviations of the distribution of parameter estimates for the extended sample of 280 clinics. The model is estimated separately for each clinic.

Table 16 Counterfactual policy outcomes

	ML redistribution	Provide ML-based τ_i	
	$\Delta dy \stackrel{!}{=} 0$	$\sigma_{\xi_j} = 0$	$\sigma_{\xi_j} = 0$ $\beta_j = \hat{\beta}_j + \kappa$
Overall prescribing, Δd, in percent of $N_d = 13,484$	-9.10 [-9.95, -8.36]	-1.62 [1.29, 2.27]	-9.47 [-10.27, -8.87]
Treated bacterial cases, Δdy, in percent of $N_{dy} = 8,199$	0	-8.20 [-8.96, -7.33]	0
Overprescribing, $\Delta d\neg y$, in percent of $N_{d\neg y} = 5,285$	-22.7 [-24.7, -21.0]	-12.67 [-13.86, -11.82]	-23.66 [-25.49, -22.19]
Mean change in payoffs, $\frac{1}{J}\sum_{j=1}^{J} W_j(\boldsymbol{d}_j)$	0.065	0.107	0.102

Counterfactual changes relative to the status quo in percent across 280 clinics. The observed absolute numbers are reported in the left column. In counterfactual three, we set $\kappa = 0.034$ $[0.032, 0.037]$ to obtain $\Delta dy = 0$ as in counterfactual one. Bootstrapped 95 percent confidence intervals are reported in brackets.