# Machine predictions and human decisions with variation in payoffs and skill[*]

Michael Allan Ribers[†] Hannes Ullrich[‡]

May 2020

## Abstract

Evaluating the policy potential of machine learning is challenging because human decisions differ due to unobserved variation in incentives and skill. We overcome this challenge addressing the pressing policy problem of mitigating one of the leading causes of antibiotic resistance, antibiotic prescribing in primary care. Focusing on urinary tract infections, we document significant variation in treatment decisions. We propose a framework that combines diagnostic prediction and a prescription choice model with heterogeneity in physician payoff functions and diagnostic skill. Counterfactual policies replacing physician decisions with prediction-based rules reduce antibiotic use by 6.0 percent while combining machine learning predictions with physician diagnostic skill achieves a three times larger reduction of 17.8 percent.

JEL codes: C10; C55; I11; I18; Q28

[*] We benefited from very helpful feedback by Jason Abaluck, David Chan, Greg Crawford, Tomaso Duso, Daniel Ershov, Mogens Fosgerau, Matthew Gentzkow, Günter Hitsch, Ulrich Kaiser, Chuck Manski, Chloé Michel, Jeanine Miklós-Thal, Ziad Obermeyer, Yeşim Orhun, Maria Polyakova, Sherri Rose, Stephen Ryan, Bertel Schjerning, Karl Schmedders, Aaron Schwartz, Stephan Seiler, Jann Spiess, André Veiga, Christoph Wolf, and participants at the Annual Health Econometrics Workshop 2018 at Johns Hopkins University, the 2019 CESifo Area Conference on the Economics of Digitization, the Digital Economy Workshop 2019 at Católica Lisbon, the 2019 NBER Conference on Machine Learning in Health Care, the NBER Summer Institute 2019 for IT and Digitization, the 13th Digital Economics Conference in Toulouse, as well as in seminars at DIW Berlin, ESMT Berlin, University of Copenhagen, University Paris-Sud, University of Zurich, and Vienna University of Economics and Business. We are deeply indebted to Lars Bjerrum and Gloria Cristina Cordoba Currea for providing their expertise on Danish primary care as well as to Jenny Dahl Knudsen, Sidsel Kyst, and Rolf Magnus Arpi for enabling us to work with the microbiological laboratory data. 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.

[†] DIW Berlin; Department of Economics, University of Copenhagen; and BCCP. michael.ribers@econ.ku.dk

[‡] DIW Berlin; Department of Economics, University of Copenhagen; BCCP; and CESifo. hullrich@diw.de.

# 1 Introduction

Machine learning methods and the increasing availability of high-quality, 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, Agrawal et al. 2018, Athey 2018). Prominent examples include bail decisions in criminal justice, hiring, detecting social service fraud, healthcare provision, and labor market assistance programs. In numerous situations, machine learning can provide a standardized, data-based risk assessment. Yet, evaluating the potential of machine learning predictions relative to the status quo is complicated by the fact that human decisions are outcomes of individual decision makers' incentives and prediction technologies. Importantly, observed heterogeneity in decisions can be a result of variation in both (Chan et al. 2019). In addition, in settings studied so far, the target outcome needed to train a prediction model is difficult to observe and often sampled selectively as a result of human decisions.[1] These challenges have limited what can be learned about the mechanisms by which policies using machine learning predictions can yield improved outcomes.

In this paper, we attempt to overcome these challenges by considering the potential of machine learning for a high stakes prediction policy problem in which outcomes are observed with high accuracy and independent of the human decision of interest. The threat of increasing antibiotic resistance due to human antibiotic overuse is a major health policy challenge and a prime example where correct prediction of bacterial vs. other causes of infections is a key difficulty (WHO 2014).[2] Physicians face two problems when treating patients. First, they need to assess the underlying cause of reported symptoms. Second, given their assessment, the decision to prescribe an antibiotic solves a trade-off between effectively treating bacterial infections and causing increased antibiotic resistance. The risk assessment of a bacterial cause of infection depends on physician diagnostic skill. The solution to the trade-off is determined by physicians' weights on patients' sickness cost and the social cost of antibiotic resistance. Understanding the distinct sources of variation in physician decisions is crucial for evaluating machine learning as an improved risk assessment technology.

[1] For example, in health settings such as the diagnosis of heart attacks considered in Mullainathan and Obermeyer (2019), a patient is defined as recovered when no subsequent return to the hospital is observed. In Kleinberg et al. (2018), the machine learning algorithm can only be trained on observed recidivism by defendants to whom judges decided to grant bail, the very decision to which machine learning predictions are being compared.

[2] Antibiotics are used to treat bacterial infections by killing or inhibiting growth of bacteria in the body. Their effectiveness is decreasing due to antibiotic resistant bacteria threatening to render simple infections, such as pneumonia or infections in wounds, a fatal risk. In the US alone, antibiotic-resistant infections result in an estimated 23,000 deaths, $20 billion in direct healthcare costs, and $35 billion in lost productivity each year (CDC 2013).

We evaluate counterfactual policies using machine learning predictions to inform antibiotic treatment decisions for urinary tract infections (UTI) in primary care in Denmark.[3] The outcome, whether bacteria are the cause of an infection or not, is accurately observed by gold standard microbiological analysis, but with a delay of several days, corresponding to nearly a complete course of antibiotic treatment. Due to the acute nature of UTI, physicians must make a treatment decision before test results arrive. This feature enables us to evaluate policies based on *ex ante* machine learning predictions, explicitly taking physician heterogeneity into account. Delayed diagnostic results with simultaneous urgency to treat are a common challenge in health care; for example, in biopsies for malignant tumors or testing for tuberculosis. Therefore, understanding the role of instant diagnostic information for treatment decisions is important (Cassidy and Manski 2019).

For predicting UTI test results, we train a machine learning algorithm, a random forest, on high-dimensional, administrative data from Denmark for the years 2010, 2011, and 2012. The outcome is an indicator variable taking the value of one when bacteria are isolated in patient urine samples submitted for microbiological laboratory testing. The covariates in the prediction model include a rich set of patient-level information, such as gender, age, detailed employment status and type, education, income, civil status and more, past antibiotic prescriptions, past microbiological test results, medical outpatient claims histories, hospitalization records, as well as the same information on each individuals' household members. Machine learning applied to these data predicts out of sample realizations of bacterial UTI well, with an area under the ROC curve of 0.70. In addition, we document large heterogeneity in physician decisions evaluated by true and false positive rates as well as by the degree of predictive information contained in their decisions. The predictive information of physicians' prescription decisions is positively associated with clinic size and the propensity to send samples to the microbiological laboratory, but negatively with physician age. These observations hint at stronger standardization and availability of diagnostic procedures in clinics with younger physicians, higher aptitude toward diagnostic technologies, and more exposure to UTI cases.

[3] UTI are one of the most common classes of bacterial infections. Foxman (2002) reports almost half of all women contract a UTI once in their lifetime. In the US, yearly UTI-related healthcare costs including workplace absences are estimated at $3.5 billion (Flores-Mireles et al. 2015). According to Bjerrum and Lindæk (2015), each year 10 percent of women receive antibiotic treatment for UTI. In the US, nearly half of all antibiotic prescriptions are made by primary care physicians (CDC 2015). In Europe, primary care accounts for 90 percent of prescriptions (Llor and Bjerrum 2014). In Denmark, general practitioners are responsible for roughly 75 percent of antibiotic prescriptions (Danish Ministry of Health 2017). While slightly imprecise we use “physician," "general practitioner," and "primary care physician" interchangeably.

To illustrate the prevalent heterogeneity in treatment decisions, we begin by following the growing literature considering policies that redistribute treatment from low-risk to high-risk cases based on machine learning predicted risk (Bayati et al. 2014, Chalfin et al. 2016, Kleinberg et al. 2018, Yelin et al. 2019, Hastings et al. 2020). Specifically, we make use of the prediction results and analysis in Ribers and Ullrich (2019). Here, complete replacement of physician decisions along the full predicted risk range, delaying prescriptions for patients below a predicted risk threshold, and instantly giving prescriptions to patients above, cannot improve prescribing without further knowledge of the social payoff function. Instead, prescription rules replacing physician decisions only for cases with high machine learning accuracy can reduce overall prescribing by up to 6.0 percent without reducing the number of treated patients suffering from a bacterial infection.

One strong underlying assumption in this literature is that decision makers adhere perfectly to policies, which is equivalent to assuming algorithms replace human discretion. To combine predictions with physicians' own risk assessment and accommodate their individual objectives, we build on Chan et al. (2019) and propose a framework that combines machine learning predictions with a binary choice model governing treatment decisions. We decompose diagnostic skill into two sources of information about patients' true sickness state: analyzing information encoded in observable data amenable to machine learning methods and information acquired in clinical practice, which is not commonly encoded and stored at scale. In clinical practice, when patients present UTI symptoms, physicians gather diagnostic information about patients' true sickness state by observing their health condition, personal characteristics, and medical histories as well as performing either one or both of the rapid diagnostic technologies available today: urine dipstick and microscopic analysis (Davenport et al. 2017). While the dipstick analysis is a standard procedure, using microscopy requires additional equipment and specific training, so variation in skill can be expected. Similarly, some physicians may explore patients' medical histories in more detail than others and differ in analytical skill such that risk assessment based on observable data may vary. This is the dimension on which the main strength of machine learning methods, forming systematic risk assessments based on observables, can be exploited.

To identify two-dimensional diagnostic skill, we build on results in Chan et al. (2019), who exploit quasi-random assignment of suspected pneumonia cases to radiologists evaluating chest X-rays. In Denmark, primary care providers are assigned by an individual's municipality of residence. Switching away from these default assignments is possible for a small fee but uncommon [SOURCE]. Therefore, physicians treating UTI are almost completely determined by location of residence. Yet, patient communities may differ in their risk of having bacterial UTI due to socioeconomic or geographic factors. To identify skill and payoff function parameters, we provide evidence for the assumption that patients are comparable across physicians, conditional on predicted risk based on a rich set of observables. We recover a vector of three parameters for every primary care clinic using simulated maximum likelihood estimation. The first two measure the accuracy of diagnostic information physicians use, while the third parameter governs the trade-off physicians solve in making treatment decisions. The mean estimated signal noise parameter on patient-type information is larger than signal noise on clinical diagnostic information, implying that, on average, physicians rely more on information from in-clinic diagnostics than on observing patient types. In addition, we find significant heterogeneity in the estimated noise parameters. One-third of physicians make no use of patient type information encoded in observable data. The mean estimated weight of the social cost of antibiotic resistance relative to an individual patient's sickness cost is 0.6 with a standard deviation of the distribution across physicians of 0.18. This implies that the mean physician weighs the social cost of increasing antibiotic resistance due to one antibiotic prescription slightly above one half the health benefit of curing one patient.

Correlating the parameter estimates with primary care clinic characteristics, we find that clinics with more patients per physician use patient type information less. The noise parameter on clinical diagnostic information is positively correlated with mean physician age and negatively correlated with the intensity of laboratory testing, suggesting that physicians with higher skill rely more on high-quality diagnostic technologies and may be younger on average.

Based on the structural model, we evaluate three counterfactual policies. The first two counterfactuals target physicians' diagnostic skill, the third targets physicians' payoff functions. Counterfactual one provides physicians with the machine learning prediction for each patient and assumes physicians combine the prediction with their own clinical diagnostic information. We find that overall prescribing decreases by 17.8 percent (3672 prescriptions) and overprescribing decreases by 33.3 percent (2663 prescriptions). Interestingly, the number of treated bacterial infections is also reduced, by eight percent (1009 prescriptions). In counterfactual two, we improve the minimum clinical diagnostic skill to the estimated median level of the physician population. Such a policy could include training skills in performing microscopic analysis or investing in technical diagnostic equipment for clinics that seemingly have difficulties acquiring clinical diagnostic information. Surprisingly, such an intervention does not affect overall prescribing. Yet, it results in a sizable shift of nearly 400 prescriptions from non-bacterial to bacterial infections, thereby significantly improving treatment decisions by approximately 4.5%. In counterfactual three, we manipulate the parameters of the payoff functions while holding patient type and clinical diagnostic information fixed. In particular, we increase the payoff parameter such that the overall reduction in prescribing is equivalent to the counterfactual reduction achieved by providing the machine learning prediction to physicians without noise. 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 individual patients' sickness cost due to foregone antibiotic treatment. The policy reduces overprescribing by 24.4 percent (1953 prescriptions). Yet, manipulating the payoff function weights without improving diagnostic information induces adverse effects. The number of treated bacterial infections decreases by 13.6 percent (1716 prescriptions). This result illustrates 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.

Using Danish data has crucial advantages. First, Danish administrative data are unique in the world in scope and in interconnectivity, covering rich information including patients' and patient household members' personal background information, detailed employment histories, as well as medical prescriptions and claims records. Second, medical education is homogenous and health care provision strongly standardized and centralized. Thus, we can expect to estimate a lower bound of the degree of skill heterogeneity. Third, Denmark is a country with an excellent record of low antibiotic use (Goossens et al. 2005). In 2017, the Danish government initiated a national action plan in which one main goal is to reduce overall antibiotic prescribing by one-third by 2020 compared to 2016 (Danish Ministry of Health 2017). Consequently, we suspect the improvements we find for Denmark are a lower bound for the attainable improvements elsewhere.

We contribute to the existing literature in several ways. Kleinberg et al. (2015) argue that prediction policy problems are important and commonplace. Existing work considers the prediction of shootings in Chicago to target a crime prevention program (Chandler et al. 2011), online reviews to predict hygiene inspections (Kang et al. 2013), prediction of household consumption responses for the targeting of a tax rebate program in Italy (Andini et al. 2018), and prediction of high-risk opioid prescriptions using administrative data from the US (Hastings et al. 2019). Ribers and Ullrich (2019) also consider antibiotic prescription policies but focus on the value of combining rich administrative data for prediction and the challenge of evaluating policies implemented out of sample. Kleinberg et al. (2018) make an important methodological contribution by using machine learning to evaluate the potential improvements of judges' bail decisions, where only crimes committed by released defendants can be observed. If judges selected released defendants based on unobservables, predicted crime rates for the jailed based on observables can be biased. Kleinberg et al. (2018) propose a solution based on assumptions of judges' random case assignment and varying leniencies, as well as homogenous risk prediction technologies. Focusing on patients for whom physicians ordered laboratory tests at initial consultations, we relax the homogenous skill assumption and assess the combination of machine learning predicted risk with the diagnostic information physicians hold. In clinical practice, microbiological analysis at the laboratory is typically used if the urine dipstick analysis is inconclusive regarding the bacterial cause of an infection (Davenport et al. 2017). Even if the majority of dipstick analyses are inconclusive (Devillé et al. 2004), by focussing on this set of consultations, the results cannot easily be generalized to the full population of prescription decisions.[4] Yet, the considered treatment decisions reflect many typical health care provision situations such that our results have relevance beyond the considered context. In the context of antibiotic prescribing our analysis holds empirical relevance covering over 70,000 consultations, out of which one-third received antibiotic treatment, over a short time period and small geographic area.

A large economic and public health literature considers demand-side mechanisms for policy interventions, including prescription surveillance and stewardship (Laxminarayan et al. 2013), general practitioner competition (Bennett et al. 2015), financial incentives for physicians (Currie et al. 2014, Das et al. 2016), and peer effects (Kwon and Jun 2015). Hallsworth et al. (2016) find a significant reduction in prescribing in a randomized intervention giving social norm feedback to high-prescribers. Yet, it is difficult to tell from this intervention whether providers aligned their preferences more closely to what is socially desired or they improved diagnostic efforts. This distinction is important for policy. The preference mechanism may have adverse effects due to increased underprescribing if treatment quality was unaffected by the policy. The potential to increase diagnostic information would suggest value of policies solving the core prediction problem. Our analysis is also connected to Currie and MacLeod (2017) who allow for heterogeneity in skill but assume homogenous preferences to evaluate the counterfactual of reassigning C-sections from low- to high-

[4] It is an interesting complementary question how physicians choose to test patients. Abaluck et al. (2016) find physicians systematically misallocate image scans for diagnosis of pulmonary embolism to low- and high-risk patients, with negative impacts on diagnostic quality and health benefits. Mullainathan and Obermeyer (2019) consider heterogenous yield of emergency room testing for heart attacks and investigate human biases as main cause. Cassidy and Manski (2019) show that diversifying test and treatment decisions for tuberculosis reduces bias and uncertainty in subgroup risk assessment.

risk pregnancies. While they focus on the effects of improving surgical and decision-making skill, we focus specifically on the potential of data-driven predictions to complement physician skill.

The remainder of the paper is organized as follows. Section 2 presents the institutional background and data. Section 3 shows the results of the prediction algorithm and describes observed physician decisions. Section 4 presents the prediction policy problem of antibiotic treatment decisions. Section 5 describes the framework for the evaluation of prediction-based policy rules replacing physician decisions and shows the corresponding counterfactual results. Section 6 develops the structural model, discusses identification and counterfactual policy evaluations. Section 7 concludes.

## 2 Danish administrative data and laboratory test results

We use Danish administrative registry data which 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. Denmark has several regulations of relevance for general practitioner decision making, which we review here to facilitate the appreciation of the validity of our results outside of the Danish context.

### 2.1 The Danish healthcare system

Denmark has a universal and tax financed single payer health care system with general practitioners as the primary gatekeepers. Every person living in Denmark is allocated to a general practitioner by a list-system within a fixed geographic radius around the home address. Patients can switch physicians from their initial assignment at a small cost but most stick with their assigned general practitioner. Although primary care 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. Importantly, physicians do not generate earnings by prescribing drugs to patients who have to purchase their prescriptions from local pharmacies. In 2012, Denmark had 2200 primary care clinics with a median size of just above one general practitioner per clinic (Møller Pederson et al. 2012). Throughout the paper, we will use physician and clinic interchangeably because most of our medical transaction data are observed at the clinic level.

Prescription drugs are subsidized but patients co-pay a fraction of the list price depending on their cumulative yearly prescription drug expenditures. The Danish market for prescription drugs is highly regulated resulting in uniform pricing at pharmacies nationwide and antibiotic treatment is in general cheap, about 100 Danish Kroner (15 US Dollars) per complete treatment. General practitioners are responsible for prescribing approximately 75 percent of the human consumed systemic antibiotics in Denmark (Danish Ministry of Health 2017).

## 2.2 Analysis sample based on clinical microbiological laboratory test results

Individual-level clinical microbiological laboratory test results comprise the central data set of our analysis. Particularly, it contains the outcome we aim to predict, $y_{it}$, a binary outcome indicating if bacteria was isolated in a urine sample acquired at time $t$ from patient $i$ consulting a general practitioner. We acquired clinical microbiological laboratory test results from Herlev hospital and Hvidovre hospital, the 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 is present in a patient's biological sample. In addition, patient and clinic identifiers and information on the biological 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. Urine samples constitute 477,609 samples out of which 156,694 are submitted by general practitioners. Bacteria was isolated in approximately one out of three urine samples, both overall and among the general practitioner submitted samples. We further restrict the number of urine sample observations by excluding tests from patients who received a systemic antibiotic prescription or were previously tested within 28 days prior to the urine sample acquisition date in order to focus on consultations that constitute a first contact with a physician within a patient's treatment spell. Lastly, we also exclude pregnant women from our analysis as both the test decision, including many mandatory tests during pregnancy, and the prescription decision cannot be compared to the typical non-emergency patient.

The final set of test observations is comprised of 74,511 test results for urine samples taken during initial consultations with men or non-pregnant women in a non-emergency setting consulting physicians in 688 primary care clinics. For this set, the majority (80%) of the test procedure lasted two to four days during which general practitioners remain uninformed. Since we know the precise timing of urine sample acquisitions and the test response date, we can determine exactly in which period physicians' prescribe antibiotics under uncertainty. By focussing on consultations during which physicians submitted a urine sample for microbiological laboratory testing, we ensure that test outcomes are observed for all patients regardless of the physicians' prescription decisions. We so avoid the selective labels problem tackled by Kleinberg et al. (2018), a common challenge when evaluating counterfactuals in machine learning applications, and advance what can be learnt about the role of machine learning predicted risk for treatment decisions conditional on laboratory testing. The cost of this approach is the lack of generalizability of our results 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). When a physician decides to test, the value of diagnostic information is presumably high so that the prediction-based policies proposed here improve upon situations in which physicians are making decisions under significant uncertainty.

## 2.3 Danish national registries

The administrative data provided by Statistics Denmark covers the entire population of Denmark between January 1, 2002, and December 31, 2012 and, hence, all individuals in our laboratory data as well as family members can be identified in the registries. For each individual, 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*) linkable across registries via unique individual and physician identifiers.

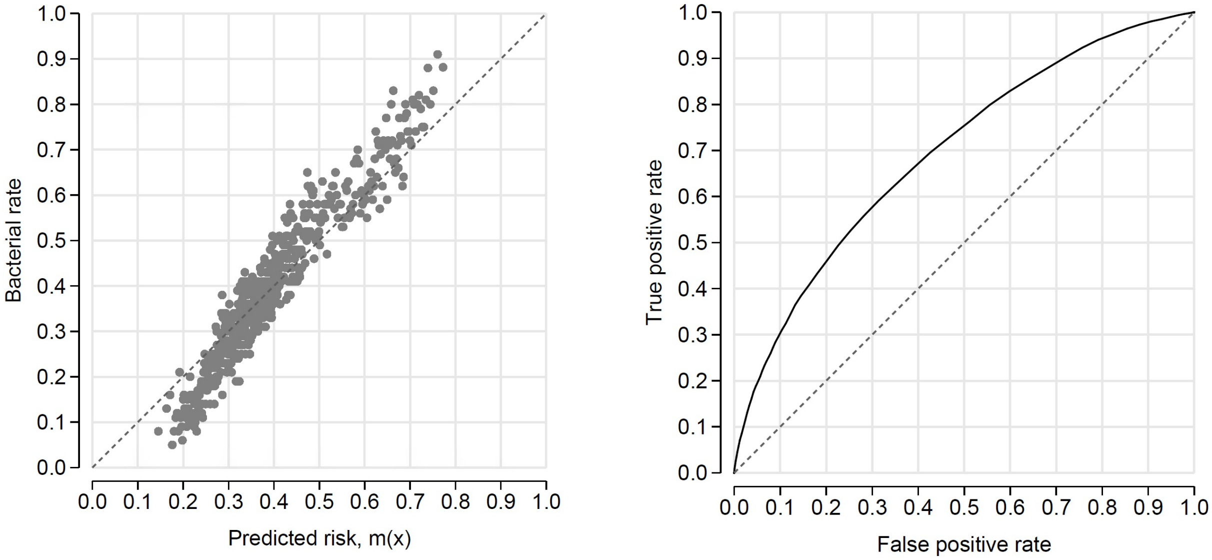
The demographic data include gender, age, education, occupation, income, marriage and family status, home municipality, immigration status and place of origin, and lastly includes household member identifiers which allows us to identify the patients' family members and add their demographic data as well as the laboratory data and the data from the following registires. The data on systemic antibiotic prescriptions contain slightly more than 35 million purchased prescriptions. We observe the date of purchase, patient and prescribing primary care clinic identifiers, anatomical therapeutic chemical drug classification, drug name, price, indication of use, purchased package size and defined daily dose. It should be noted that the indication of use is imprecise in the sense that many prescriptions are given with a UTI indication but prescriptions for UTI are also given with a more generic indication, e.g. against infection, or without indication at all. The hospitalization data comprise all patient contacts with hospitals, including ambulatory visits. The data contain observations on hospitalizations of more than 2 million unique individuals per year over since 2002 and includes information on hospitalization admittance 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 Danish general practitioner clinic services provided to the Danish population of patients. The claims data are comprised of approximately 100 million claims per year and includes physician and patient identifiers, consultation time, services provided and physician fees. Among other, the claims data allow us to identify pregnant women from mandatory pregnancy-associated examinations who we exclude from the analysis. The combination of the laboratory data and the administrative registers yields a vector $x_{it}$ of predictors for patient $i$ at time $t$ provided to the prediction algorithm for each tested patient. All contained historic data relative to the test acquisition time are, in principle, observable to the physician at the time of the consultation.

## 3 Machine learning predictions and physician decisions

As proposed in Ribers and Ullrich (2019), we train a random forest algorithm (Breiman 2001) that relates patient covariates to laboratory test outcomes in order to predict if patients suffer from bacterial UTI and, consequently, if antibiotic treatment is beneficial or should be avoided. Our implementation differs from typical machine learning practice in that we do not randomly split our data in training and out of sample validation partitions. We intend to evaluate counterfactual prediction-based prescription policies and this requires that the prediction function and the policy rules are applied to future patients relative to the training data. Standard machine learning practice assumes that outcomes $y$ and covariates $x$ are independent draws from a joint distribution of $(Y, X)$ which remains the same for the training and out of sample partitions (Athey 2018). When the prediction function is applied to future patients and not random partitions of the data, this assumption no longer holds by construction and we need to retrain the prediction model over time. For this purpose, we create 24 monthly out-of-sample evaluation partitions from January 2011 to December 2012, and use all data prior to the respective test observations as training data. After training the prediction model, we drop clinics with fewer than 10 observations, a selection step which is inconsequential for our results but provides a minimum level of statistical power. All reported results are based on these out-of-sample data accounting for 53,976 observations and 482 primary care clinics.

## 3.1 Machine learning performance

A random forest is an ensemble of regression trees applied to bootstrapped versions of the training data. A tree represents a partition of the data created as a sequence of binary splits over individual variables where each split is determined by the homogeneity of the test outcomes in the created partitions. A simple model, in our case the mean, is universally fitted to all observations in each final partition, or leaf, of the tree. The random forest prediction, $m(x)$, is then the mean prediction over all trees. We illustrate the prediction quality on the left panel of Figure 1 which plots the average test results against the average out of sample predicted risk. Every sphere represents a bin containing 100 patients where patients are assigned to bins sorted by their predicted risk. Outcomes are close to the 45 degree line throughout the risk distribution, showing that the algorithm on average correctly predicts bacterial risk. Among the riskiest 100 patients in the evaluation partitions, 88 percent are tested positive for bacteria following the initial consultation with the physician. Equivalently, the observed bacterial UTI rate for the 100 least riskiest patients is 8 percent. That the random forest predictions performs well throughout the risk range is a first necessity for improvements to physician decision making.



**(a)** Laboratory test results relative to machine predictions of bacterial test outcomes. Spheres represent averages over 100 tested patients sorted by predicted risk.

**(b)** The receiver operating curve plotting the trade-off between the true positive rate and the false positive rate for varying classification thresholds. The area under the curve equals 0.68.

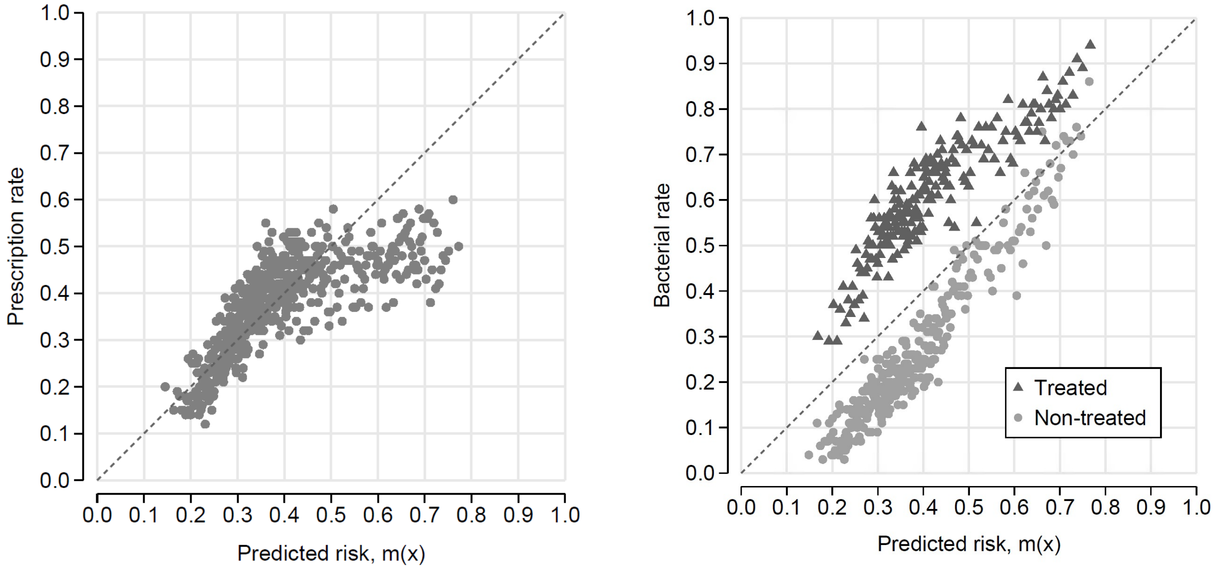
**Figure 1:** Prediction quality of the random forest algorithm out-of-sample.

Binary predictions can be accomplished by comparing the random forest prediction to a classification threshold. Typical machine learning applications sets the classification threshold to pick majority vote, i.e. a classification threshold at 0.5. However, other classification thresholds might be appropriate depending on the final application. As is standard in machine learning applications, we plot the receiver operating curve (ROC) on the out-of-sample evaluation observations as shown in Figure 1(b). The receiver operating curve plots of the true positive rate against the false positive rate as the classification threshold is varied from 0 to 1. The closer the receiver operating curve is to the top-left corner, the better the prediction quality. A common metric by which to measure overall prediction accuracy is therefore the area under the ROC, the AUC. Our bacterial UTI prediction function has an AUC equal to 0.70. Kleinberg et al. (2018) report a comparable AUC of 0.707.

It is instructive to see which variables are the most important predictors. Figure 9 in Appendix A shows the feature importance for each variable, computed as the decrease in prediction error when variable values are permuted. Variables with feature importance equal to or below zero are considered to have no impact on prediction quality. For exposition, we collect variables in groups containing (i) patient characteristics and test timing; (ii) patient past prescriptions; (iii) patient past laboratory test results; (iv) patient past hospitalizations; (v) patient past general practice insurance claims; (vi) household members' past prescriptions; (vii) household members' past laboratory test results; (viii) household members' past hospitalizations; (ix) household members' past hospitalizations; and (x) household members' past general practice insurance claims. The most notable result is that patients' past test results, while important, do not appear to be the most important features. This is likely explained by strong correlation between past test results and observed past antibiotic treatment, which cannot be captured by the feature importance metric.

## 3.2 Physician prescribing conditional on predicted risk

Figure 2a plots the physicians' prescription rate prior to obtaining a test result against the algorithm's predicted risk. Again, spheres represent averages over bins containing 100 patients sorted by the algorithm's predicted risk. Physicians seem to evaluate low risk patients correctly on average, as the prescription rate and predicted risk appear well correlated in the low risk range. However, as predicted risk increases, the physicians' prescription rate flattens out, hence, the physicians and the random forest algorithm seem to disagree on the high risk patients who it appears that physicians have difficult time identifying. It is important to note that although physicians on average prescribe at the average bacterial rate, they are not always prescribing to the patients with bacterial UTIs, in fact far from it.



**(a)** Physician prescribing relative to machine learning predictions of bacterial test outcomes. Spheres represent averages over 100 tested patients sorted by predicted risk.

**(b)** Laboratory test results conditional on physician prescribing relative to machine learning predictions. Spheres represent averages over 100 patients sorted by predicted risk.

**Figure 2:** Bacterial test outcomes, prescribing, and unobservables.

To see this more clearly, we evaluate test outcomes versus algorithm predictions conditional on physician prescribing as shown in Figure 2b. It is important to note that patient predictions are not re-computed conditional on the prescription decision to prescribe but only re-sorted, i.e. physician instant prescriptions choices are not included in the covariates. Several important insights can be observed. Conditional on the level of machine-predicted risk, physicians are on average able to prescribe to patients that more frequently show bacterial test realizations. This could be due to physician expertise and diagnostic unobservables, the latter, among other, covering in-house diagnostic tests such as nitrite dipsticks or microscopy.[5] Physicians in Denmark can in principle observe all

[5] Nitrite dipstick can detect bacteria that transform Nitrate to Nitrite. In the hold out data, the detectable genera are Escherichia, Enterobacter, Klebsiella, Citrobacter, and Proteus. The non-detectable genera are Staphylococcus, Pseudomonas, Enterococci, Acinetobacter, and Streptococcus. Inspecting prescription choices separately by dipstick-detectable and non-detectable bacterial species isolated in laboratory tests allows us to investigate whether physicians select on nitrite dipstick test results. While patients with dipstick-detectable bacteria have a higher prescription rate, 64 percent, relative to prescription rate for patients with non-dipstick-detectable bacteria, 55 percent, the difference is moderate. This suggests that dipstick test results leave significant uncertainty, which is consistent with evidence

data used by the machine learning algorithm, by accessing the medical administrative data through their IT system or by asking patients. Large-scale information acquisition would be prohibitively time-consuming for physicians. However, it would be feasible in practice for a physician to combine clinical diagnostic information with a check of a patient's prescription history. Figure 2b also shows that physicians prescribe antibiotics to a substantial number of patients for whom predicted risk is very low, the lower left triangles. We define overprescribing as any prescription to a patient with a negative bacterial test result and observe that overprescribing on average decreases among the treated as machine predicted risk increases. Further, physicians do not prescribe to a substantial number of patients for whom predicted risk is very high, the top right spheres. We define underprescribing as physicians' decisions not to prescribe to patients suffering from a bacterial infection and observe that underprescribing on average decreases as machine predicted risk decreases.

### 3.3 Physician heterogeneity

To inspect heterogeneity in physician decisions and the role of diagnostic skill further we consider the predictive value of physician decisions for laboratory test outcomes. We estimate a second random forest including physician choices as predictors and compute the clinic-specific difference in AUC between the random forest including the choice variable and the random forest excluding it. Figure 3 shows that including physician choice as a predictor increases the AUC on average and for most clinics. The observed heterogeneity suggests that combining physician information and data-based predictions at the physician level has large potential.

To learn about potential correlates with this measure, we link it to a set of clinic characteristics, which we observe for 283 out of the total of 482 clinics. Table 6 in Appendix B shows the coefficients of a linear regression of the change in AUC on clinic characteristics. The number of patients per physician in a clinic is positively associated with improvement in prediction due to information contained in treatment choices. One interpretation of this observation is that physicians with more frequent patient exposure are better at identifying bacterial infection causes, which suggests heterogeneity in physician technologies for risk prediction. Physician age, on the other hand, is negatively associated, while the number of laboratory tests ordered per patient are positively associated with the ability to identify bacterial infections. While we can give no causal interpretation to the parameters estimated in this analysis, the correlations hint towards differences in the use of diagnostic technologies across clinics mainly based on clinic size, physician age, and the propensity

reported in the medical literature (Devillé et al. 2004).

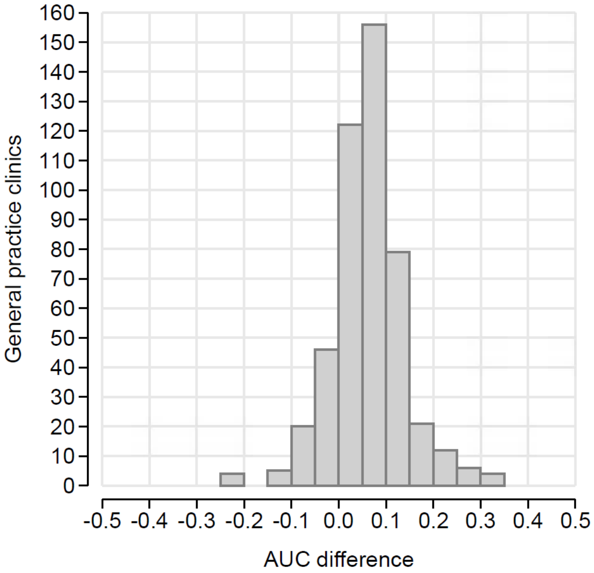


**Figure 3:** Distribution of physician-specific changes in AUC due to treatment choice as a predictor. Bins include at least three observations to ensure anonymity.

to use laboratory diagnostics.

The interpretation of the change in AUC due to the treatment choice predictor as physician expertise is confounded by the fact that physician choice is the outcome of an optimization problem according to the physician's objective function. It is therefore informative to further inspect physicians' prescription choices relative to the outcomes. Because here we consider the decision to prescribe any antibiotic or none, we can view our empirical setup through the lens of a classification problem with binary decisions and states. Using the insight in Chan et al. (2019) we then interpret the location of clinics in the ROC space as being informative about their production possibilities frontier, reflecting their diagnostic skill, and their preferences for how to weigh the antibiotic resistance externality against the individual cost of having a sick patient. To locate prescription regimes in the ROC space, Figure 4 shows a heatmap of clinics' prescription rates relative to negative and positive bacterial test outcomes. To maintain anonymity at the clinic level, we plot only areas of three or more physicians.

Physician prescribing relative to bacterial outcomes varies widely. Physicians in the top left of the plot have high prescribing rates for bacterial infections and low prescribing rates for non-bacterial test outcomes. Accordingly, physicians at the bottom right, have low prescribing rates for bacterial infections and high prescribing rates for non-bacterial test outcomes. Variation between these two extremes reflects variation in skill. An ROC curve thus reflects a physician's skill level as

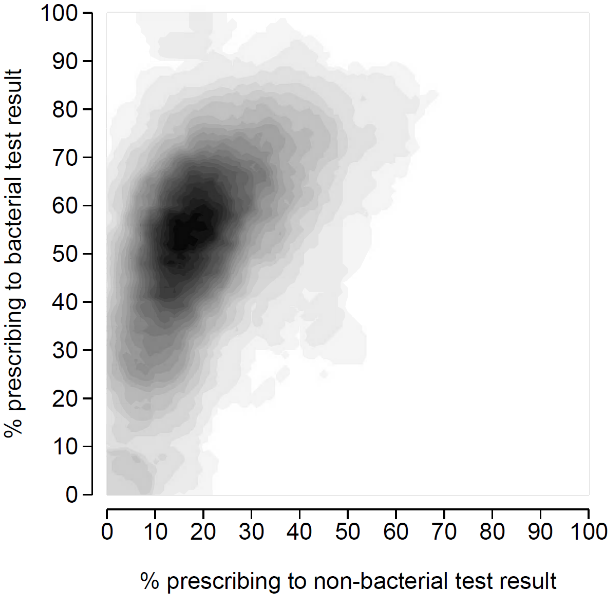


**Figure 4:** Heatmap of physician prescribing prior to knowing test results conditional on test outcomes. Due to anonymity restrictions, individual physicians cannot be shown.

it represents the set of permissible trade-offs between true positives (appropriate prescribing) and false positives (overprescribing). The location on a given ROC curve is determined by the preference weights generating a physician's optimal trade-off, given skill. Physicians close to the origin have 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 more intense prescribers with 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 UTIs. Yet, a relevant number of physicians are located towards the top right, some of which cannot be plotted due to anonymization restrictions. Physicians located to the top right of prescribing rates of (0.4,0.4) account for 8.4 percent of all physicians. Machine learning predictions have the potential to shift physicians prediction technology towards the top left in ROC space or allow policy makers to shift physicians away from areas of over- or underprescribing along their diagnostic production possibilities frontiers.

# 4 Trading off private and social cost in antibiotic prescribing

The potential of prediction-based policy needs to be evaluated relative to the quality of human decision making. In particular, it is a priori unclear to what extent machine learning can replace expertise or how they may complement each other. To explore this, we propose a series of counterfactual analyses combining machine learning predictions and an economic model of the trade-off inherent in treatment decisions at initial consultations of potential UTI patients. We assume throughout that the prescription decision, $d$, solves a trade-off between the patient suffering a sickness cost, $a$, from delaying prescribing until a test result is available, and the social cost of prescribing, $b$, 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 UTI. Antibiotic treatment is only curative and alleviates sickness if a patient suffers from a bacterially caused infection. The payoff function at a patient's initial consultation can therefore be written as

$$\pi(d; y) = -ay(1-d) - bd, \tag{1}$$

where $y$ is an indicator for the true realization of a bacterial infection. We further assume that $0 < b < a$ such that prescribing is always optimal when an infection is known to be bacterial with certainty. Although test outcomes are not observed at the initial consultations, counterfactual prescription rules can be evaluated *ex post* by computing differences in outcomes between a prediction-based policy, $\delta$, and physicians' observed prescription choices, $\delta^J$:

$$\Pi(\delta) = \mathrm{E}\big[\pi(\delta, y) - \pi(\delta^J, y)\big] = a\,\underbrace{\mathrm{E}[(\delta-\delta^J)y]}_{\Delta\delta y} - b\,\underbrace{\mathrm{E}[\delta-\delta^J]}_{\Delta\delta}, \tag{2}$$

where the expectation is over the realizations of $y$ and $\delta^J$ in all consultations. The effect of a prediction-based policy can be separated into two terms: the benefit from an increase in correctly treated bacterial UTI patients, $\Delta\delta y$, and the benefit from reducing antibiotic use, $\Delta\delta$. For all counterfactual machine learning based policy rules, $\delta$, presented in the following sections, we will report outcomes as the percentage change in observed prescribing:

$$\%\Delta\delta = \frac{\mathrm{E}[\delta-\delta^J]}{\mathrm{E}[\delta^J]}, \tag{3}$$

as the percentage change in prescriptions given to patients with bacterial UTI:

$$\%\Delta\delta y = \frac{\mathrm{E}[(\delta-\delta^J)y]}{\mathrm{E}[\delta^J y]}, \tag{4}$$

and as the percentage change in overprescribing:

$$\%\Delta\delta(1-y) = \frac{\mathrm{E}[(\delta - \delta^J)(1-y)]}{\mathrm{E}[\delta^J(1-y)]}. \tag{5}$$

In Section 5, we evaluate counterfactual prescription rules that override physician decisions based on machine learning risk predictions $m(x)$. For these counterfactuals, we interpret equation (1) as the policy maker's payoff function. In real world applications physicians retain final decision-making authority. If payoff functions, specifically the weights $a$ and $b$, and the expectation of $y$ vary across physicians and an assumption of perfect physician adherence is untenable, what we learn from these counterfactuals is limited. Thus, in Section 6, we propose counterfactual policies in which machine learning risk predictions for individual patients are provided to physicians and enter their optimization problem. For this purpose, we develop a structural model where equation (1) represents physician $j$'s individual payoff function so that parameters $a_j$ and $b_j$ and risk assessments vary across physicians.

# 5 Prescription rules based on machine learning predicted risk

In this section, we evaluate prescription policies that override physician decisions on the basis of predicted risk. That is, the prescription rules either delay potential prescribing until test results are available or assign antibiotic treatment prior to receiving test results as a function of predicted risk only. We consider two sets of prescription rules. First, prescription rules that override all physician prescription decisions, and second, prescription rules combining prediction-based decisions and physician prescriptions such that physician decisions are optimally replaced in selected risk ranges.

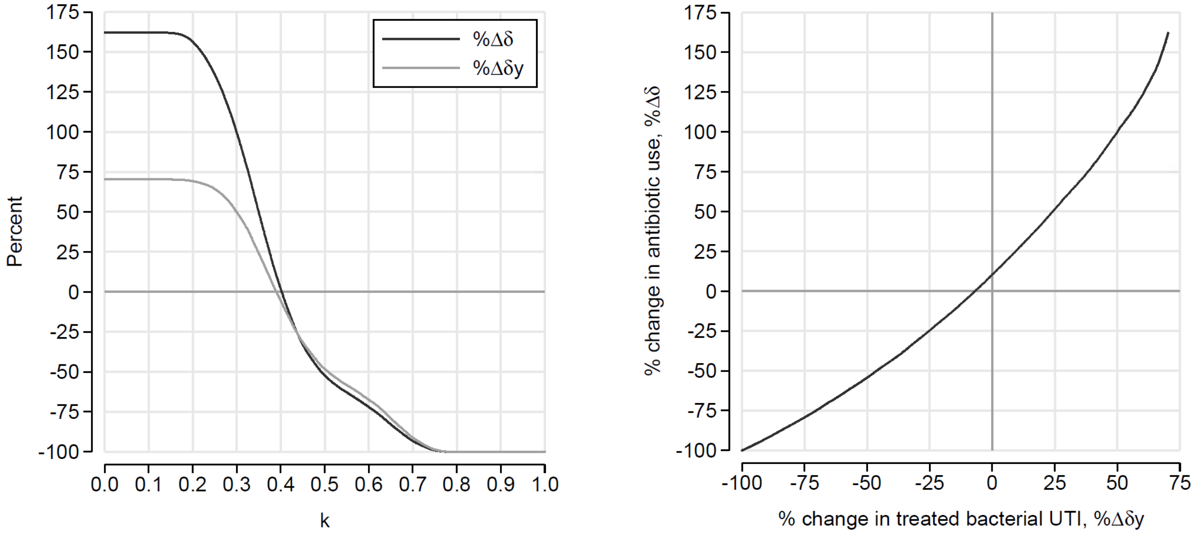
## 5.1 Prescription rule based on predicted risk only

We document in Section 3.2 that overprescribing occurs most frequently at low predicted risk and decreases on average as predicted risk increases; and similarly, that underprescribing occurs most frequently at high predicted risk and decreases as predicted risk decreases. Given such monotonicity, it is optimal to consider prediction-based policies that postpone prescriptions for patients with low predicted risk until test results are available and assign prescriptions prior to observing test results to patients with high predicted risk. Thus, the prescription rules we consider are step functions:

$$\delta(m(x);k) \quad = \quad \begin{cases} 0 & \text{if } m(x) < k, \\ 1 & \text{if } k \leq m(x), \end{cases} \tag{6}$$

where prescriptions are never given for predicted risk, $m(x)$, below a cut-off, $k$, and always given above.

Figure 5 shows counterfactual changes in antibiotic prescriptions relative to observed physician prescribing in percent, $\%\Delta\delta$, and the percentage change in treated patients with bacterial UTI, $\%\Delta\delta y$, as $k$ varies over the full risk range. For $k = 0$, all tested patients receive a prescription and overall prescribing increases by 162 percent while increasing the number of correctly treated bacterial infections by 70 percent. At the other extreme, $k = 1$, no patients are treated at the initial consultation and 100 percent of the initially prescribed antibiotics are delayed while 100 percent of the observed treated patients with bacterial UTIs would go untreated.



**(a)** Change in prescribing and treated bacterial infections by cut-off $k$

**(b)** Change in prescribing and treated bacterial infections

**Figure 5:** Counterfactual outcomes of prescription rule based on predicted risk only

For interior $k$, the function $\%\Delta\delta y$ crosses 0 at a lower threshold $k$ than $\%\Delta\delta$ so that improvements cannot be attained without further restrictions on payoff function parameters $a$ and $b$. Antibiotic use cannot be reduced without decreasing the number of treated bacterial UTIs and the number of treated bacterial UTIs cannot be increased without also increasing antibiotic use. Given knowledge of $a$ and $b$, however, there may be cases in which the prescription rule leads to payoff improvements for some $k$. Then, a reduction of antibiotic prescribing can be achieved at the expense of delaying treatment for an increased number of patients with a bacterial infection or an increase in antibiotic use is preferred for a further increase in the number of treated bacterial UTIs.

### 5.2 Prescription rule based on predicted risk and physician decisions

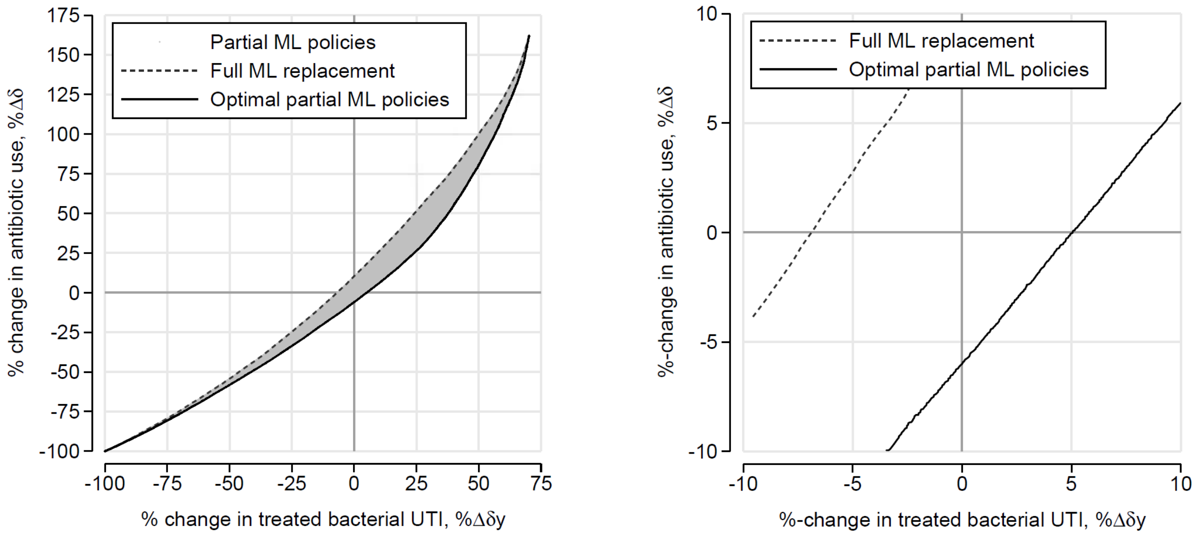
Figure 2b suggests physicians hold valuable information which is lacking in machine learning predicted risk. Such information includes observation and assessment of symptoms and rapid but error-prone diagnostics such as dipstick and microscopic analyses. If decision-relevant information is unobservable to the policy maker, a rule basing decisions on predicted risk when prediction quality is high and relying on physician discretion otherwise may combine observable and unobservable information such that general improvements are possible. Therefore, we adapt the prescription rule in equation (6) to include both predicted risk-based decisions and physician decisions such that

$$\delta\big(m(x); \delta^J, k_L, k_H\big) = \begin{cases} 0 & \text{if } m(x) < k_L, \\ \delta^J & \text{if } k_L \leq m(x) \leq k_H, \\ 1 & \text{if } k_H < m(x), \end{cases} \tag{7}$$

where $m(x)$ is the machine learning predictions as a function of patient observables, $\delta^J$ is the observed physician prescription decision and $(k_L, k_H)$ are policy thresholds. This rule delays antibiotic prescribing to patients with risk predictions below $k_L$ and assigns prescriptions prior to observing test outcomes to patients with risk predictions above $k_H$. Physician decisions for patients with risk predictions between $k_L$ and $k_H$ remain unaffected.

Figure 6a shows the set of attainable trade-offs between changes in antibiotic use and treated bacterial infections based on $\delta\big(m(x); \delta^J, k_L, k_H\big)$ for all $(k_L, k_H)$. The left bound (dashed) of this set represents policies where $k_L = k_H$, corresponding to the prescription rule using only predicted risk analyzed in Section 5.1. The right bound (solid) represent the best possible policies combining decisions based on predicted risk and physician discretion that minimize antibiotic use while treating as many bacterial infections as possible. The majority of policies cannot achieve improvements in payoffs imposing further assumptions on the values of payoff parameters $a$ and $b$. For many $(k_L, k_H)$, achieving a reduction in antibiotic use comes at the cost of reducing the number of treated bacterial infections and, analogously, increasing the number of treated bacterial infections comes at the cost of an increase in antibiotic use.

However, we identify a set of values of $(k_L, k_H)$ such that policies decrease the number of



**(a)** Set of attainable effects

**(b)** Set of attainable effects, magnified at origin

**Figure 6:** Counterfactual outcomes for combined prescription rule

antibiotic prescriptions while keeping fixed or increasing the number of treated bacterial infections. For better exposition, Figure 6b shows the two bounds in a magnified area around the origin. As opposed to prescription rules $\delta(k)$ based on predicted risk only, prescription rules incorporating physician discretion achieve payoff improvements irrespective of the values of parameters $a$ and $b$.

Table 1 reports the attainable improvements for the two policies bounding the set of payoff-improving policies, where either $\Delta\delta = 0$ or $\Delta\delta y = 0$. Confidence intervals are based on re-computation of policy results given $k_L$ and $k_H$ for 100 bootstrapped samples. The maximal reduction of antibiotic consumption while treating the same number of bacterial infections is six percent of the observed sum of antibiotic prescriptions. Similarly, the maximal increase in treated bacterial UTIs while holding antibiotic use constant is five percent of the observed number of treated bacterial UTIs.[6]

Physician decisions with full information are a useful benchmark to understand the relevance of counterfactual policies considered here. For both rules, we observe that approximately 70 percent of patients to whom counterfactual policies give prescriptions at the initial consultation receive prescriptions following the arrival of microbiological test results. Combined with an estimated 24 percent spontaneous recovery rate (Ferry et al. 2004), this suggests that prescriptions based on

[6] Quantifying the potential effects of predicted risk-based prescription policies face additional practical challenges in this setting. For example, policy makers may need to optimally update $(k_L, k_H)$ over time. Ribers and Ullrich (2019) investigate these challenges further and find that the achievable improvements are robust.

**Table 1** Partial replacement policy outcomes

| Policy rule | $(k_L, k_H)$ | Antibiotic use (%) | Treated bacterial UTI (%) |
|---|---|---|---|
| Reduce antibiotic prescribing s.t. $\Delta\delta y = 0$ | (0.31, 0.62) | $-6.01$ $[-6.75, -5.28]$ | 0.00 $[-0.90, 0.90]$ |
| Increase treated bacterial UTI s.t. $\Delta\delta = 0$ | (0.29, 0.62) | 0.00 $[-0.76, 0.76]$ | 5.08 $[4.23, 5.93]$ |

Notes: 95% confidence intervals are computed with 100 bootstrap samples holding $(k_L, k_H)$ fixed.

machine learning predictions resemble physician choices under full information.

## 6 Policies with full physician discretion and predicted risk

The evaluation of prescription rules so far overrides some or all of physicians' treatment decisions, which implies either replacing human discretion or assuming decision makers' adhere perfectly to these rules. This assumption is invoked by much of the existing literature evaluating machine learning predictions in comparison to human decisions (Bayati et al. 2014, Chalfin et al. 2016, Kleinberg et al. 2018, Ribers and Ullrich 2019, Yelin et al. 2019, Hastings et al. 2020). Yet, such an implementation is unlikely in practice so that the results reported in the previous section may offer limited insights. In addition, although replacing human discretion can avoid human error, it also discards valuable diagnostic information used by expert physicians which may be missing in machine learning predictions. Ignoring physicians' diagnostic skill can make the reported results too pessimistic. Hence, quantifying the value of machine learning predictions is difficult without further knowledge of physicians' potentially heterogenous payoff functions and diagnostic skill.

### 6.1 ROC space and production function

Here: building on the ROC space/production possibilities frontier description in Chan et al. (2019), pages 6-8 incl. Footnote 7 on selection models, to describe the distinction between the ROC curve for ML prediction and the GP ROC curve. The combination of both is what can be achieved ideally. Model formalizes how they can be combined.

### 6.2 A model of antibiotic treatment choice with variation in payoffs and skill

We propose a framework that combines machine learning predictions with a model of primary care providers' treatment choice allowing for heterogenous payoff functions and skill. The model follows Chan et al. (2019) in separating individual physicians' treatment choice problem from the preceding step of forming predictions. Specifically, we consider physician skill in two dimensions: diagnosis based on observable background information, as also used by the machine learning algorithm, and diagnosis based on unobservable clinical information available only to the physician. The distinction of these two types of diagnostic skill and payoff functions in a model of physician prescription choice provides a systematic tool to analyze the effects of counterfactual policies improving diagnostic skill and manipulating physicians' payoff functions.

**Prediction**

We model patient $i$'s sickness realization as determined by a latent index, $\nu_i$, such that the patient has a bacterially caused UTI according to

$$y_i = \mathbb{1}[\nu_i > \bar{\nu}], \tag{8}$$

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

$$\nu_i \sim \mathcal{N}(\mu_i, \sigma_\nu^2). \tag{9}$$

We do not require any assumptions on the distribution of patient types, $\mu_i$, across physicians. Instead, we recover $\mu_i$ from $m(x_i) = \mathrm{E}\{y_i \mid x_i\} \equiv \mathrm{E}\{y_i \mid \mu_i\} = 1 - \Phi(\mu_i)$, that is, by assigning patient type as the machine learning predicted risk conditional on observables $x_i$.

In clinical practice, when patients present UTI symptoms, physicians gather information about patients' true sickness state by observing $x_i$, including $i$'s personal characteristics and medical histories. Some physicians may research patients' medical histories in more detail than others so that risk assessment based on observable data depends on analytical skill. We assume that the physician receives a noisy signal about patient $i$'s type, where lower noise implies higher skill,

$$\xi_{ij} \sim \mathcal{N}(\mu_i, \sigma_{\xi_j}^2). \tag{10}$$

In addition, physicians can acquire clinical diagnostic information by observing patients' health condition and 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 may introduce substantial variation in diagnostic skill in this setting, an observation that has been documented in medical decision making more generally (Hoffrage et al. 2000, Pallin et al. 2014). Thus, physicians receive a noisy signal by observing clinical diagnostic information,

$$\eta_{ij} \sim \mathcal{N}(\nu_i, \sigma^2_{\eta_j}). \tag{11}$$

Given both patient type and clinical diagnostic signals, the physician forms her posterior beliefs about the latent patient index according to

$$\nu_i \mid \xi_{ij}, \eta_{ij} \sim \mathcal{N}(m, \sigma^2) \tag{12}$$

where the posterior mean and variance are given by

$$m = \frac{\xi_{ij}\sigma^2_{\eta_j} + \eta_{ij}(\sigma^2_{\xi_j} + \sigma^2_\nu)}{\sigma^2_{\xi_j} + \sigma^2_\nu + \sigma^2_{\eta_j}} \qquad \text{and} \qquad \sigma^2 = \frac{(\sigma^2_{\xi_j} + \sigma^2_\nu)\sigma^2_{\eta_j}}{\sigma^2_{\xi_j} + \sigma^2_\nu + \sigma^2_{\eta_j}}. \tag{13}$$

### Treatment choice

Physician $j$'s payoff function reflecting the trade off between curing a patient with a predicted probability and incurring the social cost of increased antibiotic resistance is defined analogous to equation (1) with physician-specific parameters $a_j$ and $b_j$. The physician proceeds to prescribe am antibiotic iff

$$\mathrm{E}\{\pi(0; y_i) \mid \xi_{ij}, \eta_{ij}\} < \mathrm{E}\{\pi(1; y_i) \mid \xi_{ij}, \eta_{ij}\} \quad \Leftrightarrow \quad \Phi\left(\frac{\bar{\nu} - m}{\sigma}\right) < 1 - \frac{b_j}{a_j}. \tag{14}$$

As patient types are only identified relative to their distance from the sickness threshold, $\bar{v}$, in terms of units of $\sigma_\nu$, we set $\bar{v} = 0$ and $\sigma_\nu = 1$ resulting in the prescription rule:

$$d_{ij} \mid \xi_{ij}, \eta_{ij} = \mathbb{1}[\, 0 < \underbrace{\xi_{ij}\sigma^2_{\eta_j} + \eta_{ij}(\sigma^2_{\xi_j} + 1) + \sqrt{(1 + \sigma^2_{\xi_j} + \sigma^2_{\eta_j})(\sigma^2_{\xi_j} + 1)\sigma^2_{\eta_j}}\,\Phi^{-1}(1 - \frac{b_j}{a_j})}_{g(\xi_{ij}, \eta_{ij} \mid a_j, b_j, \sigma_{\xi_j}, \sigma_{\eta_j})} \,] \tag{15}$$

We assume physicians hold correct beliefs about their own skill. Low signal variance reflects high skill. The comparative statics with respect to $\sigma_{\xi_j}$, $\sigma_{\eta_j}$, and $b_j/a_j$ are intuitive. 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 $\sigma_{\xi_j}$ and $\sigma_{\eta_j}$ is ambiguous. Low skill, reflected in large parameter values, increases the last term of the inequality and hence the probability to prescribe. On the other hand, with low skill, large signal realizations can turn the sign of the first or second terms and lead to prescriptions for non-bacterial cases or vice versa.

### 6.3 Identification

Our identification argument follows Chan et al. (2019) who show that, under the assumption of random assignment, $\sigma_{\eta_j}$ and $b_j/a_j$ are identified. This result holds in our setting if assignment is as good as random conditional on observed patient type. In Denmark, primary care providers are assigned by an individual's municipality of residence. Switching away from these default assignments is possible but uncommon. Therefore, physicians treating UTI are almost completely determined by location of residence. Yet, patient communities may differ in their risk of having bacterial UTI due to socioeconomic or geographic factors. To identify skill and payoff function parameters, we rely on the assumption that patients are comparable across physicians, conditional on predicted risk based on a rich set of observables. We estimate patient types based on the high-dimensional set $x_i$, accounting for potential selection on a rich set of observables. Figure 7 shows physician-level mean bacterial rates and mean predicted risk are both strongly centered around the value of 0.4. Yet, mean bacterial rates still vary conditional on predicted risk but orthogonal to mean predicted risk.

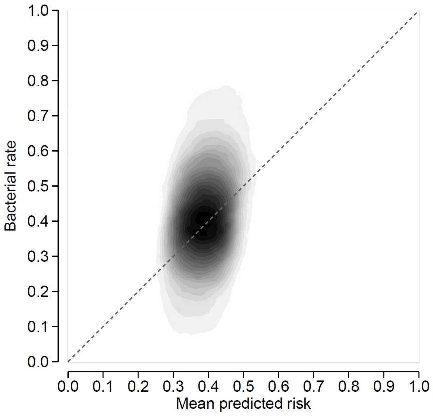
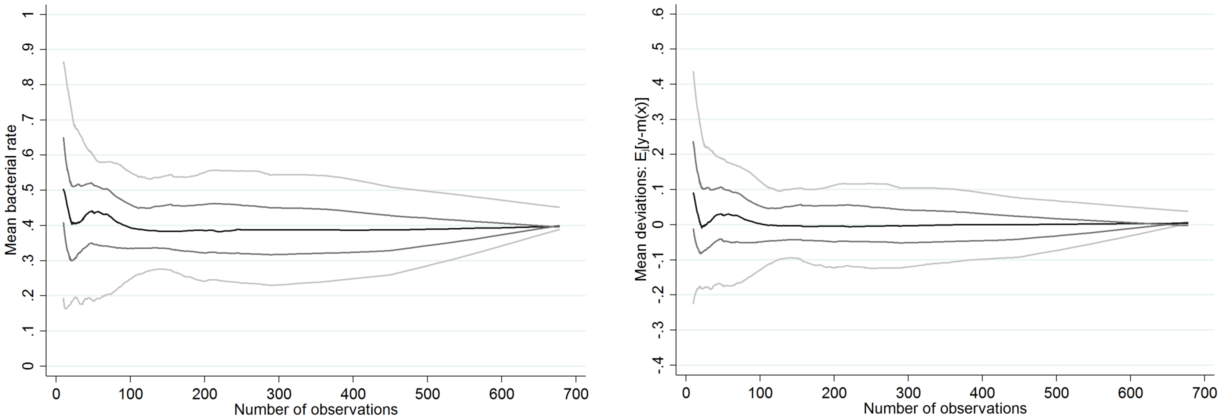


**Figure 7:** Heatmap of physician mean bacterial rate and predicted bacterial risk. Areas with three physicians or more are plotted.

To investigate whether this variation reflects differential selection into testing or noise, we analyze the distribution of physician-level mean differences between the true bacterial rate and predicted risk. Figure 8 plots the physician-level mean bacterial rates and the mean differences by the number of laboratory tests used by physicians. The number of tests reflects physicians' test decisions and determines the sample size for estimation. For identification, we are worried about how physicians' test decisions may lead to variation in unobserved but diagnosis-relevant patient characteristics. In particular, physicians might obtain varying bacterial rates if their testing behavior varied systematically. Abaluck et al. (2016) show how variation in test yield reflects variation in pre-test diagnostic skill. We observe in Figure 8a that bacterial rates vary widely in the full range between 0 and 1. This variation becomes smaller when sample sizes increase, suggesting that part of the variation is driven by noise. Substantial variation remains even for the largest primary care clinics. Figure 8b shows that conditioning on predicted risk reduces this variation across all sample sizes. The remaining variation is decreasing in sample size and is symmetric even for physicians for whom we do not observe large amounts of testing. If variation in bacterial rates was driven by physicians ability to achieve larger bacterial rates using fewer tests, we would expect a pattern of decreasing bacterial rates in sample size. We take these observations as suggestive that physicians do not systematically select tested patients based on unobservables.



**(a)** Physician-mean bacterial outcomes by number of test observations. 1st, 25th, 50th, 75th, 99th percentiles smoothed using LOWESS.

**(b)** Physician-mean deviation of bacterial outcomes from predicted risk by number of test observations. 1st, 25th, 50th, 75th, 99th percentiles smoothed using LOWESS.

**Figure 8:** Mean bacterial rate and observations per physician.

To consider the potential role of unobservable information further, we also consider the balance of the types of bacteria found in tests as well as their antibiotic resistances. Different bacteria have varying difficulty of detection by in-clinic diagnostics such as dipstick and microscopic analysis. Further, they can lead to different symptoms and disease severities. These may be important sources of information contained in unobservables, which we assume do not drive the decision to procure a microbiological test. Table 2 reports the descriptive statistics of physician-level shares of bacteria species and resistances for two groups of physicians, those above and below the median clinic-level mean bacterial rate. The balance across nearly all bacteria and molecules provides further evidence that the distributions of tested cases are balanced across physicians.

**Table 2** Balance of types of bacterial infection causes

| | $\underline{E_j[y]}$ | $\overline{E_j[y]}$ | $\Delta$ |
|---|---|---|---|
| E.coli | 0.66<br>(0.10) | 0.67<br>(0.08) | 0.014<br>(0.008) |
| E.faecalis | 0.06<br>(0.05) | 0.04<br>(0.03) | -0.013<br>(0.004) |
| Enterococcus | 0.03<br>(0.04) | 0.04<br>(0.03) | 0.005<br>(0.003) |
| K.pneumoniae | 0.04<br>(0.04) | 0.05<br>(0.04) | 0.003<br>(0.004) |
| S. agalactiae | 0.05<br>(0.04) | 0.03<br>(0.03) | -0.011<br>(0.003) |
| Others | 0.16<br>(0.07) | 0.16<br>(0.06) | 0.003<br>(0.006) |
| J01CA01 | 0.40<br>(0.10) | 0.40<br>(0.08) | 0.002<br>(0.008) |
| J01CA04 | 0.02<br>(0.03) | 0.03<br>(0.03) | 0.006<br>(0.002) |
| J01CA11 | 0.26<br>(0.09) | 0.24<br>(0.07) | -0.022<br>(0.007) |
| J01DC02 | 0.09<br>(0.06) | 0.09<br>(0.05) | 0.002<br>(0.005) |
| J01DD13 | 0.03<br>(0.03) | 0.03<br>(0.03) | 0.002<br>(0.003) |
| J01EA01 | 0.22<br>(0.08) | 0.22<br>(0.07) | -0.004<br>(0.007) |
| J01EB02 | 0.35<br>(0.09) | 0.34<br>(0.08) | -0.011<br>(0.008) |
| J01MA02 | 0.10<br>(0.06) | 0.10<br>(0.05) | 0.002<br>(0.005) |
| J01MB02 | 0.09<br>(0.06) | 0.10<br>(0.05) | 0.001<br>(0.005) |
| J01XE01 | 0.07<br>(0.05) | 0.08<br>(0.06) | 0.007<br>(0.005) |
| Number of clinics | 241 | 241 | |

Notes: This table reports weighted mean bacterial species and resistance rates above and below the median of mean bacterial rates $E_j[y]$.

### 6.4 Estimation

We estimate the model by simulated maximum likelihood using observed data on prescription decisions, $d_{it}$, sickness realizations, $y_{it}$, and patient types $\mu_i$ recovered from random forest predictions $m(x_i)$. We normalize $a_i = 1$ because only the ratio $b_j/a_j$ is identified. The simulated likelihood contribution from a single observation follows from

$$\mathcal{L}_{ij}(d_{ij}, y_{ij} \mid \Theta_j, \mu_i) = \begin{cases} \Pr\{g(\xi_{ij}, \eta_{ij} \mid b_j, \sigma_{\xi_j}, \sigma_{\eta_{ij}}) > 0, \nu_{ij} > 0 \mid \Theta_j, m(x_i)\} & \text{if } d_{ij} = 1, y_{ij} = 1, \\ \Pr\{g(\xi_{ij}, \eta_{ij} \mid b_j, \sigma_{\xi_j}, \sigma_{\eta_{ij}}) > 0, \nu_{ij} < 0 \mid \Theta_j, m(x_i)\} & \text{if } d_{ij} = 1, y_{ij} = 0, \\ \Pr\{g(\xi_{ij}, \eta_{ij} \mid b_j, \sigma_{\xi_j}, \sigma_{\eta_{ij}}) < 0, \nu_{ij} < 0 \mid \Theta_j, m(x_i)\} & \text{if } d_{ij} = 0, y_{ij} = 0, \\ \Pr\{g(\xi_{ij}, \eta_{ij} \mid b_j, \sigma_{\xi_j}, \sigma_{\eta_{ij}}) < 0, \nu_{ij} > 0 \mid \Theta_j, m(x_i)\} & \text{if } d_{ij} = 0, y_{ij} = 1. \end{cases} \tag{16}$$

where $\Theta = \{b_j, \sigma_{\xi_j}, \sigma_{\eta_j}\}$ and $\xi_{ij}$ and $\eta_{ij}$ are simulated conditional on observed $y_i$ using the distributional assumptions in equations (10) and (11). For simulations, we use 20,000 quasi-random draws created by modified latin hypercube sampling (Hess et al. 2006). Defining $\mathcal{I}_j$ as the set of patients consulting physican $j$, the joint likelihood over outcomes $\boldsymbol{d}_j = \{d_{ij}\}_{i \in \mathcal{I}_j}$ and $\boldsymbol{y}_j = \{y_i\}_{i \in \mathcal{I}_j}$ is given by

$$\mathcal{L}_j(\boldsymbol{d}_j, \boldsymbol{y}_j \mid \Theta_j, m(x_i)) = \prod_{i \in \mathcal{I}_j} \mathcal{L}_{ij}(d_{ij}, y_i \mid \Theta_j, m(x_i)). \tag{17}$$

Physician skill and preferences can now be recovered from

$$\hat{\Theta}_j = \arg\min_{\Theta_j} \sum_{i \in \mathcal{I}_j} \log \mathcal{L}_j(d_j, y_j \mid \Theta_j, m(x_i)). \tag{18}$$

Estimating $\hat{\Theta}_j$ for every physician clinic allows us to recover the nonparametric physician heterogeneity distribution.

### 6.5 Estimation results

Table 3 reports the means and standard deviations of $\hat{\Theta}_j$. The means of the noise parameters for patient type ($\sigma_{\xi_j}$) and clinical diagnostic information ($\sigma_{\eta_j}$) are large, 3.03 and 2.23. Interestingly, mean $\sigma_{\xi_j}$ is markedly larger than $\sigma_{\eta_j}$, meaning 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 UTI. 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 $\sigma_{\xi_j}$ and $\sigma_{\eta_j}$. The mean value of 0.6 of the preference parameter estimates, bounded by 0 and 1 by the assumption that $0 < b < a$, suggests conservative physicians on average. The mean physician weighs the social cost of increasing antibiotic resistance due to one antibiotic prescription slightly above one half the health benefit of curing one patient. Yet, the standard deviation of 0.18 reflects substantial heterogeneity in how physicians solve this trade-off.

**Table 3** Distribution of parameter estimates

| | Mean | (SD) |
|---|---|---|
| Type signal noise, $\sigma_{\xi_j}$ | 3.03 | (2.01) |
| Diagnostic signal noise, $\sigma_{\eta_j}$ | 2.23 | (1.39) |
| Payoff function parameter, $b_j/a_j$ | 0.60 | (0.18) |

Notes: This table reports the means and standard deviations of the distribution of parameter estimates over 482 physician clinics. The model is estimated for each.

Figures 10 to 12 in Appendix D show the distributions of parameter estimates. For anonymization we only show heatmaps and do not report values in areas containing less than three clinics. Both skill parameters $\sigma_{\xi_j}$ and $\sigma_{\eta_j}$ are concentrated in the area between 0.5 and 2.5. This concentration is more pronounced for $\sigma_{\eta_j}$, suggesting that the majority of physicians makes use of clinical diagnostic information but significant heterogeneity remains. The large estimate of $\sigma_{\xi_j}$ on average suggests that providing machine learning predictions can improve physician information significantly. In particular, we find a relevant number of physicians with very large $\sigma_{\xi_j}$ estimates. In Figure 10, physicians at the top, with $\sigma_{\xi_j} > 4$, account for one-half of all physicians. This group does not appear to use patient type information encoded in observables. Yet, the distribution of the noise parameter for clinical diagnostic information of this set of physicians, with mean 3.0 and standard deviation 2.7 of $\sigma_{\eta_j}$, is comparable to its distribution for the remaining set of physicians, with mean 2.7 and standard deviation 2.7 of $\sigma_{\eta_j}$. Therefore, combining systematic information in predictions $m(x_i)$ with valuable clinical diagnostic information used by most physicians may substantially improve decisions. A small share of physicians, around five percent, appear to have poor overall diagnostic skill reflected by both very large $\sigma_{\xi_j}$ and $\sigma_{\eta_j}$ estimates. Figures 11 and 12 do not show a systematic relationship between the estimated payoff weights and both noise parameters.

Figure 13 in Appendix E shows the distributions of the observed mean prescribing, overprescribing, and underprescribing rates and their simulated counterparts based on the parameter estimates.

The simulated distributions closely resemble the observed data, suggesting good model fit.

To investigate potential sources of heterogeneity across primary care clinics, we correlate parameter estimates with observable clinic characteristics. We aggregate individual physician characteristics to the primary care clinic level because prescriptions are observed for clinics. Due to data limitations we are able to merge characteristics for a subset of 283 out of the total of 482 clinics. Linear regression results of the parameters estimates on clinic characteristics in Table 4 show several interesting patterns. The size of clinics measured as the number of patients per physicians is positively associated with the noise in physicians' use of information encoded in observables $x_i$. One interpretation is that with a larger number of patients to care for acquiring, keeping and using background information about individual patients is more difficult. If so, physicians would rely less on such information.

**Table 4** Parameter estimates and clinic characteristics

| | Linear regression | | | | | |
|---|---|---|---|---|---|---|
| | Type signal noise | | Symptom signal noise | | Preferences | |
| $N = 283$ | $\sigma_{\xi_j}$ | | $\sigma_{\eta_j}$ | | $b_j/a_j$ | |
| Patients per physician | 0.32 | (0.15)** | -0.14 | (0.13) | 0.03 | (0.01)** |
| Laboratory tests per patient | 0.13 | (0.12) | -0.31 | (0.07)*** | -0.01 | (0.01) |
| Mean number of physicians | -0.28 | (0.26) | 0.03 | (0.16) | -0.03 | (0.02)* |
| Mean age of physicians | -0.49 | (0.80) | 1.12 | (0.52)** | 0.17 | (0.07)** |
| Share of female physicians | -0.04 | (0.12) | -0.06 | (0.08) | 0.02 | (0.01) |
| $\mathrm{R}^2$ | 0.02 | | 0.09 | | 0.05 | |

Notes: Heteroskedasticity-robust standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

The noise parameter for clinical diagnostic information is negatively correlated with the number of laboratory tests a clinic requested per patient. This may reflect that clinics with a higher testing intensity have higher skill in deciding to do additional clinical tests or the ability to better extract information from test diagnostics. Noise in clinical diagnostic information is positively associated with the mean age of physicians in a clinic. Several narratives would support this correlation. For example, older physicians might rely more on their clinical experience and personal knowledge of recurring patients than on costly diagnostic tests. Alternatively, they may be less likely to purchase new diagnostic equipment and rely on existing hardware which they are accustomed to. The results for preference parameter estimates are more difficult to interpret because only the ratio between weights $b_j$ and $a_j$ is identified. Clinics with on average older physicians and more patients treated per physician have larger weights on the antibiotic resistance externality relative to the weight on alleviating patients' sickness cost, while this ratio is negatively associated with clinic size measured as the number of physicians.

### 6.6 Counterfactual policy evaluation

To evaluate the effects of policies on prescription decisions, we report policy outcomes for three counterfactual interventions in Table 5. The first two counterfactuals target physicians' diagnostic skill, the third targets physicians payoff functions. Counterfactual 1 provides physicians with the machine learning prediction type $\mu_i$ for every patient and assumes that physicians use it without noise by setting $\sigma_{\xi_j} = 0$. In this counterfactual, the clinical diagnostic information signal and the payoff function parameter is held fixed. We find that overall prescribing decreases by 17.8 percent (3672 prescriptions) and overprescribing decreases by 33.3 percent (2663 prescriptions). Interestingly, the number of treated bacterial infections is also reduced, by 8 percent (1009 prescriptions). Improved and more precise information on patient type, holding preferences fixed, makes transparent how conservative physicians are on average. It also shows a potential limitation of machine learning predicted risk. Further improvements in predictions should push the change in treated bacterial infections towards zero. In the context of antibiotic prescribing we are more concerned with overprescribing. It is impossible to undo realized antibiotic treatments while decisions to delay prescribing can be corrected after several days, when complete test results are available.

**Table 5** Counterfactual policy outcomes

| | 1. Provide ML-based $\mu_i$ | 2. Improve diagnostic skill | 3. Manipulate payoffs |
|---|---|---|---|
| | Set $\sigma_{\xi_j} = 0$ | If $\sigma_{\eta_j} >$ median($\sigma_{\eta_j}$),<br>set $\sigma_{\eta_j}$ to median | Set $b_j/a_j + 0.08$ |
| Overall prescribing | -17.8 (-3672) | 0.0 ( 3) | -17.8 (-3669) |
| Treated bacterial infections | -8.0 (-1009) | 3.0 ( 372) | -13.6 (-1716) |
| Overprescribing | -33.3 (-2663) | -4.6 (-369) | -24.4 (-1953) |

Notes: This table reports changes to the status quo in percent. Absolute changes are reported in parentheses.

In counterfactual 2, we improve the minimum clinical diagnostic skill to the median level of the physician population. Specifically, we set $\sigma_{\eta_j}$ to the median for every clinic below the median of $\sigma_{\eta_j}$. This could be interpreted as training skills in performing microscopic analysis or investing in technical diagnostic equipment for clinics seemingly having difficulties acquiring clinical diagnostic information. Surprisingly, such an intervention does not affect overall prescribing. Yet, it results in a sizable shift of nearly 400 prescriptions from non-bacterial to bacterial infections.

In counterfactual 3, we manipulate the parameters of the payoff functions while holding fixed patient type and clinical diagnostic information. In particular, we increase the payoff parameter such that the overall reduction in prescribing is equivalent to the counterfactual reduction achieved by providing the machine learning prediction to physicians without noise. The parameter value that achieves this increases $b_j/a_j$ by 0.08. 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 foregone antibiotic treatment. By design, the overall reduction in prescribing is the same as in counterfactual 1. Overprescribing is reduced significantly by 24.4 percent (1953 prescriptions). Yet, manipulating the payoff function weights without improving diagnostic information induces adverse effects as reflected in a large decrease in treated bacterial infections by 13.6 percent (1716 prescriptions). This result illustrates 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 considered here allows for interventions in which the two aims, providing machine learning predictions to experts and incentivizing social behavior, can be implemented and evaluated as complements.

# 7 Conclusion

In this paper, we show how machine learning can create value by providing predictions based on administrative health and socioeconomic data to expert physicians. In many situations, including in health care provision, measurement and selection problems render this evaluation a challenging task. We focus on antibiotic prescribing for suspected urinary tract infections, a setting which is both typical for health care and provides high quality outcomes on which machine learning algorithms can be trained. We first evaluate prescription policies based on machine learning predicted risk relying on the assumption that physicians adhere perfectly to prescription rules. We document that physician decisions vary widely in terms of overprescribing intensity and the number of treated bacterial infections. Therefore, we find that basing prescription rules on predicted risk only for cases in which prediction quality is high maximizes the feasible payoff gains. To relax the perfect adherence assumption and achieve a richer combination of machine learning predictions and diagnostic information held by physicians, we apply a framework of two-dimensional physician diagnostic skill and payoff functions generating antibiotic treatment decisions. We find substantial heterogeneity in the extent of information physicians use based on observable data and clinical diagnostics unobservable to us. In addition, we find large heterogeneity in physician payoff functions determining the tradeoff between the value of curing a patient suffering from a bacterial UTI and the cost of increasing antibiotic resistance due to prescribing an antibiotic treatment.

Several important avenues for further research remain. It would be worthwhile to attempt encoding further clinical information such as recorded symptoms and results from in-clinic diagnostics to further improve machine learning predictions. We also omit an important dimension of antibiotic prescribing, the choice of molecule. It remains an open question for future research to what extent machine prediction of bacterial species and molecule-specific resistances are able to further inform prescription choices. Further research is needed in the analysis of experts' behavioral reactions to prediction-based prescription rules and the communication of machine learning predictions as physicians' incentives to treat and test may change due to these policies. For this, evaluation of prediction-based prescription policies would strongly benefit from interventions in the field.

One limitation is that we consider only initial consultation prescription occasions in which a laboratory test was ordered. Yet, our analysis holds empirical relevance because treatment decisions for suspected UTI, one of the most common types of infections, must be made before definitive diagnostic test results are available. This situation resembles many other health care provision problems. Ribers and Ullrich (2019) analyze the general setting of complete treatment spells in a model of endogenous diagnostic information acquisition, where forward-looking physicians decide whether to treat, including choice of molecule, or wait for laboratory diagnostic results. Extending the approach considered here to a similarly general setting is beyond the scope of this paper.

Our analysis shows the potential of using machine learning predictions for policies in common health care situations. The quality of prediction algorithms and data availability are improving at a rapid pace in many settings in and beyond health care. This paper aims to contribute to solving the important challenge of evaluating how machine learning predictions can improve human expert decisions in the presence of unobservable heterogeneity in skill and payoffs.

# Appendices

## Appendix A Machine learning performance: feature importance

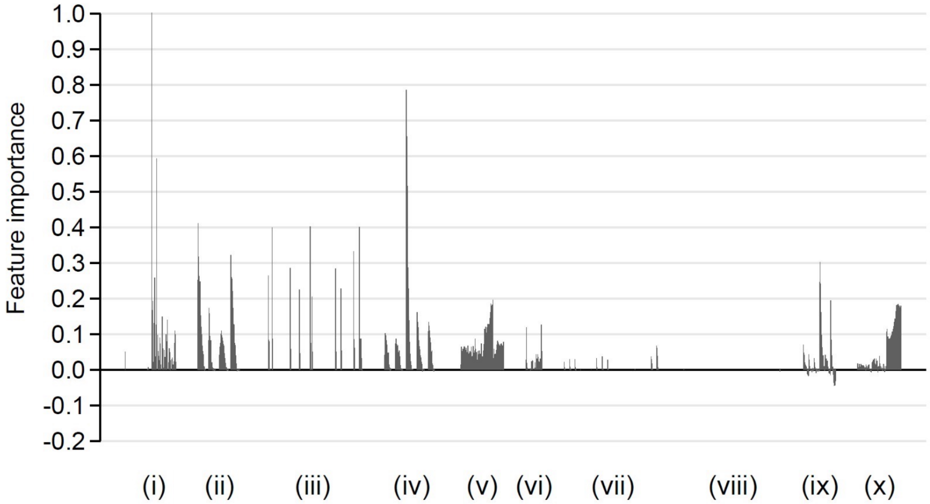


**Figure 9:** Feature importance averaged over the random forests used on each of the 24 monthly folds from January 2011 to December 2012. Variables are listed by groups containing (i) patient characteristics and test timing; (ii) patient past prescriptions; (iii) patient past laboratory test results; (iv) patient past hospitalizations; (v) patient past general practice insurance claims; (vi) household members' past prescriptions; (vii) household members' past laboratory test results; (viii) household members' past hospitalizations; (ix) household members' past hospitalizations; and (x) household members' past general practice insurance claims.

# Appendix B Physician heterogeneity

**Table 6** Physician decisions as predictors and clinic characteristics

| Outcome: $\Delta$ AUC from including treatment decisions as predictor | $\beta$, linear regression | |
|---|---|---|
| Number of clinic's unique patients per physician | 0.004 | (0.007) |
| Number of laboratory results per unique patient | 0.008 | (0.004)** |
| Mean number of physicians | 0.001 | (0.009) |
| Mean age of physicians | -0.078 | (0.032)** |
| Share of female physicians | 0.001 | (0.006) |
| $R^2$ | 0.04 | |

Notes: Heteroskedasticity-robust standard errors in parentheses. Based on 283 out of 482 clinics for which characteristics can be linked. ** $p < 0.05$

# Appendix C Simulation

## Smoothed AR simulator

Assume $a_j^1 = 1$ simulation procedure is as follows:

1. Draw signals, $\xi_{ij}^r$ and $\eta_{ij}^r$ conditional on $m(x_i)$ and observed $y_i$.

2. Compute the expected utility:

$$\mathrm{E}\{\pi \mid d_{ij} = 0, \xi_{ij}, \eta_{ij}\} = -\Pr\{y_{it} = 1 \mid \mu_i\} = \Phi\left(\frac{\bar{\nu} - m}{s}\right) - 1$$

   and

$$a\,\mathrm{E}\{\pi \mid d_{ij} = 1, \xi_{ij}^r, \eta_{ij}^r\} = -b_j,$$

   where $m$ and $\sigma^2$ are functions of $\xi_{ij}^r, \sigma_\xi, \eta_{ij}^r$ and $\sigma_\eta$ as stated in equation 13.

3. Insert the utilities into the logit formula:

$$S_0^r = \frac{e^{\frac{1}{\lambda}\left(\Phi\left(\frac{\bar{\nu}-m}{s}\right)-1\right)}}{e^{\frac{1}{\lambda}\left(\Phi\left(\frac{\bar{\nu}-m}{s}\right)-1\right)} + e^{\frac{1}{\lambda}(-b_j)}} = \frac{1}{1 + e^{\frac{1}{\lambda}\left(1-\Phi\left(\frac{\bar{\nu}-m}{s}\right)-b_j\right)}}$$

   and

$$S_1^r = \frac{e^{\frac{1}{\lambda}(-b_j)}}{e^{\frac{1}{\lambda}\left(\Phi\left(\frac{\bar{\nu}-m}{s}\right)-1\right)} + e^{\frac{1}{\lambda}(-b_j)}} = \frac{1}{1 + e^{\frac{1}{\lambda}\left(\Phi\left(\frac{\bar{\nu}-m}{s}\right)-1+b_j\right)}}.$$

4. Redo steps 1-3 until $R$ repetitions have been reached.

5. The simulated probability can be computed from:

$$\hat{P}_{n0} = \frac{1}{R}\sum_{r=1}^{R} S_0^r \qquad \text{and} \qquad \hat{P}_{n1} = 1 - \hat{P}_{n0} = \frac{1}{R}\sum_{r=1}^{R}(1 - S_0^r) = \frac{1}{R}\sum_{r=1}^{R} S_1^r$$

## Appendix D Parameter estimates

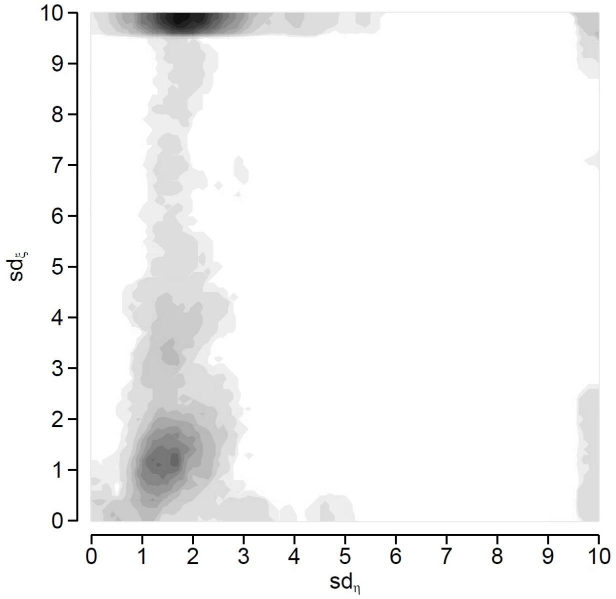


**Figure 10:** Heatmap of physician-level estimates for $\sigma_\xi$ and $\sigma_\eta$. To ensure anonymity, the figure shows a heatmap covering only areas with three physicians or more.

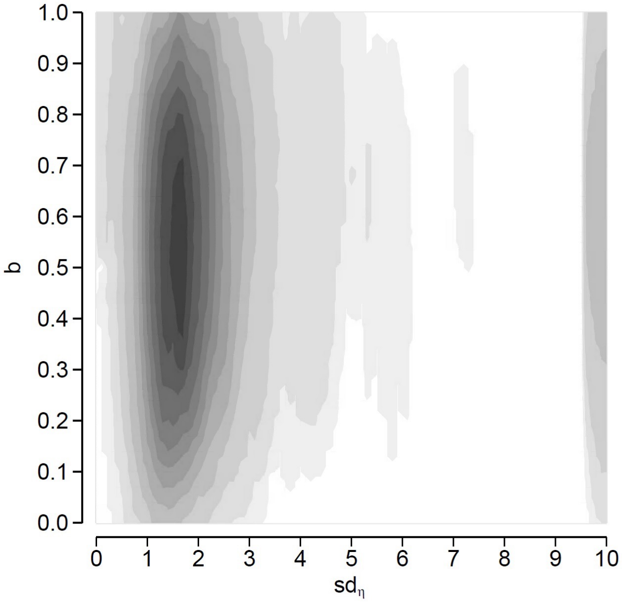


**Figure 11:** Heatmap of physician-level estimates for $\sigma_\eta$ and $b_i/a_i$. To ensure anonymity, the figure shows a heatmap covering only areas with three physicians or more.

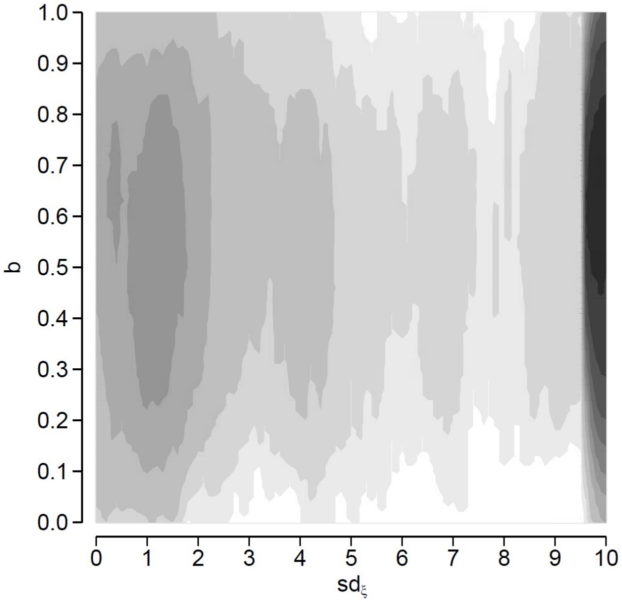


**Figure 12:** Heatmap of physician-level estimates for $\sigma_\xi$ and $b_i/a_i$. To ensure anonymity, the figure shows a heatmap covering only areas with three physicians or more.

# Appendix E Model fit

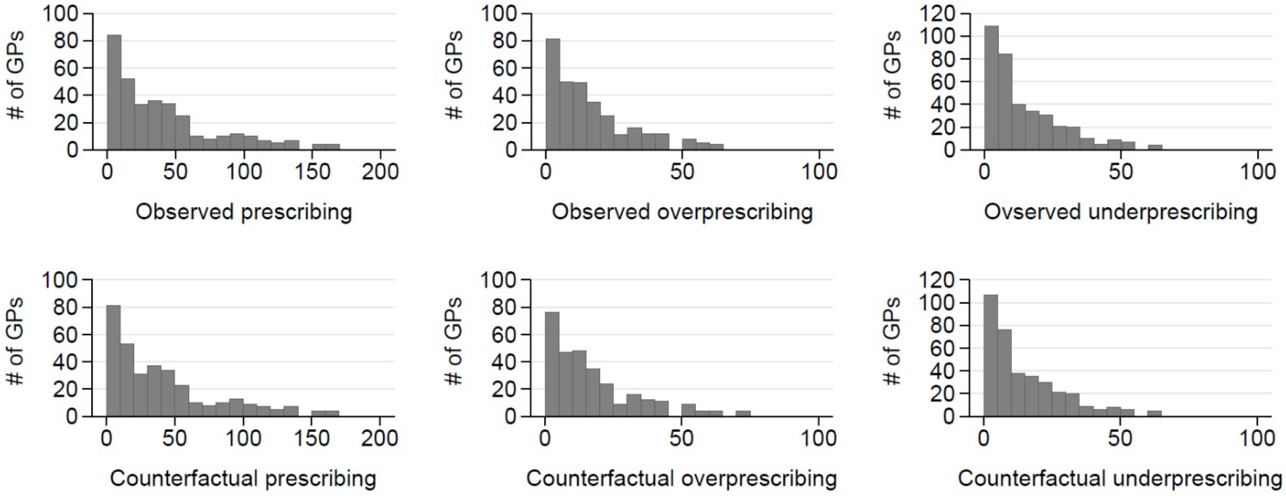


**Figure 13:** Observed and simulated moments

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

[3] 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.

[4] 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.

[5] Bayati, Mohsen, Mark Braverman, Michael Gillam, Karen M. Mack, George Ruiz, Mark S. Smith, and Eric Horvitz (2014), "Data-driven decisions for reducing readmissions for heart failure: general methodology and case study," *PLoS ONE*, 9 (10), e109264.

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

[7] Breiman, Leo (2001), "Random forests," *Machine Learning*, 45 (1), 5-32.

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

[9] 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.

[10] CDC (2015), Outpatient antibiotic prescriptions — United States, 2015, `https://www.cdc.gov/antibiotic-use/community/pdfs/Annual-Report-2015.pdf`, accessed 4/2/2019.

[11] 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.

[12] Chan, David C., Matthew Gentzkow, and Chuan Yu (2019), Selection with variation in diagnostic skill: evidence from radiologists, NBER Working Paper No. 26467.

[13] Chandler, Dana, Steven D. Levitt, and John A. List (2011), "Predicting and preventing shootings among at-risk youth," *American Economic Review*, 101 (3), 288-292.

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

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

[16] 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.

[17] 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.

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

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

[20] 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.

[21] Devillé, Walter L.J.M., Joris C. Yzermans, Nico P. van Duijn, P. Dick Bezemer, Daniëlle A.W.M. van der Windt, and Lex M. Bouter (2004), "The urine dipstick test useful to rule out infections. A meta-analysis of the accuracy," *BMC Urology*, 4 (4), 1-14.

[22] 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.

[23] 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.

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

[25] Goossens, Herman, Matus Ferech, Robert Vander Stichele, and Monique Elseviers (2005), “Outpatient antibiotic use in Europe and association with resistance: a cross-national database study”, *The Lancet*, 365 (9459), 579-587.

[26] 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.

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

[28] Hastings, J.S., M. Howison, S.E. Inman (2020), “Predicting high-risk opioid prescriptions before they are given,” *Proceedings of the National Academy of Sciences*, 117(4), 1917-23.

[29] 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.

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

[31] 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.

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

[33] 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.

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

[35] 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.

[36] 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.

[37] Mullainathan, Sendhil and Ziad Obermeyer (2019), Who is tested for heart attack and who should be: predicting patient risk and physician error, NBER Working Paper No. 26168.

[38] 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.

[39] 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.

[40] Ribers, Michael and Hannes Ullrich (2019), "Prescribing antibiotics under uncertainty about resistance," mimeo.

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

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

[43] 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.