Machine learning and physician prescribing: a path to reduced antibiotic use*

Michael Allan Ribers† Hannes Ullrich‡

April 2023

Abstract

Inefficient human decisions are driven by biases and limited information. Health care is one leading example where machine learning is hoped to deliver efficiency gains. Antibiotic resistance constitutes a major challenge to health care systems due to human antibiotic overuse. We investigate how machine learning provides new opportunities to reduce antibiotic use, with the help of physicians. We focus on urinary tract infections in primary care, a leading cause for antibiotic use, where physicians often prescribe prior to attaining diagnostic certainty. Symptom assessment and rapid testing provide diagnostic information with limited accuracy, while laboratory testing can diagnose bacterial infections with considerable delay. Linking Danish administrative and laboratory data, we optimize policy rules which base initial prescription decisions on machine learning predictions and delegate decisions to physicians where these benefit most from private information at the point-of-care. We find human-algorithm complementarity is essential to achieve efficiency gains with a potential reduction in antibiotic prescribing by 8.1 percent and in overprescribing by 20.3 percent.

*We benefited from helpful suggestions by Jason Abaluck, Rolf Magnus Arpi, Lars Bjerrum, Chiara Canta, Gloria Cristina Cordoba Currea, Greg Crawford, Tomaso Duso, Günter Hitsch, Shan Huang, Ulrich Kaiser, Reinhold Kesler, Jenny Dahl Knudsen, Sidsel Kyst, Chloé Michel, Jeanine Miklós-Thal, Maria Polyakova, Carlo Reggiani, Sherri Rose, Stephen Ryan, Karl Schmedders, Aaron Schwartz, André Veiga, participants at the Annual Health Econometrics Workshop 2018, the 2019 CESifo Area Conference on the Economics of Digitization, the Digital Economy Workshop 2019, the 2019 NBER Conference on Machine Learning in Health Care, the International Conference on Computational Social Science 2020, as well as in seminars at DIW Berlin, ESMT Berlin, Toulouse Business School, University of Copenhagen, University of Zurich, and Vienna University of Economics and Business. 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 and University of Copenhagen, Department of Economics - michael.ribers@econ.ku.dk

‡DIW Berlin and University of Copenhagen, Department of Economics - hullrich@diw.de.

1 Introduction

Professionals and domain experts frequently make costly decisions under time pressure and with limited information, often processed with a host of biases (Thaler and Sunstein 2009, Kahneman et al. 2021). Advances in computing power and rapidly increasing data availability have provided new potential solutions for high-stakes problems with prediction at their core (Kleinberg et al. 2015). Hopes are high that machine learning can help improve human decision making by offering a systematic prediction of the ground truth and guiding optimal decisions. Yet, humans often hold abstract, context-specific information which may be difficult to assess using machine learning (Autor 2015). Employers observe candidates' soft skills in job interviews, judges learn about defendants' personalities in face-to-face questioning, and physicians observe patients' ailments with potentially complex symptoms. Empirical evidence on the relevance and nature of complementarities between data-driven and human decisions is scarce but key for guiding policy-making in response to the economic transformation induced by artificial intelligence.

In this paper, we provide such evidence for a salient case in health care. Antibiotic resistance is one of the greatest threats to global health (WHO 2012, 2014).[1] Because human antibiotic consumption is considered the main driver of antibiotic resistance, reducing the use of antibiotics is a prime policy concern (Goossens et al. 2005, Adda 2020). The decision to use an antibiotic involves a prediction task in determining the cause of a patient's illness. Physicians collect and interpret clinical facts including symptoms, point-of-care test results, and maybe patient's background and medical data, requiring human judgment and curiosity. On the other hand, machine learning has shown to be an effective method to elicit predictive information from large scale data (Agrawal et al. 2018, Athey 2018). It can exploit systematic patterns in data collected across patients and health care providers such as electronic health records, administrative data, and genomics databases.

The treatment of urinary tract infections (UTI) in primary care, a leading cause for human antibiotic use (Grigoryan et al. 2014), provides a unique setting to study the potential to reduce antibiotic use by the means of machine learning-predicted risk. An accurate diagnostic for UTI can only be provided by analysis of urine samples in a microbiological laboratory outside of primary care clinics. These laboratory test results arrive with a delay of several days, corresponding to nearly a full course of antibiotic treatment. Thus, at initial consultations physicians must decide under

[1] Worldwide, 4.95 million deaths are estimated to be associated with antibiotic resistance and 1.27 million deaths are directly attributable (Murray and et al. 2022, Laxminarayan 2022). In the US alone, antibiotic-resistant infections are estimated to cause $20 billion in direct healthcare costs and $35 billion in lost productivity each year (CDC 2013).

uncertainty whether to prescribe an antibiotic or delay treatment until the test result is known.

Crucially, because *ex post* positive and negative laboratory results as well as the initial treatment decisions are observed, prescription decisions can be evaluated based on the true outcome. Hence, we avoid the common selective labels problem for the decision of interest (Lakkaraju et al. 2017, Kleinberg et al. 2018a). To achieve this, we restrict our analysis to consultations at which a laboratory test is acquired. While this restriction may limit the external validity of the quantitative results, which we inspect in robustness checks, our setting provides a unique lens to measure complementarities between physician and prediction-based decisions, providing generalizable insights.

We first apply a machine learning algorithm, XGboost, to high-dimensional, administrative data from Denmark to predict the risk of bacterial presence for 48,406 initial consultations. The outcome is a binary variable indicating when bacteria are isolated in a patient's urine sample in the laboratory. The prediction model includes patients' historical medical outpatient claims, antibiotic prescriptions, microbiological test results, personal characteristics such as gender, age, employment information, education, income, civil status, clinic identifiers, past test yield, time indicators and more. XGBoost predicts bacterial infections out-of-sample with an area under the ROC curve (AUC) of 0.72. This prediction quality is comparable with values in the literature, for example Mullainathan and Obermeyer (2022) with 0.69 for heart attacks, Kleinberg et al. (2018a) with 0.707 for risk of recidivism, and between 0.56 and 0.83 for predicting antibiotic resistance conditional on the presence of bacteria and antibiotic prescription in Yelin et al. (2019) and Kanjilal et al. (2020).

We then define the policy problem as a trade-off between the social cost of prescribing, i.e. promoting resistance, and the health benefits of antibiotic treatment. Using an objective function which reflects this trade-off, we consider policies that reassign antibiotic treatment based on risk predictions with the aim to reduce antibiotic use. Observing that physicians make the fewest errors relative to machine learning in intermediate ranges of predicted risk, we evaluate rules which delay prescriptions until test results are available for low predicted risk, prescribe an antibiotic instantly for high predicted risk, and delegate decisions to physicians for intermediate predicted risk.

Applying this policy, assuming physicians comply, antibiotic use can be reduced by 8.1 percent without reducing the number of treated patients who suffer from a UTI. The policy can reduce overprescribing, prescriptions to non-bacterial cases, by 20.3 percent. In 47.2 percent of consultations, the decision would be made by the prediction-based rule, overturning 15.0 percent of the observed decisions made by physicians. We find that only decision rules that combine machine learning and human decisions improve outcomes, even with the rich individual-specific data in this setting.

A common argument for the superiority of machine learning predictions has been that boundedly rational humans use overly simplified prediction models (Mullainathan and Obermeyer 2022). Using LASSO to select models predicting the test outcome and prescription decisions, we do not find that decisions can be predicted by a model of lower complexity than the model predicting test outcomes. Instead, we find that physicians contribute diagnostic information not encoded in data. To quantify this contribution, we compute the difference between machine learning prediction error and physician decision error. This informational advantage of physicians over machine learning is largest at intermediate ranges of predicted risk and negative at low and high predicted risk. Correlating this measure with point-of-care diagnostic claims, we find that physicians' informational advantage is largest where the use of such diagnostics is highest. Hence, physicians acquire and interpret important information at the point-of-care which is not available to the machine learning algorithm. While information is increasingly encoded for machine learning, the human informational advantage needs to be quantified to identify settings in which complementarities exist.

Finally, we shed light on challenges of a potential implementation. Existing counterfactual policy evaluations have been performed in-sample (Bayati et al. 2014, Kleinberg et al. 2018a, Yelin et al. 2019, Hastings et al. 2020).[2] However, the distribution of individuals to which the policy is applied is typically unknown and may vary. We find that adjusting policy parameters over time suffices to achieve stable out-of-sample outcomes. While the need to adjust predictions over time or other dimensions has been documented, we conclude the same is true for policy designs which automate or delegate decisions. Furthermore, the policies we consider are redistributive. For example, some patients who might need antibiotic treatment must wait under the policy while other patients who are not in need of treatment would now receive an antibiotic right away. A policy which does not redistribute prescriptions between salient socio-economic groups leads to reductions of 6.1-8.1 percent in antibiotic use and 15.3-20.5 percent in overprescribing, quantifying the trade-off between efficiency and between-group fairness.

The paper is organized as follows. Section 2 relates our work to the existing literature. Section 3 provides background information on Danish primary care and UTI and Section 4 describes our data. Section 5 shows the results of the prediction algorithm. Section 6 presents the framework for prediction-based policy rules to improve antibiotic prescribing. Section 7 presents and discusses policy outcomes. Section 8 investigates sources of machine learning and human decision complementarity. Section 9 discusses potential implementation issues and Section 10 concludes.

[2] An exception are experiments, e.g. Dubé and Misra (2023) where targeted pricing improves profits out-of-sample.

2 Related work

We contribute to a growing literature considering prediction problems in management and policy (Kleinberg et al. 2015). Existing work has studied the potential for machine learning to improve decisions such as for crime prevention programs (Chandler et al. 2011), hygiene inspections (Kang et al. 2013), worker productivity in law enforcement and education (Chalfin et al. 2016), C-sections (Currie and MacLeod 2017), tax rebate programs (Andini et al. 2018), opioid prescriptions (Hastings et al. 2020), financial stock analysis (Cao et al. 2021), testing for heart attack (Mullainathan and Obermeyer 2022). We consider professional experts, general practitioners, who reach decisions on behalf of their patients, and how their decisions may be shared between humans and an algorithm when the objective entails a private benefit and a social cost.

Recent work has focused on algorithms as a substitute for human decisions but the question of whether data-driven models can complement human decisions has been investigated at least since Blattberg and Hoch (1990). As data sets have grown in scale and advances in artificial intelligence have enabled increasingly flexible prediction models, the contribution of human intuition and information is becoming more nuanced. Valuable complementarities can arise if humans fill crucial remaining gaps where procedural expertise, subjective evaluations, highly flexible assessments, or domain-specific knowledge of rare events are required, commonly the case in abstract task-intensive occupations such as medical care (Autor 2015). Contrary to Agrawal et al. (2018) and Agrawal et al. (2019), who focus on human judgements which are difficult to encode, and Choudhury et al. (2020), who focus on machine learning biases due to input incompleteness as sources of complementarity, we identify information humans acquire, which remains difficult to encode, as a relevant factor.

Two papers use similar machine learning prediction results to study questions complementary to ours. Ribers and Ullrich (2022) propose and estimate a decision-theoretic model of utility-maximizing physicians. Relying on more restrictive structural assumptions, they analyze the welfare implications of alternative policy designs when diagnostic information and preferences vary between physicians. Instead, we provide a model-free measure of physician private information by analyzing patient-level variation in decision errors. This quantitative measure establishes a mechanism by which an algorithmic decision rule which delegates a share of decisions to physicians can improve outcomes. The risk threshold-based rule we assess allows a transparent patient-consultation-level analysis of the complementarity between human and machine learning-based decisions and can be easier to implement in practice. In a short note, Huang et al. (2022) quantify the returns to

increasing the scope of administrative data. They find that even though returns to the scope of data are decreasing for prediction quality, returns can be increasing for decision outcomes. We focus on the complementarity between machine learning and human decisions with private information as an important mechanism.

Finally, we contribute to the literature on demand-side policy interventions, which include antibiotic prescription surveillance and stewardship programs (Laxminarayan et al. 2013), general practitioner competition (Bennett et al. 2015), financial incentives for physicians (Yip et al. 2010, Currie et al. 2014, Das et al. 2016), education programs (Arnold and Straus 2005, Butler et al. 2012), peer effects (Kwon and Jun 2015), and social norm feedback (Hallsworth et al. 2016).

3 Institutional background and treatment of UTI

3.1 Primary healthcare in Denmark

Denmark has several regulations that impact decision making in primary care. General practitioners act as the primary gatekeepers in a universal and tax financed single payer health care system. Every person living in Denmark is allocated to a general practitioner by a list-system within a fixed geographic radius around the home address. General practitioners work as privately owned businesses but all fees for services are collectively negotiated between the national union of general practitioners and the public health insurer. Physicians do not generate earnings by prescribing drugs to patients who have to purchase their prescriptions from local pharmacies. General practitioners are responsible for prescribing approximately 75 percent of the human consumed systemic antibiotics in Denmark (Danish Ministry of Health 2017). Pharmacies earn a fixed fee per processed prescription regardless of price or other drug attributes, for example branded versus generic drugs. Prescription drugs are subsidized but patients co-pay a fraction of the list price. The Danish market for prescription drugs is highly regulated resulting in low and uniform prices for antibiotics nationwide, about 100 Danish Kroner (15 US Dollars) per complete treatment.

3.2 Diagnosis and treatment of UTI

UTI are among the most common types of infections and a leading reason for antibiotic treatment in primary care (Grigoryan et al. 2014, Gupta et al. 2017). UTIs occur when bacteria, most often *Escherichia coli*, enter the urethra and infect the urinary tract, the bladder, or kidneys. Left untreated, they can lead to sepsis and death. The estimated cost to the health care system attributable to

community-acquired UTI amount to $1.6-3.5 billion per year in the US alone (Foxman 2002, Flores-Mireles et al. 2015). Once diagnosed, the use of antibiotics is indicated by clinical guidelines.[3] In our setting, over 80 percent of UTI-indicated prescriptions are for pivmecillinam, belonging to the class of penicillins and recommended as first-line antibiotic for UTI, or sulfamethizole.[4]

Prevalence of UTI is highest among women. Foxman (2002) reports that nearly half of all women experience at least one UTI in their life. Many more subgroups are known to be at increased risk of UTI, such as children and the elderly, patients with certain conditions such as diabetes or immunodeficiency, or individuals with underlying urological abnormalities (Foxman 2002). Many of such subgroups are identifiable in observable data using personal characteristics such as age and gender or past health care utilization and diagnoses.

UTI symptoms require medical attention. They include dysuria, urinary frequency, urgency, new-onset incontinence, and pain. Systemic signs of an infection such as fever, shivering, or systemic unwellness can also occur. Attributing symptoms to UTI is difficult as they are also associated with other conditions, e.g. sexually transmitted urethritis or vaginitis, noninfectious urethritis, early pyelonephritis, overactive bladder, benign prostatic hyperplasia, bladder or kidney stones, or even a bladder tumor (Wilson and Gaido 2004, Gupta et al. 2017, Nik-Ahd et al. 2018, Holm et al. 2021). Less commonly, UTIs can also be caused by fungi or viruses. Notably, symptoms are difficult to encode systematically. For example, the assessment of "pain" requires contextual elicitation and judgment of its nature, severity, location, and chronology. Beyond symptoms, physicians may elicit contextual information, including behavioral factors, from speaking to patients.

Point-of-care testing such as urinary dipstick and microscopy analysis provides diagnostic results at the consultation. Both types of diagnostics can have very low specificity, the true negative rate, as low as 0.41 or sensitivity, the true positive rate, as low as zero (Devillé et al. 2004, Wilson and Gaido 2004, Chu and Lowder 2018). Further analysis can be done by urine culture which takes about one day. Finally, samples can be sent to a hospital laboratory for a reliable measure of a patient's true infection state. Laboratory testing is highly accurate, requires little human judgement, and has been established as the gold standard for diagnosis. However, test results come with a delay of about three days (Schmiemann et al. 2010). This test can confirm treatment decisions *ex post*, ensure full

[3]See *Medicinrådets behandlingsvejledning vedrørende urinvejsinfektioner* (https://medicinraadet.dk/media/ucsjy4e4/medicinrådets-behandlingsvejledning-vedr-urinvejsinfektioner-vers-1-1_adlegacy.pdf) or *Urinary Tract Infections* (https://www.mayoclinic.org/diseases-conditions/urinary-tract-infection/symptoms-causes/syc-20353447) by the Cleveland Clinic, accessed 11/2/2022.

[4]Less frequently used antibiotics are nitrofurantoin, trimethoprim, amoxicillin, fluoroquinolones, and fosfomycin.

information is available to adjust treatment later, and provide antibiotic resistance information.

4 Danish Administrative Data and Laboratory Test Results

4.1 Danish national registries

The administrative data provided by Statistics Denmark cover all citizens and residents in Denmark between January 1st, 2002, and December 31st, 2012. The demographic data from the Danish Civil Registry (*Det Centrale Personregister, CPR*) includes gender, age, municipality, immigration status and place of origin, marriage and family status. It provides a unique person identifier which facilitates accurate linkage of patients between Danish national registers. It also includes household member identifiers which allow us to link the patient's family and household members including their demographic and administrative data. We also obtain information on employment (*Integrerede Database for Arbejdsmarkedsforskning, IDA*) and education (*Uddannelseregister, UDDA*).

The prescription drug register (*Lægemiddeldatabasen, LMDB*) contains each individual's complete purchase history of systemic antibiotics, including the date of purchase, patient and prescribing physician identifiers, and product information. The hospitalization data (*Landspatientregisteret, LPR*) comprise all patient contacts with hospitals, including ambulatory visits. The data include admission and discharge information, procedures performed, type of hospitalization (ambulatory, emergency, etc), diagnoses, and the number of bed days. The claims data (*Sygesikringsregisteret, SSR*) cover all medical services provided to the population of patients in primary care, including consultation week, services provided, and physician fees. Primary care providers are identified via unique clinic identifiers which can be linked to physicians' personal identifiers (*Yderregister, YDER*).

4.2 Microbiological laboratory data

Herlev hospital and Hvidovre hospital, two major hospitals in Denmark's capital region covering a catchment area of roughly 1.7 million people, provided us with test results from their clinical microbiological laboratories between January 1st, 2010, and December 31st, 2012. The data contain patient and clinic identifiers as well as information on test type, sample date, arrival date at the laboratory, result date, isolated bacteria, and antibiotic-specific resistances of isolated bacteria.

The laboratory test data are central because they reveal bacterial presence in a urine test sample, the outcome we aim to predict. According to the Danish guidelines urinalysis should only

be performed in patients with signs and symptoms of UTI.[5] The test procedure takes 3.1 days on average, during which physicians are uninformed about the test result. Since we know the precise timing of test acquisitions, prescription purchases, and the test response date, we can determine physicians' treatment decisions prior to being informed about test outcomes.

4.3 Analysis sample

Overall, the data contain 2,579,617 biological samples submitted for testing in the capital region by both general practitioner clinics and hospitals. Urine samples constitute 477,609 samples out of which 156,694 are marked as general practitioners by the laboratory. Some clinics submit mainly specialist fee claims to the health care system. We drop these to ensure the sample includes only general practitioners. To focus on consultations that constitute a first contact with a physician, we exclude observations where a patient received a systemic antibiotic prescriptions or had laboratory test conducted within 4 weeks prior to the observed test date. In these situations, physicians are unlikely to hold prior diagnostic information and must prescribe under uncertainty. By considering such initial consultations, we exclude potentially complicated treatment spells where patients are tested in later stages. We also avoid patients in long-term treatment, potentially due to severe antibiotic resistance problems. Additionally, we exclude urine samples collected during pregnancy as the vast majority of these are mandatory routine checks and do not represent UTI consultations. The final analysis sample consists of 65,919 initial consultations where a urine sample was sent to a laboratory for testing from 583 primary care clinics.

4.4 Laboratory test outcomes and prescribing

We consider binary test outcomes that indicate whether bacteria are isolated in patients' urine samples and do not focus on specific bacterial species.[6] We observe when a test is acquired from the patient at an initial consultation and the initial prescription decision when a prescription for a systemic antibiotic is purchased at a pharmacy on the test day or the day after.[7]

[5] See https://medicinraadet.dk/media/ucsjy4e4/medicinrådets-behandlingsvejledning-vedr-urinvejsinfektioner-vers-1-1_adlegacy.pdf, accessed 11/2/2022.

[6] In the policy analysis we describe the distribution of bacterial species to consider potential reasons for disagreements between machine learning and physician decisions. *Escherichia coli* represent 71 percent of cases in our data.

[7] We only observe the purchase date of a prescription which might differ from the date the physician provided the patient with the prescription. Hence, we must define what constitutes an initial prescription and choose to do so based on the patient purchasing the antibiotic on the day of the test or the following day. Defining initial prescriptions

Table 1 shows that the bacterial rate and prescription rate remain stable at 37-39 percent over the three sample years. This suggests that physicians match antibiotic prescriptions to bacterial infections very well at the initial consultation. Yet, the prescribing rates conditional on test outcome show that this is not the case. Physicians only prescribe antibiotics at initial consultations to 61 percent of patients with bacterial infections, implying underprescribing to 39 percent. Conversely, 26 percent of patients with a negative test result receive an antibiotic at the initial consultation, defined as overprescribing. Hence, the descriptives indicate a potential for improving physician decisions in treating UTI patients.

Table 1 Summary statistics for laboratory tests and initial antibiotic prescribing.

	All tested			Positive test		Negative test	
	N	Bacterial rate	Prescribing rate	N	Prescribing rate	N	Prescribing rate
2010	17,513	0.37	0.39	6,411	0.60	11,102	0.27
2011	21,237	0.39	0.39	8,305	0.60	12,932	0.25
2012	27,169	0.39	0.39	10,510	0.61	16,659	0.25
Total	65,919	0.38	0.39	25,226	0.61	40,693	0.26

5 Machine learning and physician decisions

5.1 Predicting bacterial UTI using administrative data

We use the machine learning algorithm XGBoost (Hastie et al. 2009, Chen and Guestrin 2016) to relate patient i's covariates x_i to the binary laboratory bacterial test outcome, y_i. XGBoost is an implementation of the extreme gradient boosted regression tree method which provides a non-parametric risk prediction. The vector x_i contains 1,557 patient-specific covariates which may, in principle, be observable to the physician at the time of consultation.[8] The covariates in the prediction model include each patient's past medical outpatient claims, antibiotic purchases, microbiological test results, a rich set of characteristics such as gender, age, employment, education,

as any antibiotic purchased between the test date and the date the laboratory answer is provided to the physician does not qualitatively change the result of our analysis. We choose the shorter definition of an initial prescription for our main analysis as we want to exclude potential prescriptions that result from unobserved additional contact between the patient and the physician while awaiting the test result.

[8] Out of the 1,557 covariates 1,038 are categorical variables that are transformed into dummy variables for each category. The final number of covariates for XGBoost is 12,727.

income, civil status and more, as well as the same information on each individuals' household members. We also include clinic identifiers, clinic-level past average resistance, and regional prescribing rates to account for clinic-specific practice styles.

We use data from 2010 for hyperparameter tuning and create 24 monthly out-of-sample policy evaluation partitions from January 2011 to December 2012. For each policy evaluation partition, we retrain the XGBoost algorithm using all patient observations prior to the partition as training data. This procedure aims to strike a balance between the computational cost of frequent updating of the prediction algorithm and the desire to use the most recent historical data relative to a consultation. Figure 7 in Appendix A.1 illustrates the hyperparameter search partitions as well as the training and policy evaluation data partitions. Table 5 in Appendix A.2 reports the tuning results.

We report three measures of predictor importance for XGBoost – gain, frequency, and cover – in Figure 8 and Table 6 in Appendix A.3. Across these measures, age, gender, clinic identifier, and recent antibiotic prescriptions are among the top 30 predictors reported in Table 6 in Appendix A.4. Further important predictors include a patient's most recent antibiotic resistance results, clinic-specific resistance levels, regional prescription intensity, hospital stays, as well as a patient's education, immigration status, and origin country. While many plausible narratives may relate these predictors to bacterial outcomes, machine learning algorithms do not have causal content and so we refrain from further interpretation.

The AUC on the joint set of partitions is 0.721 with the associated ROC curve reported in Figure 9 in Appendix A.5. This AUC value falls in the ranges of prediction quality in the literature, for example Mullainathan and Obermeyer (2022) with 0.69 for heart attacks, Kleinberg et al. (2018a) with 0.707 for risk of recidivism, and between 0.56 and 0.83 for predicting antibiotic resistance conditional on the presence of bacteria in Yelin et al. (2019) and Kanjilal et al. (2020).

Figure 1 shows machine learning predicted risk, $m(x_i)$, and test outcomes for all test observations in the joint set of 24 monthly out-of-sample policy partitions. We sort all patients by their predicted risk and compute average bacterial outcomes for consecutive bins of 100 patients. One bin is represented by one sphere. Outcomes are close to the 45 degree line throughout the risk distribution, showing that the algorithm on average correctly predicts bacterial risk.

Our implementation is standard with the exception that we cannot split our data randomly into training and out-of-sample partitions using k-fold cross-validation. In practical applications the prediction function must be constructed at or prior to the clinical consultation using historic data only. Splitting the data randomly could lead to spill-overs across time as past outcomes may

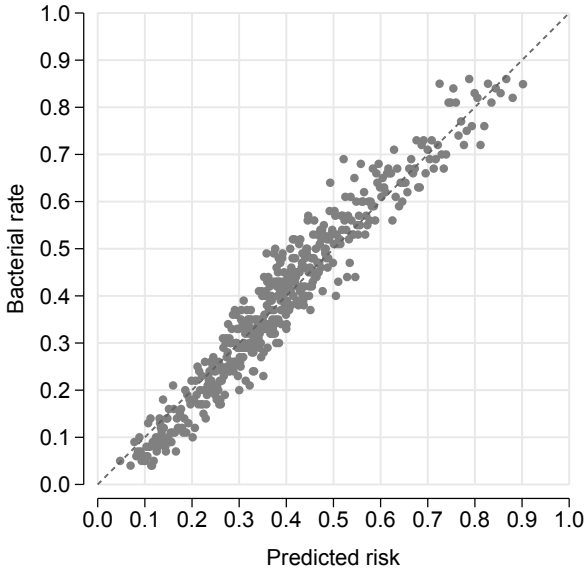


Figure 1: Laboratory test outcomes relative to predicted risk of bacterial UTI. Spheres and triangles represent bins of 100 patients sorted by predicted risk.

be predicted using a model trained on future observations. To verify that the non-random monthly partitions do not result in overfitting, we show out-of-sample AUC for all 24 policy evaluation partitions in Table 7 in Appendix A.6. Sample sizes, bacterial and prescribing rates, risk predictions, and prediction quality is stable across monthly out-of-sample predictions, suggesting that there is little distributional shift. We further inspect the potential risk of overfitting by training XGBoost using 2010 data only and predict on the complete data in 2011 and 2012. Even though we forego the use of increasing amounts of training data over time, following this static approach results in an out-of-sample AUC of 0.709 for 2011 and 2012, slightly below the value achieved using the main procedure.

A further potential source of overfitting may be that XGBoost recovers overly flexible conditional expectation functions on high-dimensional data. To insure against this risk of overfitting and inspect the relevance of model uncertainty, we reproduce our prediction exercise using parametric logistic LASSO. Using the same tuning and training procedure as described for XGBoost, we obtain an out-of-sample AUC of 0.707, which is just below the value achieved using XGBoost.[9]

Finally, while machine learning predictions cannot be expected to extrapolate beyond our sample, we can provide a partial assessment. Figure 10 (a) in Appendix A.7 shows the distribution of risk predictions for a subset of the general population sampled on a random day with no con-

[9]The optimal tuning parameter lambda is 0.0087 on the hyperparameter folds.

sultation.[10] This distribution resembles the risk distribution in the analysis sample for patients without a bacterial infection. A notable difference is the larger density at low-risk predictions for the random population sample, which is driven by a larger proportion of men who on average exhibit lower risk of UTI. Analogously, Figure 10 (b) in Appendix A.7 shows the distribution of risk predictions for patients who were prescribed a UTI-indicated antibiotic but are not in our analysis sample because no laboratory sample was collected.[11] The distribution of risk predictions closely resembles the analysis sample for patients with a bacterial infection. These observations suggest that the prediction model is also informative for patients outside of the analysis sample.

5.2 Bacterial rate conditional on predicted risk and physician prescribing

Motivated by the trade-off between the benefit and the social cost of antibiotic use, we focus is on the binary choice of prescribing an antibiotic and not on molecule choice. Figure 2 splits the sample into patients who received a prescription (treated) and those who did not receive a prescription (non-treated) at the initial consultation. Again, each group is sorted by predicted risk and arranged into bins of 100 patients. Hence, the figure shows test outcomes versus risk predictions conditional on antibiotic prescribing prior to receiving test results. Conditional on predicted risk, patients with an initial prescription have higher bacterial rates than patients without an initial prescription. Hence, physicians appear to have diagnostic information which the machine learning algorithm does not capture. For example, point-of-care testing and symptom assessment provide instant, albeit imperfect, diagnostic information which is not included in administrative data. The difference in bacterial rates is largest for intermediate predicted risk, which represents the set of patients for which machine learning predictions are the least informative.

Even though physicians appear to have important private diagnostic information, prescriptions often do not match the true test outcomes. On average, 39.6 percent of patients who received an antibiotic did not have a bacterial infection and the overprescribing rate varies drastically with predicted risk. Among the 100 treated patients with the lowest predicted risk, the left most triangle in Figure 1, only 27 patients had a bacterial infection resulting in 73 percent overprescribing. In contrast, 87.5 patients had a bacterial infection among the 100 treated patients with the highest predicted risk. Among the untreated, 25.1 percent of patients have bacterial infections. The error rate again varies with predicted risk showing an increasing bacterial rate for the non-treated patients

[10] The sample is drawn such that it has the same number of observations as the analysis sample for $y = 0$.

[11] The sample is drawn such that it has the same number of observations as the analysis sample for $y = 1$.

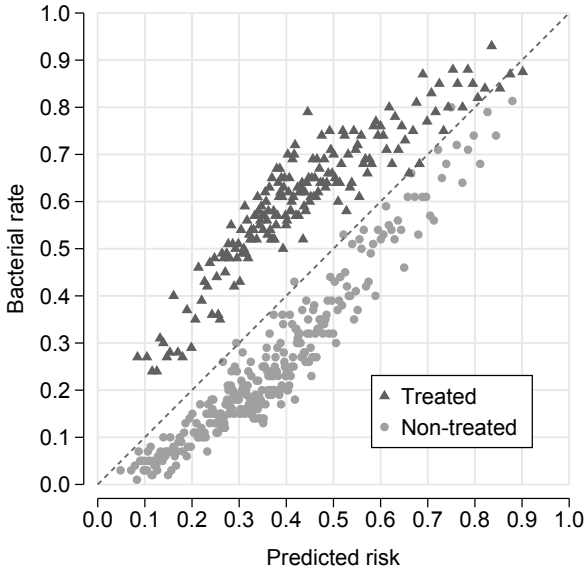


Figure 2: Laboratory test outcomes relative to predicted risk of bacterial UTI conditional on antibiotic prescribing prior to receiving test results. Spheres and triangles represent bins of 100 patients sorted by predicted risk conditional on treatment.

as predicted risk increases. Among the 100 non-treated patients with the highest predicted risk, the right most sphere on Figure 1, 81 patients had a bacterial infection. These observations indicate that the match between prescriptions and bacterial infections can be improved at the extremes of the risk prediction range where machine learning classification accuracy is high and physician decisions reflect considerable over- and underprescribing.

6 Prediction-based prescription policy

6.1 Payoff from antibiotic prescribing

Physicians face a trade-off when they prescribe antibiotics as the potential curative effect must be weighed against the cost of promoting antibiotic resistance (Adda 2020). This cost is incurred every time an antibiotic is consumed regardless of whether the patient suffers from a bacterial infection or not. In contrast, antibiotics only have a curative effect for bacterial infections. We focus on antibiotic prescription decisions at the initial consultation of a sickness spell where the patient is suspected of suffering from a UTI and a urine sample is collected for laboratory testing. Test results are on average available within 3.1 days after which patients can be treated accordingly. Yet, delayed treatment to a patient who suffers from a bacterial infection comes at a sickness cost to the patient

in the waiting period. Hence, the prescription decision at the initial consultation represents a trade-off between the social cost of prescribing and the patient sickness cost from delayed prescribing until the test result is available. We write the policy maker's realized payoff at initial consultations as

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

where $y \in \{0, 1\}$ with $y = 1$ if the patient has a UTI and $y = 0$ otherwise. The decision $d \in \{0, 1\}$ is $d = 1$ if an antibiotic is prescribed and $d = 0$ otherwise. The parameter $\alpha > 0$ is the weight on the patient's sickness cost while awaiting the test result and the parameter $\beta > 0$ reflects the social cost of prescribing.[12]

6.2 Policy rules

We document in Section 5 that overprescribing, prescriptions to patients with negative test results, occurs most frequently at low predicted risk and decreases on average as predicted risk increases. Equivalently, underprescribing, delaying prescriptions to patients with positive test results, occurs most frequently at high predicted risk and decreases as predicted risk decreases. This motivates prescription rules of the form:

$$\delta_i(k_L, k_H) = \begin{cases} 0 & \text{if } m(x_i) \leq k_L, \\ d_i & \text{if } k_L < m(x_i) < k_H, \\ 1 & \text{if } k_H \leq m(x_i), \end{cases} \tag{2}$$

where $m(x_i)$ is the machine learning prediction for patient i as a function of patient observables x_i, d_i is the observed prescription decision, and (k_L, k_H) are threshold parameters to be determined subject to $0 \leq k_L \leq k_H \leq 1$. This rule postpones prescribing until test results are available for patients

[12] An alternative payoff function that includes the potential social cost of a follow-up prescription to a patient who suffers from a bacterial UTI but did not receive antibiotic treatment at the initial consultation has the following form:

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

where $d \in (0, 1)$ is the prescription decision at the initial consultation, $y \in (0, 1)$ is the sickness state, and $\rho \in (0, 1)$ is the spontaneous natural recovery rate that occur while the patient await the test results. Hence, a similar expression to equation (1) except that the interpretation of the sickness cost differs.

with low predicted risk, $m(x_i) \leq k_L$, give antibiotic prescriptions to patients with high predicted risk, $k_H \leq m(x_i)$, and delegate decisions to physicians for intermediate risk, $k_L < m(x_i) < k_H$.[13]

6.3 Policy objective

Policies are evaluated by aggregating payoff differences between the counterfactual prescription rules in equation (2) and observed prescription choices:

$$\Pi = \sum_{i \in \mathcal{I}} [\pi(\delta_i; y_i) - \pi(d_i; y_i)] = \alpha \sum_{i \in \mathcal{I}} y_i(\delta_i - d_i) \; - \; \beta \sum_{i \in \mathcal{I}} (\delta_i - d_i). \tag{3}$$

The aggregation is over the set of patient indices $\mathcal{I}$ that cover the policy evaluation period.

The effect of a policy can be decomposed in two terms. The first term accrues from an increase in the number of correctly treated UTI patients. The second term accrues from the change in overall antibiotic use. If a prescription rule increases the number of treated UTI while reducing the overall number of antibiotics used, a policy maker will be better off regardless of weights α and β, as each term in equation (3) is positive. However, depending on a policy maker's weights it can be optimal to implement a rule that increases the number of untreated UTI to accomplish a larger reduction in antibiotic use than would otherwise be possible. Equivalently, a policy maker might prefer to increase the number of treated UTI at the expense of increasing overall antibiotic use.

As we do not observe α and β, we cannot know the trade-off a policy maker would prefer. Instead, we follow the idea in Kleinberg et al. (2018a) and focus on a particular objective that aims to lower antibiotic use while keeping the number of treated UTI unchanged. If we can show that this prescription rule can reduce overall antibiotic use then the policy maker receives a positive payoff regardless of $\alpha > 0$ and $\beta > 0$. [14] Hence, we choose k_L and k_H that solves

$$\min_{k_L, k_H} \; \sum_{i \in \mathcal{I}} \delta_i(k_L, k_H) - d_i \quad \text{s.t.} \quad \sum_{i \in \mathcal{I}} y_i(\delta_i(k_L, k_H) - d_i) = 0. \tag{4}$$

The resulting policy parameters from equation (4) also minimize overprescribing since the change

[13] An alternative policy could include the physician's decision as a predictor and evaluate a decision rule using a single threshold k. While allowing more flexible combinations of physician decisions with other variables in the prediction algorithm, an implementation would involve higher physician effort because her decision would be a required input at every consultation. Huang et al. (2022) use such a rule and find similar results as for the policy we consider.

[14] To obtain a positive payoff difference in equation (3) it is irrelevant if $\alpha > \beta$ or $\alpha < \beta$. However, for $\alpha < \beta$ it is never optimal to prescribe prior to observing test results. We observe significant prescribing which is recommended practice by prescription guidelines (Danish Health and Medicines Authority 2013), suggesting that $\alpha > \beta$.

in prescribing to non-UTI cases can be written

$$\sum_{i\in\mathcal{I}} \delta_i(1-y_i) - \sum_{i\in\mathcal{I}} d_i(1-y_i) = \sum_{i\in\mathcal{I}} (\delta_i - d_i) - \sum_{i\in\mathcal{I}} y_i(\delta_i - d_i), \tag{5}$$

which equals the change in antibiotic use when the change in treated UTI, the last term, is zero.

7 Counterfactual outcomes

7.1 Reducing antibiotic use

We measure counterfactual policy outcomes relative to observed levels in years 2011 and 2012. We construct the 95% confidence intervals by re-computing policy results over 100 bootstrap samples for fixed patient risk predictions $m(x_i)$ and policy parameters (k_L, k_H).

Table 2 reports outcomes where policy parameters are chosen to reduce antibiotic use without treating fewer UTI patients. The policy with $k_L = 0.320$ and $k_H = 0.601$ results in a reduction in overall antibiotic use of 8.1 percent and a reduction in overprescribing of 20.3 percent relative to observed decisions. Physicians' decisions are overruled and reversed for 15 percent of cases. Out of all consultations, 52.8 percent are delegated to physicians.[15]

To qualify these findings, we can relate the change in prescribing to the national action plan initiated by the Danish government in 2017 which aimed to reduce overall antibiotic prescribing by one third within three years (Danish Ministry of Health 2017). For the UTI consultations we consider, the reduction of 8.1 percent would achieve one fourth of this goal. One important limitation regarding this interpretation is that our sample comprises only initial consultations during which a urine sample was taken for laboratory testing. We cannot exclude that, at the extreme, no reductions in antibiotic use may be possible for other consultations while keeping the number of infections treated with an antibiotic fixed.

To shed some light on external validity, we can inspect the policy on the random sample drawn from the general population as described in Section 5. In this sample, one percent are above $k_H = 0.601$ and, hence, would be given an antibiotic compared to 5.6 percent of the non-UTI cases observed in our analysis sample. For the sample of UTI-indicated prescriptions without laboratory

[15]Table 8 in Appendix C reports results based on LASSO predictions showing similar policy outcomes. The lower prediction quality achieved by LASSO results in a ten percentage points larger share of decisions delegated to physicians and a reduction in antibiotic use by seven percent. While the somewhat better policy outcomes using XGBoost can justify the use of better algorithms, the general result that the combination of physician decisions and machine learning leads to improvements does not appear to depend on the choice of algorithm.

Table 2 Counterfactual outcomes for 2011 and 2012

k_L	0.320
k_H	0.601
Change in treated UTI, in %	0.0 [−1.2, 1.0]
Change in antibiotic use, in %	−8.1 [−8.9, −7.4]
Change in overprescribing, in %	−20.3 [−21.9, −18.9]
Physician decisions overruled, in %	15.0 [14.7, 15.3]
Patients delegated to physicians, in %	52.8 [52.3, 53.3]
Consultations	48,406
UTIs	18,815
Treated UTIs	11,402
Antibiotic prescriptions	18,872
Overprescribing	7,470

95% confidence intervals are based on 100 bootstrap samples of 2011 and 2012 where machine learning predictions and the policy parameter (k_L, k_H) remain fixed.

testing, 15.2 percent of observations are below $k_L = 0.320$, hence would not receive an antibiotic, compared to 17.7 percent with a bacterial infection in our analysis sample. These false negative rates are comparable between the two distributions and the lower out-of-sample rate is expected given that UTI-indicated prescriptions include (unobserved) false positives.

To provide further insights on the relevance of conditioning on laboratory testing, we investigate the sensitivity of the counterfactual results to varying intensity of test selection. In the claims data, we observe the use of rapid diagnostics such as dipstick tests and microscopy analysis. As either one of these diagnostics is typically used at UTI consultations, these claims provide an approximate number of UTI consultations for each primary care clinic. To quantify test intensity, we divide the clinic-specific number of laboratory tests in the analysis sample by the number of rapid diagnostics performed at all initial consultations with or without a laboratory test. Figure 11 in Appendix D shows the counterfactual reduction in antibiotic use conditional on varying test intensities and Figure 12 in Appendix D the associated sample sizes. The solid line shows results for all samples from clinics above or equal to the testing intensity threshold. The dashed line shows results for all samples from clinics below the threshold. Policy results are close to our main result and their confidence intervals largely overlap. These results indicate that the quantitative results are robust for a range of testing intensities and may not be strictly limited to our specific sample.

7.2 Comparing policy outcomes with prescribing under full information

The evaluated policy may give antibiotics unnecessarily, even if prescriptions are assigned to patients who test positive for UTI. For example, asymptomatic infections may be purposely left untreated even though physicians have an accurate evaluation of the risk of a positive test result. To investigate whether the policy gives antibiotics to high predicted risk patients who would not be treated even under full information about the presence of bacteria, we consider physicians' follow-up prescription choices when the definitive test outcome is known. We focus on 1,820 patients with a positive test result to whom the counterfactual policy assigns an instant antibiotic prescription but physicians did not. For these patients, we find that 71.8 percent receive an antibiotic prescription after the definitive arrival of microbiological test results. [16] With an estimated 24 percent spontaneous recovery rate (Ferry et al. 2004), this suggests that prescriptions based on machine learning predictions resemble physician choices under full information.

The converse problem is that physicians may give more than 71.8 percent follow-up prescriptions to patients with bacterial infections for whom the policy delays antibiotic use. The follow-up rate for these 1,820 UTI cases is counterfactual, hence unobserved. In a worst-case scenario, the counterfactual follow-up rate could be as large as 100 percent. In this case, the policy reduction in antibiotic use changes from 8.1 percent to 5.3 percent while the reduction in overprescribing remains unchanged at 20.3 percent. If the counterfactual follow-up rate is lower than 71.8 percent, the reduction in antibiotic use would be larger. In fact, the rate of follow-up prescribing to all patients who did not receive an initial prescription but showed a positive test result is 63.8 percent.

7.3 Waiting for molecule-specific resistance information

One reason that could lead physicians to postpone treatment to high-risk patients might be a lack of information about a patient's antibiotic resistance profile. To avoid prescribing an ineffective antibiotic, the physician may choose to wait for the test results even if predicted bacterial risk is correctly evaluated to be high. To understand the importance of this potential reason for postponing treatment, we analyze bacterial species and resistance profiles for patients with high predicted risk, $m(x_i) > k_H$, conditional on physicians' initial prescription decisions. Physicians might know with high certainty that a patient's symptoms are caused by a bacterial infection and suspect bacteria to

[16] Some times, partial test results are communicated or patients re-visit the physician before the average test delay of three days, which we do not observe. Based on follow-up prescriptions from two days after the initial consultation or later, the share is 75.5 percent.

be resistant against one or several antibiotics. They may then decide to wait for further information about which antibiotic to use. If so, for high-risk patients, we would expect resistance rates to be higher when physicians wait at the initial consultations compared to when they prescribe instantly.

Table 10 in Appendix E shows the distribution of bacterial species. We observe some differences in the detected bacteria for treated and untreated patients with high predicted bacterial risk, which can be explained by point-of-care diagnostics that provide information for identifying, for instance, E. coli bacteria, the main cause of UTI.[17] Table 11 in Appendix E shows resistance rates for E. coli bacteria. We find small differences in resistances against most antibiotics prescribed for UTI. When physicians treat instantly and bacteria are found, these have one to five percentage points lower resistance levels than when physicians decide to wait and bacteria are found. One possible explanation is that physicians have informative priors about levels of antibiotic resistance and consider them when deciding to treat instantly or to wait for complete test results. Quantitatively the differences do not appear of first-order importance. Yet, there seems to be value in addressing prediction of specific bacteria and resistances in further research. In hospital contexts, Yelin et al. (2019) and Kanjilal et al. (2020) find promising results for predicting resistance levels.

8 How does the policy achieve improvements?

8.1 Prescribing without physicians

In the counterfactual policy the majority of decisions, 52.8 percent, is delegated to physicians. A natural question is how well a policy would fare if all decisions were made by the algorithm. To explore this policy, we impose $k_L = k_H$ in equation (2), collapsing the share of decisions delegated to physicians to zero. In this restricted form, the prescription rules become step functions where prescriptions are never given below the cut-off, $k \equiv k_L = k_H$, and always given above.

Table 3 shows the policy outcomes without physician input. Here, 39.7 percent of physicians' decisions are overturned as a prescription is given to all patients with predicted risk equal to 0.405 or higher. Notably, a reduction in antibiotic use is not possible. Instead, antibiotic use increases by 7.1 percent and overprescribing increases by 17.9 percent.[18] We conclude that even with high-

[17] Nitrite dipstick detect bacteria that transform Nitrate to Nitrite, which belong to the genera Escherichia, Enterobacter, Klebsiella, Citrobacter, and Proteus. The non-detectable genera are Staphylococcus, Pseudomonas, Enterococci, Acinetobacter, and Streptococcus.

[18] Similar policy are obtained using LASSO predictions as reported in Table 9 in Appendix C. The lower prediction quality achieved by LASSO results in an increase in antibiotic use by 11.3 percent and an increase in overprescribing

dimensional individual-specific data, machine learning predictions need to be combined with physician expertise to provide policy improvements.

Table 3 Counterfactual outcomes for 2011 and 2012, no physician input

k			0.405
Change in treated UTI, in %	0.0	[−1.6,	1.7]
Change in antibiotic use, in %	7.1	[5.8,	8.6]
Change in overprescribing, in %	17.9	[16.0,	20.6]
Physician decisions overruled, in %	39.7	[39.4,	40.2]

95% confidence intervals are based on 100 bootstrap samples of 2011 and 2012 where machine learning predictions and the policy parameter k remain fixed.

8.2 Physician bounded rationality

Existing literature has associated decisions error with human bounded rationality, such that decision makers may use overly simplified prediction models (Camerer 2019). If so, machine learning can help overcome such human limitations. To assess whether physicians may be using overly simple prediction models to inform their decisions, we follow Mullainathan and Obermeyer (2022) by comparing the complexity of models that predict the test outcome y and the physician decision d. We train two separate LASSO-logit models for the outcome variables y and d using all potential predictors in our main analysis and data from 2010. We vary the regularization parameter to induce a sequence of models with increasing numbers of predictor variables selected up to a maximum of 10'000 variables. We then predict both outcomes using each of the selected models on all data in 2011 and 2012 to assess potential differences in the complexity of models predicting y and d.

Figure 3 shows the AUC values for all prediction models and the associated number of selected predictors. The light gray line represents all models predicting d and the dark gray line shows all models predicting y. Confidence intervals at the 95% level are computed using bootstrap holding the selected model fixed. The best selected model uses 369 variables to predict y and 406 variables to predict d but the highlighted maximum AUC-values lie within the AUC confidence intervals over large ranges of model complexity. This finding indicates that similar complexity is required for predicting either outcome. Thus, we do not find evidence consistent with observations in other

by 28.7, even though the share of overruled physician decisions of 41.2 is close to the XGBoost results. While XGBoost predictions achieve better policy outcomes, its much higher flexibility does not improve policy outcomes such that machine learning alone could be used to make decisions.

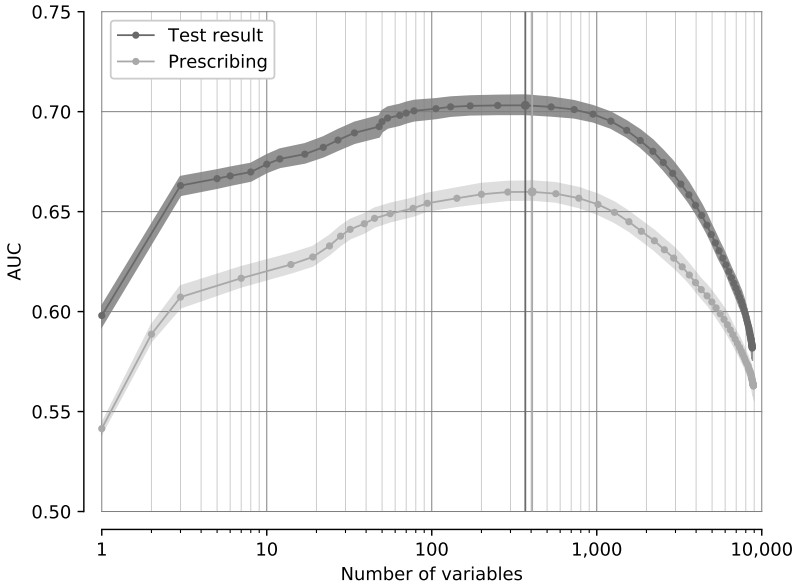


Figure 3: Model complexity for predicting bacterial outcomes and physician decisions

contexts that gains from machine learning are due to overly simple human prediction models using observable variables.

8.3 Private diagnostic information

Physician decisions and machine learning may be complements if physicians hold private information. Private information can be acquired during consultations and is not fully amenable to inclusion in data sets for machine learning. To explore the nature of private diagnostic information, we analyze the distribution of a measure of physician information relative to machine learning predictions. We relate this measure to two potentially important sources of private diagnostic information: rapid point-of-care dipstick tests and microscopy analysis. In the administrative claims data, we do not observe the outcomes of point-of-care diagnostics but we observe whether dipstick and microscopy diagnostics were used during a consultation.

We define private diagnostic information as the difference between machine learning prediction errors, $|y_i - m(x_i)|$, and physician prescription errors, $|y_i - d_i|$, which yields

$$\iota_i = |y_i - m(x_i)| - |y_i - d_i| = (d_i - m(x_i))y_i + (m(x_i) - d_i)(1 - y_i). \tag{6}$$

This measure represents physicians' diagnostic information relative to information recovered by machine learning predictions. The left panel of Figure 4 shows the distribution of private diagnostic

information ι_i for bins of 100 patients sorted on predicted risk. In line with our discussion of over- and under-prescribing, private information follows an inverted U-shape with low information in the low and high risk range but high private information in the intermediate risk range.

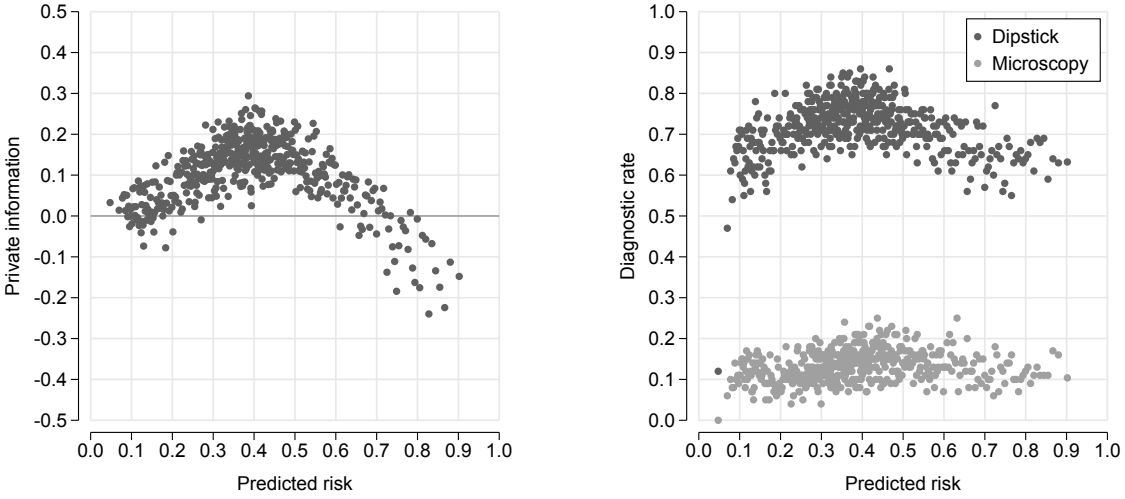


Figure 4: Physician private information relative to machine learning predictions (left) and the dipstick and microscopy diagnostic rate (right) as a function of predicted risk.

The right panel of Figure 4 shows the dipstick and microscopy rate across the risk range. A dipstick diagnostic is used in 72 percent and microscopy in 13 percent of all consultations. The variation in the use of both diagnostics over predicted risk corresponds to the inverted U-shape observed in the private diagnostic information measure, albeit more pronounced for dipsticks. This is indicative that rapid diagnostics are a decision-relevant source of private information.

Table 4 shows regression results of private diagnostic information on dummy variables indicating the use of dipstick and microscopy diagnostics at initial consultations. Dipstick diagnostic claims are significantly positively correlated with private information while microscopy claims show no significant correlation. With a mean ι_i of 0.12, private information is on average one fourth larger in consultations with a dipstick diagnostic.

The average effect masks heterogeneity in the information rapid tests provide. Regression results conditional on the laboratory test outcome in Table 4 reveal that private information is derived from patients with bacterial infections. When no bacterial infection is detected, dipstick and microscopy diagnostics may lead to a loss in private information. Accuracy decreases when a patient is withheld a prescription because of a false negative diagnostic or when a prescription is given due to a false positive diagnostic. Appendix H provides a description of how the use of an imperfect diagnostic

Table 4 Regression analysis of private diagnostic information ι_i

		Laboratory test			Laboratory test	
	All	Negative	Positive	All	Negative	Positive
Dipstick	0.030	-0.075	0.204			
	(0.005)	(0.006)	(0.009)			
Microscopy				-0.002	-0.088	0.112
				(0.006)	(0.008)	(0.011)
Observations	48,406	29,591	18,815	48,406	29,591	18,815

This table reports coefficients for linear regressions of private diagnostic information ι_i on binary indicators of the use of rapid diagnostic tools at initial consultations, conditional on positive and negative test outcomes.

may lead to decision errors. Specifically, overprescribing may increase if the physicians' prescribing rate without the diagnostic to non-bacterial patients is lower than the false positive rate of the diagnostic test. In our data, 19 percent of patients without a bacterial infection receive an antibiotic prescription when no dipstick is used, and 24 percent when no microscopy analysis is carried out. When rapid diagnostics are used for non-bacterial patients, poorer prescription decisions would be expected as the diagnostic false positive rates are typically estimated above these levels in the medical literature (Devillé et al. 2004).

The change in prescribing to negative cases due to rapid diagnostics is only one effect of rapid diagnostics. For patients with bacterial infections in our data, dipstick diagnostics have a clear beneficial effect with 65 percent prescribing to the sick when a dipstick is used relative to 48 percent when a dipstick is not used. This result illustrates how physicians collect context-specific, yet imperfect, information which can complement to information recoverable from administrative data using machine learning.

9 Policy implementation

We have focused on the potential to reduce antibiotic use when there is no uncertainty in choosing the thresholds that delegates decisions between the machine learning algorithm and physicians. In practice, there is uncertainty in selecting the delegation thresholds because they have to be determined ahead of implementation based on historical information. In addition, policy makers may have objectives other than reducing antibiotic use. In this section we provide evidence that both issues can be resolved in practice and discuss further potential implementation issues.

9.1 In-sample vs. out-of-sample policy parameters

For the main results in Table 2, policy parameters k_L and k_H are optimized in-sample. That is, we solve equation (4) after observing machine learning predictions, prescription choices and test outcomes for 2011 and 2012. In a real world application, both policy parameters need to be determined ahead of time. There are many potential ways to go about this task, see for example Hazan (2022). We purposely avoid adding structural assumptions on the evolution of the policy parameters over time because this would imply fundamental knowledge and expectations on the underlying joint evolution of health outcomes, predictions, and physician decisions. Instead, we explore simple ways to determine and update the policy parameters out-of-sample, as outlined in Appendix B, to show that even given parameter uncertainty the policy results may be realizable.

Specifically, we determine k_L and k_H out-of-sample based on historic data relative to the observations to which the policy parameters are applied. We implement this on intervals of the lengths of one year, i.e. using 2011 to determine policy parameters for 2012, as well as one half-year, one quarter, and one month. The longest out-of-sample period where all methods overlap and can be compared is the full year 2012. Figure 5 shows the counterfactual results for the differently updated policy parameters.

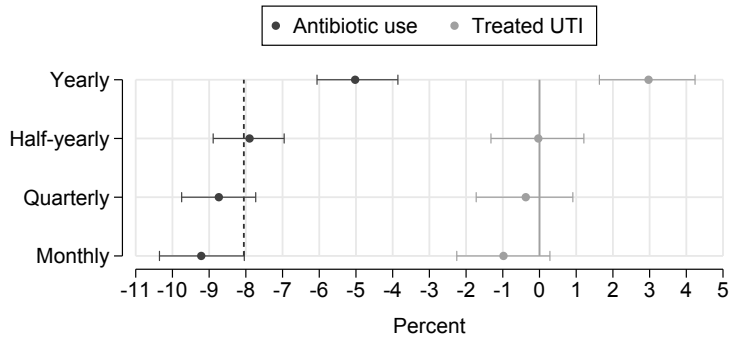


Figure 5: Prediction-based prescription policy outcomes for 2012 updating policy parameters out-of-sample at the yearly, half-yearly, quarterly and monthly level. The dashed lines show the main results in Table 2.

The dashed vertical lines show our main results from Table 2. Yearly policy parameters cannot reproduce the in-sample results. The number of correctly treated bacterial UTIs increases while the reduction in antibiotic use is substantially lower than the in-sample results. Yet, by updating policy parameters at half-yearly, quarterly, and monthly intervals, the out-of-sample results correspond to the main results. Table 12 in Appendix F shows all in-sample and out-of-sample 2012 policy

results for each update method. The different update intervals do not result in significantly varying in-sample policy outcomes.

9.2 Alternative policy objectives

Motivated by common public health policy considerations, we have focused on the policy objective of reducing antibiotic use without treating fewer patients with bacterial UTI (WHO 2012, 2014). Alternative policy objectives can be attained. Figure 6 shows the set of attainable changes in antibiotic use and the number of treated bacterial UTIs for all possible policy parameters $0 \leq k_L \leq k_H \leq 1$. The full range can be seen in Figure 13 in Appendix G. The upper bound of this set represents the payoff-maximizing trade-offs between antibiotic use and treated UTIs.

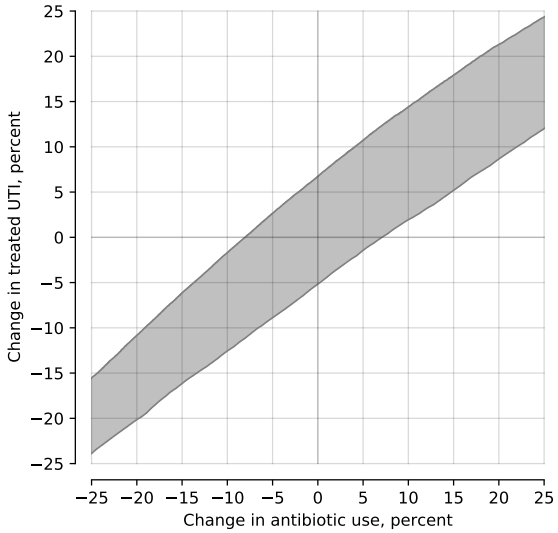


Figure 6: Policy outcomes as a function of policy parameters (k_L, k_H)

In the upper left quadrant, antibiotic use is reduced while the number of treated bacterial UTIs is increased. In this region, any policy maker will prefer the counterfactual policy outcomes relative to the status quo regardless of policy maker preferences $\alpha > 0$ and $\beta > 0$. Our main result lies at the boundary of this region where the upper bound intercepts the horizontal axis. Here, the change in the number of treated bacterial infections is zero and the change in antibiotic use is -8.1 percent. Where the upper bound intersects the vertical axis, the counterfactual policy keeps the number of antibiotic prescriptions at initial consultations constant but increases the number of treated bacterial infections by 7.0 percent. Although the overall use of antibiotics is unchanged, the more efficient use of antibiotics still leads to a reduction in overprescribing by 10.5 percent.

Larger reductions in antibiotic use can be obtained, but not without decreasing the number of treated bacterial infections. For instance, reducing antibiotic use by 20.0 percent would require 10.8 percent of patients with bacterial infections, who were given antibiotics, to delay treatment until test results are available. Analogously, a 20 percent increase in treated UTI could be attained but only with a 17.9 percent increase in antibiotic use. Ultimately, policy maker preferences unobserved to us determine the optimal trade-off and implementation of machine learning based policy.

The line on the lower bound of the set in Figure 6 represents the changes in antibiotic use and treated UTIs for policies that do not delegate any decisions to physicians. Non-cooperative policies are inferior throughout the risk range which generalizes the findings in Section 8.1.

9.3 Efficiency and equity

The policies we consider are redistributive, following much of the literature (Kleinberg et al. 2015, 2018a, Hastings et al. 2020, Mullainathan and Obermeyer 2022). Fairness concerns become salient, perhaps more so than for human biases, when machine learning predictions form the basis of decision outcomes (Kleinberg et al. 2018b, Cowgill and Tucker 2019, Rambachan et al. 2020, Coston et al. 2021). A growing literature has pointed out that excluding sensitive predictors in pursuit of fairness and equity can be detrimental for aggregate outcomes as well as for disadvantaged groups (Kleinberg et al. 2018b, Cowgill and Tucker 2019, Manski et al. 2022). Hence, to cast light on potential fairness concerns and the cost for alleviating them, we assess and adapt our policy function on subgroups of patients but keep risk predictions unchanged.

We take a pragmatic approach and assess groups divided by age, gender, income, and immigration status. Equity concerns are salient for these patient characteristics but they are also important predictors of UTI reported in Table 6 in Appendix A.4. We first quantify redistribution between subgroups based on the main policy parameters reported in Section 7. Panel A in Tables 13 to 16 in Appendix I shows that antibiotic use decreases more strongly for young, male, immigrant, and high income patients. All subgroups, except for income-based groups, deviate significantly from the aggregate outcome of the main policy. The main policy increases the number of treated UTI for women while lowering the number of treated UTIs for men, and fails to lower overall antibiotic use for women. Similarly, it reduces the number of treated UTI for patients with immigrant status. Hence, the main policy achieves reductions at the cost of discrepancies between patient subgroups and violates the constraint on the number of treated UTI.

To maintain the number of treated UTI for each group, we solve equation (4) separately by

subgroup and evaluate policy outcomes. Panel B of Tables 13 to 16 in Appendix I reports group-specific and aggregate outcomes for policy parameters k_L^j and k_H^j optimized for patient group j. With this policy, reductions in antibiotic use are similar across groups with the exception that males have a larger reduction than women. Throughout, fewer physician decisions are overruled in aggregate and more decisions are delegated to physicians compared to the main policy. The group-specific policies reduce aggregate antibiotic use by 6.1 to 8.1 percent compared to the 8.1 percent reduction attained by the main policy, illustrating the trade-off between efficiency and group equity.

9.4 Discussion

While we investigate the complementarity between human and machine learning-based decisions, the evaluated decision rules are simple enough for a potential implementation.[19] As only a share of decisions is delegated to physicians, the policy we evaluate does not confer full agency to physicians as opposed to, for example, in De-Arteaga et al. (2020), Donahue et al. (2022), or Ribers and Ullrich (2022). This design could be implemented in telemedicine services or pharmacies. Similarly, in 2019, the UK National Health Service trialed a smartphone app where an antibiotic could be obtained without seeing a physician, based on symptom reports and a dipstick result.[20] In this study, patients received nitrofurantoin, the first-line antibiotic for UTI in the UK. In cases where this antibiotic was considered clinically unsuitable, the second-line option trimethoprim was given.[21]

The policy we consider could be implemented analogously by giving pivmecillinam, the recommended first-line antibiotic in Denmark.[22] Sulfamethizole or nitrofurantoin could be given where a penicillin is clinically unsuitable. Upon reception of the full resistance profile results, the treatment could be adjusted accordingly. Such a policy is consistent with the observation in our data that pivmecillinam and sulfamethizole account for over 80 percent of initial, UTI-indicated antibiotics.

The prediction of antibiotic resistance can play an additional role in policy implementation to improve the efficacy of antibiotic treatment. Yelin et al. (2019) and Kanjilal et al. (2020) provide retrospective analyses of such predictions in hospital settings. By conditioning their analysis on observing positive bacterial test results and antibiotic use, these studies can not assess the extensive

[19] An implementation would be feasible in Danish primary care because IT systems are interconnected nation-wide.

[20] See Thornley et al. (2020) and https://www.bbc.com/news/uk-england-derbyshire-49031625, accessed 12/7/2022.

[21] In this study, administered prescriptions could not be evaluated because the true sickness condition was not assessed. Hence, only the change in prescriptions was documented, lacking an evaluation of patients' health outcomes.

[22] See UTI guidelines by the Danish Medical Council at https://medicinraadet.dk/anbefalinger-og-vejledninger/behandlingsvejledninger/urinvejsinfektioner-uvi, accessed 12/5/2022.

margin of initial antibiotic prescribing, an important dimension for reducing overall antibiotic use.

After implementation, physicians may adjust their behavior. For instance, they may avoid using laboratory testing for certain patients to maintain decision authority. On the other hand, they may test more often to save the effort and costs of obtaining information. Physicians may also put in more effort in cases where the policy delegates, and less effort where the algorithm makes decisions. In addition, they may attempt to manipulate policy parameters by improving point-of-care diagnostic information, which can expand their decision authority over time. Alternatively, they may decrease their diagnostic efforts to save time by letting the algorithm make more decisions. Although recent studies have focused on how humans use and may manipulate machine learning recommendations (Björkegren et al. 2020, De-Arteaga et al. 2020, Stevenson and Doleac 2022), research on equilibrium behavior in high-stakes decisions is still limited. We leave this area for future research.

10 Conclusion

The quality of prediction algorithms and available data are improving at a rapid pace. In this paper, we document the complementary role of machine learning methods for decision making in a typical context of primary health care provision. We show that decision rules based on machine learning predictions using administrative data may provide a path to improve antibiotic prescribing. Antibiotic prescribing has important societal implications due to increasing antibiotic resistance driven by inefficient antibiotic use. While counterfactual policies based on machine learning predictions alone do not deliver improvements, antibiotic use can be reduced by delegating decisions between physicians and machine learning where each are most certain.

We consider the specific case of UTI in primary care in Denmark, a country with a record of low antibiotic use (Goossens et al. 2005). While our analysis may be challenging to implement in other countries due to the lack of linked data, we suspect the potential reductions we find present a lower bound of what may be achievable in other institutional settings. One limitation is that we consider only initial consultations in which a laboratory test was used. This restriction enables us to observe the ground truth irrespective of physicians' initial treatment decisions, allowing us to evaluate physicians' decisions. We provide some evidence that our results may not be limited to this specific sample but further research is needed on new data from varying contexts.

While we focus on human-AI complementarity for efficient decision outcomes, the considered policy may also help increase productivity. Because share of decisions does not require human input,

physicians and patients may save time and effort. These valuable resources may instead be used on more productive physician-patient interactions and other diagnostic tools at the point-of-care.

One promising avenue for further research is the use of human-acquired information such as recorded symptoms and point-of-care diagnostics in machine learning to capture further nuances of potential complementarities. Another important area in which further research is needed is the analysis of experts' behavioral reactions to prediction-based policies. Physicians' incentives to exert effort in gathering information are likely to change. Potential equilibrium effects of decision rules call for careful evaluation of interventions in the field.

References

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

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

Agrawal A, Gans JS, Goldfarb A (2019) Exploring the impact of artificial Intelligence: Prediction versus judgment. *Information Economics and Policy* 47:1–6.

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

Arnold SR, Straus SE (2005) Interventions to improve antibiotic prescribing practices in ambulatory care. *Cochrane Database of Systematic Reviews* 4.

Athey S (2018) The impact of machine learning on economics. *The Economics of Artificial Intelligence: An Agenda* (Joshua Gans, and Avi Goldfarb, University of Chicago Press: Ajay K. Agrawal).

Autor DH (2015) Why Are There Still So Many Jobs? The History and Future of Workplace Automation. *Journal of Economic Perspectives* 29(3):3–30.

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

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

Björkegren D, Blumenstock JE, Knight S (2020) Manipulation-Proof Machine Learning. arXiv 2004.03865.

Blattberg RC, Hoch SJ (1990) Database Models and Managerial Intuition: 50% Model + 50% Manager. *Management Science* 36(8):887–899.

Butler CC, Simpson SA, Dunstan F, Rollnick S, Cohen D, Gillespie D, Evans MR, Health SLiEaP, Alam MF, Bekkers MJ, Evans J, Moore L, Howe R, Hayes J, Hare M, Hood K (2012) Effectiveness of

multifaceted educational programme to reduce antibiotic dispensing in primary care: Practice based randomised controlled trial. *BMJ* 344:d8173.

Camerer CF (2019) 24. Artificial Intelligence and Behavioral Economics. *24. Artificial Intelligence and Behavioral Economics*, 587–610 (University of Chicago Press).

Cao S, Jiang W, Wang JL, Yang B (2021) From Man vs. Machine to Man + Machine: The Art and AI of Stock Analyses. NBER Working Paper No. w28800.

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

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

Chandler D, Levitt SD, List JA (2011) Predicting and preventing shootings among at-risk youth. *American Economic Review* 101(3):288–292.

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

Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, Chen K, Mitchell R, Cano I, Zhou T, Li M, Xie J, Lin M, Geng Y, Li Y, Yuan J, XGBoost contributors (2022) Package 'xgboost'.

Choudhury P, Starr E, Agarwal R (2020) Machine learning and human capital complementarities: Experimental evidence on bias mitigation. *Strategic Management Journal* 41(8):1381–1411.

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

Coston A, Rambachan A, Chouldechova A (2021) Characterizing Fairness Over the Set of Good Models Under Selective Labels. *Proceedings of the 38th International Conference on Machine Learning*, 2144–2155.

Cowgill B, Tucker CE (2019) Economics, Fairness and Algorithmic Bias. *In preparation for The Journal of Economic Perspectives* .

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

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

Danish Health and Medicines Authority (2013) *Guidelines on Prescribing Antibiotics: For Physicians and Others in Denmark*.

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

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

De-Arteaga M, Fogliato R, Chouldechova A (2020) A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12.

Devillé WL, Yzermans JC, van Duijn NP, Bezemer PD, van der Windt DA, Bouter LM (2004) The urine dipstick test useful to rule out infections. A meta-analysis of the accuracy. *BMC Urology* 4(1):4.

Donahue K, Chouldechova A, Kenthapadi K (2022) Human-Algorithm Collaboration: Achieving Complementarity and Avoiding Unfairness. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1639–1656.

Dubé JP, Misra S (2023) Personalized Pricing and Consumer Welfare. *Journal of Political Economy* 131(1).

Ferry SA, Holm SE, Stenlund H, Lundholm R, Monsen TJ (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.

Flores-Mireles AL, Walker JN, Caparon M, Hultgren SJ (2015) Urinary tract infections: Epidemiology, mechanisms of infection and treatment options. *Nature Reviews Microbiology* 13:269–284.

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

Goossens H, Ferech M, Vander Stichele R, Elseviers M (2005) Outpatient antibiotic use in Europe and association with resistance: A cross-national database study. *The Lancet* 365(9459):579–587.

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

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

Hallsworth M, Chadborn T, Sallis A, Sanders M, Berry D, Greaves F, Clements L, Davies SC (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.

Hastie T, Tibshirani R, Friedman J (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (New York: Springer), second edition.

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

Hazan E (2022) *Introduction to Online Convex Optimization* (MIT Press).

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

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

Kahneman D, Sibony O, Sunstein CR (2021) *Noise: A Flaw in Human Judgment* (New York: Little, Brown Spark).

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

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

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

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

Kleinberg J, Ludwig J, Mullainathan S, Rambachan A (2018b) Algorithmic Fairness. *AEA Papers and Proceedings* 108:22–27.

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

Lakkaraju H, Kleinberg J, Leskovec J, Ludwig J, Mullainathan S (2017) The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 275–284.

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

Laxminarayan R, Duse A, Wattal C, Zaidi AKM, Wertheim HFL, Sumpradit N, Vlieghe E, Hara GL, Gould IM, Goossens H, Greko C, So AD, Bigdeli M, Tomson G, Woodhouse W, Ombaka E, Peralta AQ, Qamar FN, Mir F, Kariuki S, Bhutta ZA, Coates A, Bergstrom R, Wright GD, Brown ED, Cars O (2013) Antibiotic resistance – the need for global solutions. *The Lancet Infectious Diseases Commission* 13(12):1057–1098.

Manski CF, Mullahy J, Venkataramani A (2022) Using Measures of Race to Make Clinical Predictions: Decision Making, Patient Health, and Fairness. NBER Working Paper No. 30700.

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

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

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

Rambachan A, Kleinberg J, Ludwig J, Mullainathan S (2020) An Economic Perspective on Algorithmic Fairness. *AEA Papers and Proceedings* 110:91–95.

Ribers MA, Ullrich H (2022) Machine predictions and human decisions with variation in payoff and skills: The case of antibiotic prescribing. Working paper.

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

Stevenson MT, Doleac JL (2022) Algorithmic Risk Assessment in the Hands of Humans. Working paper.

Thaler RH, Sunstein CR (2009) *Nudge: Improving Decisions About Health, Wealth, and Happiness* (New York: Penguin Books).

Thornley T, Kirkdale CL, Beech E, Howard P, Wilson P (2020) Evaluation of a community pharmacy-led test-and-treat service for women with uncomplicated lower urinary tract infection in England. *JAC-Antimicrobial Resistance* 2(1):dlaa010.

WHO (2012) The evolving threat of antimicrobial resistance: Options for action. Technical report, World Health Organization.

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

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

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

Yip WCM, Hsiao W, Meng Q, Chen W, Sun X (2010) Realignment of incentives for health-care providers in China. *The Lancet* 375(9720):1120–1130.

Appendices

Appendix A Machine learning

A.1 Overview of machine learning data partitions

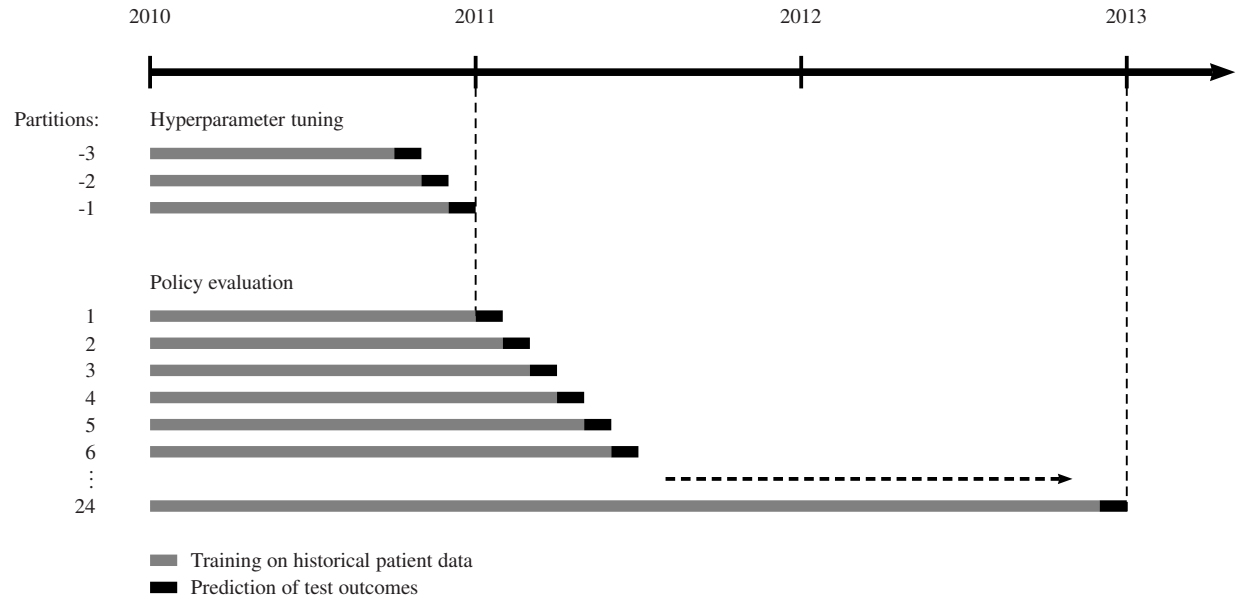


Figure 7: Outline of the data partitions used for hyperparameter tuning as well as the month-by-month progressing training and out-of-sample prediction partitions

A.2 Hyperparameters

Table 5 Top 5 hyperparameter search results

Rank	Rounds	Learning rate	Tree depth	Avg. AUC
1	446	0.04	3	0.69997
2	353	0.05	3	0.69956
3	604	0.02	4	0.69949
4	434	0.04	4	0.69932
5	739	0.03	3	0.69913

We restrict the hyperparameter search space to the learning rate, the number of boosting rounds and the tree depth. The AUC is averaged over the three hyperparameter partitions.

A.3 Predictor importance

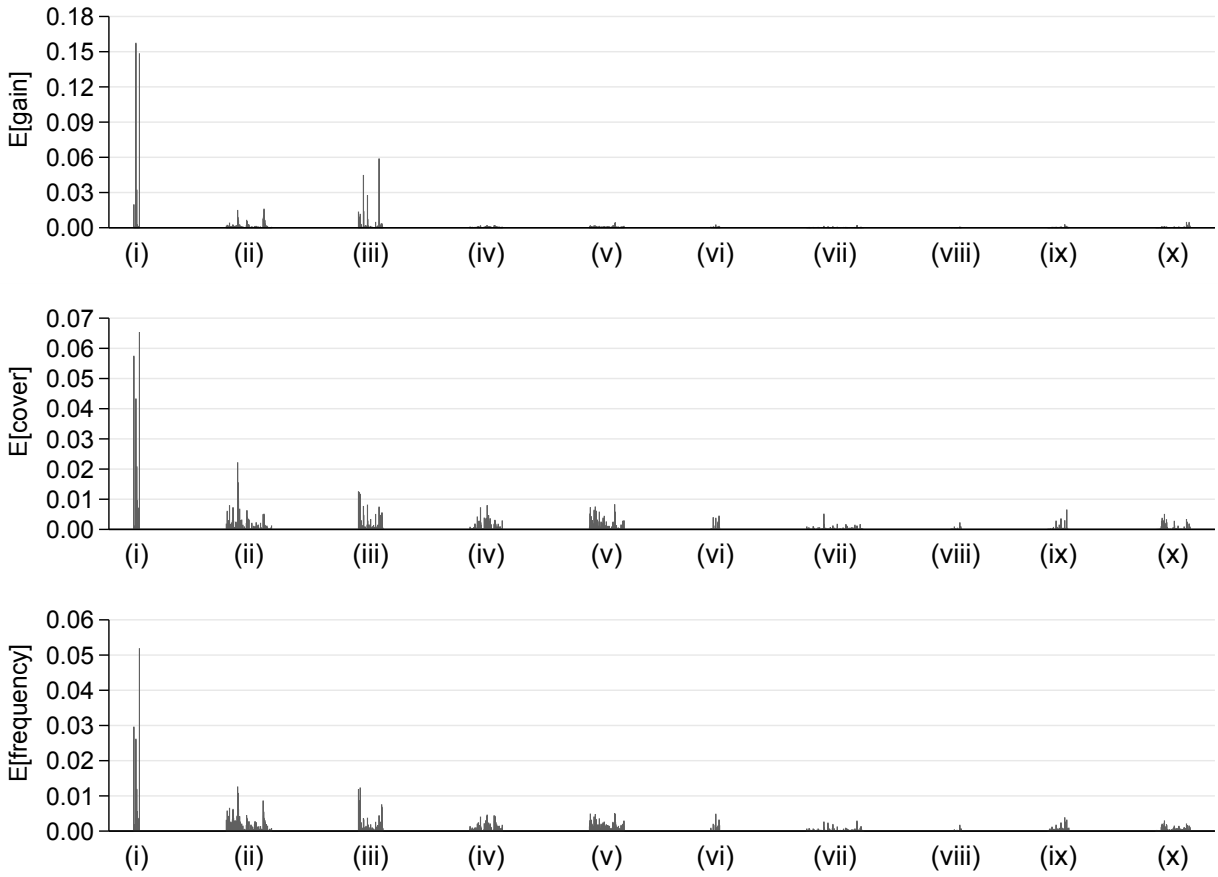


Figure 8: Average gain, cover and frequency over the 24 monthly XGBoost models.

Gain, cover, and frequency provide measures of predictor importance (Chen et al. 2022). Variables in Figure 8 are listed by groups based on their administrative data sources:

(i) patient demographics, test timing and assigned physician identifier
(ii) patient prescriptions and assigned physician's average antibiotic use
(iii) patient laboratory tests and assigned physician's average test results
(iv) patient hospitalizations
(v) patient primary care claims
(vi) Household characteristics
(vii) Household member prescriptions
(viii) Household member laboratory tests
(ix) Household member hospitalizations
(x) Household member primary care claims

A.4 Top 30 predictors by gain, cover, and frequency and gain

Table 6 Top 30 predictors by gain, cover, and frequency

	Sorted by gain			Sorted by cover			Sorted by frequency		
	Predictor	Group	E[gain]	Predictor	Group	E[cover]	Predictor	Group	E[frequency]
1	Gender	(i)	0.1572	Age	(i)	0.0652	Age	(i)	0.0518
2	Age	(i)	0.1482	Clinic identifier	(i)	0.0574	Clinic identifier	(i)	0.0296
3	Resistance to J01XE01 (1)	(iii)	0.0588	Gender	(i)	0.0433	Gender	(i)	0.0261
4	Resistance to J01CA11 (1)	(iii)	0.0446	Prescription ATC code (1)	(ii)	0.0221	Prescription ATC code (1)	(ii)	0.0125
5	Immigration status	(i)	0.0321	Immigration status	(i)	0.0208	GP 6 months mean resistance	(iii)	0.0123
6	Resistance to J01DD13 (1)	(iii)	0.0274	Prescription ATC code (3)	(ii)	0.0155	Immigration status	(i)	0.0118
7	Clinic identifier	(i)	0.0197	Prescription ATC code (2)	(ii)	0.0146	GP all previous mean resistance	(iii)	0.0118
8	Days since prescription (4)	(ii)	0.0159	GP all previous mean resistance	(iii)	0.0125	Prescription ATC code (3)	(ii)	0.0107
9	Prescription ATC code (1)	(ii)	0.0147	GP 1 year mean resistance	(iii)	0.0122	GP 1 year mean resistance	(iii)	0.0086
10	Resistance to J01CA11 (2)	(iii)	0.0138	GP 6 months mean resistance	(iii)	0.0116	Days since prescription (1)	(ii)	0.0086
11	GP all previous mean resistance	(iii)	0.0133	Origin country	(i)	0.0096	Prescription ATC code (2)	(ii)	0.0083
12	GP 6 months mean resistance	(iii)	0.0115	Prescription ATC code (4)	(ii)	0.0095	Days since lab test (1)	(iii)	0.0075
13	Days since prescription (3)	(ii)	0.0104	Education	(i)	0.0092	Days since lab test (1)	(iii)	0.0068
14	GP 1 year mean resistance	(iii)	0.0091	Weeks since specialist (28)	(iv)	0.0082	Municipal DID of J01CF01	(ii)	0.0065
15	Days since prescription (2)	(ii)	0.0090	Resistance to J01DD13 (1)	(iii)	0.0081	Municipal DID of J01FA01	(ii)	0.0062
16	Prescription ATC code (2)	(ii)	0.0088	Hospital bed days (7)	(iv)	0.0079	Prescription ATC code (4)	(ii)	0.0061
17	Days since prescription (1)	(ii)	0.0076	Municipal DID of J01CF01	(ii)	0.0079	Municipal DID of J01EB02	(ii)	0.0058
18	Resistance to J01DD13 (2)	(iii)	0.0071	Resistance to J01CA11 (1)	(iii)	0.0076	Municipal DID of J01AA07	(ii)	0.0057
19	Prescription ATC code (3)	(ii)	0.0065	Claim of non-GP specialist (21)	(v)	0.0075	Education	(i)	0.0056
20	Prescription indication (2)	(ii)	0.0063	Resistance to J01XE01 (1)	(iii)	0.0074	Days since prescription (3)	(ii)	0.0054
21	Days since prescription (7)	(ii)	0.0063	Municipal DID of J01FA01	(ii)	0.0072	Origin country	(i)	0.0054
22	Prescription ATC code (4)	(ii)	0.0057	Claim of non-GP specialist (4)	(v)	0.0072	Weeks since specialist (28)	(v)	0.0050
23	Resistance to J01XE01 (2)	(iii)	0.0055	Hospital diagnose (9)	(iv)	0.0072	Claim of non-GP specialist (4)	(v)	0.0048
24	Prescription indication (3)	(ii)	0.0055	Employment industry	(i)	0.0071	Mother's age	(vi)	0.0048
25	Prescription indication (4)	(ii)	0.0052	Prescription ATC code (8)	(ii)	0.0067	Weeks since specialist (30)	(v)	0.0048
26	Resistance to J01MA02 (1)	(iii)	0.0046	Prescription ATC code (7)	(ii)	0.0066	Claim of non-GP specialist (21)	(v)	0.0047
27	Weeks since GP visit, family (8)	(x)	0.0045	Days since hospital, family (8)	(ix)	0.0065	Hospital bed days (7)	(iv)	0.0046
28	Weeks since GP visit, family (17)	(x)	0.0045	Claim of non-GP specialist (17)	(v)	0.0064	Prescription indication (2)	(ii)	0.0044
29	Weeks since specialist (30)	(v)	0.0044	Prescription indication (3)	(ii)	0.0063	Resistance to J01XE01 (1)	(iii)	0.0044
30	Municipal DID of J01CF01	(ii)	0.0039	Claim of non-GP specialist (24)	(v)	0.0061	Days since hospital (1)	(iv)	0.0044

All variables are measured relative to the laboratory test date and refer to the patient unless otherwise specified by family relation, region or clinic. Numbers in brackets indicate the recency of the observation. For instance, "prescription ATC code (3)" contains the ATC code (The Anatomical Therapeutic Chemical) of the patient's 3rd most recent prescription relative to the test date. DID stands for defined daily dose per 1000 inhabitants per day and codes of the form J01**** are the ATC code of a specific antibiotic.

A.5 Receiver operating characteristic curve

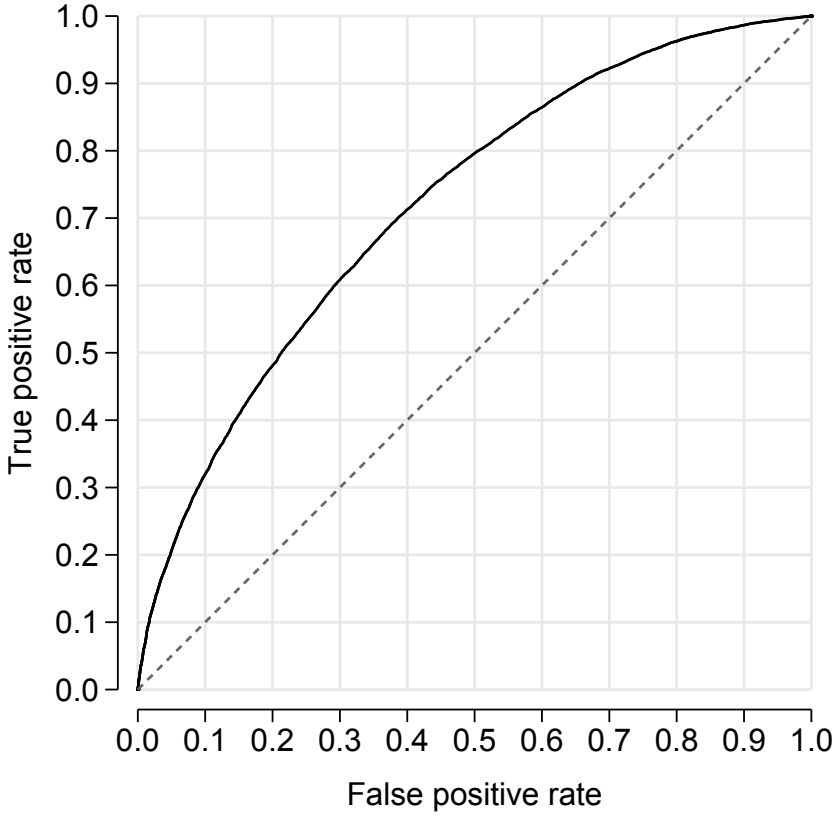


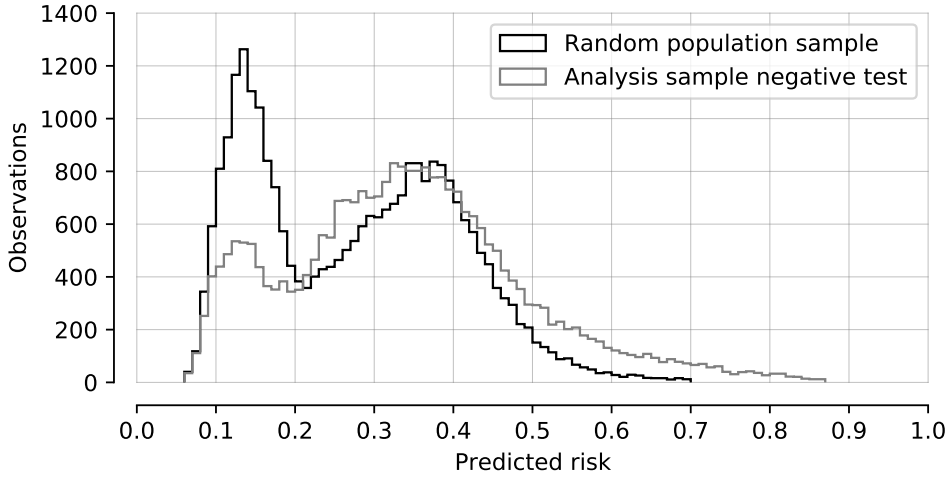
Figure 9: Receiver operating characteristic (ROC) curve for XGBoost. The ROC plots all trade-offs between true positive and false positive rates which are achievable by a prediction technology for a binary outcome. A technology with perfect predictions achieves a true positive rate of one and a false positive rate of zero. The dashed diagonal represents the ROC curve of a prediction technology which is as good as random draws, i.e. providing no information.

A.6 Data partitions

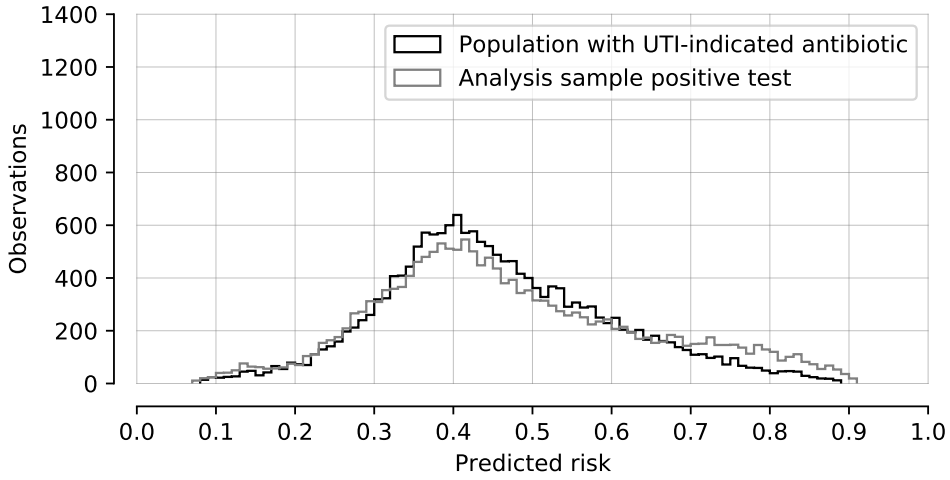
Table 7 Summary statistics for data partitions

	Training					Prediction						
Partition	N	E[y]	E[d]	E[d|y=1]	E[d|y=0]	N	E[m(x)]	E[y]	E[d]	E[d|y=1]	E[d|y=0]	AUC
-3	12,867	0.37	0.39	0.61	0.27	1,618		0.37	0.39	0.58	0.27	
-2	14,485	0.37	0.39	0.60	0.27	1,705		0.37	0.38	0.60	0.26	
-1	16,190	0.37	0.39	0.60	0.27	1,323		0.36	0.42	0.62	0.30	
1	17,513	0.37	0.39	0.60	0.27	1,755	0.36	0.36	0.37	0.58	0.25	0.71
2	19,268	0.37	0.39	0.60	0.27	1,510	0.37	0.37	0.38	0.59	0.26	0.73
3	20,778	0.37	0.39	0.60	0.27	1,811	0.37	0.38	0.37	0.57	0.25	0.71
4	22,589	0.37	0.39	0.60	0.26	1,413	0.37	0.40	0.40	0.60	0.27	0.70
5	24,002	0.37	0.39	0.60	0.26	1,864	0.38	0.40	0.37	0.55	0.24	0.71
6	25,866	0.37	0.39	0.60	0.26	1,753	0.40	0.41	0.38	0.58	0.24	0.73
7	27,619	0.37	0.39	0.59	0.26	1,257	0.41	0.41	0.45	0.68	0.29	0.69
8	28,876	0.37	0.39	0.60	0.26	1,936	0.40	0.40	0.38	0.61	0.23	0.70
9	30,812	0.38	0.39	0.60	0.26	2,092	0.39	0.39	0.40	0.62	0.26	0.72
10	32,904	0.38	0.39	0.60	0.26	2,027	0.39	0.39	0.40	0.61	0.26	0.70
11	34,931	0.38	0.39	0.60	0.26	2,166	0.39	0.39	0.37	0.58	0.24	0.71
12	37,097	0.38	0.39	0.60	0.26	1,653	0.39	0.41	0.40	0.61	0.25	0.72
13	38,750	0.38	0.39	0.60	0.26	2,244	0.40	0.39	0.39	0.61	0.24	0.74
14	40,994	0.38	0.39	0.60	0.26	1,914	0.40	0.38	0.37	0.62	0.23	0.72
15	42,908	0.38	0.39	0.60	0.26	2,202	0.39	0.36	0.36	0.59	0.24	0.71
16	45,110	0.38	0.39	0.60	0.26	1,683	0.40	0.40	0.41	0.63	0.25	0.73
17	46,793	0.38	0.39	0.60	0.26	2,064	0.40	0.37	0.38	0.60	0.25	0.74
18	48,857	0.38	0.39	0.60	0.26	2,410	0.39	0.38	0.38	0.59	0.26	0.73
19	51,267	0.38	0.39	0.60	0.26	1,645	0.41	0.43	0.44	0.65	0.28	0.72
20	52,912	0.38	0.39	0.60	0.26	2,759	0.40	0.40	0.41	0.62	0.27	0.72
21	55,671	0.38	0.39	0.61	0.26	2,506	0.38	0.40	0.39	0.60	0.25	0.73
22	58,177	0.38	0.39	0.61	0.26	2,770	0.39	0.39	0.40	0.62	0.27	0.72
23	60,947	0.38	0.39	0.61	0.26	3,018	0.39	0.37	0.37	0.60	0.24	0.74
24	63,965	0.38	0.39	0.61	0.26	1,954	0.39	0.39	0.41	0.62	0.27	0.73

A.7 Risk predictions beyond the analysis sample



(a) In-sample non-UTI and out-of-sample random population



(b) In-sample UTI and out-of-sample initial UTI-indicated prescriptions

Figure 10: In-sample and out-of-sample predicted risk distributions. Bars with fewer than 10 patients have been removed due to anonymity restrictions. The samples without laboratory tests are drawn such that they have the same number of observations as the corresponding analysis sample for $y = 0$ and $y = 1$.

Appendix B Policy algorithm

We use the following algorithm to compute in-sample and out-of-sample counterfactual policy evaluations for 2012 with policy parameters updated yearly, half-yearly, quarterly and monthly. We define the start date $\underline{t} = 01/01/2011$ and end date $\bar{t} = 31/12/2012$.

Policy algorithm

1. Train a prediction model and predict test outcomes following Appendix A.

2. For fixed update period Δt, define partitions of patients in the policy period

$$I_j^{\Delta t} = \{i \mid t_i \in [t_j, t_j + \Delta t)\} \quad \text{for} \quad j \in \{1, \ldots, \frac{\bar{t} - \underline{t}}{\Delta t}\}$$

   where t_i is patient i's test date and $t_j = \underline{t} + (j-1) \times \Delta t$.

3. Compute $k_L^{\Delta t}(j)$ and $k_H^{\Delta t}(j)$ on $I_j^{\Delta t}$ using equation (4) for each j.

4. Evaluate in-sample policy outcomes for patients in $I_j^{\Delta t}$ by

$$\sum_{i \in I_j^{\Delta t}} y_i(\delta_i(k_L^{\Delta t}(j), k_H^{\Delta t}(j)) - d_i)$$

   and

$$\sum_{i \in I_j^{\Delta t}} \delta_i(k_L^{\Delta t}(j), k_H^{\Delta t}(j)) - d_i.$$

5. Evaluate out-of-sample policy outcomes for patients in $I_j^{\Delta t}$ for $j \geq 2$ by

$$\sum_{i \in I_j^{\Delta t}} y_i(\delta_i(k_L^{\Delta t}(j-1), k_H^{\Delta t}(j-1)) - d_i)$$

   and

$$\sum_{i \in I_j^{\Delta t}} \delta_i(k_L^{\Delta t}(j-1), k_H^{\Delta t}(j-1)) - d_i.$$

6. Aggregate results across all partitions used for policy evaluation.

Appendix C Policy outcomes using LASSO for prediction

Table 8 Counterfactual policy outcomes

k_L			0.300
k_H			0.633
Change in treated UTI, in %	0.0	[−1.3,	1.1]
Change in antibiotic use, in %	−7.0	[−7.9,	−6.3]
Change in overprescribing, in %	−17.6	[−18.8,	−16.2]
Physician decisions overruled, in %	11.4	[11.1,	11.8]
Patients delegated to physicians, in %	62.3	[61.9,	62.8]
Consultations			48,406
UTIs			18,815
Treated UTIs			11,402
Antibiotic prescriptions			18,872
Overprescribing			7,470

95% confidence intervals are based on 100 bootstrap samples of 2011 and 2012 where Lasso predictions and the policy parameter (k_L, k_H) remain fixed.

Table 9 Counterfactual outcomes for 2011 and 2012 using Lasso, no physician input

k			0.389
Change in treated UTI, in %	0.0	[−1.8,	1.8]
Change in antibiotic use, in %	11.3	[9.7,	13.0]
Change in overprescribing, in %	28.7	[25.5,	32.0]
Physician decisions overruled, in %	41.2	[40.8,	41.7]

95% confidence intervals are based on 100 bootstrap samples of 2011 and 2012 where machine learning predictions and the policy parameter k remain fixed.

Appendix D Policy outcomes and sample selectivity

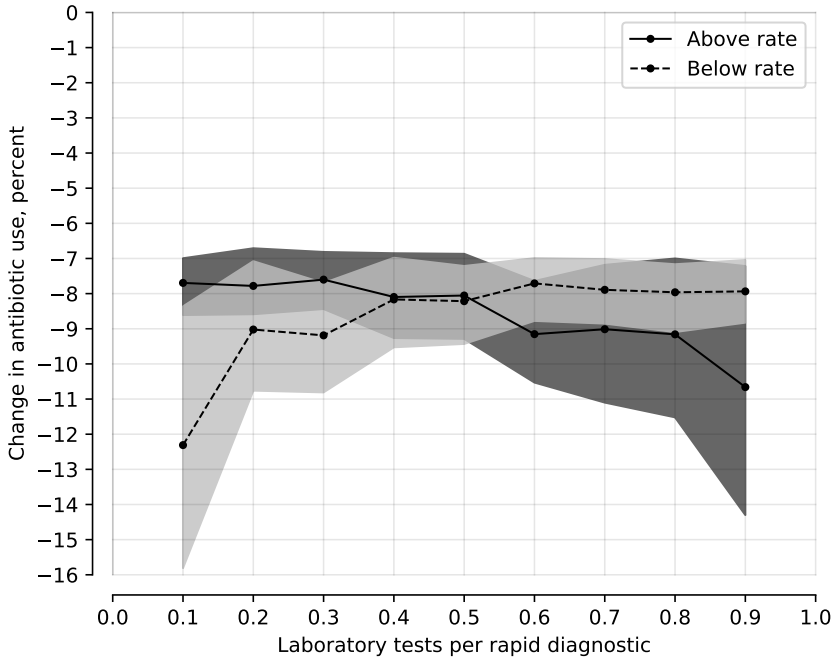


Figure 11: Policy outcomes by laboratory testing intensity

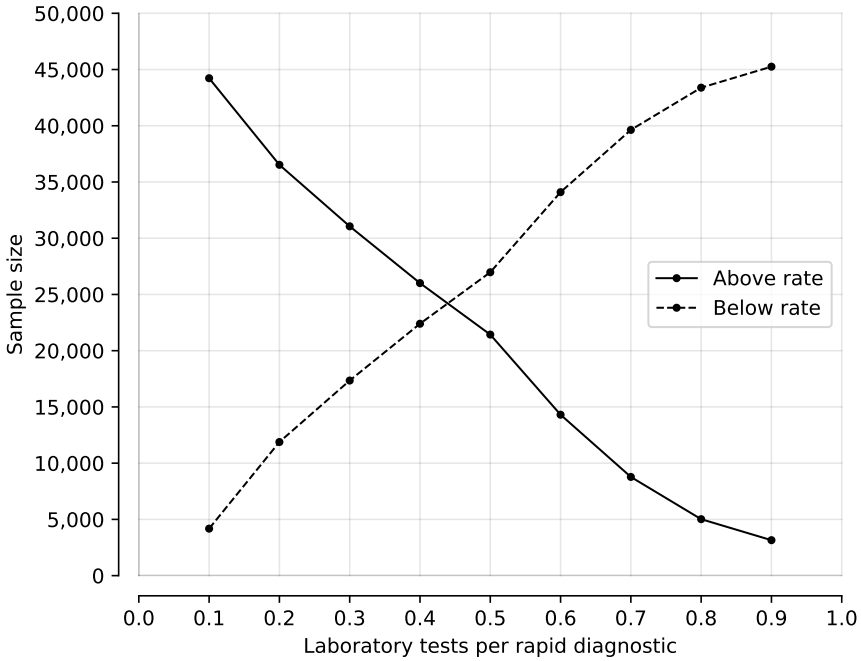


Figure 12: Sample size by laboratory testing intensity

Appendix E Laboratory results for high predicted risk patients

Table 10 Distribution of bacterial species in positive test results for high predicted risk patients conditional on physician antibiotic prescription decisions at the initial consultation

	$d = 0,\ k_H \leq m(x)$		$d = 1,\ k_H \leq m(x)$	
Species/genus	Obs	Pct	Obs	Pct
E. coli	1,237	68.0	1,907	79.9
K. pneumoniae	166	9.1	145	6.1
E. faecalis	98	5.4	59	2.5
Enterococcus	45	2.5	33	1.4
P. mirabilis	33	1.8	31	1.3
Other	241	13.2	213	8.9
Total	1,820	100.0	2,388	100.0

Table 11 Antibiotic resistance for positive E. coli test results among high predicted risk patients conditional on physician antibiotic prescription decisions at the initial consultation

	$d = 0,\ k_H \leq m(x)$		$d = 1,\ k_H \leq m(x)$		
Antibiotic (ATC-code)	Obs	Resistance	Obs	Resistance	Difference
Ampicillin (J01CA01)	1,237	0.437	1,907	0.390	0.047 [0.012 ,0.083]
Mecillinam (J01CA11)	1,237	0.058	1,907	0.036	0.023 [0.007 ,0.038]
Trimethoprim (J01EA01)	1,237	0.310	1,907	0.261	0.050 [0.017 ,0.082]
Sulfamethizole (J01EB02)	1,237	0.373	1,907	0.331	0.042 [0.007 ,0.077]
Ciprofloxacin (J01MA02)	1,237	0.089	1,907	0.056	0.033 [-0.003 ,0.068]
Nitrofurantion (J01XE01)	1,237	0.042	1,907	0.027	0.015 [0.002 ,0.029]

Appendix F In-sample and out-of-sample policy results

Table 12 Policy results for 2012 with policy parameters set in-sample and out-of-sample

	Change in antibiotic use (%)		Change in treated UTI (%)	
k_L, k_H computed	In-sample	Out-of-sample	In-sample	Out-of-sample
Yearly	-8.6 [-9.8, -7.4]	-5.0 [-6.1, -3.9]	0.0 [-1.3, 1.5]	3.0 [1.6, 4.2]
Half-yearly	-8.7 [-9.8, -7.6]	-7.9 [-8.9, -7.0]	0.0 [-1.1, 1.3]	-0.0 [-1.3, 1.2]
Quarterly	-8.8 [-9.9, -7.7]	-8.7 [-9.7, -7.7]	0.0 [-1.0, 1.3]	-0.4 [-1.7, 0.9]
Monthly	-9.2 [-10.1, -8.0]	-9.2 [-10.4, -8.0]	0.0 [-1.1, 1.2]	-1.0 [-2.3, 0.3]

95% confidence intervals are based on 100 bootstrap samples where machine learning predictions and policy parameters (k_L, k_H) remain fixed.

Appendix G Alternative policy objectives

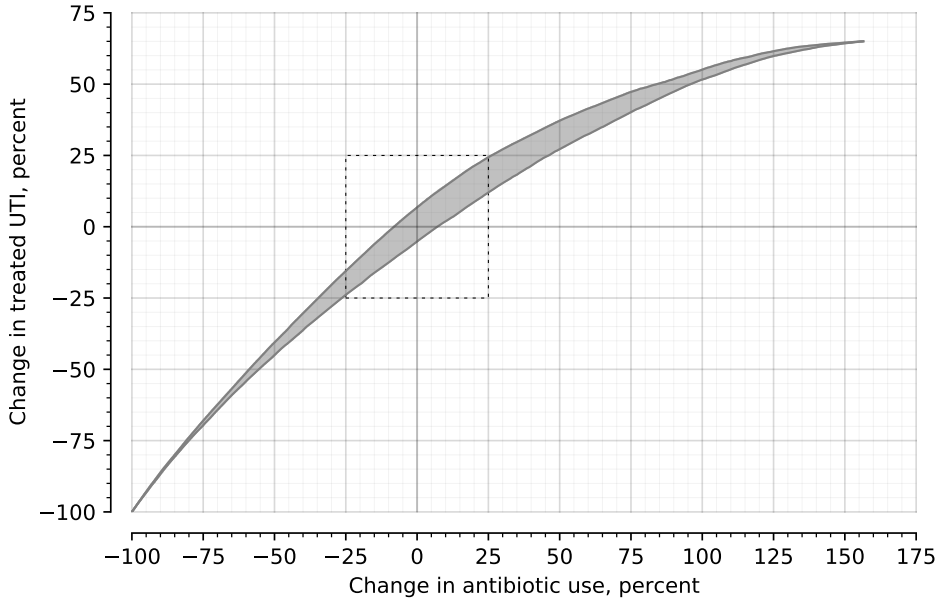


Figure 13: The set of all policy outcomes as a function of the policy parameters (k_L, k_H) for 2011 and 2012. The dashed rectangle shows the policy outcomes highlighted in Figure 6 in the main text.

Appendix H The effect of urine dipstick rapid diagnostics

If a prescription decision is based on an imperfect diagnostic tool such as dipstick rapid tests, more frequent use of these tests can increases antibiotic prescribing to patients without a bacterial infection. To see this, assume a diagnostic test is applied at a rate γ unconditional on the patient sickness state $y \in \{0, 1\}$. The diagnostic induced change in physician prescribing to patients without a bacterial infection is then given by

$$\Pr(d = 1 \mid y = 0, \gamma > 0) - \Pr(d = 1 \mid y = 0, \gamma = 0) \tag{7}$$

$$= [(1 - \gamma)\Pr(d = 1 \mid y = 0, \gamma = 0) + \gamma \Pr(d = 1 \mid y = 0, \gamma = 1)] \tag{8}$$

$$- \Pr(d = 1 \mid y = 0, \gamma = 0) \tag{9}$$

$$= \gamma(\Pr(d = 1 \mid y = 0, \gamma = 1) - \Pr(d = 1 \mid y = 0, \gamma = 0)) \tag{10}$$

Assuming physician compliance with the diagnostic tool, overprescribing will increase if the false positive rate of the diagnostic is higher than the physician false positive rate without the diagnostic. Further, the diagnostic rate γ only affects the size of the effect but does not affect the direction of the change.

An analogous argument can be made for the patients with a positive test result:

$$\Pr(d = 1 \mid y = 1, \gamma > 0) - \Pr(d = 1 \mid y = 1, \gamma = 0) \tag{11}$$

$$= [(1 - \gamma)\Pr(d = 1 \mid y = 1, \gamma = 0) + \gamma \Pr(d = 1 \mid y = 1, \gamma = 1)] \tag{12}$$

$$- \Pr(d = 1 \mid y = 1, \gamma = 0) \tag{13}$$

$$= \gamma(\Pr(d = 1 \mid y = 1, \gamma = 1) - \Pr(d = 1 \mid y = 1, \gamma = 0)) \tag{14}$$

Underprescribing will decrease if the true positive rate of the diagnostic is higher than the physician true positive rate without the diagnostic. Once again, the diagnostic rate γ only affects the size of the effect but does not affect the direction of the change.

Appendix I Results for patient group analysis

Table 13 Counterfactual policy outcomes by age

	age < 48	$48 \leq$ age	Aggregated
Panel A: Main algorithm			
k_L	0.320	0.320	0.320
k_H	0.601	0.601	0.601
Change in treated UTI, in %	-19.9 $[-21.1, -18.7]$	15.2 $[-13.5, -17.3]$	0.0 $[-1.2, 1.0]$
Change in antibiotic use, in %	-28.7 $[-29.5, -27.7]$	10.7 $[9.4, 12.1]$	-8.1 $[-8.9, -7.4]$
Change in overprescribing, in %	-39.3 $[-40.7, -37.7]$	2.3 $[0.0, 4.6]$	-20.3 $[-21.9, -18.9]$
GP decisions overruled, in %	12.8 $[12.4, 13.3]$	17.1 $[16.6, 17.6]$	15.0 $[14.7, 15.3]$
Patients delegated to GPs, in %	51.4 $[50.8, 52.1]$	54.2 $[53.6, 54.9]$	52.8 $[52.3, 53.3]$
Panel B: Sub-group algorithm			
k_L^j	0.211	0.378	-
k_H^j	0.535	0.653	-
Change in treated UTI, in %	0.0 $[-1.0, 1.0]$	0.0 $[-1.2, 1.4]$	0.0 $[-1.0, 0.8]$
Change in antibiotic use, in %	-4.8 $[-5.6, -4.0]$	-7.2 $[-8.5, -6.0]$	-6.1 $[-6.8, -5.4]$
Change in overprescribing, in %	-10.7 $[-12.0, -9.1]$	-20.9 $[-23.6, -18.8]$	-15.3 $[-16.7, -14.1]$
GP decisions overruled, in %	6.2 $[5.9, 6.6]$	18.6 $[18.1, 19.1]$	12.4 $[12.2, 12.7]$
Patients delegated to GPs, in %	74.0 $[73.5, 74.4]$	48.2 $[47.7, 49.0]$	61.0 $[60.6, 61.4]$
Consultations	24,047	24,359	48,406
Bacterial UTIs	7,744	11,071	18,815
Treated UTIs	4,935	6,467	11,402
Antibiotic prescriptions	9,004	9,868	18,872
Overprescribing	4,069	3,401	7,470

95% confidence intervals are based on 100 bootstrap samples of 2011 and 2012 where machine learning predictions and the policy parameters (k_L, k_H) and (k_L^j, k_H^j) remain fixed.

Table 14 Counterfactual policy outcomes by gender

	Female	Male	Aggregated
Panel A: Main algorithm			
k_L	0.320	0.320	0.320
k_H	0.601	0.601	0.601
Change in treated UTI, in %	5.0 [4.0, 6.1]	−33.1 [−36.7, −29.1]	0.0 [−1.2, 1.0]
Change in antibiotic use, in %	0.6 [−0.2, 1.4]	−51.9 [−54.0, −49.7]	−8.1 [−8.9, −7.4]
Change in overprescribing, in %	−6.9 [−8.3, −5.7]	−69.3 [−71.8, −66.9]	−20.3 [−21.9, −18.9]
GP decisions overruled, in %	13.4 [13.1, 13.7]	20.0 [19.4, 20.7]	15.0 [14.7, 15.3]
Patients delegated to GPs, in %	63.5 [63.0, 64.1]	18.1 [17.3, 18.8]	52.8 [52.3, 53.3]
Panel B: Sub-group algorithm			
k_L^j	0.326	0.207	-
k_H^j	0.650	0.520	-
Change in treated UTI, in %	0.0 [−1.0, 1.1]	0.0 [−3.8, 3.7]	0.0 [−1.1, 1.1]
Change in antibiotic use, in %	−5.4 [−6.1, −4.6]	−20.4 [−22.8, −18.4]	−7.9 [−8.7, −7.2]
Change in overprescribing, in %	−14.5 [−15.8, −13.4]	−39.4 [−42.4, −36.2]	−19.9 [−21.3, −18.8]
GP decisions overruled, in %	12.2 [11.9, 12.6]	15.2 [14.6, 15.8]	12.9 [12.7, 13.3]
Patients delegated to GPs, in %	65.5 [65.0, 66.0]	37.1 [36.4, 38.1]	58.8 [58.3, 59.2]
Consultations	36,960	11,446	48,406
Bacterial UTIs	16,101	2,714	18,815
Treated UTIs	9,905	1,497	11,402
Antibiotic prescriptions	15,761	3,111	18,872
Overprescribing	5,856	1,614	7,470

95% confidence intervals are based on 100 bootstrap samples of 2011 and 2012 where machine learning predictions and the policy parameters (k_L, k_H) and (k_L^j, k_H^j) remain fixed.

Table 15 Counterfactual policy outcomes by immigration status

	Immigrant	Non-immigrant	Aggregated
Panel A: Main algorithm			
k_L	0.320	0.320	0.320
k_H	0.601	0.601	0.601
Change in treated UTI, in %	−33.7 [−36.6, −30.7]	4.4 [3.1, 5.7]	0.0 [−1.2, 1.0]
Change in antibiotic use, in %	−46.9 [−48.7, −44.8]	−1.5 [−2.6, −0.7]	−8.1 [−8.9, −7.4]
Change in overprescribing, in %	−59.5 [−62.0, −56.4]	−11.4 [−13.2, −10.0]	−20.3 [−21.9, −18.9]
GP decisions overruled, in %	20.0 [19.2, 20.8]	14.0 [13.7, 14.3]	15.0 [14.7, 15.3]
Patients delegated to GPs, in %	32.8 [31.6, 33.8]	56.7 [56.3, 57.3]	52.8 [52.3, 53.3]
Panel B: Sub-group algorithm			
k_L^j	0.245	0.332	-
k_H^j	0.459	0.627	-
Change in treated UTI, in %	0.0 [−3.2, 2.3]	0.0 [−1.3, 1.3]	0.0 [−1.2, 1.1]
Change in antibiotic use, in %	−9.2 [−11.1, −7.3]	−6.6 [−7.6, −5.7]	−7.0 [−7.9, −6.3]
Change in overprescribing, in %	−18.0 [−21.4, −14.9]	−17.6 [−19.2, −16.2]	−17.7 [−19.1, −16.4]
GP decisions overruled, in %	14.0 [13.3, 14.7]	13.9 [13.6, 14.3]	13.9 [13.7, 14.2]
Patients delegated to GPs, in %	53.8 [52.6, 55.0]	55.8 [55.3, 56.3]	55.5 [55.0, 56.0]
Consultations	7,934	40,472	48,406
Bacterial UTIs	2,269	16,546	18,815
Treated UTIs	1,322	10,080	11,402
Antibiotic prescriptions	2,708	16,164	18,872
Overprescribing	1,386	6,084	7,470

95% confidence intervals are based on 100 bootstrap samples of 2011 and 2012 where machine learning predictions and the policy parameters (k_L, k_H) and (k_L^j, k_H^j) remain fixed.

Table 16 Counterfactual policy outcomes by income

	Income < 175.000	$175.000 \leq$ Income	Aggregated
Panel A: Main algorithm			
k_L	0.320	0.320	0.320
k_H	0.601	0.601	0.601
Change in treated UTI, in %	1.2 $[-0.2, 2.8]$	-1.1 $[-2.6, 0.2]$	0.0 $[-1.2, 1.0]$
Change in antibiotic use, in %	-6.7 $[-7.8, -5.3]$	-9.4 $[-10.8, -8.3]$	-8.1 $[-8.9, -7.4]$
Change in overprescribing, in %	-19.0 $[-21.0, -16.4]$	-21.5 $[-23.6, -19.9]$	-20.3 $[-21.9, -18.9]$
GP decisions overruled, in %	15.6 $[15.1, 15.9]$	14.4 $[14.0, 14.8]$	15.0 $[14.7, 15.3]$
Patients delegated to GPs, in %	49.7 $[49.2, 50.5]$	56.0 $[55.3, 56.6]$	52.8 $[52.3, 53.3]$
Panel B: Sub-group algorithm			
k_L^j	0.326	0.303	-
k_H^j	0.601	0.626	-
Change in treated UTI, in %	0.0 $[-1.4, 1.7]$	0.0 $[-1.3, 1.4]$	0.0 $[-1.2, 1.1]$
Change in antibiotic use, in %	-8.1 $[-9.5, -6.7]$	-8.1 $[-9.5, -6.9]$	-8.1 $[-8.9, -7.4]$
Change in overprescribing, in %	-21.0 $[-22.9, -18.4]$	-20.1 $[-21.8, -18.2]$	-20.5 $[-21.8, -19.1]$
GP decisions overruled, in %	16.2 $[15.7, 16.6]$	12.3 $[11.9, 12.7]$	14.3 $[14.0, 14.6]$
Patients delegated to GPs, in %	48.1 $[47.5, 48.9]$	61.1 $[60.5, 61.8]$	54.5 $[54.1, 55.1]$
Consultations	$24,603$	$23,803$	$48,406$
Bacterial UTIs	$9,728$	$9,087$	$18,815$
Treated UTIs	$5,576$	$5,826$	$11,402$
Antibiotic prescriptions	$9,108$	$9,764$	$18,872$
Overprescribing	$3,532$	$3,938$	$7,470$

95% confidence intervals are based on 100 bootstrap samples of 2011 and 2012 where machine learning predictions and the policy parameters (k_L, k_H) and (k_L^j, k_H^j) remain fixed.