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

Michael Allan Ribers† Hannes Ullrich‡

February 2022

Abstract

Inefficient human decisions are driven by biases and limited information. The health care sector is one leading example where machine learning is hoped to deliver efficiency gains. Antibiotic resistance constitutes a major challenge to health care systems largely due to human antibiotic overuse. In this paper, we investigate how machine learning provides new opportunities to design decision rules that reduce antibiotic use. We focus on urinary tract infections in primary care which constitute one of the leading causes for antibiotic use. Laboratory testing can diagnose bacterial infections but with considerable delay such that physicians often prescribe prior to attaining diagnostic certainty. Using Danish administrative and laboratory data, we find that machine learning methods are capable of predicting the presence of bacteria out-of-sample and complement physician prescribing. We optimize policy rules which delegate a share of prescription decisions to physicians and base the remaining decisions on machine learning predictions. The policy shows a potential to reduce antibiotic prescribing by 8.4 percent and overprescribing by 20.9 percent without assigning fewer prescriptions to patients with bacterial infections. We find that human-algorithm cooperation is essential to achieve efficiency gains.

*We benefited from helpful feedback by Jason Abaluck, Rolf Magnus Arpi, Lars Bjerrum, Gloria Cristina Cordoba Currea, Greg Crawford, Tomaso Duso, Günter Hitsch, Shan Huang, Ulrich Kaiser, Jenny Dahl Knudsen, Sidsel Kyst, Chloé Michel, Jeanine Miklós-Thal, Maria Polyakova, 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, 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

Human decision making is prone to noise. Managers and professionals more generally need to make costly decisions under 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). Hence, hopes are high that machine learning will eventually help improve human decision making by offering a systematic prediction of the ground truth and guiding optimal decisions. Yet, the practical and scientific verdict about its value is still open.

In this paper, we turn to a salient case in health care. Antibiotic resistance is considered one of the greatest threats to global health (WHO 2012, 2014). Worldwide, 4.95 million deaths are estimated to be associated with antibiotic resistance and 1.27 million deaths are directly attributable (Murray 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). Because human consumption of antibiotics is considered the main driver of antibiotic resistance, limiting misuse of antibiotics is a prime policy concern (Goossens et al. 2005, Adda 2020).

Physician decisions to prescribe antibiotics involve a prediction task due to limited diagnostic information about the cause of an infection. To reduce uncertainty, physicians interpret reported symptoms, perform point-of-care tests, and may also consider a patient's background and medical data. However, lack of time and analytic tools may hinder effective use of such data. Machine learning has shown to be an effective method to elicit predictive information from large scale data (Agrawal et al. 2018, Athey 2018). In particular, machine learning can exploit systemic patterns in data collected across many patients and health care providers. Analyzing whether policy rules based on algorithmic risk predictions can improve decisions made by medical professionals is central in identifying the general value of machine learning for experts and the advancement of efficient antibiotic use in particular.

Specifically, we evaluate the potential of a machine learning-based prescribing rule to reduce antibiotic use for the treatment of bacterial urinary tract infections (UTI) in primary care. Primary care is responsible for the bulk of human antibiotic use and UTI is one of the most common types of infections requiring antibiotic treatment. An accurate diagnostic can only be provided by inspection of urine samples in a microbiological laboratory but laboratory testing has the important and general limitation that results arrive with a delay corresponding to nearly a complete course of antibiotic

treatment. Thus, at the initial consultation physicians must decide under uncertainty whether to prescribe an antibiotic or delay until the test result is known. Using machine learning predictions of individual patients' risk of bacterial infection, we evaluate policies which assign initial antibiotic prescriptions according to risk thresholds and delegate decisions with high remaining uncertainty to physicians.

We apply a machine learning algorithm, XGboost, to high-dimensional, administrative data from Denmark to predict the risk of bacterial presence for 53,219 initial consultations. The observed outcome variable, a binary variable indicating when bacteria are isolated in a patient sample, is based on the microbiological test result. Laboratory test results for urine samples provide a reliable measure of a patient's true infection state because they are highly accurate and require little human judgement. The covariates in the prediction model include each individual patient's medical outpatient claims histories, past antibiotic prescriptions, past microbiological test results, a rich set of personal characteristics such as gender, age, detailed employment status and type, education, income, civil status and more, as well as the same information on each individuals' household members. We find that machine learning applied to these data predicts realizations of bacterial UTI in out of sample test results well, with an area under the ROC curve (AUC) of 0.72.

We 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 the algorithmic risk prediction. Analyzing machine learning predictions against laboratory test outcomes conditional on physicians' prescription decisions, we find that physicians hold valuable private diagnostic information, specifically in intermediate ranges of predicted risk. Therefore, 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 the decision to the physician for intermediate predicted risk.

We find that overall prescribing can be reduced by 8.4 percent without reducing the number of treated patients who suffer from a bacterial UTI. At the same time, the policy rules can reduce overprescribing, prescriptions to non-bacterial cases, by 20.9 percent. In 45.7 percent of consultations, the decision was made by the prediction-based rule, overturning 14.4 percent of the observed decisions made by physicians. Out of the patients who receive an antibiotic based on the algorithm's decision, 71-75 percent received a prescription following the test result in our sample. Combined with a spontaneous recovery rate of 24 percent reported in the literature, we conclude that the the algorithm's decisions resemble physician choices under full information. Finally, decision rules

that do not include physicians fail to improve prescription outcomes. We conclude that even with high-dimensional individual-specific data, machine learning alone does not improve on human decisions but that the combination of machine learning and human experts leads to much improved outcomes.

Implementing prediction-based policies is challenging because the distribution of patients to which policy rules are applied is unknown and may vary over time. Even if the quality of risk predictions was unaffected by sample variation, variation in the underlying risk distribution of arriving patients may affect policy outcomes. We focus on shifting antibiotic prescriptions from low to high risk individuals such that the number of treated bacterial UTIs remains constant. As the distribution shifts, so does the share of low and high risk patients targeted by the policy. Therefore, the policy rule is not guaranteed to achieve the objective out-of-sample. While distribution shift is a general challenge for any policy optimized based on an empirical risk distribution, existing counterfactual policy evaluations of machine learning predictions have been performed in-sample (Bayati et al. 2014, Kleinberg et al. 2018, Yelin et al. 2019, Hastings et al. 2020).[1] To gain confidence that our policy objective can be attained in a real world setting, we evaluate out-of-sample policy outcomes by updating policy parameters over time. We show how failure to adjust policy parameters over time can lead to failure of such policies. In our setting, updating policy parameters at a half-yearly or quarterly frequency suffices to achieve in-sample policy outcomes. We extend the cautious note by Rose (2018), who discusses the potential need to adjust prediction algorithms over time or other dimensions, by showing the same is true for the design of policies which delegate decisions between an algorithm and human decision makers.

The paper is organized as follows. Section 2 relates our work to the existing literature. Section 3 presents the institutional background and data. Section 4 shows the results of the prediction algorithm. Section 5 presents the framework for the design and evaluation of prediction-based policy rules to improve antibiotic prescribing. Section 6 presents the potential policy outcomes. Section 7 shows how policies without physician input fail to achieve improvements. Section 8 concludes.

[1] An exception is work using randomized controlled experiments such as Dubé and Misra (2021) who show that a price targeting scheme successfully improves firm profits out-of-sample.

2 Related work

We contribute to a growing literature considering prediction problems in management and policy decisions. Kleinberg et al. (2015) argue these problems are important and commonplace. Existing work has studied the potential role of machine learning in improving decisions 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), and opioid prescriptions (Hastings et al. 2020). Huang et al. (2022) quantify the value of combining administrative data for prediction and policy outcomes in antibiotic prescribing but do not consider the problem of delegating decisions to physicians. We consider a costly decision making process by a particular kind of manager, the general practitioner who reaches possibly costly decisions on behalf of her patients, and how these decisions may be shared between humans and an algorithm to achieve a common goal.

Because the ground truth is often observed selectively, Kleinberg et al. (2018) propose a design based on random assignment and varying leniency of judges to evaluate the potential improvements of bail decisions using machine learning predictions of recidivism risk. By focusing on patients for whom physicians ordered laboratory tests at initial consultations, we can evaluate machine learning predictions and physician decisions based on observed diagnostic test results, which were not available to physicians at the time of the initial prescription decisions. It is an interesting, complementary question how physicians choose to test patients. Abaluck et al. (2016) find that physicians' systematic misallocations of image scans for diagnosis of pulmonary embolism to low- and high-risk patients impact diagnostic quality and health benefits. While our quantitative results may not be easily generalized beyond this specific context, our considerations about how to delegate decisions to physicians and a prediction-based rule are general. For policy efforts targeting inefficient antibiotic prescribing our analysis holds empirical relevance covering over 70,000 consultations in a relatively short time period and small geographic area.

In medicine, machine learning has driven hopes that electronic health records, insurance claims, administrative data, and genomics databases will help inform medical decision making. Obermeyer and Emanuel (2016) describe how machine learning will improve prognosis in clinical practice in the near future. For example, machine learning methods have been used and are expected to be used increasingly to predict organ failure or mortality, to automatize the interpretation of medical imaging such as mammograms, to automatize 24-hour monitoring in critical care, and to improve

diagnostics. Chen and Asch (2017) highlight the challenges arising from complicating factors in medicine. For example, predicting cancer treatment success in the distant future is a difficult task based on historical health data. Google Flu Trends received much attention but had little success when attempting to predict the prevalence of influenza by combining online search data with a small sample of influenza cases. One lesson from this is that combining relevant types of data and outcomes is crucial (Lazer et al. 2014). Focusing on UTI and diagnosis by urine samples provides a reliable ground truth, with little human judgement in the laboratory diagnostic step, and clear implications for recommended treatment using antibiotics (Grigoryan et al. 2014).

In studying how to improve the efficiency of antibiotic prescribing, we also contribute to the literature considering demand-side mechanisms for policy interventions. These 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). Ribers and Ullrich (2020) model individual physicians' antibiotic prescriptions in the Danish context to study the role of diagnostic skill and physician preferences as mechanisms prediction-based policies can leverage.

3 Danish Administrative Data and Laboratory Test Results

We use Danish administrative data which cover a vast array of information including patient and patient household members' detailed socioeconomic data, antibiotic prescription histories, general practice insurance claims, and hospitalizations. Unique personal identifiers enable us to merge the administrative data to individual laboratory test results acquired from two Danish hospitals responsible for the majority of laboratory testing in the Capital region. Before describing the data in more detail, we provide an overview of the institutional setting, as Denmark has several regulations that impact decision making in general practice and are therefore important for understanding the scope of our results outside the Danish context.

3.1 The Danish healthcare system

Denmark has a universal and tax financed single payer health care system with general practitioners as the primary gatekeepers. Every person living in Denmark is allocated to a general practitioner by a list-system within a fixed geographic radius around the home address. 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 prescription processed regardless of the prescription drug 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 depending on their cumulative yearly prescription drug expenditures. The Danish market for prescription drugs is highly regulated resulting in uniform pricing at pharmacies nationwide and antibiotic treatment is in general cheap, about 100 Danish Kroner (15 US Dollars) per complete treatment.

3.2 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. For each individual, we observe a comprehensive set of socioeconomic and demographic variables combined with the complete prescription history of systemic antibiotics, hospitalizations, and general practitioner insurance claims. The demographic data from the the Danish Civil Registry (*Det Centrale Personregister, CPR*) includes gender, age, home municipality, immigration status and place of origin, marriage and family status. It provides a unique personal identifier which facilitates accurate linkage of patients between all Danish national registers. It also includes household member identifiers which allow us to link the patient's family and household members and add their demographic and administrative data as well. We additionally obtain detailed background information on employment (*Integrerede Database for Arbejdsmarkedsforskning, IDA*) and education (*Uddannelseregister, UDDA*).

Our data on prescription drugs (*Lægemiddeldatabasen, LMDB*) contain the complete history on purchases of systemic antibiotics for every individual in Denmark. We observe the date of purchase, patient and prescribing physician identifiers, anatomical therapeutic chemical drug classification, drug name, price, indication of use, purchased package size, and defined daily dose. The hospitalization data (*Landspatientregisteret, LPR*) comprise all patient contacts with hospitals, including ambulatory visits. The data include information on admission and discharge dates, procedures performed, type of hospitalization (ambulatory, emergency, etc), primary and secondary diagnoses, and the number of total bed days. The insurance claims data (*Sygesikringsregisteret, SSR*) cover all

Danish general practitioner clinic services provided to the population of patients, including physician and patient identifiers, consultation time, consultation type, services provided, and physician fees. Primary care providers are identified via unique clinic identifiers which can be linked to individual physician personal identifiers (*Yderregister, YDER*).

3.3 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 and information on the test type, test acquisition date, sample arrival date at the laboratory, test result response date, isolated bacteria, and a list of antibiotic-specific resistances if bacteria were isolated. The laboratory test data are central because they reveal bacterial presence in a urine test sample, the outcome we aim to predict. 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.

3.4 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 submitted by general practitioners. While we focus on predicting test results for urine samples, we use data from all 2,579,617 observed tests as input for the machine learning algorithm. We restrict the number of test observations by excluding tests where a patient received a systemic antibiotic prescriptions or had another test conducted within 4 weeks prior to the respective test date. We make this restriction to focus on consultations that constitute a first contact with a physician within a patient's treatment spell. In these situations, physicians do not hold current test result information and must prescribe under uncertainty. By considering only 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. Finally, we exclude urine samples collected during pregnancy as the vast majority of these are mandatory routine checks and do not indicate the beginning of a urinary tract infection spell. The final resulting sample we use for the analysis consists of 72,685 initial consultations where a urine

sample was sent to a laboratory for testing from 688 primary care clinics.

3.5 Laboratory test outcomes and prescribing

We consider binary test outcomes that indicate whether bacteria are isolated in patients' urine samples or not. Hence, we do not distinguish between bacterial species but only focus on whether a UTI is caused by bacteria. Yet, in our counterfactual analysis we describe the distribution of bacterial species to consider potential reasons for disagreements between machine learning and physician decisions. 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.[2] Our main focus is on the binary choice of prescribing an antibiotic and not on the antibiotic molecule choice. In the main results section, we provide evidence that resistance complications are not driving initial prescription decisions.

Table 1 Summary statistics for laboratory tests and antibiotic prescribing

				Conditional prescribing rate	
	N	Bacterial rate	Prescribing rate	Positive test	Negative test
2010	19,466	0.36	0.38	0.59	0.26
2011	23,351	0.38	0.38	0.59	0.25
2012	29,868	0.38	0.38	0.60	0.24
Total	72,685	0.38	0.38	0.59	0.25

Table 1 shows that the bacterial rate is stable over the three sample years as is the prescribing rate. In fact, the prescribing rate of 38 percent equals the bacterial rate, which may suggest that physicians match antibiotic prescriptions to bacterial infections very well at the initial consultation. Yet, the final two columns show that this is not the case. Physicians prescribe antibiotics to 59 percent of patients with bacterial infections, that is, they delay prescribing to 41 percent until the test result is available. Accordingly, the rate of overprescribing, prescriptions given to patients without

[2]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 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.

a bacterial infection, must be larger than zero. Indeed, on average 25 percent of patients with a negative test result receive an antibiotic at the initial consultation, although this rate slightly declines over time. Overprescribing provides patients with no realized benefit from antibiotic treatment. Hence, these descriptives indicate that there is potential for improving decisions.

4 Machine learning

We use the machine learning algorithm XGBoost (Hastie et al. 2009, Chen and Guestrin 2016) to relate patient covariates, x_i, at the time patient i's urine sample is acquired to the laboratory bacterial test outcome, y_i. The vector x_i contains 1,557 patient-specific covariates from our Danish administrative data which may, in principle, be observable to the physician at the time of the consultation. We use the 2010 data 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. Figure 4 in Appendix A.1 schematically illustrates the hyperparameter search partitions as well as the training and policy evaluation data partitions.

Table 4 in Appendix A.2 reports the top 5 hyperparameter tuning results and AUC for each tuning and out-of-sample partition. Table 5 in Appendix A.3 shows the sample sizes, bacterial and prescribing rates, risk predictions, and out-of-sample AUC for all 24 policy evaluation partitions. The AUC on the joint set of partitions is 0.72.

While we explicitly treat the machine learning algorithm as a black box, we report three measures of predictor importance for the XGBoost algorithm – gain, frequency, and cover – in Figure 5 and Table 6 in Appendix A.4. Across these three measures, age and gender are among the top predictors. Laboratory data is prominent with regard to the number of days since the patient's last test prior to the consultation, previous patient-specific antibiotic resistance test results, and average bacterial rates at a clinic. Prior patient-specific prescriptions and regional prescription levels are also important predictors, and so is patient immigrant status and origin country.

Our implementation is standard with the one exception that we can not randomly split our data in training and out-of-sample validation partitions due to the dependency of physician-patient interactions over time. Specifically, in practical applications the prediction function must be applied to patients at the moment of clinical consultation while making use of only historic training data. Standard machine learning practice assumes that outcomes and covariates are independent draws

from a joint distribution which remains the same for the training and out of sample partitions (Athey 2018). This assumption does not hold by construction without random splits. Hence, we validate our predictions over time which we do by calculating AUC on a monthly basis. The AUC values, provided in Table 5 in Appendix A.3, are reasonably similar across monthly samples throughout 2011 and 2012.

The left panel of Figure 1 shows the average machine learning predicted risk, $m(x_i)$, and average 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 generate bins of 100 patients, where 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. Although unbiased, the intermediate risk predictions are less informative for individual diagnoses. Prediction accuracy increases for spheres that are closer to either $(0, 0)$ or $(1, 1)$ as the share of correct classifications is higher towards the boundaries. Among the 100 patients with the lowest predicted risk, the average predicted risk is 5.9 percent and only 5 percent are in fact tested positive for bacterial UTI. Equivalently, among the 100 patients with the highest predicted risk 82.4 percent are tested positive for bacteria and the average predicted risk is 89.5 percent.

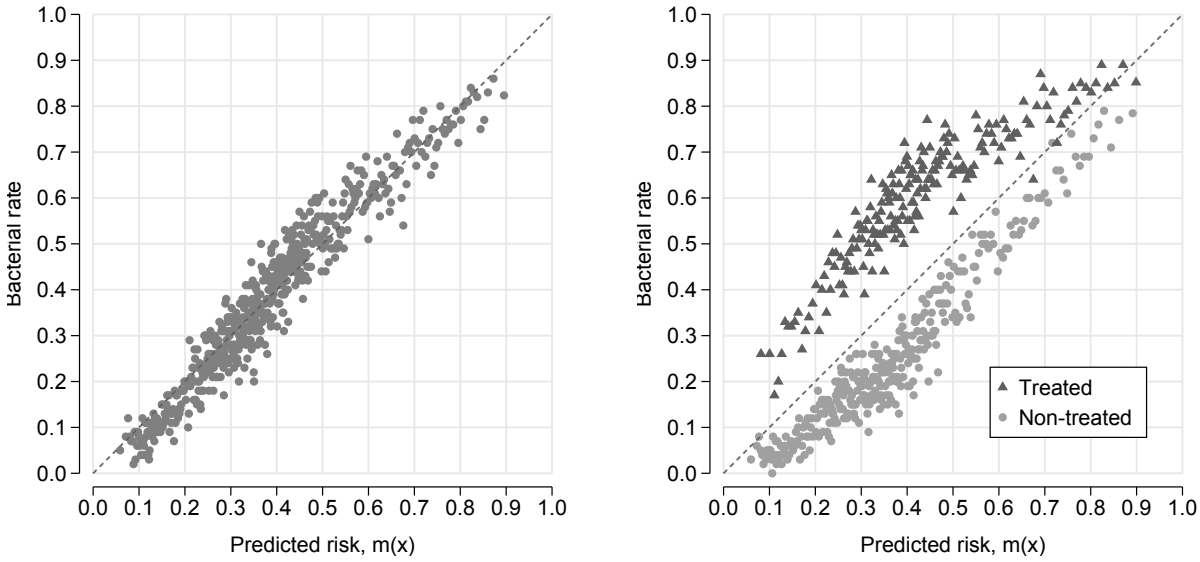

Figure 1: Laboratory test outcomes relative to predicted risk of bacterial UTI unconditional (left) and conditional (right) on antibiotic prescribing prior to receiving test results. Spheres and triangles represent bins of 100 patients after sorting by predicted risk.

The right panel of Figure 1 splits the sample into patients who received a prescription (treated) and those who did not receive a prescription (non-treated) at the initial consultation. Hence, it illustrates test outcomes versus risk predictions conditional on physician antibiotic prescribing prior to receiving laboratory test results. Several important insights emerge. Conditional on predicted risk, the sets of patients with an initial prescription have higher bacterial rates than the sets of 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 provides some, albeit imperfect, diagnostic information on the same day 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.

Yet, physicians' prescription decisions often do not match the true test outcomes, resulting in significant overprescribing. Among the 100 patients with the lowest predicted risk who received an antibiotic, only 26 patients had a bacterial infection. In contrast, 85 patients had a bacterial infection among the 100 patients with the highest predicted risk who received an antibiotic. Hence, we conclude that the risk predictions show potential for reducing physician overprescribing when the machine learning classification accuracy is high. The right panel of Figure 1 also shows an increasing bacterial rate for the non-treated patients as predicted risk increases. Among the 100 non-treated patients with the lowest predicted risk, only 3 patients had a bacterial infection. In contrast, 78 patients had a bacterial infection among the 100 untreated patients with the highest predicted risk. Hence, we equivalently conclude that the risk predictions may be a useful tool in targeting the non-treated patients with high bacterial rate.

These observations indicate that efficiency in the match between prescriptions and bacterial infections can be improved at the extremes of the risk prediction range. At intermediate ranges of predicted risk, physician private information may remain an important source of information that is beneficial to include in a policy rule making use of both algorithm and human decision inputs.

5 Prediction-based prescription policy

5.1 Policymaker payoff

Physicians face a trade-off when they prescribe antibiotics as the potential curative effect must be weighed against the social cost of prescribing. Antibiotic use imposes a social cost because it promotes antibiotic resistance which in turn negates future antibiotic effectiveness (Adda 2020). The

social cost of prescribing 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 and do not provide any benefit against non-bacterial caused infections. Our focus of attention is 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 from the patient for laboratory testing. Test results are on average available within 3.1 days after which patients can be treated in accordance with the laboratory test outcomes. However, delayed treatment to a patient who suffers from bacterial UTI comes at a sickness cost to the patient in the waiting period and so the initial prescription decision can be perceived as a trade-off between the social cost of prescribing and the patient sickness cost from delayed prescribing until the test result is available. Hence, we write a policy maker's realized payoff from initial consultation prescription decisions 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 bacterial UTI and $y = 0$ otherwise, and $d \in \{0, 1\}$ is an indicator for the antibiotic prescription decision with $d = 1$ if an antibiotic is prescribed and $d = 0$ otherwise. The parameter $\alpha > 0$ is the social planner preference weight on avoidance of the patient's sickness cost while awaiting the laboratory test result and the parameter $\beta > 0$ reflects the preference weight on the social cost of prescribing.[3]

5.2 Policy rules

We document in section 4 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

[3] 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.

prediction-based prescription rules that postpone prescribing until test results are available for patients with low predicted risk, give antibiotic prescriptions to patients with high predicted risk, and delegate the decision to the physician for an intermediate risk range. Thus, the prediction-based prescription rules we consider are functions 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 and k_L, k_H are policy threshold parameters to be determined subject to $0 \leq k_L \leq k_H \leq 1$. In the intermediate risk range $k_L < m(x_i) < k_H$, the proposed policy rules require physician-patient consultations. Because initial consultations where $m(x_i) \leq k_L$ or $k_H \leq m(x_i)$ do not require any human input, the share of decisions made without delegation to the physician indicates additional potential efficiency gains in terms of physician and patient time and effort.[4] We discuss a variation of this rule in section 7 where no prescription decisions are delegated to physicians by setting $k_L = k_H$.

5.3 Policy objectives

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 which, unless otherwise stated, covers 2011 and 2012.

The effect of a prescription policy can be separated into two terms. The first term is the part of the joint payoff that accrues from an increase in the number of correctly treated bacterial UTI patients. The second term is the part of joint payoff that accrues from the change in overall antibiotic use. If a prediction-based prescription policy increases the number of treated bacterial UTIs while

[4] An alternative policy design could include the physician's decision as a predictor and evaluate a decision rule using a single threshold k. While such a rule may allow a more flexible combination of physician decisions with administrative data via the prediction algorithm, an implementation would involve higher physician effort because her decision would be a required input at every consultation. For convenience, Huang et al. (2022) use such a rule and find similar results as for the policy we consider.

simultaneously reducing the overall number of antibiotics used, a policy maker will be better off, regardless of the values of the preference weights α and β, as each term in equation (3) is positive. However, depending on the policy maker's preference weights it might be optimal to a policy maker to implement a policy rule that increases the number of untreated bacterial UTIs relative to the status quo in order to accomplish an even larger reduction in antibiotic use than would otherwise be possible. Equivalently, a policy maker might prefer to increase the number of treated bacterial UTIs at the expense of also increasing overall antibiotic use relative to the status quo. The preferred objective depends on the policy maker's weights α and β.

As we do not observe policy maker preferences and therefore do not know the exact trade-off a policy maker would prefer, we instead focus on the particular policy objective that aims to lower antibiotic use while keeping the number of treated bacterial UTIs unchanged similar to the considerations in Kleinberg et al. (2018). If we can show that this prediction-based policy rule can reduce overall antibiotic use then any policy maker will receive a positive payoff from such a policy parameter regardless of the policy maker preference parameters $\alpha > 0$ and $\beta > 0$. Hence, we choose k_L and k_H that solves

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

Note the t index on the patient set I_t which indicates that we can determine k_L and k_H on subsets of the full policy evaluation period. Our main application determines a single k_L and a single k_H for the entire period 2011 and 2012 but shorter periods become important when we discuss how to set the policy parameters out-of-sample which we return to in section 6.3.

The policy parameters resulting from equation (4) also minimize overprescribing under the condition that the number of treated bacterial UTIs remain constant. To realize this note that the difference between counterfactual and observed overprescribing can be written:

$$\sum_i \delta_i(1-y_i) - \sum_i d_i(1-y_i) = \sum_i (\delta_i - d_i) - \sum_i y_i(\delta_i - d_i), \tag{5}$$

which equals the change in antibiotic use when the number of treated bacterial UTIs, the last term above, is zero. We report the change in overprescribing throughout.

6 Policy outcomes

6.1 Reducing antibiotic use

We measure counterfactual policy outcomes relative to observed levels, reporting the percentage change in antibiotic use, the percentage change in treated UTI, as well as the percentage change in overprescribing. Table 2 reports counterfactual outcomes when policy parameters are chosen to reduce antibiotic use without treating fewer bacterial UTI patients. The 95% confidence intervals are based on re-computation of percentage policy results over 100 bootstrap samples for fixed patient risk predictions, $m(x_i)$, and policy parameters (k_L, k_H) in the original sample. The policy with $k_L = 0.31$ and $k_H = 0.62$ results in a reduction in overall antibiotic use of 8.4 percent and a reduction in overprescribing of 20.9 percent relative to observed prescribing and overprescribing, respectively. There is no change to the number of treated UTI by definition of the policy parameters although 95% confidence intervals for fixed k_L, k_H show that the constraint in equation (4) are potentially violated when the policy parameters are set out of sample, a point we return to in section 6.3.

Table 2 Counterfactual outcomes for 2011 and 2012

k_L	0.310
k_H	0.622
Change in treated UTI, in %	0.0 [−1.0, 0.7]
Change in antibiotic use, in %	−8.4 [−9.2, −7.6]
Change in overprescribing, in %	−20.9 [−22.4, −19.5]
GP decisions overruled, in %	14.4 [14.1, 14.6]
Patients delegated to physician, in %	54.3 [45.2, 46.0]
Total consultations	53,219
Total bacterial UTIs	20,411
Total treated UTIs	12,114
Total antibiotic prescriptions	20,154
Total overprescribing	8,040

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.

In 2017, the Danish government initiated a national action plan in which one main goal is to reduce overall antibiotic prescribing by one third by 2020 compared to 2016 (Danish Ministry of Health 2017). For the initial consultations we consider, combining machine learning and physician

decisions attains one third of this goal.

6.2 Discussion

We may worry that antibiotics are given unnecessarily by our policy rules, even if prescriptions are assigned to patients who are truly suffering from bacterial UTIs. For example, asymptomatic infections may be left untreated even though physicians have a very accurate evaluation of the high 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 in their sample, it is informative to consider physicians' prescription choices when the definitive test outcome is known. We look closer at patients with a positive test result to whom the counterfactual policy assigns an antibiotic prescription but physicians did not give a prescription at the initial consultation. For these 1,907 patients, we find that 71 percent receive a follow-up antibiotic prescription after the definitive arrival of microbiological test results.[5] This is higher than 64 percent rate of follow-up prescribing when test results are available to patients who in general did not receive an initial prescription but showed a positive test result regardless of the counterfactual prescription rule. 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.

One further 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. This would imply that information only about high predicted bacterial risk is not useful. 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. If physicians know with high accuracy whether an infection is bacterial and suspect resistance to be high, we expect to observe that resistance is higher for bacteria found in patients for whom physicians did not give a prescription at initial consultations than for bacteria found in patients receiving a prescription instantly.

Table 7 in Appendix C shows the distribution of bacterial species. We observe some differences

[5]For some patients, partial test results may be communicated before the average delay of three days for the definitive test result, which we do not observe. Based on follow-up prescriptions from two days after the initial consultation or later, the share is 71 percent.

in the detected bacteria for treated and untreated patients with high predicted bacterial risk. When physicians decide to treat instantly, E. coli account for roughly 67 percent, K. pneumoniae for 9 percent, E. faecalis for 6 percent, Enterococcus for 3 percent, P. mirabilis for 2 percent while other species account for 13 percent of the bacteria found. When physicians decide to delay prescribing, E. coli account for roughly 77 percent, K. pneumoniae for 6 percent, E. faecalis for 3 percent, Enterococcus for 2 percent, P. mirabilis for 1 percent while other species account for 11 percent of the bacteria found. The significant difference in the fraction of detected bacteria can be explained by in-clinic diagnostics done by some physicians, which can for instance identify E. coli bacteria relatively well.[6]

Table 8 in Appendix C shows the resistance rates for samples with E. coli bacteria which are the main cause of UTI. We find small differences in resistances against the bulk of antibiotics prescribed for UTI. When physicians decide to treat instantly and bacteria are found, these have 2-6 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 in anticipation of complete test results. Given these differences, there seems to be value to address prediction of specific bacteria and resistances in further research. Yelin et al. (2019) and Kanjilal et al. (2020) find promising results for predicting resistance levels using electronic health records in hospital contexts. Yet, the differences in our analysis do not appear of such a magnitude that decisions to postpone prescribing are driven by uncertainty about bacterial resistance.

6.3 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 sickness realizations. In reality, both policy parameters need to be determined ahead of time. There are potentially many ways to go about this task, see for example Hazan (2021). However, we explore

[6]Nitrite dipstick can detect bacteria that transform Nitrate to Nitrite. In the hold out data, the detectable genera are Escherichia, Enterobacter, Klebsiella, Citrobacter, and Proteus. The non-detectable genera are Staphylococcus, Pseudomonas, Enterococci, Acinetobacter, and Streptococcus. Inspecting prescription choices separately by dipstick-detectable and non-detectable bacterial species isolated in laboratory tests allows us to investigate whether physicians select on nitrite dipstick test results. While patients with dipstick-detectable bacteria have a higher prescription rate, 64 percent, relative to prescription rate for patients with non-dipstick-detectable bacteria, 55 percent, the difference is moderate. This is consistent with the significant lack of accuracy of dipstick test results (Devillé et al. 2004).

simple ways to determine and update the policy parameters out-of-sample as outlined in Appendix B to show that the policy results can likely be realized in a real world application but care needs to be exercised.

Specifically, we determine k_L and k_H based on historic data relative to the observations which the policy parameters are applied to. We do this at the yearly level, using 2011 to determine policy parameters for 2012, as well as the half-yearly, quarterly, and monthly level. The longest policy evaluation period where all methods can be jointly evaluated and compared is the full year 2012. Figure 2 shows the out-of-sample counterfactual results with each different update frequency of the policy parameters.

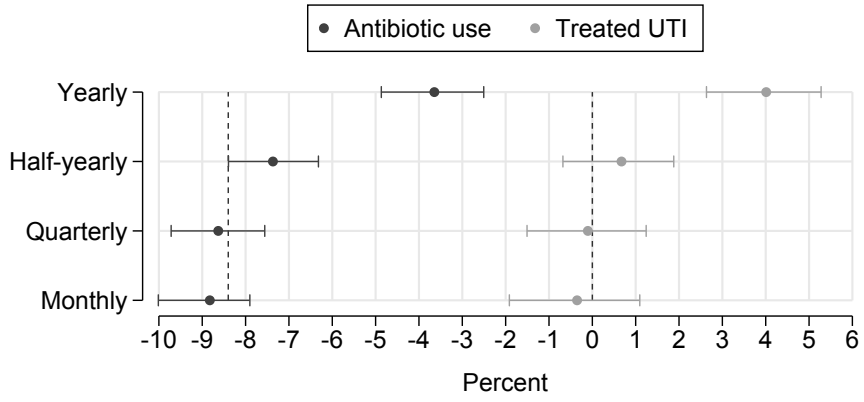

Figure 2: Prediction-based prescription policy outcomes for 2012 updating policy parameters out-of-sample at the yearly, half-yearly, quarterly and monthly level

The dashed vertical lines show our main results from Table 2. The yearly policy parameters cannot reproduce the in-sample results. Instead the number of correctly treated bacterial UTIs increases while the reduction in antibiotic use is only nearly one third of the in-sample results. However, the half-yearly, quarterly and monthly update frequencies generate out-of-sample results that compare to the in-sample results. Table 9 Appendix D shows all in-sample and out-of-sample policy results for each update method for 2012. The different update frequencies result in slightly different in-sample policy outcomes compared to our main result in Table 2 but these differences are are not significant.

6.4 The policy possibility frontier

Motivated by common public health policy considerations, so far we have focused on the policy objective of reducing antibiotic use without treating fewer patients with bacterial UTI (WHO 2012,

2014). Alternative policy objectives could be attained by varying the policy parameters (k_L, k_H). Figure 3 shows the relevant subset of attainable changes in antibiotic use and the number of treated bacterial UTIs for policy parameters $0 \leq k_L \leq k_H \leq 1$. Figure 6 in Appendix E shows the full set of attainable policy outcomes. We refer to the upper-left bound of this set as the policy possibility frontier. This bound shows the efficient combinations of the two terms in the objective function that can be achieved by tracing out all possible decision rules implied by equation (2).

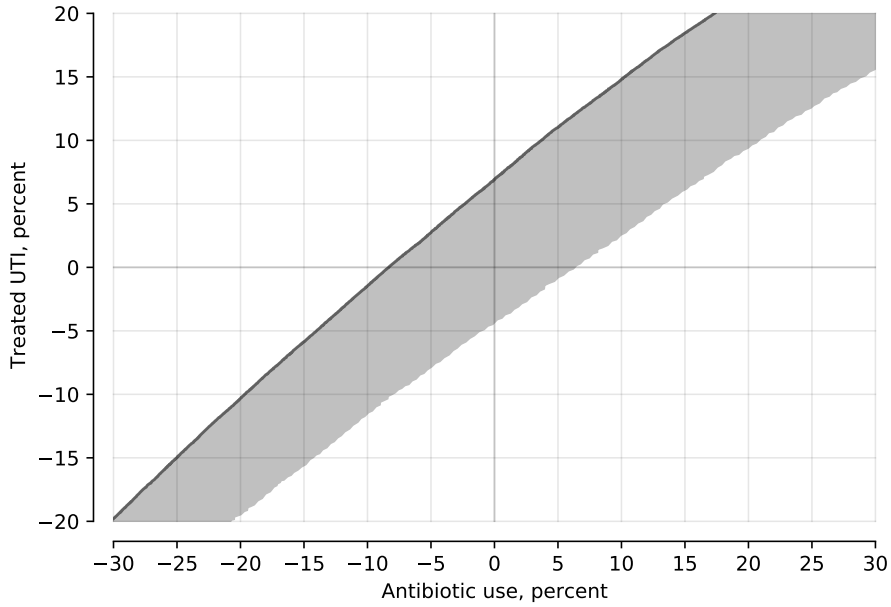

Figure 3: Policy outcomes as a function of policy parameters (k_L, k_H)

In the upper left quadrant, antibiotic use is always reduced or unchanged while the number of treated bacterial UTIs is positive or zero. In this area, any policy maker will prefer the counterfactual policy outcomes relative to the status quo regardless of the policy parameters $\alpha > 0$ and $\beta > 0$. Our main result lies at the limit of this set where the policy possibility frontier intercepts the horizontal axis. Here, the change in the number of treated bacterial infections is zero and the change in antibiotic use is -8.4 percent. At the other limit of this set, where the policy possibility frontier 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. At the same time, overprescribing is reduced by 10.5 percent.

The prediction-based policy rules can also provide larger reductions in antibiotic use compared to our main result, but not without decreasing the number of treated bacterial infections. For instance, a 20 percent reduction in antibiotic use can be attained. However, this would require 10.3

percent of patients with bacterial infections to have their treatment delayed until test results are available.

7 Prescribing without physicians

In the counterfactual policy with delegation, physicians make the majority of treatment decisions (54.3 percent) without use of machine learning risk predictions. A natural question that arises is how well a policy would fare by which all prescription decisions were determined by the algorithm. We explore this policy by imposing the restriction $k_L = k_H$ in equation (2), collapsing the decisions delegated to the physician in the intermediate risk range to an empty set. 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:

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

Table 3 shows that this rule, holding the number of prescriptions to patients with a bacterial infection constant, gives a prescription to all patients with predicted risk equal to 0.4 or higher. Here, 39.9 percent of physicians' decisions are overturned. A reduction in antibiotic use is not possible. Instead, antibiotic use increases by 6.3 percent. In addition, overprescribing increases by 15.9 percent. Hence, without physician input in decision-making, no improvements are feasible.

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

k		0.40
Change in treated UTI, in %	0.0	[−1.7, 1.6]
Change in antibiotic use, in %	6.3	[4.7, 7.6]
Change in overprescribing, in %	15.9	[12.7, 18.3]
GP decisions overruled, in %	39.9	[39.5, 40.2]

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

Figure 7 in Appendix E shows counterfactual percentage changes in antibiotic prescriptions versus the percentage changes in treated patients with bacterial UTIs as k varies over the full risk range. For $k = 0$, all tested patients receive a prescription and overall prescribing increases by 164

percent while increasing the number of correctly treated bacterial infections by 68 percent. At the other extreme, $k = 1$, no patients are treated at the initial consultation and 100 percent of the initially prescribed antibiotics are delayed while 100 percent of the observed treated patients with bacterial UTIs go untreated.

Notably, for any negative percentage change in antibiotic use the change in treated UTIs is also negative, and so we can conclude that algorithmic prescribing alone cannot reduce antibiotic use without decreasing the number of treated bacterial UTIs. Equivalently, algorithmic prescribing alone cannot increase the number of treated UTIs without additional use of antibiotics as no positive change in treated UTIs corresponds to a negative change in antibiotic use. In particular, the rules in equation (6) fail to deliver improvements regardless of policy makers' preference parameters a and b. We conclude that even with high-dimensional individual-specific data, machine learning predictions need to be combined with physician expertise to provide policy improvements.

8 Conclusion

The quality of prediction algorithms and available data are improving at a rapid pace. In this paper, we document the potential 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 under uncertainty about the cause of infection is a common decision problem for expert physicians. The aggregate of these decisions has important societal implications due to the empirical relevance of increasing antibiotic resistance driven by inefficient antibiotic prescribing. We show that counterfactual decision rules lead to improvements over physician prescribing. While policies based on machine learning predictions alone do not deliver improvements, antibiotic use can be reduced by delegating decisions to physicians when machine learning predictions are most uncertain.

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). Hence, while our results may be challenging to reproduce or implement in other countries due to the lack of linked data, we suspect the potential reductions we find here present a lower bound of what may be achievable in other institutional settings. Promoting efforts for similar evaluations in other health care systems would therefore be worthwhile.

One limitation is that we consider only initial consultation prescription occasions in which a

laboratory test was used. UTI typically must be treated quickly and laboratory testing takes considerable time. Hence, many UTI cases do not make use of elaborate testing even though its use is promoted to improve antibiotic prescribing in Denmark. Therefore, our conclusions hold for the set of consultations which were likely more difficult to diagnose, requiring an advanced diagnostic.

One promising avenue for further research is the use of clinical information such as recorded symptoms and results from in-clinic diagnostics in machine learning predictions. Another important area in which further research is needed is the analysis of experts' behavioral reactions to the prediction-based policies proposed here as physicians' incentives to treat and test are likely to change. Finally, our machine prediction results could be used to assist physicians in their decision-making, for example by providing physicians with the machine predicted risk evaluation at every prescription occasion. Ribers and Ullrich (2020) attempt to decompose prescription decisions into two mechanisms, physicians' diagnostic information and payoff functions. A more complete assessment of equilibrium effects of such recommendations or information provision is likely possible by trialling interventions in the field.

References

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

Adda J (2020) Preventing the spread of antibiotic resistance. *AEA Papers and Proceedings* 110:255–259, URL http://dx.doi.org/10.1257/pandp.20201014.

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

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).

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, URL http://dx.doi.org/10.1371/journal.pone.0109264.

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

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, URL http://dx.doi.org/10.1136/bmj.d8173.

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

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, URL http://dx.doi.org/10.1257/aer.101.3.288.

Chen JH, Asch SM (2017) Machine learning and prediction in medicine-beyond the peak of inflated expectations. *New England Journal of Medicine* 376(26):2507–2509.

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, KDD '16 (New York, NY, USA: Association for Computing Machinery), URL http://dx.doi.org/10.1145/2939672.2939785.

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'.

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 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, URL `http://dx.doi.org/10.1257/aer.20151138`.

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, URL `http://dx.doi.org/10.1186/1471-2490-4-4`.

Dubé JP, Misra S (2021) Personalized Pricing and Consumer Welfare. *Journal of Political Economy* Forthcoming, URL `http://dx.doi.org/10.2139/ssrn.2992257`.

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.

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, URL `http://dx.doi.org/10.1016/S0140-6736(05)17907-0`.

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.

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, URL `http://dx.doi.org/10.1073/pnas.1905355117`.

Hazan E (2021) Introduction to online convex optimization.

Huang S, Ribers MA, Ullrich H (2022) Assessing the value of data for prediction policies: The case of an-

tibiotic prescribing. *Economics Letters* 110360, ISSN 0165-1765, URL http://dx.doi.org/10.1016/j.econlet.2022.110360.

Kahneman D, Sibony O, Sunstein CR (2021) *Noise: A Flaw in Human Judgment* (New York: Little, Brown Spark), ISBN 978-0-316-45140-6.

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), URL http://dx.doi.org/10.1126/scitranslmed.aay5067.

Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S (2018) 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.

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

Laxminarayan R (2022) The overlooked pandemic of antimicrobial resistance. *The Lancet* URL http://dx.doi.org/10.1016/S0140-6736(22)00087-3.

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* 1–42.

Lazer D, Kennedy R, King G, Vespignani A (2014) The parable of Google Flu: Traps in big data analysis. *Science* 343(6176):1203–1205, URL http://dx.doi.org/10.1126/science.1248506.

Murray CJ, Ikuta KS, Sharara F, Swetschinski L, Aguilar GR, Gray A, Han C, Bisignano C, Rao P, Wool E, Johnson SC, Browne AJ, Chipeta MG, Fell F, Hackett S, Haines-Woodhouse G, Hamadani BHK, Kumaran EAP, McManigal B, Agarwal R, Akech S, Albertson S, Amuasi J, Andrews J, Aravkin A, Ashley E, Bailey F, Baker S, Basnyat B, Bekker A, Bender R, Bethou A, Bielicki J, Boonkasidecha S, Bukosia J, Carvalheiro C, Castañeda-Orjuela C, Chansamouth V, Chaurasia S, Chiurchiù S, Chowdhury F, Cook AJ, Cooper B, Cressey TR, Criollo-Mora E, Cunningham M, Darboe S, Day NPJ, Luca MD, Dokova K, Dramowski A, Dunachie SJ, Eckmanns T, Eibach D, Emami A, Feasey N, Fisher-Pearson N, Forrest K, Garrett D, Gastmeier P, Giref AZ, Greer RC, Gupta V, Haller S, Haselbeck A, Hay SI, Holm M, Hopkins S, Iregbu KC, Jacobs J, Jarovsky D, Javanmardi F, Khorana M, Kissoon N,

Kobeissi E, Kostyanev T, Krapp F, Krumkamp R, Kumar A, Kyu HH, Lim C, Limmathurotsakul D, Loftus MJ, Lunn M, Ma J, Mturi N, Munera-Huertas T, Musicha P, Mussi-Pinhata MM, Nakamura T, Nanavati R, Nangia S, Newton P, Ngoun C, Novotney A, Nwakanma D, Obiero CW, Olivas-Martinez A, Olliaro P, Ooko E, Ortiz-Brizuela E, Peleg AY, Perrone C, Plakkal N, Ponce-de-Leon A, Raad M, Ramdin T, Riddell A, Roberts T, Robotham JV, Roca A, Rudd KE, Russell N, Schnall J, Scott JAG, Shivamallappa M, Sifuentes-Osornio J, Steenkeste N, Stewardson AJ, Stoeva T, Tasak N, Thaiprakong A, Thwaites G, Turner C, Turner P, van Doorn HR, Velaphi S, Vongpradith A, Vu H, Walsh T, Waner S, Wangrangsimakul T, Wozniak T, Zheng P, Sartorius B, Lopez AD, Stergachis A, Moore C, Dolecek C, Naghavi M (2022) Global burden of bacterial antimicrobial resistance in 2019: A systematic analysis. *The Lancet* 399(10325):629–655, ISSN 0140-6736, 1474-547X, URL `http://dx.doi.org/10.1016/S0140-6736(21)02724-0`.

Obermeyer Z, Emanuel EJ (2016) Predicting the future – big data, machine learning, and clinical medicine. *New England Journal of Medicine* 375(13):1216–1219.

Ribers MA, Ullrich H (2020) Machine predictions and human decisions with variation in payoffs and skills. DIW Discussion Paper No. 1911.

Rose S (2018) Machine learning for prediction in electronic health data. *JAMA Network Open* 1:4.

Thaler RH, Sunstein CR (2009) *Nudge: Improving Decisions About Health, Wealth, and Happiness* (New York: Penguin Books), revised & expanded edition edition, ISBN 978-0-14-311526-7.

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.

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, URL `http://dx.doi.org/10.1016/S0140-6736(10)60063-3`.

Appendices

Appendix A Machine learning

A.1 Overview of machine learning data partitions

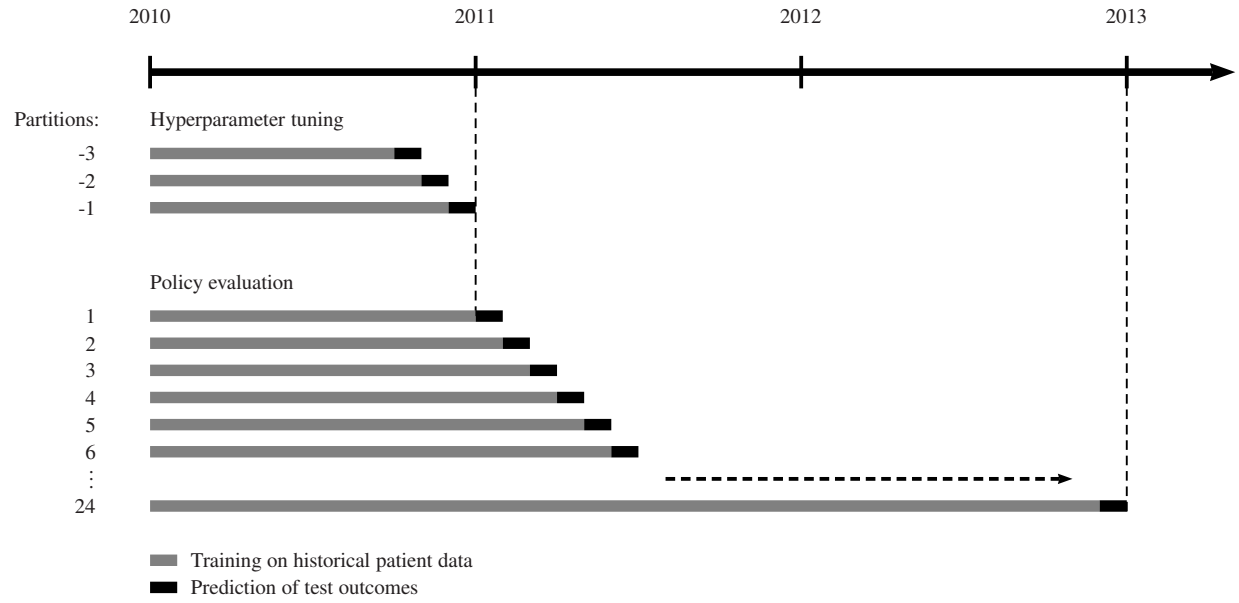

Figure 4: Outline of the data partitions used for hyperparameter tuning as well as the month-by-month progressing training and out-of-sample prediction partitions

A.2 Hyperparameters

Table 4 Top 5 hyperparameter search results

Rank	Rounds	Learning rate	Tree depth	Avg. AUC
1	599	0.03	3	0.70404
2	582	0.03	4	0.70398
3	330	0.04	4	0.70387
4	210	0.06	3	0.70383
5	616	0.04	3	0.70378

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 Data partitions

Table 5 Data partitions summary statistics

	Training					Prediction						
Partition	N	E(y)	E(d)	E(d|y=1)	E(d|y=0)	N	E(mx)	E(y)	E(d)	E(d|y=1)	E(d|y=0)	AUC
-3	14,349	0.36	0.38	0.59	0.26	1,777		0.36	0.38	0.57	0.27	-
-2	16,126	0.36	0.38	0.59	0.26	1,888		0.37	0.37	0.58	0.25	-
-1	18,014	0.36	0.38	0.59	0.26	1,452		0.36	0.41	0.60	0.30	-
1	19,466	0.36	0.38	0.59	0.26	1,934	0.36	0.35	0.36	0.57	0.24	0.71
2	21,400	0.36	0.38	0.59	0.26	1,669	0.37	0.36	0.37	0.58	0.25	0.72
3	23,069	0.36	0.38	0.59	0.26	2,018	0.36	0.37	0.36	0.57	0.24	0.70
4	25,087	0.36	0.38	0.59	0.26	1,558	0.37	0.39	0.39	0.59	0.26	0.71
5	26,645	0.36	0.38	0.59	0.26	2,025	0.38	0.39	0.35	0.54	0.24	0.72
6	28,670	0.36	0.38	0.58	0.26	1,917	0.39	0.40	0.37	0.57	0.24	0.73
7	30,587	0.37	0.38	0.58	0.26	1,396	0.40	0.41	0.44	0.65	0.29	0.71
8	31,983	0.37	0.38	0.59	0.26	2,104	0.40	0.39	0.37	0.60	0.23	0.70
9	34,087	0.37	0.38	0.59	0.26	2,315	0.39	0.39	0.39	0.62	0.26	0.72
10	36,402	0.37	0.38	0.59	0.26	2,212	0.38	0.39	0.39	0.61	0.25	0.71
11	38,614	0.37	0.38	0.59	0.26	2,370	0.39	0.39	0.36	0.57	0.24	0.71
12	40,984	0.37	0.38	0.59	0.25	1,833	0.39	0.40	0.39	0.61	0.24	0.72
13	42,817	0.37	0.38	0.59	0.25	2,467	0.40	0.39	0.37	0.60	0.23	0.75
14	45,284	0.37	0.38	0.59	0.25	2,093	0.40	0.38	0.37	0.61	0.22	0.72
15	47,377	0.37	0.38	0.59	0.25	2,404	0.38	0.36	0.35	0.57	0.23	0.71
16	49,781	0.37	0.38	0.59	0.25	1,860	0.39	0.39	0.38	0.61	0.24	0.73
17	51,641	0.37	0.38	0.59	0.25	2,300	0.39	0.36	0.36	0.58	0.24	0.74
18	53,941	0.37	0.38	0.59	0.25	2,668	0.38	0.37	0.37	0.59	0.25	0.73
19	56,609	0.37	0.38	0.59	0.25	1,812	0.41	0.43	0.43	0.64	0.27	0.72
20	58,421	0.38	0.38	0.59	0.25	2,996	0.40	0.40	0.40	0.60	0.26	0.72
21	61,417	0.38	0.38	0.59	0.25	2,767	0.39	0.39	0.38	0.59	0.24	0.72
22	64,184	0.38	0.38	0.59	0.25	3,043	0.39	0.39	0.39	0.61	0.26	0.72
23	67,227	0.38	0.38	0.59	0.25	3,288	0.39	0.37	0.36	0.58	0.23	0.74
24	70,515	0.38	0.38	0.59	0.25	2,170	0.39	0.38	0.40	0.61	0.27	0.73

A.4 Predictor importance

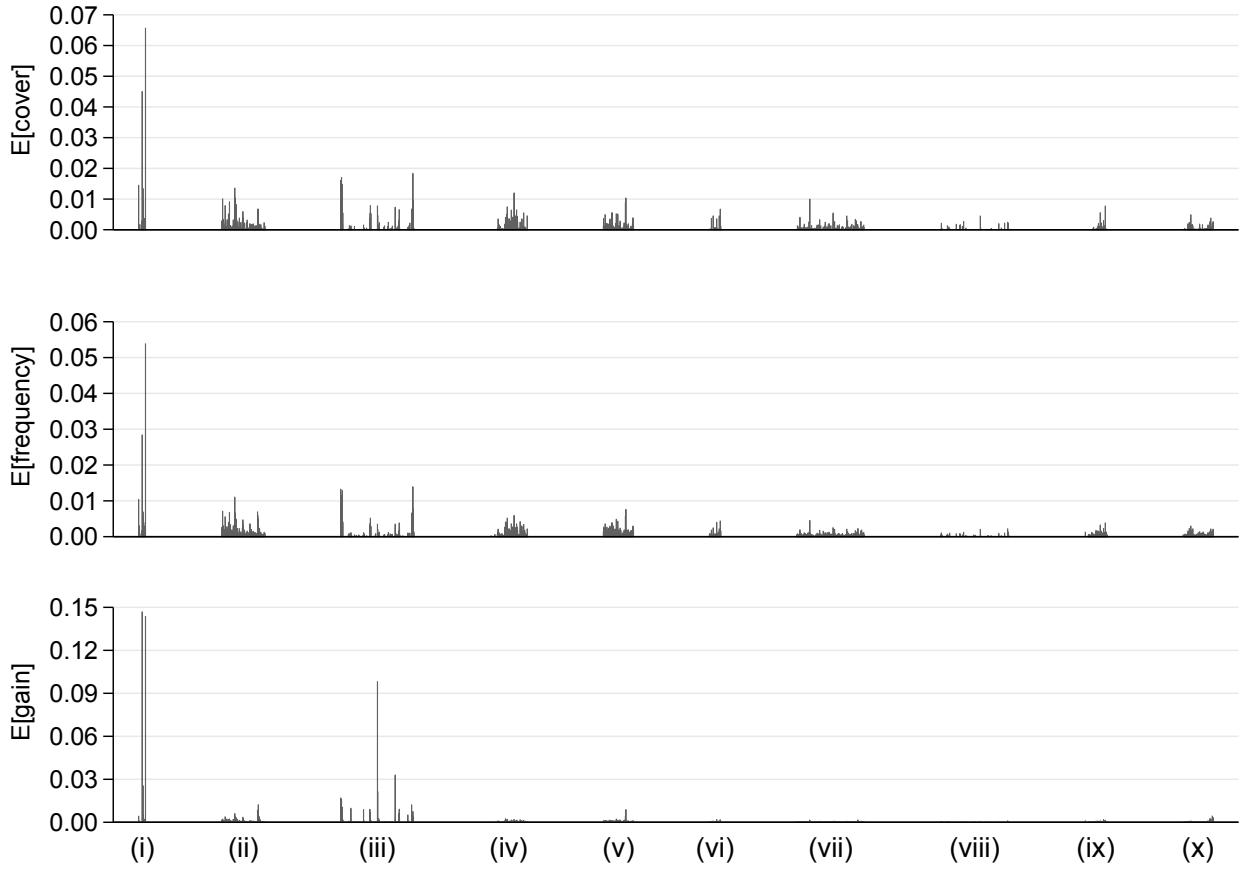

Figure 5: Mean gain, cover and frequency averaged over the 24 monthly out-of-sample policy partitions from January 2011 to December 2012.

Gain, cover, and frequency provide measures of predictor importance (Chen et al. 2022). Variables in Figure 5 are listed by groups based on their administrative data sources:

(i) patient demographics and test timing
(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 hospitalizations
(xi) Household member primary care claims

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

	Cover			Frequency			Gain		
	Predictor	Group	Cover	Predictor	Group	Frequency	Predictor	Group	Gain
1	Age	i	0.0656	Age	i	0.0538	Gender	i	0.1468
2	Gender	i	0.0450	Gender	i	0.0284	Age	i	0.1436
3	Laboratory1 days	iii	0.0183	Laboratory1 days	iii	0.0139	Laboratory1 J01EA01	iii	0.0980
4	Physician 12M avg(y)	ii	0.0167	Physician avg(y)	iii	0.0132	Laboratory1 J01MA02	iii	0.0329
5	Physician avg(y)	iii	0.0161	Physician 6M avg(y)	iii	0.0129	Immigrant	i	0.0256
6	Physician 6M avg(y)	iii	0.0147	Physician 12M avg(y)	iii	0.0115	Laboratory2 J01EA01	iii	0.0213
7	Clinic identifier	i	0.0145	Prescription1 atc	ii	0.0110	Physician avg(y)	iii	0.0169
8	Prescription1 atc	ii	0.0135	Clinic identifier	i	0.0104	Physician 12M avg(y)	iii	0.0160
9	Immigrant	i	0.0134	Laboratory2 days	iii	0.0078	Prescription4 days	ii	0.0122
10	Hospital bed days	iv	0.0120	Claim30	v	0.0076	Laboratory1 species	iii	0.0122
11	Prescription3 atc	ii	0.0103	Regional 3M DID J01AA07	ii	0.0071	Physician 6M avg(y)	iii	0.0105
12	Claim30	v	0.0103	Immigrant	i	0.0069	Laboratory1 J01CA11	iii	0.0097
13	Regional 3M DID J01AA07	ii	0.0100	Prescription1 days	ii	0.0069	Laboratory2 J01MB02	iii	0.0091
14	Father prescription	vii	0.0099	Regional 3M DID J01FA01	ii	0.0067	Laboratory1 J01DD13	iii	0.0090
15	Origin country	i	0.0094	Prescription3 atc	ii	0.0066	Claim30	v	0.0088
16	Laboratory2 days	iii	0.0094	Laboratory1 species	iii	0.0066	Laboratory1 J01DC02	iii	0.0088
17	Regional 3M DID J01FA01	ii	0.0091	Prescription2 atc	ii	0.0065	Prescription1 days	ii	0.0084
18	Claim2	v	0.0087	Hospital bed days	iv	0.0059	Prescription2 days	ii	0.0080
19	Prescription6 atc	ii	0.0082	Prescription2 days	ii	0.0058	Laboratory1 days	iii	0.0075
20	Regional 3m DID J01CF01	i	0.0079	Claim28	v	0.0056	Laboratory2 J01DD13	iii	0.0068
21	Laboratory3 J01DD13	iii	0.0078	Regional 3M DID J01CF01	ii	0.0056	Prescription3 days	ii	0.0068
22	Prescription2 atc	ii	0.0078	Hospital diagnose	iv	0.0052	Laboratory1 J01MB02	iii	0.0063
23	Laboratory1 J01EA01	iii	0.0077	Laboratory3 J01DD13	iii	0.0051	Prescription1 atc	ii	0.0061
24	Family hospitalization	ix	0.0077	Origin country	i	0.0050	Origin country	i	0.0055
25	Hospital diagnose	iv	0.0074	Prescription4 days	ii	0.0050	Laboratory3 J01DD13	iii	0.0053
26	Laboratory1 J01MA02	iii	0.0073	Prescription6 atc	ii	0.0049	Laboratory1 J01XE01	iii	0.0051
27	Prescription2 days	ii	0.0068	Claim22	v	0.0048	Family claim	x1	0.0043
28	Laboratory1 species	iii	0.0067	Prescription3 days	ii	0.0047	Clinic identifier	i	0.0042
29	Partner age	vi	0.0067	Prescription3 indication	ii	0.0047	Laboratory2 days	iii	0.0042
30	Prescription4 days	ii	0.0066	Laboratory avg res	iii	0.0045	Prescription2 atc	ii	0.0042

Appendix B Policy algorithm

The following algorithm computes policy parameters k_L and k_H and evaluates policy results in-sample and out-of-sample:

Policy algorithm

1. Train a prediction model and predict test outcomes following the structure outlined in appendix A.1.
2. Compute $k_L(t)$ and $k_H(t)$ for all patients $i \in I_t$ tested in the period $[t, t + \Delta t)$ using equation (4).
3. Repeat step 2 moving t forward in time in steps of Δt each iteration.
4. Evaluate aggregated policy outcomes **in-sample** using

$$\sum_{t=2}^{T} \sum_{i \in I_t} y_i(\delta_i(k_L(t), k_H(t)) - d_i)$$

and

$$\sum_{t=2}^{T} \sum_{i \in I_t} \delta_i(k_L(t), k_H(t)) - d_i.$$

5. Evaluate aggregated policy outcomes **out-of-sample** using

$$\sum_{t=2}^{T} \sum_{i \in I_t} y_i(\delta_i(k_L(t-1), k_H(t-1)) - d_i)$$

and

$$\sum_{t=2}^{T} \sum_{i \in I_t} \delta_i(k_L(t-1), k_H(t-1)) - d_i.$$

We show results with update frequency set such that Δt has a two-yearly, yearly, half-yearly, quarterly and monthly length. We purposely avoid adding structural assumption on the evolution of the policy parameters in real time in order to avoid over-fitting.

Appendix C Laboratory results for high predicted risk patients

Table 7 Distribution of bacterial species in positive test results for high predicted risk patients conditional on physician antibiotic prescription decisions at the initial consultation

	$d = 1$		$d = 0$	
Species/genus	Obs	Pct	Obs	Pct
E. coli	1,270	66.6	1,790	77.4
K. pneumoniae	167	8.8	146	6.3
E. faecalis	115	6.0	63	2.7
Enterorcuccus	52	2.7	37	1.6
P. mirabilis	40	2.1	32	1.4
Other	263	13.8	244	10.6
Total	1.907	100	2.312	100

Table 8 Antibiotic resistance for positive E. coli test results among high predicted risk patients conditional on physician antibiotic prescription decisions at the initial consultation

	$d = 1$		$d = 0$	
Antibiotic (ATC-code)	Obs	Resistance	Obs	Resistance
Ampicillin (J01CA01)	1,787	0.40	1,268	0.46
Mecillinam (J01CA11)	1,789	0.04	1,270	0.06
Trimethoprim (J01EA01)	1,788	0.28	1,269	0.33
Sulfamethizole (J01EB02)	1,788	0.34	1,267	0.39
Ciprofloxacin (J01MA02)	1,640	0.17	1,165	0.21
Nitrofurantion (J01XE01)	1,790	0.03	1,269	0.05

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

Table 9 Policy results for 2012 with policy parameters set in-sample and out-of-sample

	Change in antibiotic use (%)		Change in treated UTI (%)	
	In-sample	Out-of-sample	In-sample	Out-of-sample
- Yearly set k_L, k_H	-8.7 $[-9.7, -7.6]$	-3.6 $[-4.9, -2.5]$	0.0 $[-1.4, 1.3]$	4.0 $[2.6, 5.3]$
- Half-yearly set k_L, k_H	-8.9 $[-9.9, -8.0]$	-7.4 $[-8.4, -6.3]$	0.0 $[-1.2, 1.1]$	0.7 $[-0.7, 1.9]$
- Quarterly set k_L, k_H	-9.1 $[-10.1, -8.0]$	-8.6 $[-9.7, -7.6]$	0.0 $[-1.2, 1.2]$	-0.1 $[-1.5, 1.2]$
- Monthly set k_L, k_H	-9.7 $[-10.9, -8.5]$	-8.8 $[-10.0, -7.9]$	0.0 $[-1.5, 1.4]$	-0.4 $[-1.9, 1.1]$

95% confidence intervals are based on 100 bootstrap samples where machine learning predictions remain fixed at the patient level and policy parameters (k_L, k_H) are optimized for each bootstrapped sample.

Appendix E Alternative policy objectives

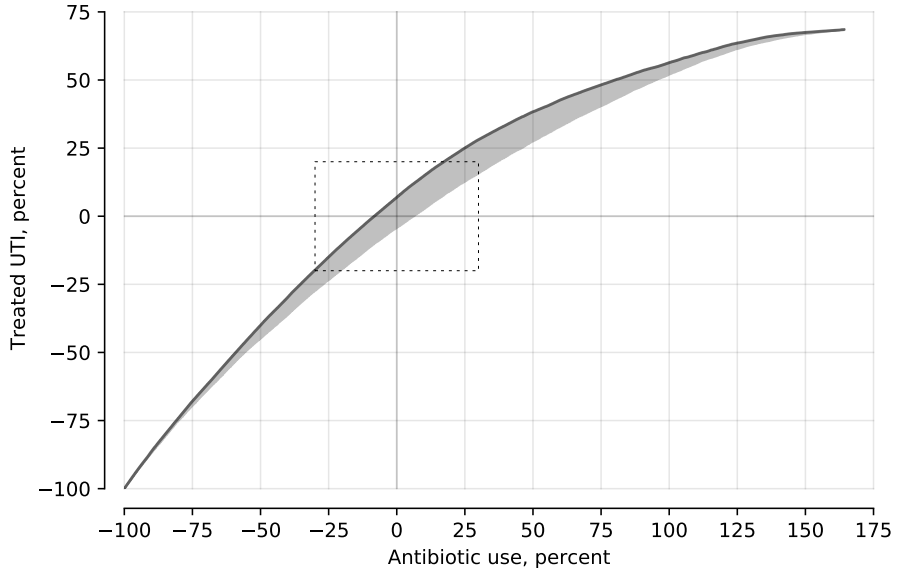

Figure 6: The set of all policy outcomes as a function of the policy parameters (k_L, k_H) for 2011 and 2012. The dashed rectangle shows the policy outcomes highlighted in Figure 3 in the main text.

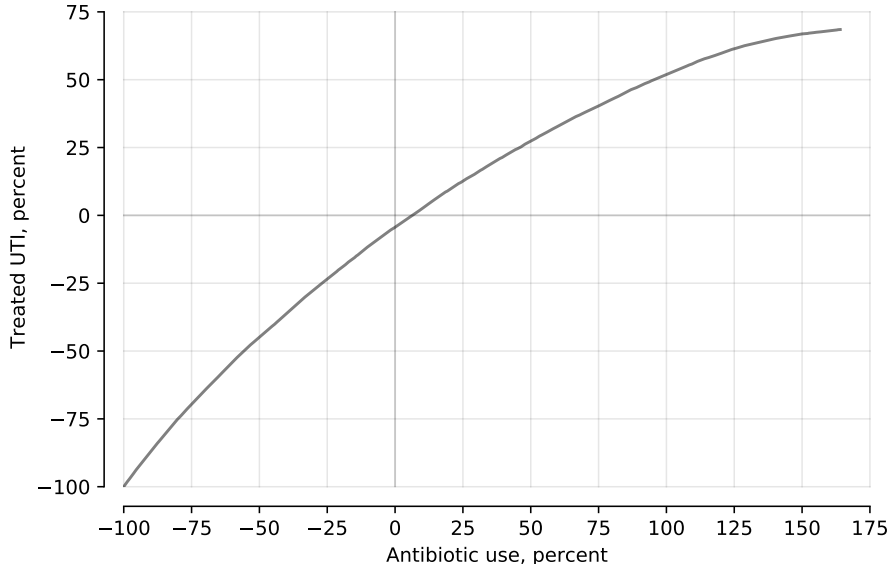

Figure 7: The set of all policy outcomes as a function of the policy parameters k for 2011 and 2012.