DIW BERLIN

1939

Discussion Papers

Deutsches Institut für Wirtschaftsforschung 2021

The Value of Data for Prediction Policy Problems: Evidence from Antibiotic Prescribing

Shan Huang, Michael Allan Ribers and Hannes Ullrich

IMPRESSUM

DIW Berlin
German Institute for Economic Research
Mohrenstr. 58
10117 Berlin

Tel. +49 (30) 897 89-0
Fax +49 (30) 897 89-200
http://www.diw.de

ISSN electronic edition 1619-4535

Papers can be downloaded free of charge from the DIW Berlin website:
http://www.diw.de/discussionpapers

Discussion Papers of DIW Berlin are indexed in RePEc and SSRN:
http://ideas.repec.org/s/diw/diwwpp.html
http://www.ssrn.com/link/DIW-Berlin-German-Inst-Econ-Res.html

The Value of Data for Prediction Policy Problems: Evidence from Antibiotic Prescribing*

Shan Huang† Michael Allan Ribers‡ Hannes Ullrich§

March 2021

Abstract

Large-scale data show promise to provide efficiency gains through individualized risk predictions in many business and policy settings. Yet, assessments of the degree of data-enabled efficiency improvements remain scarce. We quantify the value of the availability of a variety of data combinations for tackling the policy problem of curbing antibiotic resistance, where the reduction of inefficient antibiotic use requires improved diagnostic prediction. Focusing on antibiotic prescribing for suspected urinary tract infections in primary care in Denmark, we link individual-level administrative data with microbiological laboratory test outcomes to train a machine learning algorithm predicting bacterial test results. For various data combinations, we assess out of sample prediction quality and efficiency improvements due to prediction-based prescription policies. The largest gains in prediction quality can be achieved using simple characteristics such as patient age and gender or patients' health care data. However, additional patient background data lead to further incremental policy improvements even though gains in prediction quality are small. Our findings suggest that evaluating prediction quality against the ground truth only may not be sufficient to quantify the potential for policy improvements.

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

Keywords: prediction policy; data combination; machine learning; antibiotic prescribing

*We benefited from helpful feedback by seminar participants at DIW Berlin. Financial support from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 802450) is gratefully acknowledged.

†DIW Berlin; Department of Economics, University of Copenhagen; and BCCP. shuang@diw.de.

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

§DIW Berlin; Department of Economics, University of Copenhagen; and BCCP. hullrich@diw.de.

1 Introduction

Health care systems worldwide are undergoing fundamental changes by the broad and increasing adoption of digital tools and infrastructures. In recent years, the growing digitization and availability of detailed patient data has lead to a surge in digital health applications to improve information and efficiency in medical care provision. However, providing individualized data to policy makers can in practice have significant operative, legal, and social costs. Often data collection, storage, and ownership is highly fragmented across health care providers, insurance systems, and regional administrations. Moreover, linking personal data from different sources may raise privacy concerns. It is therefore important to quantify the potential benefits of combining different data sources in order to weigh and discuss administrative and political costs and benefits. Given the lack of linked personal data in many settings, evidence on the value of combining data is scarce.

In this paper, we quantify the value of combining data in the context of an important prediction policy problem in health care using rich linked administrative data from Denmark. In prediction policy problems, causal mechanisms are typically well established such that the core challenge is posed by a prediction task (Kleinberg, Ludwig, Mullainathan, and Obermeyer 2015; Agrawal, Gans, and Goldfarb 2018). We analyze the pressing challenge of increasing efficiency in antibiotic prescribing in primary care due to the wide-spread rise in antibiotic resistance, a major global health policy problem.[1] Specifically, we consider the antibiotic treatment of urinary tract infections at initial consultations in primary care. Prescription decisions pose a prediction problem because a conclusive diagnosis takes time, on average three days, but the medical benefits and costs of treatment are known. If a bacterial infection is diagnosed, antibiotics will cure the patient. If no bacterial cause is diagnosed, antibiotics are ineffective. The decision involves an important trade-off because untreated urinary tract infections can lead to pain and severe complications, while antibiotic treatment can cause side effects and is considered the main source of increasing antibiotic resistance of common bacteria. Machine learning predictions can provide a probabilistic but instant diagnostic. We quantify the incremental value of administrative health and non-health data for the prediction task itself and for prediction-based prescription policies.

Quantifying the value of data in the setting of urinary tract infections in Denmark has several advantages. Focusing on urinary tract infections is useful because the outcome is measured without relying on human expert labeling. Relying on gold standard laboratory test results with

[1] CDC (2013) reports that in the US alone, antibiotic-resistant infections result in an estimated 23,000 deaths, $20 billion in direct healthcare costs, and $35 billion in lost productivity each year. Primary care accounts for 90 percent of prescriptions in Europe and for 75 percent of prescriptions in Denmark (Llor and Bjerrum (2014); Danish Ministry of Health (2017)).

limited human judgment allows us to focus on the role of input data for prediction outcomes. The treatment of urinary tract infections also serves as a promising example for analyzing the value of data because prevalence of such infections are known to vary across different groups in the population. However, the nuanced associations of prevalence with a large number of observable factors has not been exploited systematically. This offers an opportunity to help tackle the problem of antibiotic resistance where policy action is needed but ways to improve efficiency are not yet fully understood. Finally, Denmark provides a useful setting for our analysis due to the availability of linked administrative data from broad sources covering the entire population. Denmark has a tax-financed single-payer health care system which is accessible to the entire population. Therefore, contrary to existing studies using data from individual health care providers or insurances, our data cover the population of patients. Such rich administrative data are found only in a small set of countries in the world, allowing us to quantify to what extent improvements are due to the scope and interconnectedness of these data. If the scope of data is crucial, improvements may be difficult in countries in which data availability is inferior due to lacking data collection and infrastructure.

We use Extreme Gradient Boosting (XGBoost) to predict whether a laboratory test indicates bacteria in a patient sample. The XGBoost algorithm relies on an ensemble of classification trees and is best suited for structured data with a large number of variables. It can efficiently recover complex interactions in the prediction of subgroup-specific probabilities of observing a positive bacterial test result. The out of sample prediction quality, measured as area under the ROC curve, ranges from 0.52 for the smallest set of predictor variables to 0.79 for the combination of all data sets. The largest gains in prediction quality are achieved by using the basic personal characteristics age and gender or data from health care registers. Combining these two and adding further detailed personal characteristics results only in small increases in prediction quality.

To analyze the value of data for potential improvements in prescribing efficiencies we define an objective which trades off the curative benefit of antibiotic treatment with the cost of increasing antibiotic resistance. Because the policy maker's costs and benefits are unknown, we specify prescription rules that guarantee improvements regardless of a policy maker's preferences. Consistent with the findings in Ribers and Ullrich (2019) policy improvements can only be achieved when physician decisions are used by the machine learning algorithm. As for prediction quality we find that basic personal characteristics and health care data provide the largest incremental improvements. However, we find that advanced demographics yield further large additional policy improvements relative to improvements in prediction quality. Using all

available data, antibiotic prescribing can be reduced by 10.22 percent without reducing the number of prescriptions to bacterial infection cases. Alternatively, prescriptions to bacterial cases can be increased by 8.44 percent without increasing the total number of prescriptions. These improvements correspond to about one fourth of the improvements which could be achieved with perfect prediction.

Our analysis illustrates an important point. Overall prediction quality improves less with increasingly detailed personal data than the quality of prescription decisions. This suggests that basing analyses only on prediction quality may be insufficient to draw policy conclusions. For these, the scope for improvements depends also on the quality of human experts' risk assessment as well as on their objective functions.

We contribute to a growing literature aiming to assess the value of linking data for prediction. While we investigate varying scope of administrative data, Zeltzer et al. (2019) analyze the incremental prediction quality gained from using electronic medical records for predicting health outcomes such as 30-day readmission and 1-year mortality after hospitalizations. They find that for short-term health outcomes administrative claims data achieve similar prediction quality as a combination of claims data with electronic medical records, concluding that claims data, which are typically easier to obtain, can be useful for such predictions. Hastings, Howison, and Inman (2020) also show that administrative data can help predict adverse outcomes such as dependence, abuse, or poisoning after initial opioid prescriptions in the US. Combining demographic data and antibiotic purchase histories from a large health care provider in Israel, Yelin et al. (2019) find that machine learning predictions of antibiotic resistance may help improve the choice of molecules while keeping the overall distribution of molecules prescribed fixed. We employ a similar strategy but focus on the extensive margin, the decision whether to prescribe an antibiotic at all at an initial consultation.

We also add to a small existing literature on the value of data which has focused on marketing problems. For the quality of recommendations in online search and retail, Schaefer, Sapi, and Lorincz (2018) and Yoganarasimhan (2019) find the amount and detail of data can confer important competitive advantage while Bajari, Chernozhukov, Hortaçsu, and Suzuki (2019) argue that improved forecasting technologies are key. Claussen, Peukert, and Sen (2019) quantify the economic returns to data for the recommendation of online news. Measuring user engagement, they find that news recommendations can outperform human editors, if the observed personal data is sufficiently large. We add the insight that evaluating prediction quality may not be enough when the aim is to improve human decision-making. In these situations, the decision outcomes generated by predictions must be compared to outcomes generated by human decisions.

The remainder of the paper is organized as follows. Section 2 presents the data and various segments used and Section 3 describes the machine learning prediction results. Section 4 describes the policy problem, prescription rules, and presents results. Section 5 concludes.

2 Data

2.1 Laboratory and administrative data

We use a combination of Danish health care and administrative data unique in the world in scope and interconnectedness. They cover a vast array of information including patients' detailed socioeconomic data as well as antibiotic prescription histories, general practice insurance claims and hospitalizations, all of which are essential for conducting our analysis. In addition, the coherent use of unique personal identifiers enables us to merge these data to individual laboratory test results that were acquired from the clinical microbiological laboratories at Herlev hospital and Hvidovre hospital for the period January 1st, 2010, to December 31st, 2012. These two major hospitals in Denmark's capital region cover a catchment area of 1.7 million people, around one third of the Danish population. 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 administrative data provided by Statistics Denmark and the Danish Health Data Authority (*Sundhedsdatastyrelsen*) cover all citizens in Denmark between January 1, 2002, and December 31, 2012. All observations can be linked across registers via unique patient (*Det Centrale Personregister, CPR*) and physician identifiers (*Yderregister, YDER*). The central personal registry contains a core set of socioeconomic and demographic variables. Further background information can be added from employment (*Integrerede Database for Arbejdsmarkedsforskning, IDA*) and education (*Uddannelseregister, UDDA*) registers. In addition, for each individual we observe the complete purchase history of systemic antibiotics (*Lægemiddeldatabasen, LMDB*), hospitalizations including ambulatory visits (*Landspatientregisteret, LPR*), and primary care insurance claims (*Sygesikringsregisteret, SSR*).

2.2 Analysis sample

Overall, 2,579,617 biological samples were submitted for testing in the capital region by both general practitioner clinics and hospitals in the years 2010, 2011, and 2012. Urine samples constitute 477,609 samples out of which 156,694 are submitted by general practitioners. We keep 152,011 observations for which general practitioner identifiers are not missing. 84,855 observa-

tions remain after excluding tests where the patient received a systemic antibiotic prescription or had another test conducted within 4 weeks of time of the test in question. 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. This also excludes potentially complicated and long-term treatment spells where patients are tested repeatedly or in a later stage having potentially received multiple antibiotic treatments. Lastly, we exclude 10,344 test observations of pregnant patients, for whom specific guidelines on diagnostic testing apply. The resulting sample we use for the analysis consists of 74,511 initial consultations, during which a sample was sent to a laboratory for testing.

2.3 Laboratory test results

The laboratory test result we aim to predict is a binary indicator, which equals one if any bacteria are isolated in a patient urine sample and zero otherwise. We observe the date of a patient sample acquisition, delivery date of the sample at the laboratory, and the date results are made available to the physician. In our sample period, the mean waiting time until a test result is received by a physician is 3.1 days and bacteria are isolated in 39 percent of the samples.

2.4 Segments of predictor variables

Patient information from the administrative registers can be linked to each laboratory test result. We subset all available data into segments of predictor variables representing different types of administrative data. The data are split according to data source but also with regard to availability, for instance the requirement of personal identifiers. Table 1 describes which variables and databases are used in the construction of each data segment.

Segment 1. Time and location contain information of the date of the test and the patient's municipality. From the test date, we construct variables that indicate weekday, week, month, quarter in the year, and whether the test took place within Danish national holidays. Information about timing and location of a test is primarily associated with the individual test sample and does not require detailed personal data from the patient or physician.

Segment 2. Basic demographics only contain age and gender of the patient. Hence basic demographics can be thought of as information that could easily be submitted joint with test samples and therefore do not require linkage through personal identifiers.

Table 1 Predictor variables, data segments, and required databases

	Databases[a]							
Data segments and variables	LAB	CPR	IDA	UDDA	YDER	LMDB	SSR	LPR
1 Time and location								
Time[b]	✓							
Municipality		✓						
2 Basic demographics								
Age		✓						
Sex		✓						
3 Advanced demographics								
Civil status		✓						
Family type		✓						
Living constellation		✓						
Migration background		✓						
Origin country		✓						
Employment status			✓					
Occupation			✓					
Industry			✓					
Income			✓					
Highest educational degree				✓				
4 Health								
Antibiotic prescriptions[c]						✓		
Antibiotic resistance tests[d]	✓							
Hospitalizations[e]							✓	
Claims[f]								✓
Clinic identifier					✓			
Clinic-level testing[g]	✓				✓			
Municipality-level prescribing[h]		✓				✓		
5 Physician decisions								
Instantaneous prescription					✓	✓		
Molecule					✓	✓		
Volume					✓	✓		

[a] Patient identifiers are always contained in order to link different databases. *LAB* indicates laboratory test data, *CPR* indicates population data (Det Centrale Personregister), *IDA* indicates employment data (Integrerede Database for Arbejdsmarkedsforskning), *UDDA* indicates education data (Uddannelseregister), *YDER* indicates physician identifiers (Yderregister), *LMDB* indicates prescriptions data (Lægemiddeldatabasen), *SSR* indicates claims data (Sygesikringsregisteret), and *LPR* indicates hospitalization data (Landspatientregisteret).

[b] *Time* includes Weekday, Week, Month, Quarter, and School holidays.

[c] *Antibiotic prescriptions* includes Molecule, Volume, Indication, and Number of days until current test, for the patient's past 30 prescriptions.

[d] *Antibiotic resistance tests* include Species sampled, Resistances found, and Number of days until current test, for the patient's past 5 tests.

[e] *Hospitalizations* include Length of stay, Diagnoses, Patient type, and Number of days until current test, for the patient's past 30 hospitalizations.

[f] *Claims* include Procedure code and Number of weeks until current test, separated by general practitioners and specialists, for the patient's past 30 claims.

[g] *Clinic-level testing* includes Number of antibiotic resistance tests and Bacterial test rate, separated over all previous, past 1, past 3, past 6, and past 12 months, at the general practitioner clinic.

[h] *Municipality-level prescribing* is measured by Daily Define Dose per inhabitant over past three months.

Segment 3. Advanced demographics contain detailed information obtained from the administrative registries. Specifically, it includes household information, employment and income, education, and migration background. The complexity of information implies that this data is

too burdensome to collect at each tested instance and personal patient identifiers are required in order to link to registries that contain the information.

Segment 4. Health contains extensive information from the Danish health registries as well as laboratory test records. As information on time and location of the test (Segment 1) can be expected to also be included in health data, we will only include this segment of the data with basic information on time and location of the test. Hence, in addition to patient-specific information on prescriptions, resistance test results, hospitalizations, and claims, the health segment allows for the computation of municipality and clinic level averages such as the clinic number of test and bacterial test rate, as well as municipality-level antibiotic use. The addition of the health segment as predictors in the XGBoost algorithm requires personal patient identifiers. In many countries, linking health data to other data sources is viewed particularly challenging and costly.

Segment 5. Physician decisions provide information on the physicians' treatment choice on the day of a consultation. This includes whether an antibiotic was prescribed and the specific molecule and volume. Physician decisions at the test instance serve as a noisy measure of the total information that the physician has which includes unobserved diagnostic information revealed during patient assessment such as point-of-care tests. The set of variables constructed from physician decisions relies on detailed health records for each test instance in order to reconstruct physicians' situational behavior. In many countries, such records for individual encounters between a patient and the health care system are unavailable for technological or privacy reasons.

3 Machine learning

3.1 Prediction algorithm

We employ Extreme Gradient Boosting (XGBoost) to construct the risk predictions used in our main analysis. XGBoost is an ensemble method using a sequential collection of classification trees to form a prediction model (Chen and Guestrin 2016). The advantage of this off-the-shelf machine learning algorithm is that it can accommodate a very large number of predictors while remaining computationally light. Because XGBoost uses classification trees, it allows for flexible interactions and complex non-linear relationships between the predictors. Moreover, it allows for penalization of model complexity to avoid overfitting. We search for optimal hyperparameters based on the mean area under the receiver operator curve (AUC) across three separate tuning

evaluation partitions. We separately apply the tuning procedure and determine the learning rate, the number of boosting rounds (i.e. number of trees), and the maximum depth of each tree for each combination of the data segments.

The three monthly tuning evaluation partitions consist of October, November and December 2010, where we use data from January 2010 up to the respective evaluation partition as training data. Our tuning procedure differs from regular k-fold cross validation Hastie, Tibshirani, and Friedman (2009) as the time dependency in our data prohibits random splits. Particularly, a random split of the data could result in constructing predictions for a patient based on training test results observed for the same patient but *after* the test we aim to predict.

Given the respective hyperparameters, we train the prediction model separately for each combination of data segments. We create 24 monthly out-of-sample evaluation partitions from January 2011 to December 2012 and use all observations from up to twelve months prior until the respective test observations as training data. That is, to evaluate predictions for July 2011 for instance, we use data from July 2010 to June 2011 as training data. Sample sizes and the bacterial rate of each partition are shown in Table 4 of Appendix A.

3.2 Out of sample prediction quality

We evaluate prediction quality out of sample for the different data segments using several statistics. Table 2 reports AUC values for the combinations of data segments we consider. Ribers and Ullrich (2019) document that the physician prescription decisions alone contain important predictive information. Including only physician decisions as a predictor results in an AUC of 0.69 which is similar to what can be achieved using the full segment of health care data but excluding physician decisions. Using only the basic demographic information about patient age and gender yields an AUC of 0.65, whereas adding rich personal background information in the advanced demographics segment increases the AUC to 0.67. All data segments combined result in an AUC of 0.73, which is slightly lower than what can be achieved by using both physician decisions and basic demographics, an AUC of 0.75. Figure 2 in Appendix B provides a graphical presentation of the AUC values across segments.

In Table 2, we also report the relative gains over a baseline data segment. For each data combination we subtract the baseline AUC value and divide by the difference between the AUC attained from the combination of all data segments and the baseline AUC. In panel A, the baseline data segment is time and location. In panel B, the baseline data segment is the physician decision. For example, in panel B the gain for the combination of segments time and location, basic demographics, health, and physician decisions is calculated as: $(0,768 - 0,686)/(0,786 -$

$0,686) = 0,82$. The largest incremental gains are achieved from adding Basic demographics (64%-4%=60%) and Health data (82%-4%=78%) to Physician decisions and Time and location. Only small incremental gains of 2% can be made by adding advanced demographics.

Table 2 Predictive performance: area under the receiver operating curve

Predictive performance[a]		**Data segments included**[b]				
AUC	Gain	Time and location	Basic demographics	Advanced demographics	Health	Physician decisions
A: Models excluding physician decisions						
0.517 [0.511, 0.522]		✓				
0.652 [0.647, 0.658]	64.59%	✓	✓			
0.668 [0.663, 0.673]	72.25%	✓	✓	✓		
0.689 [0.684, 0.694]	82.30%	✓			✓	
0.719 [0.714, 0.724]	96.65%	✓	✓		✓	
0.726 [0.721, 0.731]	100.00%	✓	✓	✓	✓	
B: Models including physician decisions						
0.686 [0.680, 0.692]						✓
0.690 [0.684, 0.695]	4.00%	✓				✓
0.750 [0.746, 0.755]	64.00%	✓	✓			✓
0.756 [0.751, 0.761]	70.00%	✓	✓	✓		✓
0.768 [0.763, 0.773]	82.00%	✓			✓	✓
0.784 [0.780, 0.789]	98.00%	✓	✓		✓	✓
0.786 [0.782, 0.791]	100.00%	✓	✓	✓	✓	✓

[a] **Predictive performance** shows the prediction quality for each prediction model. *AUC* shows the out-of-sample area under the receiver operating curve. 95% confidence intervals in parentheses are computed based on 1000 bootstrap samples. *Gain* shows the cumulative improvements in AUC, relative to the overall increase within a panel. That is, it shows the share of the difference in AUC between a given specification and the first specification (*Time & place* in Panel A; *Physician decisions* in Panel B), relative to the difference in AUC between the last and the first specification in the panel ($0.726 - 0.517$ in Panel A; $0.786 - 0.686$ in Panel B).

[b] **Data segments included** shows the sets of predictor variables included in each specification. We consider as data segments Time and location (*Time & place*), Basic demographics (*Basic demo.*), Advanced demographics (*Adv. demo.*), Health, and Physician decisions.

We also measure prediction quality in terms of further statistics: accuracy, positive predictive value (PPV), negative predictive value (NPV), sensitivity, and specificity. Figures 3 and 4 in Appendix B show that improvements from additional data appear mainly for low levels of either accuracy, PPV, sensitivity, or specificity. Overall, similar to the results for AUC,

the improvement in performance is relatively small after basic demographic characteristics are included. Our prediction models are not trained to target one single measure reported here but rather overall prediction quality, which is best measured by the AUC.

4 Prediction policy outcomes

While the evaluation of the gains of data combination for out-of-sample prediction quality is straightforward, it is not sufficient for learning about the potential improvements they offer compared to human decisions. The problem of antibiotic prescribing for suspected urinary tract infections provides a useful setting to assess machine learning prediction-based decision rules against expert decisions. Due to the pressing global health problem of increasing antibiotic resistance, it is important to identify policies improving the efficiency of antibiotic use. In the same primary care context in Denmark, Ribers and Ullrich (2019) show that machine learning predictions, combined with physician expertise, can lead to significant improvements in antibiotic prescribing for urinary tract infections. In this Section, we first define an objective function reflecting the trade-off between the costs and benefits of antibiotic use. Based on this objective function and prediction results for varying degrees of data combinations, we then compare counterfactual prescription rules to observed physician decisions.

4.1 Policy objective

To assess the potential for improvements in antibiotic prescribing, we define the policy maker's objective function such that a prescription decision, d, resolves the trade-off between the patient potentially suffering a sickness cost, α, from delaying prescribing until a test result is available, and the social cost of prescribing, β, that is, promoting antibiotic resistance via antibiotic use. While the social cost of prescribing is incurred for every antibiotic prescribed, the cost of waiting is only incurred by patients suffering from a bacterial infection. Likewise, antibiotic treatment is only curative, i.e. alleviating the sickness cost, if a patient suffers from a bacterial infection. Hence, we model the payoff at a patient's initial consultation as

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

where y is an indicator for whether the patient has a bacterial infection, i.e. $y = 1$ if the test outcome is positive and zero otherwise. We assume that $0 < \beta < \alpha$ such that prescribing is always optimal when an infection is known to be bacterial with certainty.[2]

[2]Our application does not strictly require $\beta < \alpha$, only that $0 < \alpha$ and $0 < \beta$. However, if $\alpha < \beta$, it is never optimal to prescribe prior to observing test results, making the policy rules redundant. In our data we observe

4.2 Prescription rules based on machine learning predicted risk

We define a prescription rule δ as a set of decisions d for all consultations in our data that either delay prescribing until test results are available or assign an antibiotic treatment prior to receiving test results. The change in payoffs induced by δ compared to the observed set of physician decisions δ^J can the be written as

$$\begin{aligned} \Pi &= \mathrm{E}\big[\pi(\delta, y) - \pi(\delta^J, y)\big] \\ &= \alpha \mathrm{E}\big[y\big(\delta - \delta^J\big)\big] - \beta \mathrm{E}\big[\delta - \delta^J\big]. \end{aligned} \tag{2}$$

The first term in Equation (2) is the change in prescriptions to bacterial urinary tract infections weighted by α. The second term is the change in the total number of prescriptions weighted by β. We cannot determine the prescription rule that achieves the maximum Π without further knowledge about the payoff weights α and β. However, we can determine those prescription rules that guarantee equation (2) to be positive for any set of payoff weights. All prescription rules that increase prescriptions to bacterial infections without increasing the total number of prescriptions, and vice versa, result in payoff improvements. We evaluate Π for the complete set of rules given as a function of machine learning predicted risk $m(x)$, where x is a vector of predictors:

$$\delta(m(x); k) \;=\; \begin{cases} 0 & \text{if } m(x) < k, \\ 1 & \text{if } k \leq m(x), \end{cases} \tag{3}$$

by varying the risk threshold $k \in [0, 1]$ above which an antibiotic is prescribed.

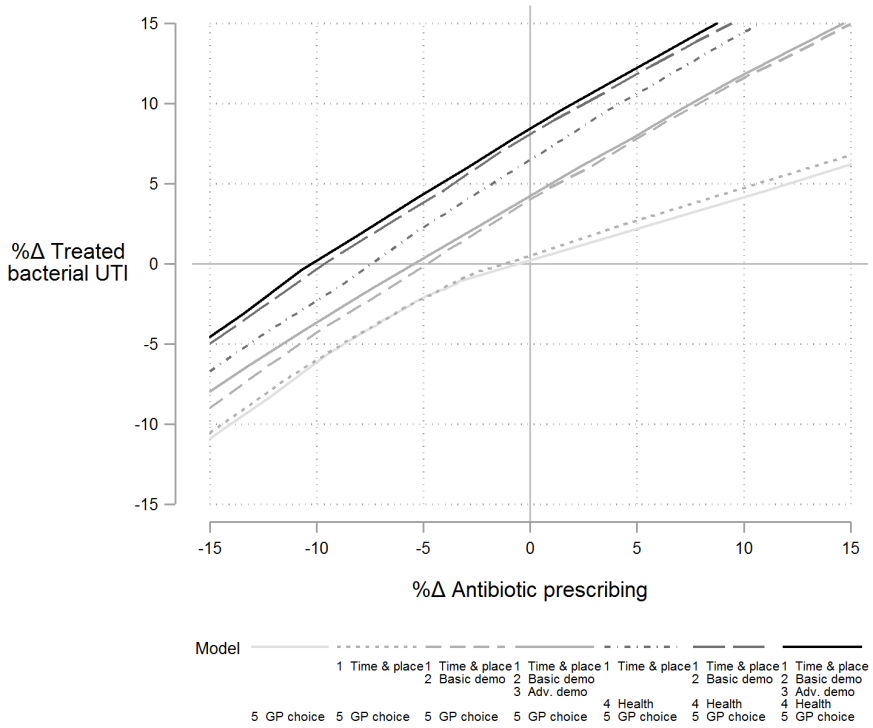
4.3 Policy results

For all machine learning prediction-based prescription rules, we report outcomes as the percentage change in the total number of antibiotic prescriptions, $\big(\delta - \delta^J\big)/\delta^J$, and the percentage change in prescriptions given to patients with bacterial infections, $\mathrm{E}\big[y\big(\delta - \delta^J\big)\big]/\mathrm{E}\big[y\delta^J\big]$. Figure 1 shows the achievable policy outcomes for the relevant range of threshold values k where each line represents different data combinations. Using physicians' initial prescription decision as the only predictor in the machine learning algorithm, the set of achievable outcomes lies just slightly above the origin. Adding information on the time and region a clinic is located in does not lead to a noticeable expansion of the set of achievable improvements. Providing the machine learning algorithm with the basic demographics age and gender significantly shifts the boundary upwards. Adding further more detailed personal background information does not appear to

significant prescribing prior to observing test results.

yield further improvements. Combining health care administrative data with physician decisions results in the largest improvements that can be achieved by a single additional data segment. Given these data include past laboratory test results, medical prescription histories, as well as in- and outpatient claims data, we would expect these to provide the most valuable information related to test outcomes. However, departing from the results on prediction quality, combining these data with basic demographics yields further important improvements. Further personal background information expands the set of achievable to a small extent.

Figure 1: Prediction-based policy outcomes



Notes: The lines show along the range of prediction threshold values the possibility frontier of achievable policy outcomes. The relevant policy outcomes are, as percentage change relative to the status quo, correctly treated bacterial urinary tract infections (*%Δ Treated bacterial UTI*) and antibiotic prescriptions (*%Δ Antibiotic prescribing*). Each line corresponds to a model trained on a different combination of data segments. As data segments, we consider Time and location (*Time & place*), Basic demographics (*Basic demo*), Advanced demographics (*Adv demo*), Health, and Physician decisions (*GP choice*).

Consistent with the results in Ribers and Ullrich (2019), our findings in Figure 5 in Appendix C shows that no positive payoff improvements can be achieved without using diagnostic information encoded in physician decisions. All achievable outcomes, even using the combination of all data segments, are located outside of the area in which the change in prescribing is negative and the change in treated bacterial urinary tract infections is positive. Therefore, payoff improvements can only be achieved for a subset of payoff weights α and β that gives rise to preferences that find decreasing antibiotic preferable even at the cost of decreasing prescriptions

given to patients with bacterial infections or, vice versa, finds increasing the number of treated preferable even at the expense of increased antibiotic use.

While Figure 1 summarizes all relevant policy improvements, it is instructive to analyze potential policy outcomes for two special prescription rules. The first prescription rule sets k such that the number of treated bacterial infections is unchanged while using fewer prescriptions relative to observed antibiotic use. The second sets k such that the total number of antibiotics used is unchanged while treating more patients with bacterial infections relative to observed outcomes. In Figure 1 the policy outcomes for these two special cases are represented by the intersections of the possibility frontiers with the horizontal and vertical axes. This follows the existing literature evaluating the potential of machine learning predictions where an objective function is maximized under constraints to guarantee increases in payoff, for example in Bayati et al. (2014), Chalfin et al. (2016), Kleinberg et al. (2018), Ribers and Ullrich (2019), Yelin et al. (2019), and Hastings et al. (2020).

Table 3 shows the percentage changes in antibiotic prescribing and correctly treated bacterial infections for these two special prescription rules. Using physicians' initial prescription decision as the only predictor in the machine learning algorithm, prescriptions can be reduced by 0.56 percent without prescribing to fewer bacterial infections. And without increasing the number of prescriptions, 0.22 percent more bacterial infections can receive a prescription. As observed in Figure 1, the largest gains are achieved by adding basic demographic data and health care data. Using age and gender information, total prescriptions can be decreased by 5.18 percent without correctly treating fewer patients. Alternatively, 4.04 percent more cases with bacterial infections can receive an antibiotic without changing the total number of prescriptions. Combining just health care administrative data with physician choice results in a decrease in total prescribing by 7.42 percent, or alternatively a 6.48 percent increase in correct prescriptions. Combining these data with basic and detailed demographic information can reduce prescriptions by 10.22 percent, or increases correct prescriptions by 8.44 percent. We can also compare this result to the maximum achievable reductions in antibiotic prescribing. Table 5 of Appendix C shows that the maximum achievable reduction is 39 percent if diagnostic prediction was perfect. With the full data, one fourth of this maximum can be achieved. Table 6 in Appendix C shows that without the physician decision as a predictor, even with the full set of administrative background data on patients, the total number of prescriptions cannot be decreased without reducing the number of correctly treated cases. Likewise, more patients with bacterial infections cannot be treated without also increasing the total number of prescriptions.

These results provide an important general insight. While prediction quality increases in

Table 3 Policy outcomes for two special cases of prescription rules, incl. physician decisions

Treated UTI constant[a]			**Antibiotic use constant**[b]			**Data segments included**[c]				
Change in antibiotic use	Gain	Risk threshold	Change in treated UTI	Gain	Risk threshold	Time & place	Basic demo	Adv. demo	Health	GP choice
−0.56%		0.28	0.22%		0.28					✓
[−0.99, −0.12]			[0.04, 0.40]							
−1.18%	6.46%	0.31	0.50%	3.41%	0.30	✓				✓
[−1.65, −0.71]			[0.26, 0.73]							
−4.83%	44.21%	0.50	3.99%	45.86%	0.48	✓	✓			✓
[−5.55, −4.11]			[3.35, 4.64]							
−5.42%	50.36%	0.49	4.22%	48.66%	0.47	✓	✓	✓		✓
[−6.23, −4.62]			[3.55, 4.90]							
−7.42%	70.99%	0.52	6.48%	76.16%	0.49	✓			✓	✓
[−8.28, −6.56]			[5.73, 7.23]							
−9.64%	94.02%	0.52	8.08%	95.62%	0.48	✓	✓		✓	✓
[−10.53, −8.76]			[7.31, 8.85]							
−10.22%	100.00%	0.52	8.44%	100.00%	0.48	✓	✓	✓	✓	✓
[−11.15, −9.30]			[7.61, 9.28]							

[a] **Treated UTI constant** shows policy outcomes for a prescription rule that does not decrease the number of treated bacterial urinary tract infections, using a given prediction model. *Antibiotic use change* shows the maximum achievable percentage decrease in antibiotic use under this rule, relative to the status quo. 95% confidence intervals in parentheses are computed based on 1000 bootstrap samples. *Gain* shows the cumulative improvements in policy outcomes, relative to the overall improvements within a panel. That is, it shows the share of the difference in policy outcomes between a given specification and the first specification (only *GP choice* as predictors), relative to the difference in policy outcomes between the last specification (all data segments as predictors) and the first. *Risk threshold* shows which threshold value k yields a prediction-based prescription rule that does not decrease the number of treated bacterial urinary tract infections.

[b] **Antibiotic use constant** shows policy outcomes for a prescription rule that does not increase the number of antibiotic prescriptions, using a given prediction model. *Treated UTI change* shows the maximum achievable percentage increase in treated urinary tract infections under this rule. 95% confidence intervals in parentheses are computed based on 100 bootstrap samples. The sub-columns *Gain* and *Risk threshold* are analogous to the above.

[c] **Data segments included** shows the sets of predictor variables included in each specification. We consider as data segments Time and location (*Time & place*), Basic demographics (*Basic demo*), Advanced demographics (*Adv. demo*), Health, and Physician decisions (*GP choice*).

small steps after adding patients' age and gender information, when benchmarked against human decisions adding more detailed data leads to sizeable further improvements. Therefore, evaluating policy potential based only on prediction quality relative to the ground truth is not sufficient. Instead, it is important to analyse the scope for improvements relative to observed human decisions, which depend on the quality of human experts' risk assessment and on their objective functions.

5 Conclusion

We provide evidence that combining administrative socioeconomic and health care data may contribute to solve an important prediction policy problem, the efficient prescription of antibiotics to curb antibiotic resistance. Health care data and basic personal information lead to the largest incremental increase in prediction quality relative to the baseline using only infor-

mation contained in physician decisions. To achieve an efficiency-improving prescription rule, using detailed personal background information in addition to health care data leads to a larger added value compared to the gains in prediction quality. It is therefore important to evaluate predictions using human decisions as the benchmark.

Our analysis has some limitations. We focus on a specific medical treatment situation, which may be more appropriate for the use of data-based predictions than others. While tackling the problem of antibiotic resistance is an important policy problem, the treatment situation we consider here resembles many other medical diagnostic settings such as biopsies for malignant tumors, testing for tuberculosis, or testing for SARS-CoV-2 where results typically arrive with significant delays, leaving physicians and patients to make decisions under diagnostic uncertainty. We consider hypothetical policy scenarios using observational data so that the policy results rely on assumptions with relevant implications for real world implementations. For example, for ethical and legal reasons one may not want to leave the final prescription decision to a strict prediction-based rule. The rule could instead provide recommendations to physicians who would then make the final decision. However, Ribers and Ullrich (2020) show that improvements due to machine learning predictions arise not only due to better information but also due to changes in physicians' payoff functions. Providing recommendations is easy and rather uncontroversial but nudging or incentivizing physicians into changing their decisions may be less so.

Finally, the use of administrative and other transaction-based data may perpetuate biases and systematic errors. One advantage in our analysis is that the outcome is directly observable without relying on noisy expert judgement. For example, radiological image classification algorithms are trained on expert judgements, which may themselves contain bias or error. Further, Pierson et al. (2021) show that physicians' knee pain diagnostics based on radiographic measures of severity, developed historically in white British populations, lead to biased treatment decisions that underserved black patients compared to when subjective pain scores were used as outcome measure. In estimating the policy potential of new applications, the target outcome for which machine learning predictions are intended to be used must be carefully assessed to avoid or even reduce biases and errors.

Appendices

Appendix A Data for prediction algorithm

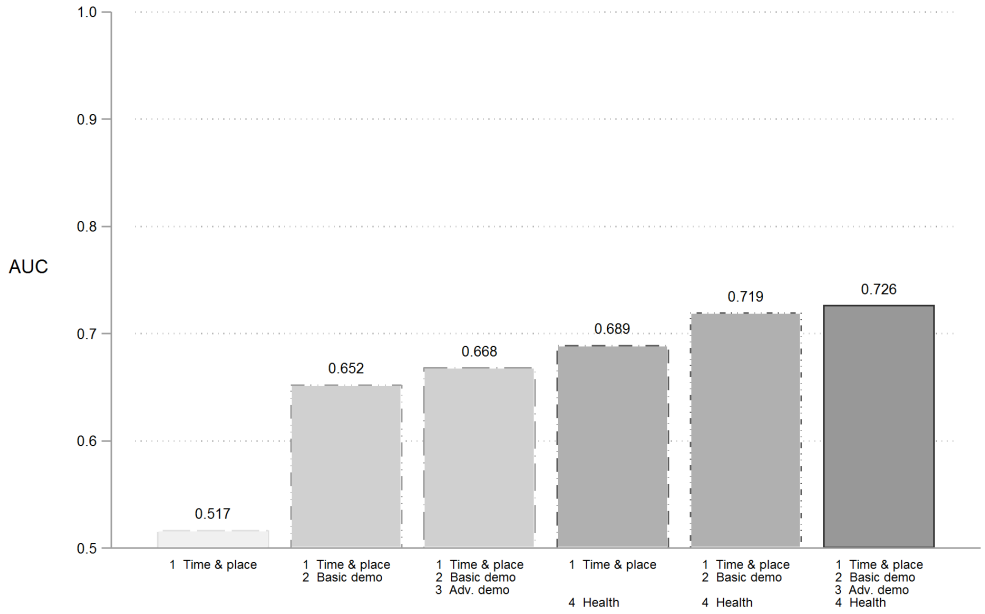
Table 4 Sample sizes used for training and prediction

	Prediction data		**Training data**	
Evaluation month	Observations	Bacterial rate	Observations	Bacterial rate
Jan 11	1,609	0.36	19,896	0.37
Feb 11	1,372	0.36	20,235	0.37
Mar 11	1,691	0.38	20,498	0.37
Apr 11	1,323	0.41	20,799	0.37
May 11	1,691	0.40	20,950	0.38
Jun 11	1,596	0.41	21,460	0.38
Jul 11	1,130	0.42	21,827	0.39
Aug 11	1,762	0.40	22,047	0.39
Sep 11	1,942	0.39	22,291	0.39
Oct 11	1,828	0.39	22,617	0.39
Nov 11	1,939	0.40	23,088	0.39
Dec 11	1,495	0.42	23,567	0.40
Jan 12	2,033	0.40	23,974	0.40
Feb 12	1,722	0.38	24,533	0.40
Mar 12	1,938	0.36	24,986	0.40
Apr 12	1,492	0.40	25,376	0.40
May 12	1,823	0.36	25,672	0.40
Jun 12	2,015	0.39	25,928	0.40
Jul 12	1,426	0.43	26,663	0.40
Aug 12	2,299	0.39	27,095	0.40
Sep 12	2,081	0.40	28,004	0.40
Oct 12	2,302	0.39	28,476	0.40
Nov 12	2,415	0.38	29,328	0.40
Dec 12	1,556	0.39	30,295	0.40

Appendix B Prediction quality

Figure 2: Predictive performance: area under the receiver operating curve (AUC)

(a) Prediction models excluding physician decisions



(b) Prediction models including physician decisions

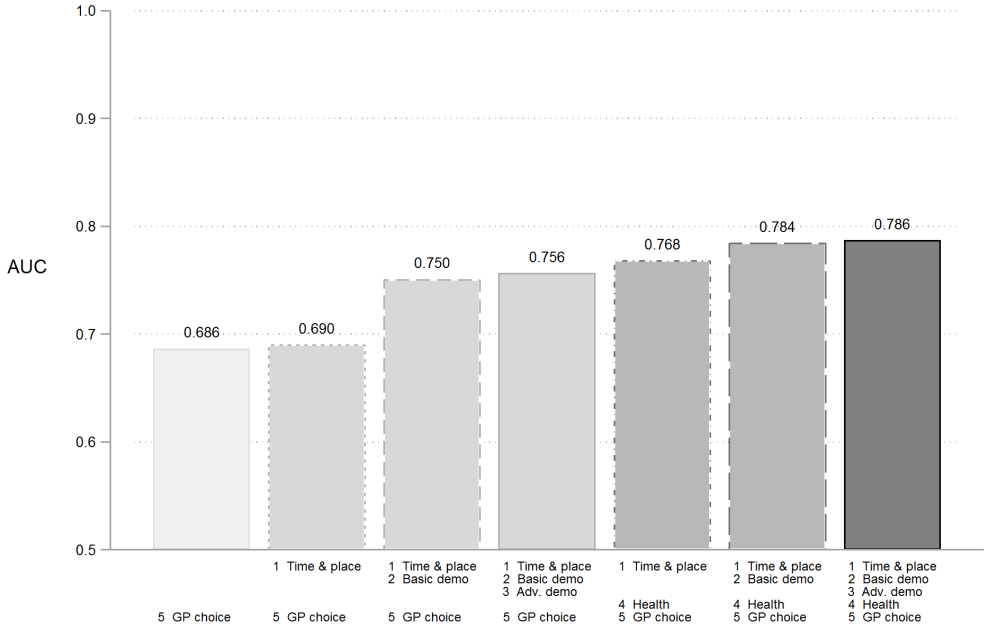
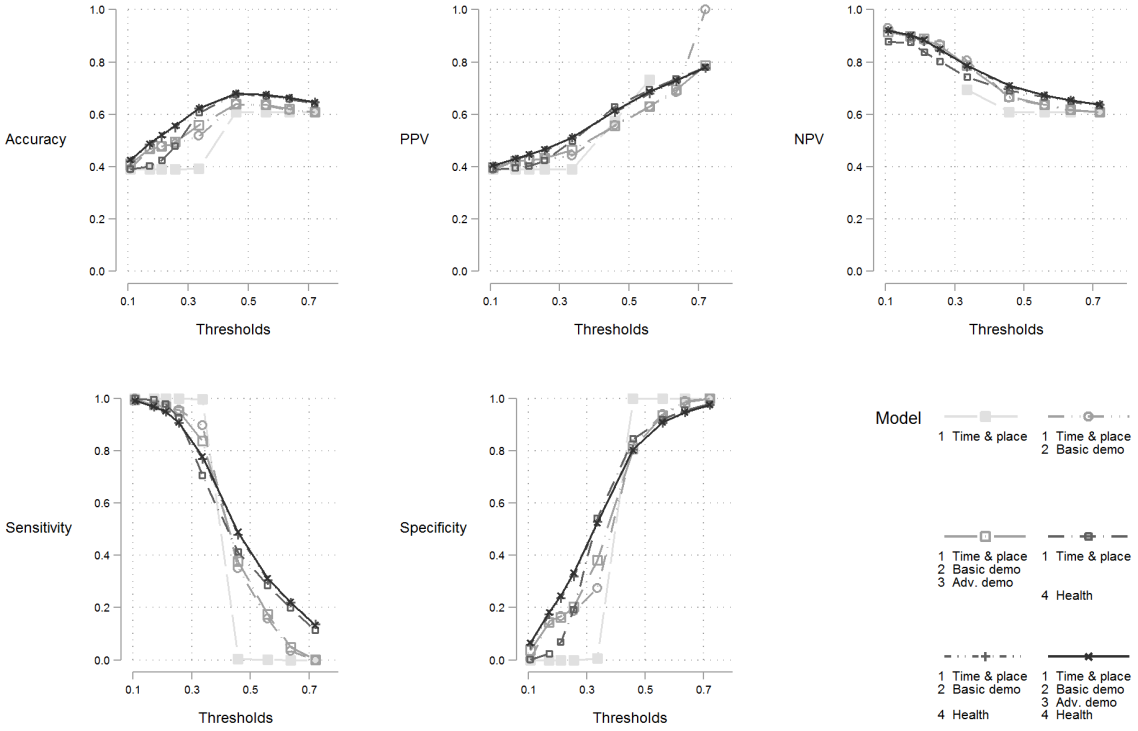
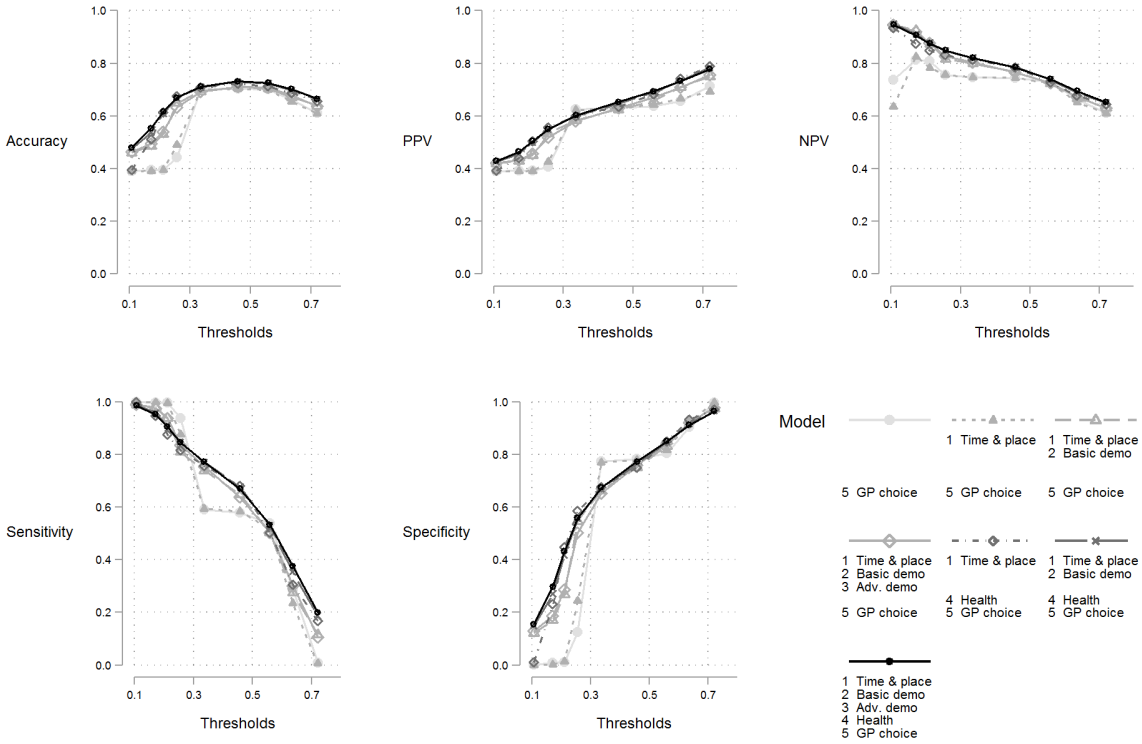


Figure 3: Predictive performance measured by accuracy, sensitivity, specificity, negative predictive value (NPV), and positive predictive value (PPV)



Notes: All measures of predictive performance are computed for different thresholds of predicted probability. The thresholds are selected as deciles of the empirical distribution of the predicted probability in the model including all data segments. This procedure follows Zeltzer et al. (2019). As data segments, we consider Time and location (*Time & place*), Basic demographics (*Basic demo*), Advanced demographics (*Adv demo*), Health, and Physician decisions (*GP choice*).

Figure 4: Predictive performance: accuracy, sensitivity, specificity, negative predictive value (NPV), and positive predictive value (PPV)



Notes: All measures of predictive performance are computed for different thresholds of predicted probability. The thresholds are selected as deciles of the empirical distribution of the predicted probability in the model including all data segments. This procedure follows Zeltzer et al. (2019). As data segments, we consider Time and location (*Time & place*), Basic demographics (*Basic demo*), Advanced demographics (*Adv demo*), Health, and Physician decisions (*GP choice*).

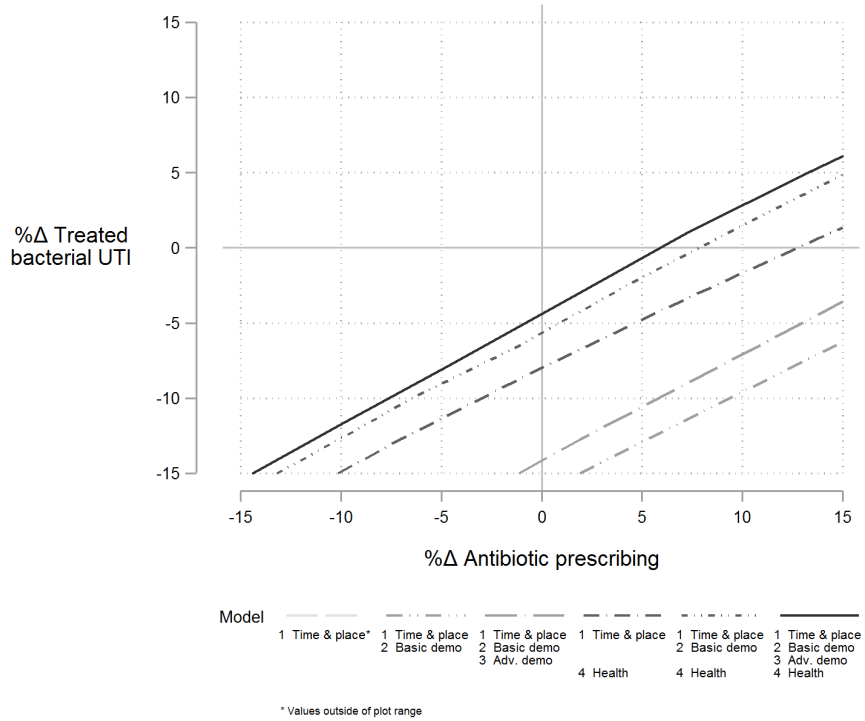
Appendix C Policy outcomes

Table 5 Confusion matrix of physicians' antibiotic prescriptions for all prediction data

	Instantaneous prescription		
Bacterial UTI	No	Yes	*Total*
No	19,510	6,377	*25,887*
Yes	6,636	9,957	*16,593*
Total	*26,146*	*16,334*	**42,480**

The share of antibiotic over-prescriptions is 39.04% given by the share of instantaneous prescriptions with no bacterial urinary tract infection (6,377) among all instantaneous prescriptions (16,334). The share of under-treated urinary tract infections is 39.99% given by the share of bacterial urinary tract infection with no instantaneous prescription (6,636) among all bacterial infections (16,593).

Figure 5: Prediction-based policy outcomes, no physician decision



Notes: The lines show along the range of prediction threshold values the possibility frontier of achievable policy outcomes. The relevant policy outcomes are, as percentage change relative to status quo, correctly treated bacterial urinary tract infections (*%Δ Treated bacterial UTI*) and antibiotic prescriptions (*%Δ Antibiotic prescribing*). Each line corresponds to a model trained on different combination of data segments. As data segments, we consider Time and location (*Time & place*), Basic demographics (*Basic demo*), Advanced demographics (*Adv demo*), Health, and Physician decisions (*GP choice*).

Table 6 Policy outcomes for two special cases of prescription rules, specifications excluding physician decisions

Treated UTI constant[a]			**Antibiotic use constant**[b]			**Data segments included**[c]				
Change in antibiotic use	Gain	Risk threshold	Change in treated UTI	Gain	Risk threshold	Time & place	Basic demo	Adv. demo	Health	GP choice
54.00% [51.65, 56.34]		0.39	−33.73% [−34.84, −32.61]		0.40	✓				
24.45% [22.32, 26.58]	61.45%	0.39	−16.29% [−17.57, −15.01]	59.44%	0.40	✓	✓			
20.00% [17.85, 22.15]	70.70%	0.41	−14.16% [−15.54, −12.79]	66.70%	0.43	✓	✓	✓		
12.77% [15.12, 10.42]	85.74%	0.37	−7.99% [−9.33, −6.65]	87.73%	0.39	✓			✓	
7.87% [5.97, 9.76]	95.94%	0.40	−5.67% [−6.93, −4.41]	95.64%	0.42	✓	✓		✓	
5.91% [4.15, 7.68]	100.00%	0.41	−4.39% [−5.63, −3.14]	100.00%	0.42	✓	✓	✓	✓	

[a] **Treated UTI constant** shows policy outcomes for a prescription rule that does not decrease the number of treated bacterial urinary tract infections, using a given prediction model. *Change in antibiotic use* shows the maximum achievable percentage decrease in antibiotic use under this rule, relative to the status quo. 95% confidence intervals in parentheses are computed based on 1000 bootstrap samples. *Gain* shows the cumulative improvements in policy outcomes, relative to the overall improvements within a panel. That is, it shows the share of the difference in policy outcomes between a given specification and the first specification (only *Time & place* as predictors), relative to the difference in policy outcomes between the last specification (all data segments except for *GP choice* as predictors) and the first. *Risk threshold* shows which threshold value k yields a prediction-based prescription rule that does not decrease the number of treated bacterial urinary tract infections.

[b] **Antibiotic use constant** shows policy outcomes for a prescription rule that does not increase the number of antibiotic prescriptions, using a given prediction model. *Change in treated UTI* shows the maximum achievable percentage increase in treated urinary tract infections under this rule. 95% confidence intervals in parentheses are computed based on 100 bootstrap samples. The sub-columns *Gain* and *Risk threshold* are analogous to the above.

[c] **Data segments included** shows the sets of predictor variables included in each specification. We consider as data segments Time and location (*Time & place*), Basic demographics (*Basic demo*), Advanced demographics (*Adv. demo*), Health, and Physician decisions (*GP choice*).

References

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

P. Bajari, V. Chernozhukov, A. Hortaçsu, and J. Suzuki. The impact of big data on firm performance: An empirical investigation. *AEA Papers and Proceedings*, 109:33–37, 2019.

M. Bayati, M. Braverman, M. Gillam, K. M. Mack, G. Ruiz, M. S. Smith, and E. Horvitz. Data-Driven Decisions for Reducing Readmissions for Heart Failure: General Methodology and Case Study. *PLoS ONE*, 9(10):e109264, Oct. 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0109264.

CDC. Antibiotic resistance threats in the United States, 2013.

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

T. Chen and C. Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785.

J. Claussen, C. Peukert, and A. Sen. The Editor and the Algorithm: Targeting, Data and Externalities in Online News. CESifo Working Paper No. 8012, 2019.

Danish Ministry of Health. National handlingsplan for antibiotika til mennesker. Tre m\aa lbare m\aa l for en reduktion af antibiotikaforbruget frem mod 2020., 2017.

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

J. S. Hastings, M. Howison, and S. E. Inman. Predicting high-risk opioid prescriptions before they are given. *Proceedings of the National Academy of Sciences*, 117(4):1917–1923, Jan. 2020. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1905355117.

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

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

C. Llor and L. Bjerrum. Antimicrobial resistance: Risk associated with antibiotic overuse and initiatives to reduce the problem. *Therapeutic Advances in Drug Safety*, 5(6):229–241, Dec. 2014. ISSN 2042-0986, 2042-0994. doi: 10.1177/2042098614554919.

E. Pierson, D. M. Cutler, J. Leskovec, S. Mullainathan, and Z. Obermeyer. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nature Medicine*, 27(1):136–140, Jan. 2021. ISSN 1546-170X. doi: 10.1038/s41591-020-01192-7.

M. A. Ribers and H. Ullrich. Battling antibiotic resistance: Can machine learning improve prescribing? DIW Discussion Paper No. 1803, 2019.

M. A. Ribers and H. Ullrich. Machine Predictions and Human Decisions with Variation in Payoffs and Skills. DIW Discussion Paper No. 1911, 2020.

M. Schaefer, G. Sapi, and S. Lorincz. The effect of big data on recommendation quality: The example of internet search. DIW Discussion Paper No. 1730, DIW, 2018.

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

H. Yoganarasimhan. Search Personalization Using Machine Learning. *Management Science*, 66 (3):1045–1070, Aug. 2019. ISSN 0025-1909. doi: 10.1287/mnsc.2018.3255.

D. Zeltzer, R. D. Balicer, T. Shir, N. Flaks-Manov, L. Einav, and E. Shadmi. Prediction accuracy with electronic medical records versus administrative claims. *Medical Care*, 57(7): 551–559, 2019.