Personalized Antibiograms: Machine Learning for Precision Selection of Empiric Antibiotics

Conor K. Corbin [1], Richard J. Medford, MD [2], Kojo Osei [1], Jonathan H. Chen, MD, PhD[1]
[1]Stanford University, Stanford, California; [2]University of Texas Southwestern, Dallas, Texas

Abstract

Up to 50% of antibiotic use in hospital settings is suboptimal. We build machine learning models trained on electronic health record data to minimize wasteful use of antibiotics. Our classifiers flag no growth blood and urine microbial cultures with high precision. Further, we build models that predict the likelihood of bacterial susceptibility to sets of antibiotics. These models contain decision thresholds that separate subgroups of patients whose susceptibility rates to narrow-spectrum antibiotics equal overall susceptibility rates to broader-spectrum drugs. Retroactively analyzing these thresholds on our one year test set, we find that 14% of patients infected with Escherichia coli and empirically treated with piperacillin/tazobactam could have been treated with ceftriaxone with coverage equal to the overall susceptibility rate of piperacillin/tazobactam. Similarly, 13% of the same cohort could have been treated with cefazolin - a first generation cephalosporin.

Introduction

Over 700,000 people die a year from antibiotic resistant infection - a figure that is rapidly growing[1]. If nothing alters its trajectory, the annual death rate by 2050 will exceed 10,000,000[1]. Lack of working antibiotics would push many modern day medical practices into extinction. Antibiotics are used prophylactically before surgery to prevent surgical site infection, and in conjunction with chemotherapy and HIV treatment when a patient's immune system is compromised. Antibiotics used improperly needlessly expedite the rate at which microbes develop resistance. Up to 50% of antibiotics prescribed in hospitals are either inappropriate or suboptimal[2]. Clinicians use their expertise to prescribe antibiotics of the proper type, duration, and route of intervention; but, prescribing the antibiotic that maximizes the likelihood of coverage while minimizing overkill is challenging.

The Joint Commission requires American hospitals to implement antibiotic stewardship programs that educate health-care workers on best prescribing practices. Clinicians are trained to order microbial cultures for presumably infected patients before beginning a course of antibiotics. Microbial cultures are sent to a microbiology lab for testing, and after about two days, the identity of the infecting agent is determined. Most microbial cultures fail to grow bacteria. If bacteria is isolated, further drug susceptibility analysis is performed, and after another day results return showing whether a set of commonly prescribed antibiotics will cover the infecting agent. In critical cases, clinicians cannot wait for susceptibility results. They start empiric treatment, a euphemism for guessing the diagnosis and optimal therapy. Empiric treatment typically consists of broad-spectrum drugs like piperacillin/tazobactam (Pip/Tazo) that maximize likelihood of microbial coverage even though something more targeted likely would suffice. Broad-spectrum antibiotics can devastate patient microbiota, and lead to the emergence of often deadly *Clostridium difficile* (*C. difficile*) infections[3]. When susceptibility results are returned, clinicians de-escalate to targeted treatment. An illustration of this clinical workflow is shown in Figure 1.

Electronic Health Records (EHRs) have been used extensively for clinical decision support[4,5]. Several of these studies address the inefficiencies of empiric antibiotic treatment. Ribers and Ullrich use a random forest to predict microbial culture results of primary care patients in Denmark suspected of urinary tract infection[6]. They report that use of their classifier could reduce antibiotic prescription by 7.42 percent without reducing the number of infections treated with antibiotics. Hernandez et al fit a series of classifiers to microbial culture data to predict the likelihood of bacterial growth[7]. Their study results were promising, but the potential for generalized performance across time and unseen patients remains unclear. Yelin et al use personal clinical histories to predict the likelihood of microbial resistance to sets of antibiotics[8]. Their prediction models and algorithmic prescribing policies show promise that machine learning can help decrease ineffective antibiotic prescriptions for patients with UTIs. More work is needed to address stewardship goals; that is, how we can use machine learning models to maximize patient coverage with narrower-spectrum empiric drugs.

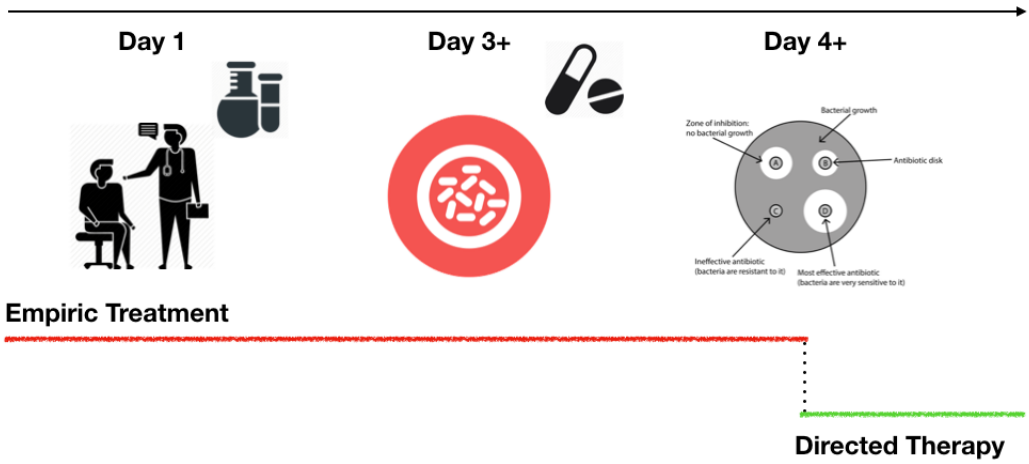


Figure 1: Graphic illustration of antibiotic delivery workflow. On Day 1 a patient shows up to the hospital with a possible infection. A clinician prescribes empiric antibiotics (broad-spectrum) to maximize the likelihood the patient responds. At the same time, a microbial culture is ordered. After Day 3 the identity of the infecting agent, if one exists, is confirmed. At Day 4 susceptibility results return detailing a set of appropriate antibiotics, and the patient is de-escalated to directed therapy.

Antibiograms are standard of care tools that summarize bacterial susceptibility patterns of commonly administered antibiotics[9] - see Table 1. Each cell represents the proportion of microbes isolated in cultures susceptible to a particular antibiotic. Antibiograms are used by clinicians administering empiric treatment. Our objective is to develop an approach that allow clinicians to prescribe narrower-spectrum antibiotics with likelihood of coverage equivalent to coverage rates of broader-spectrum drugs reported in antibiograms. Specifically, the aim of this study is to 1) predict whether bacteria will grow at all in microbial cultures and 2) predict which antibiotics an infecting bacteria will be susceptible to - creating personalized antibiograms for individual patients using patient EHR histories and machine learning models.

Table 1: 2014 inpatient antibiogram constructed using Stanford EHR data. Each cell indicates the proportion of microbial isolates susceptible to the corresponding antibiotic. Clinicians use these tables to inform empiric antibiotic treatment. Probabilities are based on prevalence, and are not adjusted on an individual patient basis. R indicates the microbe is assumed resistant to the corresponding antibiotic, and therefore not tested. NA indicates the microbe is not tested due to our microbiology labs's selective reporting guidelines. *E. coli* = *Escherichia Coli*, *K. pneumoniae* = *Klebsiella pneumoniae*, *P. aeruginosa* = *Pseudomonas aeruginosa*, Pip/Tazo = piperacillin/tazobactam.

Organism Name	Pip/Tazo	Ampicillin	Cefepime	Ceftriaxone	Cefazolin	Levofloxacin	Ciprofloxacin
E. coli	0.94	0.46	0.91	0.81	0.75	0.67	0.67
K. pneumoniae	0.88	0.0	0.91	0.87	0.81	0.87	0.84
P. aeruginosa	0.89	R	0.87	R	R	NA	0.81

Methods

Dataset

We use the Stanford Medicine Research Data Repository (STARR) inpatient clinical data warehouse which contains de-identified patient EHRs between 2008 and 2014. This dataset includes the set of all microbial cultures ordered at Stanford University Hospital, a tertiary academic medical center. Microbiology data includes presence of bacteria, type of bacteria, and antibiotic susceptibility analyses. The minimum inhibitory concentration (MIC) technique is used to assign labels (Susceptible, Intermediate, Resistant) to each antibiotic tested against the infecting agent grown in the microbial culture[10]. Our entire dataset includes patient demographics, comorbidities, lab orders, vital signs, medications, and treatment teams.

Prediction Tasks

No Bacterial Growth Predictions

We train binary classifiers that predict lack of bacterial growth in blood and urine cultures. Cultures with no bacterial growth are assigned to the positive class. Cultures that grow bacteria are assigned to the negative class. Patient medical timelines often contain multiple blood and urine culture orders. We include only the first blood and urine culture a patient receives in our analysis.

Bacteria Susceptibility Predictions

Consistent with our microbiology labs selective reporting guidelines, we infer susceptibility to Pip/Tazo if the organism demonstrates susceptibility to ampicillin. We infer susceptibility to newer generation cephalosporins if the agent demonstrates susceptibility to older generation cephalosporins. We train binary classifiers for each microbe/antibiotic combination. Cultures whose susceptibility results come back Susceptible are assigned to the positive class. Intermediate and Resistant labels are assigned to the negative class. Clinicians would not treat patients with an antibiotic labelled Intermediate. The full list of microbes and antibiotics for which we make susceptibility predictions on is as follows.

Escherichia coli (*E. coli*) and *Klebsiella pneumoniae* (*K. pneumoniae*)

- Ampicillin
- Cephalosporins (Cefepime, Ceftriaxone, Cefazolin - 4th, 3rd, and 1st generation)
- Fluoroquinolones (Levofloxacin, Ciprofloxacin - 3rd and 2nd generation)

Pseudomonas aeruginosa (*P. aeruginosa*)

- Cefepime
- Ciprofloxacin

Train Validation and Test Splits

We split our training, validation, and test sets based on the years cultures were drawn. Our training set consists of cultures ranging from 2009 to 2012, our validation set contains cultures ordered in 2013, and our test set contains those ordered in 2014. To preserve model generalizability on new patients, the sets of patients in our training, validation, and test sets are disjoint.

Feature Engineering and Re-sampling

We use patient demographics, comorbidities, prior lab tests, vital signs, medications, and treatment teams to make predictions. Prediction time for our no growth classifiers is the point at which microbial cultures are ordered. Prediction time for our susceptibility classifiers is the point at which infecting agents are known. Categorical features are represented as counts over the past 1, 2, 4, 7, 14, 30, 90, 180, 365, 730, and 1460 days. We also include the total count of occurrences over a patient's entire medical history, and the number of days since the last occurrence. Numerical features are represented with summary statistics over the past 14 day window. These summary statistics include the minimum, maximum, median, mean, standard deviation, first, last, and slope over the window. Models are fit on 4261 features. Missing values are imputed by taking the mean over columns. Features are standardized, and the Synthetic Minority Over-Sampling Technique (SMOTE) is used to address extreme class imbalance[11].

Model Selection

We train three machine learning models for each prediction task: a logistic regression with L1 regularization (LASSO)[12], a random forest[13], and a gradient boosted tree model using the extreme gradient boosting (XGBoost) implementation[14].

We tune the LASSOs regularization coefficient with a cross validation grid search sweeping over values 10^{-8} to 10^{8} in power of ten intervals using only our training set. The number of trees in our random forest models is set to 100. We tune the maximum number of features each tree is able to look at and its max depth. We set the number of boosting rounds for our XGBoost models to 100, the learning rate to 0.3, and tune the max depth of each tree, the percent of data each tree sees, the maximum features each tree uses, and the gamma, alpha, and lambda regularization parameters.

For each task, we select the model type that performs best on our 2013 validation set with respect to the area under the receiver operator curve (AUROC), retrain on the union of our training and validation sets, and evaluate the final performance on our 2014 test set. Technical performance of each final model is evaluated using AUROC and AUPRC (area under the precision recall curve). 95% confidence intervals are computed by bootstrapping the 2014 test set.

Estimating Clinical Relevance

We estimate the clinical utility of our models by retroactively computing the fraction of patients that could have been given less broad-spectrum drugs at the same susceptibility rate as Pip/Tazo shown in the 2014 antibiogram. Pip/Tazo is a commonly used broad spectrum antibiotic for empiric treatment because its efficacy against gram negative bacteria is high. In this analysis we restrict our one year test set to patients empirically treated with Pip/Tazo. We say a patient was empirically treated with Pip/Tazo if the order timestamp of Pip/Tazo was in between the order and result time of the microbial culture. For each model, we retroactively look for a decision threshold where recall is maximized and precision is at or above the 2014 antibiogram susceptibility rate for Pip/Tazo. We then report the fraction of patients in our test set whose predicted probabilities were at or above this decision threshold.

Results

Predicting No Growth Blood and Urine Cultures

Here we show the technical evaluation of our no growth blood and urine culture models. 95% of blood cultures fail to grow bacteria. 76% of urine cultures show no signs of growth. Table 2 shows AUROC and AUPRC values for each of our three models (LASSO, Random Forest, and XGBoost) on our validation set, as well as final model performance on our one year test set.

Predicting Bacterial Susceptibility

Next we show the technical evaluation of our bacterial susceptibility classifiers. Bacterial susceptibility over the 2009-2014 time period is mostly stationary over time, as seen in Figure 2. Table 3 shows the AUROC and AUPRC of the final model, model type, and corresponding 2014 antibiogram susceptibility rate for each prediction task.

Table 2: Model performance for our two no growth prediction tasks on both our validation and test sets. Area under the precision recall curve (AUPRC) and area under the receiver operator curve (AUROC) are shown for each of our three model types trained using the training set and evaluated with the 2013 validation set. AUPRC and AUROC of our final models are shown for the 2014 test set. 95% confidence intervals are estimated by bootstrapping the test set. RF = random forest, XGBoost = extreme gradient boosting, LASSO = logistic regression with L1 regularization.

Culture Type	**2013 Validation**			**2014 Test**		
	Model Type	AUPRC	AUROC	Model Type	AUPRC	AUROC
Blood [N = 19,938]	*LASSO*	*0.98*	*0.68*	LASSO	0.97 [0.96, 0.97]	0.64 [0.60, 0.68]
	RF	0.97	0.61			
	XGBoost	0.97	0.64			
Urine [N = 16,765]	*LASSO*	*0.90*	*0.70*	LASSO	0.87 [0.86, 0.88]	0.69 [0.67, 0.71]
	RF	0.88	0.68			
	XGBoost	0.88	0.66			

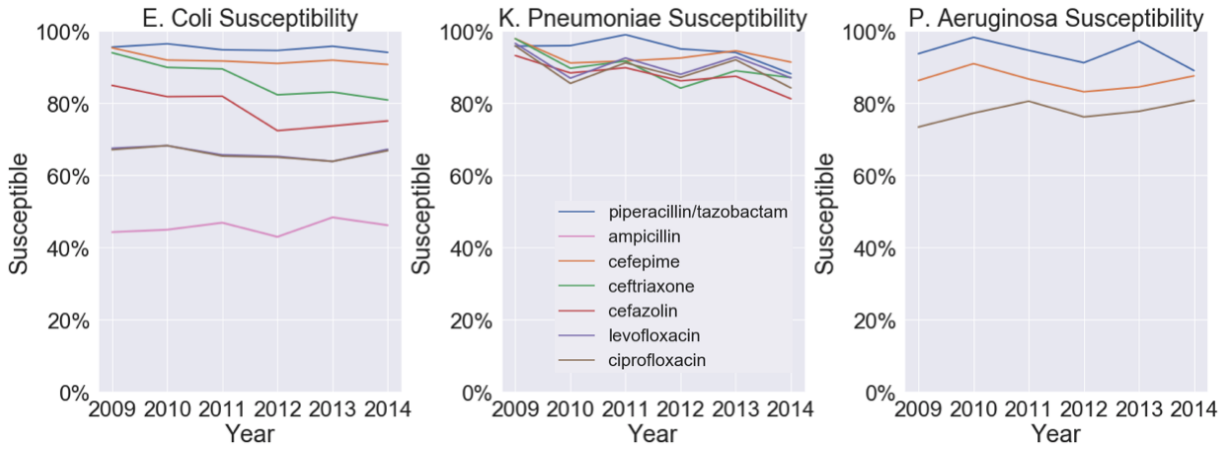


Figure 2: Bacterial susceptibility to a set of commonly prescribed antibiotics by year. Susceptibility is mostly stationary over our time window.

Estimating Clinical Relevance

Here we show the clinical relevance of our microbe susceptibility classifiers. Figure 3 shows how we choose clinically useful decision thresholds, and how they are analyzed to retroactively estimate the subset of patients that could have been given narrower-spectrum antibiotics at the same rate of coverage as Pip/Tazo in our one year test set.

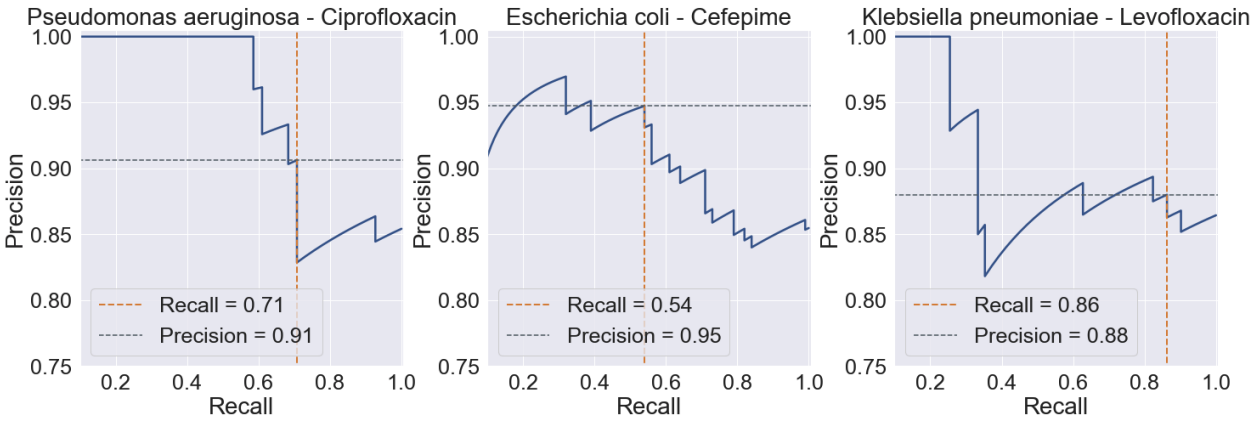
Discussion

No Growth Predictions

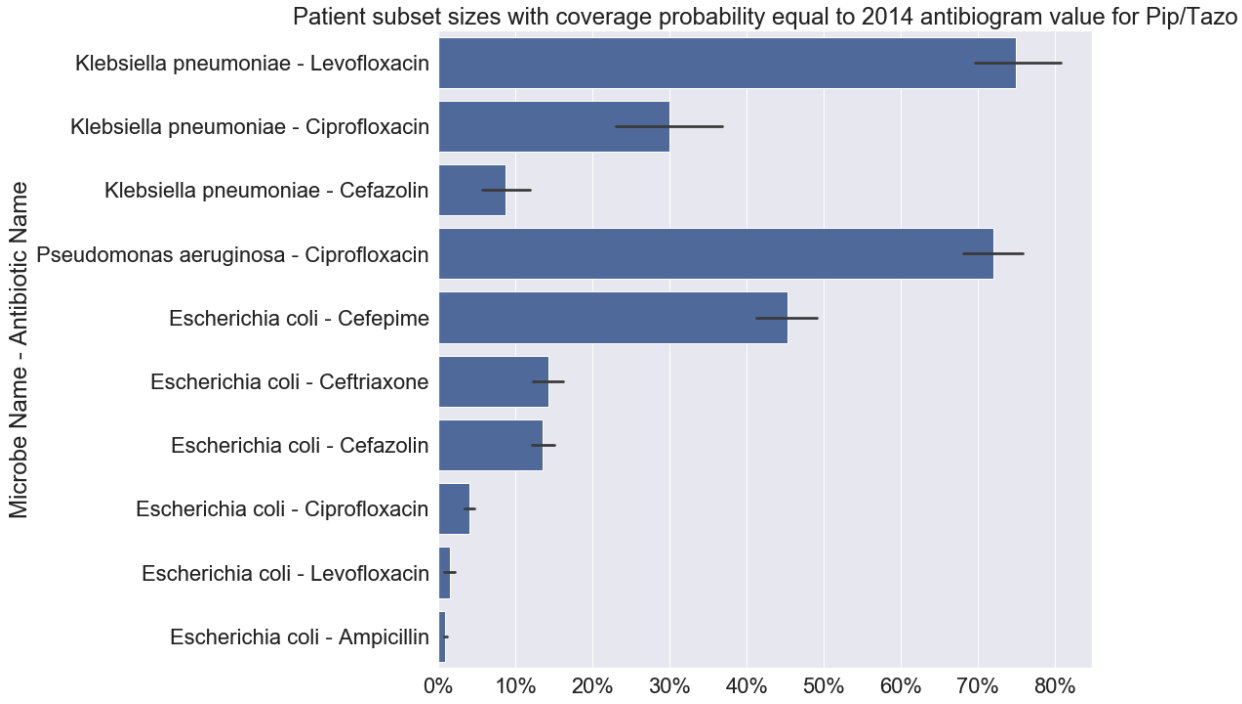
Each of our no growth classifiers contain operating regions that flag no growth cultures with high precision. Our blood culture classifier is able to predict 30% of all no growth blood cultures with $> 98\%$ precision. Our urine culture classifier predicts 36% of all no growth urine cultures with $> 90\%$ precision. While lack of bacterial growth in blood

Table 3: Technical performance of bacterial susceptibility classifiers on our one year test set. Best model indicates the type of model that had the highest area under the receiver operator curve (AUROC) on our 2013 validation set. For each model we show area under the precision recall curve (AUPRC), AUROC, and the 2014 antibiogram value - which is simply baseline prevalence. RF = random forest, XGBoost = extreme gradient boosting, LASSO = logistic regression with L1 regularization. Small variations of N within each microbe type exist due to our microbiology lab's selective reporting guidelines. We report the minimum N for each microbe type.

Organism	Antibiotic	Best Model	2014 Antibiogram Susceptibility	AUPRC	AUROC
E. coli [N= 2,424]	Ampicillin	RF	0.46	0.53 [0.48, 0.60]	0.60 [0.54, 0.65]
	Cefepime	XGBoost	0.91	0.96 [0.93, 0.98]	0.73 [0.65, 0.78]
	Ceftriaxone	RF	0.81	0.86 [0.83, 0.90]	0.63 [0.56, 0.69]
	Cefazolin	RF	0.75	0.81 [0.77, 0.86]	0.60 [0.54, 0.67]
	Ciprofloxacin	RF	0.67	0.79 [0.75, 0.84]	0.67 [0.63, 0.72]
	Levofloxacin	RF	0.67	0.76 [0.71, 0.82]	0.67 [0.64, 0.72]
K. pneumoniae [N = 671]	Cefepime	RF	0.91	0.97 [0.94, 0.99]	0.74 [0.62, 0.84]
	Ceftriaxone	RF	0.87	0.95 [0.92, 0.97]	0.70 [0.61, 0.78]
	Cefazolin	RF	0.81	0.86 [0.81, 0.93]	0.61 [0.52, 0.70]
	Ciprofloxacin	LASSO	0.84	0.86 [0.77, 0.92]	0.56 [0.44, 0.67]
	Levofloxacin	XGBoost	0.87	0.90 [0.84, 0.95]	0.56 [0.44, 0.67]
P. aeruginosa [N = 693]	Cefepime	RF	0.88	0.92 [0.83, 0.97]	0.66 [0.50, 0.76]
	Ciprofloxacin	XGBoost	0.81	0.90 [0.83, 0.94]	0.66 [0.53, 0.75]



(a) Precision Recall Curves



(b) Patient Subsets

Figure 3: **(a)** precision recall curves for classifiers evaluated on patients empirically treated with Piperacillin/Tazobactam (Pip/Tazo) in our one year test set. Cross-hairs highlight decision thresholds where recall is maximized such that precision exceeds or equals the 2014 antibiogram value for Pip/Tazo. 2014 Pip/Tazo antibiogram values for *P. aeruginosa*, *E. coli*, and *K. pneumoniae*, are 0.89, 0.94, 0.88 respectfully. Test set sample sizes are $N = 48$, $N = 119$, and $N = 59$. **(b)** percentage of patients whose predicted probabilities exceed these decision thresholds. This corresponds to subsets of patients treated with Pip/Tazo with susceptibility rates to narrower-spectrum antibiotics equal to the 2014 antibiogram values for Pip/Tazo. Microbe-antibiotic combinations for patients given Pip/Tazo whose baseline coverage rates exceed Pip/Tazo in 2014 are not shown, as 100% of these patients could have been given the antibiotics at the Pip/Tazo coverage rate without using machine learning models. These combinations are *K. pneumoniae* - Cefepime, *K. pneumoniae* - Ceftriaxone, and *P. aeruginosa* - Cefepime

or urine cultures does not necessarily mean lack of infection, there is a high likelihood that a large portion of these patients are not infected. These patients often receive empiric antibiotic treatment when it is not needed. We do not recommend that clinicians neglect treatment for individuals whose cultures are predictably negative. However we may suggest that these patients be placed on less broad-spectrum antibiotics - especially when paired with predictions made by our susceptibility classifiers that indicate high likelihood of susceptibility to narrower-spectrum drugs.

Bacterial Susceptibility Predictions

Our bacterial susceptibility classifiers contain decision thresholds that separate subsets of patients that could have been given narrower-spectrum antibiotics at the same coverage rate as Pip/Tazo. Clinicians often empirically prescribe Pip/Tazo because its likelihood of covering the suspected infecting agent is high. Use of broad-spectrum antibiotics when something more targeted would have sufficed has severe implications on patient health, cost of care, and the development of antibiotic resistance. Broad-spectrum agents can negatively impact patients' microbiota, and lead to an increased risk of *C. difficile* infection. Cost of care is increased not only because broad-spectrum agents are more expensive, but because less broad-spectrum cephalosporins and fluoroquinolones can be administered orally and reduce length of hospitalization[15]. And finally, overuse of broad-spectrum drugs severely affects the efficacy of these drugs in the decades to come due to emerging resistance.

Antibiograms guide adequate coverage and are the current standard of care in leading health institutions. They summarize microbe susceptibility patterns but do not leverage patient specific prior knowledge. We have shown that machine learning models can better discriminate bacterial susceptibility results to sets of antibiotics. Personalized antibiograms would allow clinicians to empirically prescribe less broad-spectrum drugs at higher rates of susceptibility.

Limitations

Predictive performance was variable for our different prediction tasks. Nevertheless, we demonstrate that many of our models contain useful operating regions that would allow clinicians to prescribe less broad-spectrum empiric treatment at high likelihood of coverage.

We note that lack of bacterial growth in a microbial culture does not infer lack of infection. Thus even though we are able to predict with high precision that a microbial culture will not grow bacteria, we are not able to draw the conclusion that these patients are not infected and should not be given antibiotics. Our predictions do however suggest that smaller antibiotics may be more appropriate, especially when paired with predictions suggesting the likelihood of susceptibility to these drugs is high.

Lastly due do the fact that our clinical relevance analysis was based on decision thresholds found in our one year test set, generalizability of coverage rates at these thresholds on new data remains unclear. This combined with our relatively small test set sample sizes after filtering for patients empirically treated with Pip/Tazo (Figure 3a) means that more work is needed to analyze prescribing polices that retroactively use these thresholds to optimize coverage on unseen data. Nevertheless we do show that decision thresholds exist that separate subsets of patients whose rates of coverage for narrower-spectrum antibiotics match antibiogram values for Pip/Tazo.

Conclusion

Machine learning tools can predict lack of growth in microbial cultures and likelihood of bacterial susceptibility to sets of antibiotics based on readily available EHR data. Personalized antibiograms better discriminate bacterial susceptibility compared with antibiograms (current standard of care) by leveraging patient specific medical histories. Personalized antibiograms have the potential to improve antibiotic stewardship by allowing clinicians to choose narrower-spectrum empiric antibiotics with high levels of confidence.

Acknowledgments

This research was supported in part by the NIH Big Data 2 Knowledge initiative via the National Institute of Environmental Health Sciences under Award Number K01ES026837, the Gordon and Betty Moore Foundation through

Grant GBMF8040, and a Stanford Human-Centered Artificial Intelligence Seed Grant. Patient data were extracted and de-identified by Stanford Medicines Research IT department as part of the Stanford Medicine Research Data Repository (STARR) project with support from the Stanford Clinical and Translational Science Award (CTSA) to Spectrum (UL1 TR001085), as led by the National Center for Advancing Translational Sciences at the National Institutes of Health. Additional support is from the Stanford NIH/National Center for Research Resources CTSA award number UL1 RR025744, and the National Science Foundation Graduate Research Fellowship Program.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or Stanford Healthcare.

References

[1] O'Neill J. Tackling drug resistance globally: final report and recommendations. 2016 May;.

[2] CDC. Antibiotic use in the United States, 2017: progress and opportunities. 2017;.

[3] Yoon MY, Yoon SS. Disruption of the Gut Ecosystem by Antibiotics. Yonsei Med J. 2018 Jan;59(1):4–12.

[4] Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med. 2018 May;1:18.

[5] Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature. 2019 Aug;572(7767):116–119.

[6] Ribers MA, Ullrich H. Battling Antibiotic Resistance: Can Machine Learning Improve Prescribing? 2019 Jun;.

[7] Hernandez B, Herrero P, Rawson TM, Moore LSP, Evans B, Toumazou C, et al. Supervised learning for infection risk inference using pathology data. BMC Med Inform Decis Mak. 2017 Dec;17(1):168.

[8] Yelin I, Snitser O, Novich G, Katz R, Tal O, Parizade M, et al. Personal clinical history predicts antibiotic resistance of urinary tract infections. Nat Med. 2019 Jul;25(7):1143–1152.

[9] Joshi S, et al. Hospital antibiogram: a necessity. Indian journal of medical microbiology. 2010;28(4):277.

[10] Jorgensen JH, Ferraro MJ. Antimicrobial susceptibility testing: a review of general principles and contemporary practices. Clin Infect Dis. 2009 Dec;49(11):1749–1755.

[11] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. 1. 2002 Jun;16:321–357.

[12] Tibshirani R. Regression Shrinkage and Selection Via the Lasso. J R Stat Soc Series B Stat Methodol. 1996 Jan;58(1):267–288.

[13] Breiman L. Random Forests. Mach Learn. 2001 Oct;45(1):5–32.

[14] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. 2016 Mar;.

[15] Cyriac JM, James E. Switch over from intravenous to oral therapy: a concise overview. Journal of pharmacology & pharmacotherapeutics. 2014;5(2):83.