Predicting Microculture Results for Optimized Antibiotic Treatment

Conor K. Corbin [1]

Abstract

Over 700,000 patients die a year due to antibiotic resistant infections. Antibiotic resistance is a growing public health concern. It is estimated that up to 50% of antibiotic use in medical settings is either inappropriate or sub-optimal (CDC). We target a specific inpatient clinical workflow involving antibiotic distribution, and design machine learning classifiers to intervene at various instances. We build classifiers to predict microcultures labs that will not grow bacteria (a proxy for predicting lack of infection) and classifiers that predict bug susceptibility to an array of antibiotics. Our classifiers achieve an AUROC of 0.67 and 0.70 when predicting no growth blood and urine cultures respectively. Our susceptibility classifiers contain operating regions that allow de-escalation to targeted antibiotics in cases where de-escalation would not occur using current state of the art tools.

1. Introduction

Antibiotic resistance is one of the most serious public health concerns, and threatens to return us to an age where infections were the leading cause of death in the United States. Antibiotics grant enormous utility, but when used improperly enable bacteria to develop immunity. In an effort to combat this issue, the Joint Commission has developed antimicrobial stewardship standards that hospitals around the country are expected to follow. These standards establish clinical workflows designed to optimize antibiotic treatment. When a patient presents with infection like symptoms, microcultures are ordered. Within a week, results of these cultures are returned, detailing the organism causing the infection (if one exists) and a set of antibiotics to which it is susceptible. In the meantime, empiric antibiotics (usually broad-spectrum) are administered to the patient to maximize the likelihood that they positively respond. Once results are returned, physicians transition their patients to a more targeted antibiotic therapy. A graphical illustration of the this workflow is depicted in Figure 1. Some institutions augment this workflow with the use of antibiograms - tables depicting drug/bug susceptibility patterns using local hospital data (Hindler & Stelling, 2007). In some cases, using this tool allows clinicians to de-escalate their patients to less broad-spectrum drugs as soon as the infecting agent is known. Most of the time, de-escalation occurs only after susceptibility results are returned. An example antibiogram is shown in Figure 2. During empiric treatment, patients are exposed to broad-spectrum drugs that 1) contribute to the growing problem of antibiotic resistance and 2) often cause serious side effects (Giuliano et al., 2016). We experiment with predictive models designed to intervene at several instances of this workflow in an attempt to maximize patient safety and minimize wasteful antibiotic use.

2. Related Work

Several studies have looked at inefficiencies of empiric antibiotic treatment in hospital settings. These studies break down into roughly three categories: inference studies, ML studies, and ML + policy studies. On the inference side, (Anstey et al., 2018) compute likelihood ratios of infection given various predictors and find significant associations with transient hypotension and central line presence. On the ML side, (Hernandez et al., 2017) fit a series of classifiers (Naive Bayes, Decision Tree, Random Forest, and Support Vector Machine) to microculture data to predict the presence of infection. They limit their analysis to six predictors: alanine aminotransferase, alkaline phosphatase, bilirubin, creatinine, C-Reactive protein and white blood cell count, and achieve an AUROC of 0.8. Another group looks specifically at predicting the presence of urinary tract infection from urine cultures taken by primary care doctors in Denmark (Ribers & Ullrich, 2019). They fit a Random Forest classifier and use a variety of different features including prior lab tests and values, medications and comorbidities to aid their predictions. The authors work with a dataset containing 95,594 urine samples and achieve an overall AUROC of 0.73 when predicting bacterial growth on out of bag samples. The authors take a step further and use the predicted probabilities that their machine learning classifier generates to implement an antibiotic prescribing policy. They compare the policy to the prescribing patterns of physicians. The TREAT decision support system uses a causal probabilistic network to estimate the probability of infection, its severity, source, pathogen distribution, and antibiotic coverage (Paul et al., 2006). A prospective cohort study testing

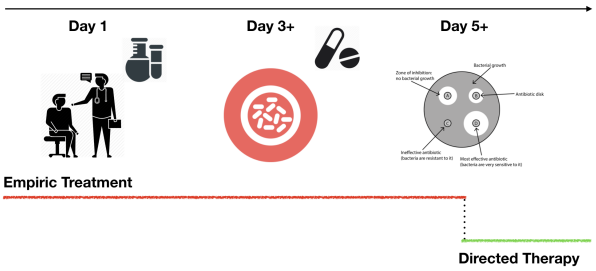

Figure 1. Graphic illustration of microculture workflow in an inpatient setting. On Day 1 a patient shows up to the hospital with signs of infection. Their physician prescribes empiric antibiotics (broad-spectrum) to maximize the likelihood the patient responds. At the same time, a microculture is ordered. At Day 3 microculture results come back detailing the infecting agent if one exists. At Day 5 susceptibility results come back detailing an array of antibiotics for which the bacteria is susceptible to, and the patient is de-escalated to directed therapy.

this technology on 1203 patients revealed an increased rate of appropriate empiric antibiotic prescriptions compared to physicians (70% vs 57%) while using less broad-spectrum antibiotics. Antibiotic resistance patterns are non-stationary, and new analysis and support systems are needed to leverage patient data to optimize antimicrobial treatment pathways.

3. Dataset and Features

We fit predictive models using data from STRIDEs (Stanford Translational Research Integrated Database Environment) inpatient clinical data warehouse which contains de-identified patient EMRs (electronic medical records) that span between 2008-2017. Our dataset contains 115k unique patient records, and 200k unique encounters. Patient information stored in EMRs include patient demographics, comorbidities, lab orders and results, vital signs, medications, and treatment teams. Of particular interest in this study is STRIDE's microculture table, which includes microculture orders, results, and bacteria susceptibility testing information on over 80,000 blood and 30,000 urine cultures respectively.

4. Methods

We implement predictive models to intervene at both day 1 - the point at which microcultures are ordered, and day 3 - the point at which the identity of the infecting agent is observed. For our first level of analysis, we train binary classifiers to predict whether or not bacteria will grow at all. We do this for two types of microcultures - blood cultures and urine cultures. Our dataset includes 80k blood cultures of which 95% come back negative, and 30k urine cultures of which 75% come back negative. Clinically, the ability to

Microbe	Isolates Tested	ALL DRUGS	Meropenem	Piperacillin-Tazobactam
ALL BUGS	5224	91	90	82
Klebsiella pneumoniae	741	100	100	95
Proteus mirabilis	262	100	100	100
Enterobacter cloacae	224	99	99	82
Klebsiella oxytoca	113	100	100	94
Pseudomonas aeruginosa	545	91	90	91
Escherichia coli	3339	100	100	95

Figure 2. Antibiogram constructed from Stanford Hospital data. Rows are bugs, columns are drugs. Each value indicates the percent of time a particular bug was susceptible to a drug. Clinicians use these tables to inform their antibiotic prescription decisions. Use of an antibiogram sometimes allows clinicians to de-escalate patients to more targeted therapy at Day 3 - when the name of the infecting agent is known. Often the drop in percent effectiveness from a broad spectrum drug to something more targeted is too much to risk.

predict negative cultures with high confidence is interesting because it suggests that a machine learning tool could help prevent needless antibiotic treatment. Our positive labels are thus negative cultures. Our second level of analysis predicts bacteria susceptibility to a set of antibiotics - independent of the type of infecting agent. Our last set of classifiers predict susceptibility to a set of antibiotics given the bacteria type. We contrast the utility of these classifiers with antibiograms - the current state of the art tool informing clinician antibiotic prescription decisions.

4.1. Feature Engineering

We use patient medical information stored in EMRs to make predictions. We leverage patient demographics, comorbidities, prior lab tests (including prior microcultures and bug susceptibility screenings), vital signs (including heart beat and blood pressure), medication use (including prior antibiotic prescriptions) and treatment teams. Categorical variables (comorbidities, treatment teams, and lab orders) are treated as counts over varying time windows in a patient history. Each categorical features is represented as counts over the past (1, 2, 4, 7, 14, 30, 90, 180, 365, 730, 1460) patient days. We also include the total number of occurrences over the patient's entire medical history - and the time in days since the last occurrence. For features that contain numerical values (heart beat, blood pressure, white blood cell count), we compute 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. Missing values are imputed by taking the mean over columns using our training set.

4.2. Feature & Model Selection

Our resulting design matrix above is sparse and contains many columns with redundant information. We utilize a recursive feature selection algorithm in an attempt to reduce

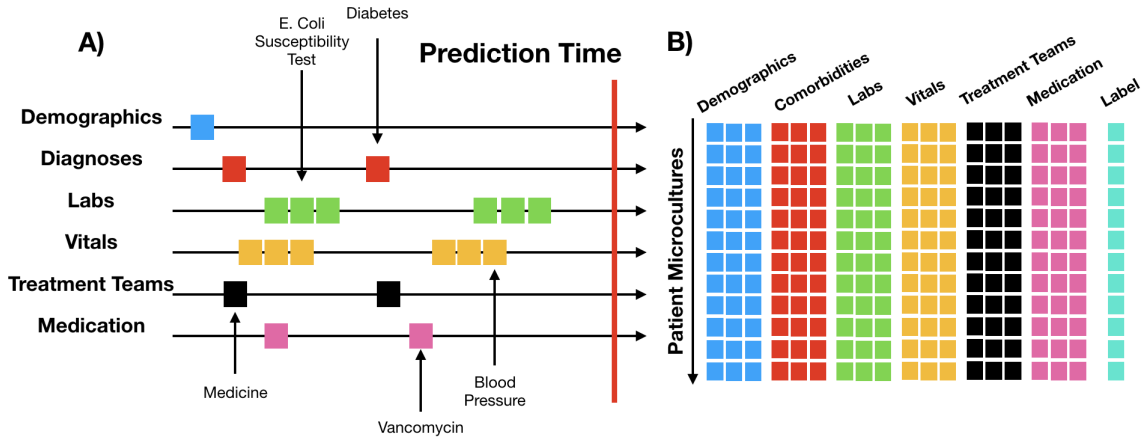

Figure 3. Patient data available in an EMR illustrated as a medical timeline (A). Data include patient demographics, diagnosis, labs, vitals, treatment teams and prior medications. The data is time series, and needs to be transformed into a patient feature matrix (B) before it can be fed into downstream machine learning models. This is done by taking counts of categorical events and summary statistics of real valued events over varying time windows.

the variance of our resulting classifiers and eliminate useless features. We do this by recursively training a Random Forest classifier on our current feature space, ranking each feature based on how well it decreases node impurity (GINI score) (Menze et al., 2009) and removing the least important features until a desired number remain (we empirically choose 5%). This is done using only our training data. We use the resulting design matrix to construct our models. We train both a linear Logistic Regression and non-linear Random Forest. We add L1-regularization to our Logistic Regression to further reduce the variance of our model. We tune the regularization coefficient with a k=10 cross-validation grid-search using only our training data. We similarly tune the number of trees, max depth, min samples required for split, and max features of our Random Forest model. We subsequently train our best linear and non-linear model on our entire training set, and evaluate on our test set. Our training and test sets are obtained with an 80/20% split. We split over the patient id column ensuring that microcultures from the same patient do not exist in both the training and test set.

5. Experiments & Results

5.1. Predicting No Growth Cultures

Our first set of analysis involves predicting whether or not bacteria will grow at all in microcultures ordered at Stanford Hospital. We look at two types of microcultures, blood cultures (N=80,000) and urine cultures (N=30,000). 96% of all blood cultures at Stanford come back negative, 75% of all urine cultures are negative. Our L1 Logistic Regression achieved an AUROC of 0.64 and 0.65 predicting no growth blood and urine cultures respectively. The Random Forest model achieved an AUROC of 0.66 and 0.70. We show an ROC and precision recall curve for the Random Forest model in Figure 4.

5.2. Predicting Bug Susceptibility

We next predict bug susceptibility to an array of commonly prescribed antibiotics, given that bacteria has grown in the extracted culture. Our positive labels are cultures that grew bacteria susceptible to the antibiotic in question. Negative labels are cultures that grew bacteria not susceptible to the antibiotic. Because we rely on the prior of bacterial growth, these classifiers are currently designed to intervene at Day 3 of the workflow. It is worth noting however that we only use data up until the point at which the microculture is ordered - and we do not make use of the name of the infecting agent when making our predictions. To instead intervene at order time (Day 1), we could more carefully design our test set such that both positive and negative growth cultures are appropriately represented. Here we train a Random Forest classifier for 16 binary prediction tasks, corresponding to 16 different antibiotics. In Table 1 we show the AUROC, AUPRC, and baseline prevalence for the top and bottom performers. Our top two performers (Vancomycin and Zosyn) motivate a path forward we illustrate in our discussion.

5.3. Personalized Antibiograms

Our last analysis focuses on classifiers specifically designed to intervene at Day 3 - the point at which the infecting agent is returned. Here we use data up until the point the results of the culture are returned, including the name of the infecting agent, to predict a set of antibiotics for which

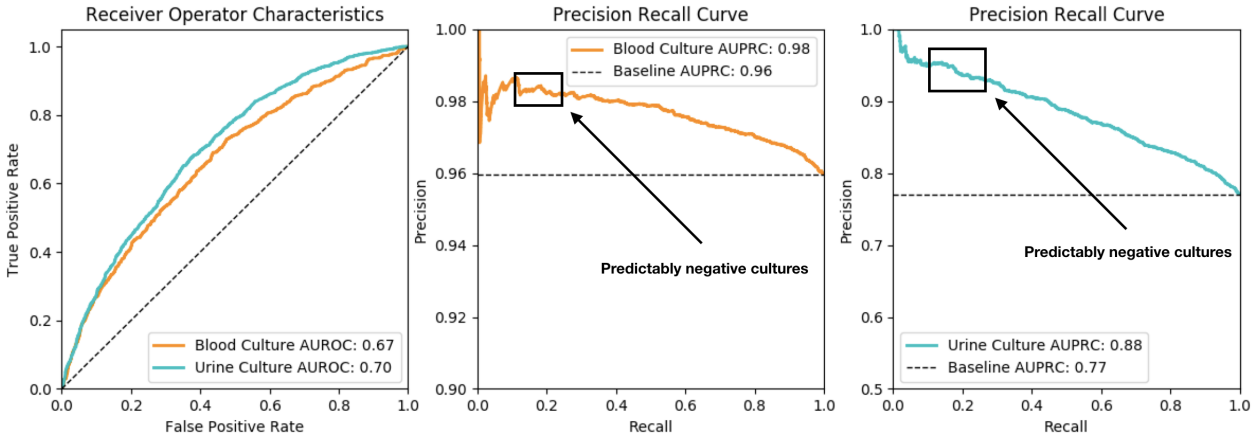

Figure 4. ROC and precision recall curves for our Random Forest Model predicting no growth blood and urine cultures. Our model achieves an AUROC of 0.66 and 0.70 on blood and urine cultures respectively. They achieve an AUPRC of 0.98 and 0.88 respectively. In both precision recall curves, we box an operating region where our classifiers can predict no growth with > 95 % confidence. While a no growth culture does not completely imply no infection, it is fair to assume that a large portion of these patients are receiving inappropriate antibiotics.

Table 1. AUROC, AUPRC and baseline positive class prevalence for best and worst classifiers predicting bug susceptibility.

DRUG	AUROC	AUPRC	PREVALENCE
VANCOMYCIN	0.86	0.96	0.92
ZOSYN	0.70	0.96	0.93
DAPTOMYCIN	0.67	0.97	0.95
...	...	...	...
AMP/SULBACTAM	0.55	0.61	0.56
CEFTRIAXONE	0.50	0.88	0.87
OXACILLIN	0.49	0.71	0.71

a specific type of bacteria (E. Coli) is susceptible to. Because we use the name of the infecting agent as a feature in these classifiers, we can directly compare the utility of our predictions to the utility of antibiograms - the tool seen in Figure 2 clinicians currently use to help make proper prescription choices. We ask the question - do personalized antibiograms allow clinicians to de-escalate to less broad-spectrum drugs when antibiograms alone do not? We focus on the following de-escalation pathway - a set of 4 antibiotics listed from broadest-spectrum to most targeted: Meropenem - Piperacillon/Tazobactam - Ceftriaxone - Cefazolin. An antibiogram revealing baseline susceptibility of E.Coli to these four drugs constructed from Stanford Hospital data is seen in Table 2. We train classifiers to predict E.Coli susceptibility to Piperacillon/Tazobactam, Ceftriaxone, and Cefazolin. Precision recall curves are shown for each of these three classifiers in Figure 4. Note the baseline for each precision recall curve perfectly corresponds to the value in the antibiogram - because the antibiogram depicts percent susceptible.

Table 2. Antibiogram depicting E. Coli susceptibility to a set of antibiotics in a common de-escalation pathway.

DRUG	SUSCEPTIBILITY	# ISOLATES
MEROPENEM	0.99	1560
PIP/TAZO	0.92	3129
CEFTRIAXONE	0.76	883
CEFAZOLIN	0.75	3632

6. Discussion & Future Work

6.1. Predicting No Growth Cultures

As seen from Figure 3 we are able to predict no growth microcultures with reasonable performance. We highlight on operating region for both blood and urine culture classifiers where we are able to predict no growth with $> 95\%$ certainty. Although a no growth blood or urine culture is not entirely indicative of the fact that a patient is not infected (infection may not be in the blood stream, or may not be a urinary tract infection), it is highly likely that a large portion of the patients depicted in this operating region are receiving antibiotics without actually being infected. A path forward would be to generate more sophisticated class labels to further separate infected from not infected, and get a true number of how many times the misuse of antibiotics on non-infected patients could have been prevented.

6.2. Predicting Bug Susceptibility

Our second set of classifiers had high variance in predictive performance. We were able to predict bug susceptibility to Vancomycin, Zosyn, and Daptomycin with fairly high

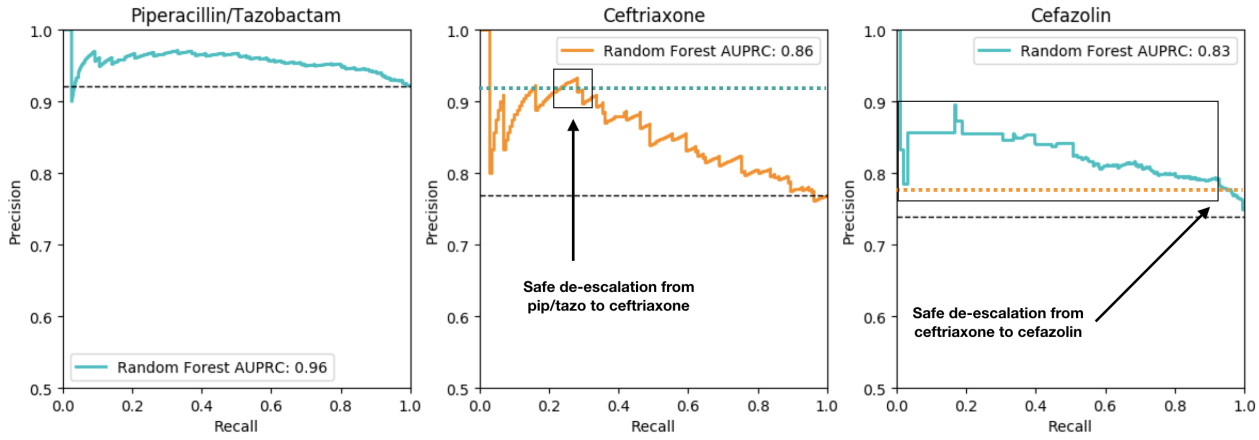

Figure 5. Precision recall curves showing how well we can predict E. Coli susceptibility to Piperacillon/Tazobactam, Ceftriaxone, and Cefazolin using EMR data. These classifiers allow us to make patient level predictions. Baseline positive class prevalence is shown with the dotted black line. Dotted colored lines indicate baseline prevalence of the more broad-spectrum drug immediately to the left of the drug in question. We box operating regions in these precision recall curves where the classifiers would allow clinicians to de-escalate to a less broad-spectrum drug in cases where an antibiogram alone would not.

performance - refer to Table 1 for values. Other drugs (Ampicillin/Sulbactam, Ceftriaxone, and Oxacillin) we did essentially no better than random chance. Our top classifiers (Vancomycin and Zosyn) are interesting because these two drugs are often prescribed together as an empiric regiment. Zosyn is a broad-spectrum drug adept at killing gram negative bacteria (bacteria with an outer cell wall). Zosyn does not cover MRSA (Methecillin Resistant Staph Aureus). Vancomycin does. 85% of the time Vancomycin and Zosyn are prescribed together, MRSA is not the infecting agent and Zosyn alone would have sufficed. This is worrisome not just because it needlessly contributes to antibiotic resistance, but also because Vancomycin and Zosyn taken together is associated with acute kidney disease (Giuliano et al., 2016). An interesting path forward would be to train a classifier to predict the absence of MRSA, and attempt to find an operating region where certain patients could safely be de-escaleted from Vancomycin + Zosyn to Zosyn.

6.3. Personalized Antibiogram

Our last set of classifiers (operating at Day 3) allow clinicians to safely de-escalate to less broad-spectrum drugs where antibiograms alone could not. This is illustrated in Figure 3. As an example, the drop in percent of E.Coli susceptible to piperacillon/tazpbactam vs. ceftriaxone (data available with an antibiogram) is quite large - 92% to 76%. In most cases a clinician will not be willing to take the added risk and de-escalate their patient infected with E.Coli to ceftriaxone. By leveraging a classifier trained on EMR data, we are able to find an operating region where patients infected with E.Coli are just as likely to be susceptible to ceftriaxone as they are piperacillon/tazobactam given the antibiogram data - 92%. In these cases, a clincian would be willing to de-escalate their patient to ceftriaxone - the more targeted therapy.

7. Conclusions

Antibiotic resistance is a growing public health concern and is only accelerated by systemic misuse of antibiotics in medical settings. We target a specific clinical workflow and build various machine learning classifiers designed to intervene at particular instances. We demonstrate an ability to predict no growth cultures at order time, and an ability to leverage ML to de-escalate patients to more targeted therapy in cases where this could not be done with the current state of the art. Machine learning shows promise in its ability to help physicians optimize antibiotic prescriptions.

8. Contributions

Though I did not have a CS229 partner, I would like to acknowledge Kojo Osei, Rich Medford, and Song Xu from the Stanford HealthRex lab. Kojo began the project as a Master's student last year, and developed the original framework to predict no growth cultures. Rich Medford is a infectious disease specialist at the UT Southwest Medical center, and gave useful insight when I became project lead. Song Xu was a post doc at Stanford's HealthRex Lab and helped develop the machine learning pipeline. I'd also like to thank my research advisor Jonathan Chen for guidance throughout the quarter.

9. Code

MicroCulture Project

References

CDC: Antibiotic use in the united states. "https://www.cdc.gov/antibiotic-use/stewardship-report/pdf/stewardship-report.pdf". Accessed: 2019-04-26.

Anstey, J. E., Murray, S., and Yim, J. W. Predicting bacteremia in hospitalized patients: An analysis of electronic health record data. In *JOURNAL OF GENERAL INTERNAL MEDICINE*, volume 33, pp. S297–S297. SPRINGER 233 SPRING ST, NEW YORK, NY 10013 USA, 2018.

Giuliano, C. A., Patel, C. R., and Kale-Pradhan, P. B. Is the combination of piperacillin-tazobactam and vancomycin associated with development of acute kidney injury? a meta-analysis. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, 36(12):1217–1228, 2016.

Hernandez, B., Herrero, P., Rawson, T. M., Moore, L. S., Evans, B., Toumazou, C., Holmes, A. H., and Georgiou, P. Supervised learning for infection risk inference using pathology data. *BMC medical informatics and decision making*, 17(1):168, 2017.

Hindler, J. F. and Stelling, J. Analysis and presentation of cumulative antibiograms: a new consensus guideline from the clinical and laboratory standards institute. *Clinical infectious diseases*, 44(6):867–873, 2007.

Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., and Hamprecht, F. A. A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC bioinformatics*, 10(1):213, 2009.

Paul, M., Andreassen, S., Tacconelli, E., Nielsen, A. D., Almanasreh, N., Frank, U., Cauda, R., Leibovici, L., and Group, T. S. Improving empirical antibiotic treatment using treat, a computerized decision support system: cluster randomized trial. *Journal of Antimicrobial Chemotherapy*, 58(6):1238–1245, 2006.

Ribers, M. A. and Ullrich, H. Battling antibiotic resistance: Can machine learning improve prescribing? 2019.