

Large-scale Prediction of Patient-Level Antibiotic Resistance: Towards Clinical Decision Support for Improved Antimicrobial Stewardship

by

Helen Zhou

Submitted to the
Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science
at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2018

© Massachusetts Institute of Technology 2018. All rights reserved.

Author
Department of Electrical Engineering and Computer Science
August 8, 2018

Certified by.....
David Sontag
Associate Professor, MIT
Thesis Supervisor

Certified by.....
Sanjat Kanjilal
Instructor of Medicine, Harvard Medical School
Thesis Supervisor

Accepted by
Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Large-scale Prediction of Patient-Level Antibiotic Resistance: Towards Clinical Decision Support for Improved Antimicrobial Stewardship

by

Helen Zhou

Submitted to the
Department of Electrical Engineering and Computer Science
on August 8, 2018, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Antibiotics are critical to modern medicine. However, levels of resistance have been rising, exacerbated by over-prescription and misuse of antibiotics. One major reason for this inappropriate usage is that doctors often must decide treatment without the results of microbiologic testing, a setting known as the *empiric treatment setting*.

Thus, this work aims to provide clinical decision support through patient-specific predictions of resistance *at the point of care*. Combining information from diagnoses, procedures, medications, clinicians' notes, and other modalities present in electronic medical records, various machine learning models such as logistic regression and decision trees are used to predict patients' probabilities of resistance to various antibiotics.

The full dataset consists of electronic medical records from patients presenting to the Massachusetts General Hospital and the Brigham & Women's Hospital between 2007 and 2016. On samples from the urinary tract (UTIs), which comprise approximately 48% of microbiology samples, the models achieve test AUCs ranging from 0.665 to 0.955 (depending on the antibiotic).

To evaluate the practical utility of these models, we extract the uncomplicated UTI cohort. Combining model predictions with well-defined treatment guidelines, a decision algorithm is constructed to recommend antibiotic treatments. For uncomplicated UTIs, the algorithm reduces test set prescriptions of broad-spectrum antibiotics by about 6.6%, while retaining similar levels of inappropriate antibiotic therapy.

Thesis Supervisor: David Sontag
Title: Associate Professor, MIT

Thesis Supervisor: Sanjat Kanjilal
Title: Instructor of Medicine, Harvard Medical School

Acknowledgments

First and foremost, I would like to thank my advisor Prof. David Sontag for providing consistent guidance and kind support over the past year. His wisdom and enthusiasm were invaluable in navigating the vast waters of research, and through his example I have learned so much about how to ask the right questions and do rigorous science.

Additionally, this work would not have been possible without the mentorship of Dr. Sanjat Kanjilal and Michael Oberst. Sanjat patiently and tirelessly sifted through countless results with me, and his expertise provided clinical insight that we would have been lost without. Mike was incredibly supportive and always willing to get in the weeds with me, from collaborating on code, to discussing both the details and presentation of our analyses. I'm thankful to have been a part of this amazing team, and I'm looking forward to potential future collaborations.

I would also like to thank the clinical machine learning group as a whole—you're all brilliant, inspiring people, and I really appreciate not only the academic but also the life advice you've shared with me. I've gotten a glimpse of what's ahead as I move on into my PhD, and I'm very excited.

Outside of this lab, there are many people to whom I am eternally grateful. To Soroush and Deb, my first UROP advisors, thank you for welcoming me into your lab during my freshman year, and for exposing me to machine learning research in the first place. To the wonderful friends I've made during my time at MIT, thank you for being there for me. We've had so many great adventures, with memories I'll cherish for a lifetime. To my family, thank you for your loving support, and for always believing in me. This thesis is dedicated to you.

Contents

1	Introduction	15
1.1	Motivation	15
1.1.1	Empiric Antibiotic Treatment	16
1.1.2	The Importance of Knowing Resistance	16
1.2	Thesis Organization	18
2	Background	19
2.1	Antibiotic Resistance	19
2.1.1	Treatment Setting	20
2.1.2	Microbiology Lab Testing	21
2.1.3	Mechanisms of Antibiotic Resistance	21
2.1.4	Predicting Antibiotic Resistance	22
2.1.5	Other Related Work in Machine Learning for Healthcare	24
3	Dataset Construction	27
3.1	Data Sources	27
3.1.1	Structured Data	27
3.1.2	Unstructured Data	33
3.2	Preprocessing	34
3.2.1	Extracting Labels from the Microbiology Data	34
3.2.2	Filtering the Microbiology Data	38
3.2.3	Feature Extraction	39

4	Exploratory Analysis of Micro Data	43
4.1	Sample Volume	43
4.1.1	Volume by Hospital Ward	43
4.1.2	Volume by Site of Infection	44
4.2	Levels of Resistance	45
4.2.1	Pathogens' Resistance	46
4.2.2	Correlations in Drug Resistance	48
4.3	Summary	50
5	General Experiment Setup	51
5.1	The Prediction Task	51
5.2	Train and Test Splits	52
5.3	Defining Empiric Prescriptions	53
5.4	Models	55
5.4.1	Logistic Regression	55
5.4.2	Decision Trees	56
5.4.3	Random Forests	56
6	Predicting Resistance in Urinary Tract Infections	57
6.1	Cohort Definitions	57
6.1.1	General UTI Cohort	57
6.1.2	Uncomplicated UTI Cohort	58
6.1.3	General vs. Uncomplicated UTI Cohort	59
6.2	Antibiotics of Interest	61
6.2.1	Drugs Relevant to UTIs	61
6.2.2	Drugs Relevant to Uncomplicated UTIs	61
6.3	Experiment Setup	63
6.3.1	Training Set	64
6.3.2	Feature Set	64
6.3.3	Model Classes	64
6.3.4	Hyperparameter Tuning	65

6.4	Model Performance	66
6.4.1	Additional Experiments	69
6.5	Model Interpretation	70
6.5.1	General UTIs	70
6.5.2	Uncomplicated UTIs	73
6.6	Evaluation in Clinical Context	77
6.6.1	Treatment Decision Algorithm	77
6.6.2	Tuning Thresholds and Model Classes	79
6.6.3	Experiment Setup	80
6.6.4	Experiment Results	81
6.7	Summary	84
7	Discussion	85
7.1	Challenge 1: High-Dimensional, Sparse Data	85
7.2	Challenge 2: Non-Stationarity	87
7.3	Challenge 3: Interpretability	88
7.4	Challenge 4: Retrospective Analysis	89
7.5	Limitations and Implications	90
A	Dataset Construction	91
A.1	Feature Extraction	91
A.1.1	Procedures	91

List of Figures

2-1	Clinical Decision Algorithm for Antibiotic Prescription	20
3-1	Coding systems used by MGH and BWH from 2000 to 2016.	31
3-2	An example history of present illness, common in clinical notes.	33
3-3	Levels of resistance based on different breakpoints.	36
3-4	Raw minimum inhibitory concentration values for cefepime.	37
3-5	Antimicrobial susceptibility testing card.	37
3-6	Illustration of filtering for samples from separate infections.	38
3-7	Illustration of windowed features.	39
4-1	Yearly sample volume, by hospital ward.	44
4-2	Yearly volume of samples for each infection site, by hospital ward.	45
4-3	Test methods over the years.	46
4-4	Resistance in MGH ER.	46
4-5	Organisms present in various sites of infection.	47
4-6	Levels of resistance for various bugs and drugs of interest.	47
4-7	Pearson correlations of resistance among drugs for UTIs.	49
5-1	Train and test set splits.	52
5-2	Quantity of prescriptions around the time of specimen collection.	54
5-3	Antibiotic prescriptions in various time windows.	54
6-1	General UTI and Uncomplicated UTI cohort definitions.	59
6-2	Clinical guidelines for uncomplicated UTIs.	62
6-3	Train and test set splits for the UTI cohort.	63

6-4	Average dev set AUCs for the general UTI cohort.	66
6-5	Average dev set AUCs for the uncomplicated UTI cohort.	66
6-6	Development set ROC curves for (one split of) each cohort.	68
6-7	Decision tree for predicting resistance to NIT.	75
6-8	Decision tree for predicting resistance to CRO.	76
6-9	Using an ROC curve to get binary predictions of resistance.	78

List of Tables

3.1	Microbiology table.	32
3.2	Encounters table.	32
3.3	Demographics table.	32
3.4	Diagnoses table.	32
3.5	Procedures.	32
3.6	Labs.	32
3.7	Medications.	32
6.1	Comparison between the general and uncomplicated UTI cohorts. . .	60
6.2	Basic information about uncomplicated UTI drugs of interest.	61
6.3	Performance on general UTIs.	67
6.4	Performance on uncomplicated UTIs.	67
6.5	AMC Features	71
6.6	CRO Features	71
6.7	FEP Features	71
6.8	ATM Features	71
6.9	TZP Features	71
6.10	IPM Features	71
6.11	MEM Features	72
6.12	CIP Features	72
6.13	LVX Features	72
6.14	NIT Features	72
6.15	SXT Features	72

6.16 GEN Features	72
6.17 CIP Features	74
6.18 LVX Features	74
6.19 NIT Features	74
6.20 SXT Features	74
6.21 CRO Features	74
6.22 Development set IAT and spectrum.	82
6.23 Test set IAT and spectrum.	82

Chapter 1

Introduction

In recent years, the adoption of electronic medical records (EMRs) has increased significantly. As a result, it has become possible to perform large-scale analyses on a variety of medical health datasets. This work aims to predict antibiotic resistance at the patient level, leveraging information from EMRs to predict resistance in an accurate and timely fashion. In this chapter, we introduce the empiric antibiotic treatment setting, and motivate the need for knowing patients' resistance to antibiotics. The chapter concludes with an explanation of how the thesis is organized.

1.1 Motivation

Antibiotics are key to many of the achievements of modern medicine. They protect us from infection, safeguard our surgeries, and can be the difference between life and death. Across the world, however, antibiotic/antimicrobial resistance (AMR) has been rising to dangerously high levels. Multi-drug resistant organisms are responsible for more than 23,000 deaths annually in the United States alone [1], and 700,000 deaths worldwide [2]. The situation is exacerbated by over-prescription and misuse of antibiotics, with 20-50% of the antibiotics prescribed by acute care hospitals being either unnecessary or inappropriate [3]. With this context in mind, we choose to narrow in on the setting in which these antibiotics are usually prescribed: the *empiric antibiotic treatment setting*.

1.1.1 Empiric Antibiotic Treatment

Typically, when a patient sees a doctor for an infection, the doctor must make an immediate treatment decision without knowledge of precisely which medications the patient is currently resistant or susceptible to. While the doctor may take a culture from the infected site and order microbiologic testing (see Section 2.1.2), testing usually requires at least two days to yield results. **Thus, the goal of this work is to provide predictions of antibiotic resistance and susceptibility** (for the remainder of the work, this is called ‘resistance’ unless otherwise specified) **for *individual patients* to assist the doctor *at the point of care*.**

1.1.2 The Importance of Knowing Resistance

Knowledge of resistance is useful for many reasons. These include (1) being able to give the patient a treatment that is maximally effective and tolerable, (2) avoiding prolonged antibiotic exposure, and (3) minimizing selection for resistance at a population level.

Effectiveness vs. Tolerability Trade-off

Clearly, the doctor would like to treat the patient with something that is *effective* against the infection. If possible, the treatment should also be *tolerable*, with minimal known negative side effects. Unfortunately, effectiveness does not always align with tolerability, and when there are multiple antibiotic treatments under consideration, doctors must manage a balance of tolerability and uncertainty about effectiveness.

For example, let us consider colistin, a last-resort antibiotic. While colistin is toxic to the kidney, resistance to it is rare. Thus, if a doctor already knew that a patient in critical condition was resistant to everything except colistin, he/she would be aware that tolerability must be sacrificed for effectiveness, thereby avoiding prolonged inappropriate antibiotic therapy.

Avoiding Antibiotic Exposure

At the patient level, prolonged antibiotic therapy is associated with an increased likelihood of resistance in the future. This is even more so than the levels of resistance in a patient's surroundings, also known as colonization pressure [4].

Beyond impact on future resistance, unnecessary antibiotic usage places patient at risk for serious adverse events with no health benefit [5][6][7]. One of the primary concerns is the risk of *Clostridium difficile* (*C. difficile*) infections [8]. Every year, *C. difficile* infections hospitalize almost 250,000 people, and claim the lives of at least 14,000 people in the United States. As asserted by the Center for Disease Control, many of these infections could have been prevented [1].

Population-level Concerns

At a higher level, knowing resistance can help clinicians choose drugs with greater societal benefit. Often times, in the absence of information, clinicians will recommend broad-spectrum antibiotics. However, overuse of these antibiotics increases the risk of bacteria developing resistance to them. In the 1980s, when fluoroquinolones (widely used antibiotics for urinary tract infections caused by *E. coli*) were first introduced, resistance was virtually zero. Today, many parts of the world are finding this treatment ineffective in more than half of patients [9].

By targeting treatment more specifically and opting for narrower-spectrum antibiotics when appropriate, doctors might even be able to limit selection for resistance in the general population. With many factors at play, predictions of resistance would significantly help reduce uncertainty and allow doctors to optimize treatment for both the patient and the population.

1.2 Thesis Organization

The remainder of this thesis is organized as follows. Chapter 2, provides background about both antibiotic resistance and relevant work in machine learning. Next, Chapter 3 discusses how the dataset was constructed from the raw data. From there, Chapter 4 gives a high-level overview of exploratory analyses with the microbiology data. Chapter 5 describes the general experimental setup, followed by Chapter 6 which dives into predicting resistance in urinary tract infections. Finally, Chapter 7 discusses at a high-level various technical challenges and interesting findings that arose, and propose future directions for the project.

Chapter 2

Background

This chapter provides background information about antibiotic resistance, as well as previous work that relates to the goal of the thesis.

2.1 Antibiotic Resistance

Antibiotics have been pivotal in shaping modern medicine, saving millions of lives from bacterial infections that were once lethal [10]. From Sir Alexander Fleming’s 1928 discovery of penicillin to the modern day, however, there has been an ongoing arms race between humans developing novel antibiotics and bacteria¹ co-evolving to develop resistance. Today, we observe alarming rates of resistance [1][9][11], along with a serious lack of new antibiotics under development [12][10].

In their global action plan to combat antimicrobial resistance [11], the World Health Organization lists five strategic objectives:

1. to improve awareness and understanding of antimicrobial resistance
2. to strengthen knowledge through surveillance and research
3. to reduce the incidence of infection
4. **to optimize the use of antimicrobial agents**
5. to develop the economic case for, and to increase sustainable investment in technologies for addressing antibiotic resistance worldwide

¹Note: in this thesis we will use the terms ‘bacteria,’ ‘pathogen,’ and ‘organism’ synonymously.

The main goal of this work is to target the fourth objective (optimized use of antibiotics), which takes place at the point of care in what is called the *empiric treatment setting*.

2.1.1 Treatment Setting

As mentioned in the introduction, *empiric antibiotic treatment* is the common practice of antibiotic prescription before receiving the results of microbiologic testing. Typically, the presenting patient has symptoms of a bacterial infection, and the clinician must make an immediate treatment decision. Although the clinician might order microbiologic testing (discussed in 2.1.2), these results usually take at least 48 hours to return.

Thus, the clinician must rely on presently available vital signs, observations about the patient, patient history, and environmental factors. Based on this information, the clinician hypothesizes about the infectious syndrome, which implies certain pathogens, which have their characteristic antibiotic susceptibilities. Using these antibiotic susceptibilities and prior knowledge such as hospital-level antibiograms (population-level resistance profiles) or antibiotic stewardship program guidelines [3], the clinician makes a treatment decision. Figure 2-1 illustrates the clinical decision-making process for antibiotic prescription. Since this work targets the empiric antibiotic treatment setting, it avoids using any information that the clinician would not have access to at the time of the treatment decision.

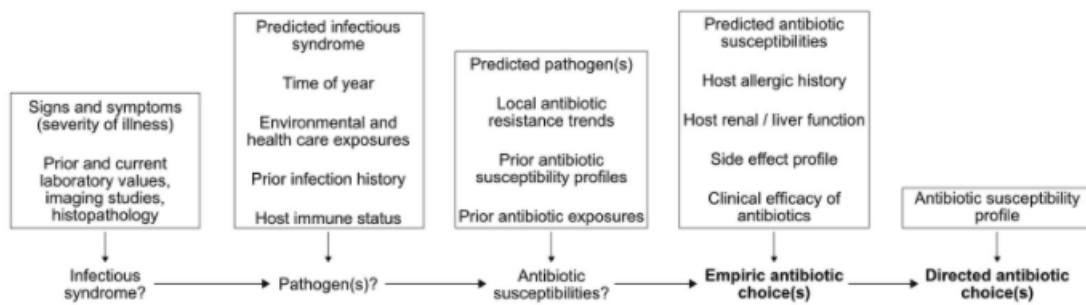


Figure 2-1: Clinical Decision Algorithm for Antibiotic Prescription

2.1.2 Microbiology Lab Testing

Pain, inflammation, swelling, and heat are the four cardinal signs of infection [13]. Upon suspecting an infection, a clinician orders a *culture*, or a specimen taken from a body site to be sent to the microbiology lab. Based on the body site of infection, the clinician makes a hypothesis about the causative pathogen, and the microbiology lab has standard procedures about which media to grow the specimen in.

To measure susceptibility of the provided *isolate* (sample of tissue, blood, etc.), lab technicians use methods which measure organism growth in different concentrations of antibiotic [14], a process which usually takes around two days. In our data, the most common measurement of resistance is **minimum inhibitory concentration (MIC)**, which is the minimum concentration of antibiotic needed to inhibit organism growth. Another commonly used metric in our data is **disk diameter (DD)**, which measures the diameter of no bacterial growth surrounding antibiotic disks placed on plated lawns of bacteria. Finally, these numerical measures are converted into a verdict of **susceptible (S)**, **intermediate (I)**, or **resistant (R)** using clinically determined breakpoints [15]. Although this data reveals the causative pathogen and its antimicrobial susceptibility profile, the results can take days to become available. This information serves as an effective ground truth for our prediction problem.

2.1.3 Mechanisms of Antibiotic Resistance

Below is a brief overview of antibiotic resistance mechanisms at multiple levels: individual bacteria, populations of bacteria, and patients.

Individual Bacteria

Depending on their type, individual bacteria can be *intrinsically resistant* to certain antibiotics; this is often due to the bacteria missing a susceptible target of the specific antibiotic. For example, a drug might target an enzyme or certain cell membrane chemical composition that the bacteria does not have. Bacteria can also *acquire resistance* through mutations in chromosomal genes and horizontal gene transfer [1][16].

Populations of Bacteria

In a population of bacteria, antibiotic overuse introduces a selective pressure, leaving behind the bacteria that are resistant to those antibiotics. As a result, our usage of antibiotics has propagated a vicious cycle of bacterial co-evolution with our antibiotic development [1][10]. However, by reducing the selective pressure, we can encourage the relative growth of susceptible (non-resistant) populations. Development of resistance has a fitness cost to the bacteria, and so susceptible bacteria may be able to out-compete resistant bacteria if the selective pressure from antibiotics were reduced [17]. Thus, beyond keeping resistance from getting worse, reducing unnecessary antibiotic usage might even *reduce* levels of resistance.

Patients

At the patient level, there are many factors that can expose individuals to various pathogens. We simplify and functionally characterize three main types of dynamics:

1. **Acquisition** of a resistant organism: Previous exposure to certain hospital environments or locations may increase the patient's chance of acquiring resistance from an external source.
2. **Emergence** of resistance: Selective pressures may influence the development of resistance. This could be quantified by examining previous exposures to antibiotics, as well as the length antibiotic therapy.
3. **Carriage** of a previously resistant isolate: Previous resistance to other drugs or at other sites of the body might affect resistance at the site of interest.

2.1.4 Predicting Antibiotic Resistance

Much of the previous work in antibiotic resistance prediction is based on statistical genomic studies [18][19][20]. Recently, there have also been promising studies using models such as logistic regression and decision trees to predict resistance in specific subsets of patients, antibiotics, and pathogens.

Sullivan et. al. focused on predicting carbapenem resistance in a cohort of 613 cases of *Klebsiella pneumoniae* bacteremia based on electronic medical records, using logistic regression to find that *Klebsiella* colonization, location, age, previous exposure, and inpatient days were significant features in models that achieved the highest positive predictive values [21].

Guillamet et. al. focused on predicting resistance to three antibiotics of interest (piperacillin-tazobactam, cefepime, and meropenem) in a cohort of 1,618 septic patients, using multivariable logistic regression and decision tree models to predict resistance with area under the receiver operating characteristic (AUROC) curves of 0.6 to 0.8. Their models selected for features such as age, nursing home admission, transfer from outside hospital, prior hospitalization, prior antibiotics, hemodialysis, parenteral nutrition, surgery, presence of a central venous catheter, duration of hospital stay before infection, mechanical ventilation, APACHE II (disease severity) score, the offending pathogen, and the infection source [22]. While these papers assume knowledge of the bacteria species which is not known at the time of empiric prescription, knowing the role of pathogen in these models is interesting from an exploratory standpoint.

There has also been prior work in predicting organism identity in urine cultures; with a cohort of 4,351 patients, MacFadden et. al. conducted various statistical analyses and used a logistic regression model trained on a patient's prior urine culture results to predict organism identity, as well as susceptibility to ciprofloxacin given previously susceptible positive cultures [23].

2.1.5 Other Related Work in Machine Learning for Healthcare

With the increasing adoption of electronic health records, there is an increasing amount of data for clinicians to process. While a clinician may find it overwhelming to scan through all of a patient's vital signs, previous clinicians' notes, and other similar patients' records, machine learning algorithms have performed well at synthesizing a variety of types of data to make specific predictions.

Natural Language Processing

Various methods in natural language processing (NLP) have been used for summarizing and extracting topics from medical texts. Topic modeling techniques such as Latent Dirichlet Allocation [24] have been used to summarize clinicians' notes [25], as well as scientific topics from journal abstracts [26]. There has also been a lot of previous work [27][28] using grammars and vocabulary from sources such as the Unified Medical Language System [29][30] to extract medical concepts from clinical text, achieving precision and recall comparable to multiple experts [31]. For this project, in addition to extraction of specific desired information from clinicians' notes, one could use these notes as a corpus for extracting topics that might be informative of antibiotic resistance.

Irregular Observations

For patients presenting with infection, it is often the case that there are large irregular gaps between medical records, versus patients staying in the intensive care unit (ICU), who might have more regular, frequent measurements taken. Several different approaches have been taken to attempt to fill in missing values, ranging from simply repeating the last value, to multiple imputation [32], to more complicated approaches involving recurrent neural networks that incorporate a balance a decay of the most recent value, and the average value of the variable [33]. In cases where patient data is missing, these techniques may prove helpful to improve predictive accuracy.

Interpreting Complex Models

Especially in the healthcare domain, it is important to be able to interpret why a model is making its decisions. This not only increases confidence in the system, but also could allow clinicians to correct a model when it is based on faulty reasoning. While simple models such as decision trees and logistic regression are quite transparent due to directly observable splits in the tree or coefficients of the model, more complex models such as neural nets are less straightforward to interpret. Text explanation, sensitivity analysis, propagating relevance scores, feature occlusion, saliency maps, and other attention models have all been used towards this goal of interpretability [34][25]. As explained by Lipton in his discussion of model interpretability, however, it is important to define what interpretability means for the particular application. For our application, physicians should be able to interpret the model enough to verify the underlying logic or correlations, and to have the model's insights be helpful in the process of antibiotic prescription.

Chapter 3

Dataset Construction

This chapter covers how the dataset was constructed. First, it describes the available sources of data. Then, it discusses the process of transforming the raw data into a cleaned dataset ready for our learning algorithms.

3.1 Data Sources

Our data comes from the Massachusetts General Hospital (MGH) and Brigham and Women’s Hospital (BWH), whose electronic medical records (EMRs) are managed by the Partners HealthCare Research Patient Data Registry (RPDR) system. Our exported dataset, which spans from 2000 until 2016, includes all of the medical data for anyone who has ever had a culture sent to the microbiology lab within that time range.

3.1.1 Structured Data

From the moment a patient walks into a medical center, to when his/her doctor writes up a discharge summary, the hospital stores a significant amount of multi-modal data into electronic medical records (EMRs). Some of the more structured forms of data are microbiology test results, encounters, demographics, diagnoses, procedures, lab values, and medications.

Microbiology Test Results

As previously explained in Section 2.1.1, microbiology results are usually returned after an empiric prescription has already been made. These results include:

- date that the culture was taken
- location (ward and floor) within the hospital
- underlying pathogen (e.g: *E. coli*, *Klebsiella*, *Salmonella*, etc.)
- phenotype; whether specimen was susceptible, intermediate, or resistant (S/I/R)
- type of test (e.g: MIC, DD) to assess S/I/R, and corresponding test value

In our prediction task, the S/I/R labels serve as the ground truth. For more information about the task setup, see Chapter 5. Chapter 4 discusses trends in the microbiology data.

Encounters

An *encounter* refers to a distinct patient visit. The encounters table contains:

- whether the visit was inpatient (stayed in the hospital) or outpatient
- the attending physician
- admission date
- discharge date (for outpatients, equal to admit date)
- where they were admitted from (e.g: nursing home, emergency room, referral)
- where they were discharged to (e.g: nursing home, rehab, deceased)

Sometimes, a patient may see doctors in multiple parts of the hospital for the same problem, and each of these gets logged as a separate *sub-encounter*. When the patient is finally discharged, someone logs a *master encounter* which encapsulates these sub-encounters. In this work, master encounters are detected and kept, while sub-encounters are discarded.

Demographics

Demographics includes basic patient information, such as gender, birth date, death date (if any), veteran status, language, race, marital status, religion, and location.

Diagnoses

The diagnoses table consists of the codes used to bill for patient care. In addition to the coding system and billing code itself, the table includes a human-readable name of the medical condition, the provider, the clinic, and the date of diagnosis. A patient will often have more than one entry per visit, as it is common to have multiple comorbidities. This work focuses on the well-recognized International Classification of Diseases (ICD-9 and ICD-10) coding systems, but our data also includes diagnoses from additional coding schemes (Figure 3-1a), many of which were home-grown within the Partners HealthCare system.

Procedures

Like diagnoses, procedures are coded for using various billing systems which have evolved over the years (the most common are shown in Figure 3-1b). The procedures table includes the coding system and billing code itself, a human-readable name of the procedure, the provider, the clinic, and the date of procedure.

Lab Values

The dataset also includes lab tests ordered by physicians. These labs capture a wide variety of signals about the patient's bodily state, such as blood counts, lymphocyte levels, and neutrophil levels. Specifically, the data fields include the date at which the specimen was received in the lab, the Logical Observation Identifiers Names and Codes (LOINC) universally used for lab tests, the physician who ordered the tests, and the test results.

Medications

The final source of structured data is the medications. For each medication prescribed to a patient, the EMR system logs the medication code and coding system (Figure 3-1c), medication date, name of the medication, quantity prescribed, provider, clinic, and hospital.

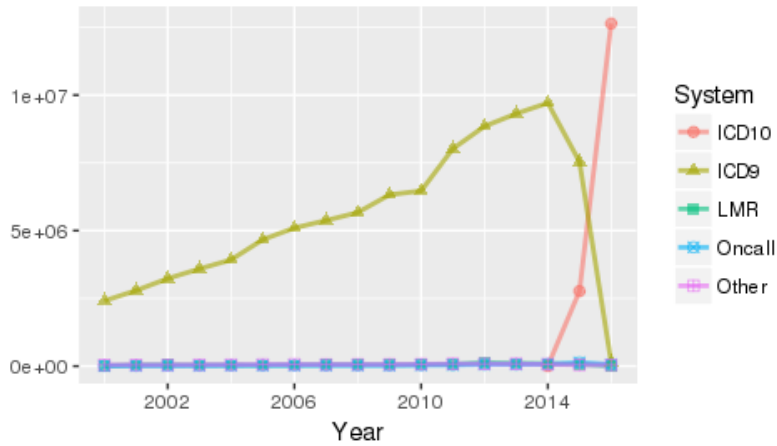
Coding Systems Over the Years

As EMR systems have evolved over the years, so have coding systems. The codes logged in medical records are often used for billing purposes, and might follow some nationally established standards. Other times, the hospital might have its own internal way of coding things. Figure 3-1 shows how usage of different diagnosis, procedure, and medication coding systems has changed over the years in different hospitals.

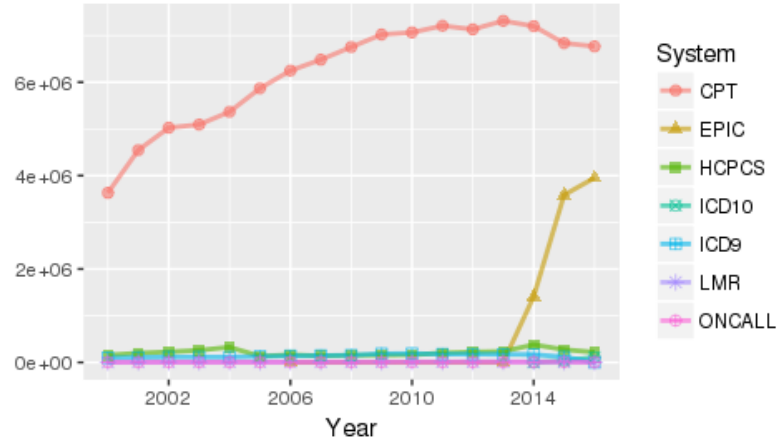
Despite these diverse coding systems and many sources of data, we wanted to derive a large unified set of features. One option was to map all of the raw data into one common schema such as the Observational Medical Outcomes Partnership (OMOP) Common Data Model, or the Unified Medical Language System (UMLS). However, mapping to these coding systems requires significant work from domain experts, and is outside the scope of this work. Instead, we examined each data source and determined the most effective method of extracting the desired information:

- Diagnoses data was restricted to ICD-9 and ICD-10 codes, which encode approximately 97.7% of diagnoses in our dataset. As shown in Figure 3-1a, the transition from ICD-9 to ICD-10 codes began around 2014.
- Procedures were extracted through a combination of CPT code matching, ICD9 code matching, and string matching with manual verification of the matches (see Appendix A.1.1). Figure 3-1b shows that CPT codes are the most commonly used, but that in recent years the EPIC coding system has become more popular.
- Medications varied greatly in the codes used. Many different names and codes often corresponded to the same medication, and we were unable to obtain access to the underlying hierarchy of the most prevalent codes. In this work it is very important to process the medications correctly, and so our clinical collaborator Dr. Sanjat Kanjilal manually and exhaustively reviewed all medication names, selecting those which corresponded to antibiotics of interest.
- Lab codes have been relatively consistent, so no special treatment is discussed in this section.

(a) Number of diagnoses encoded by each system vs. year



(b) Number of procedures encoded by each system vs. year



(c) Number of medications encoded by each system vs. year

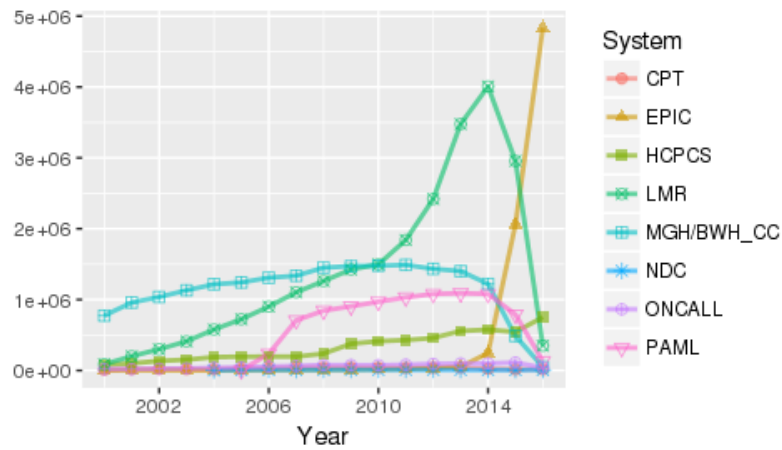


Figure 3-1: Coding systems used by MGH and BWH from 2000 to 2016. Note that medication notes are encoded much more heterogeneously than diagnoses or procedures.

Summary of Structured Data

In summary, there are many sources of structured data for our predictive models to process and combine. Tables 3.1 through 3.7 outline the major fields in each structured data table.

micro
specimen date
location in hospital
pathogen
S/I/R phenotype
micro test type / test value

Table 3.1: Microbiology table.

master_encounters
inpatient/outpatient
attending physician
admit date
discharge date
admit from / discharge to

Table 3.2: Encounters table.

patient_lookup
gender
date of birth & death (if any)
language, race, religion
marital status, is veteran
location

Table 3.3: Demographics table.

diagnoses
name of condition
billing system & code
diagnosis date
inpatient/outpatient
provider, clinic

Table 3.4: Diagnoses table.

procedures
name of procedure
billing system & code
procedure date
quantity
provider, clinic

Table 3.5: Procedures.

labs
specimen receipt time
lab description
lab code
lab result
ordering physician

Table 3.6: Labs.

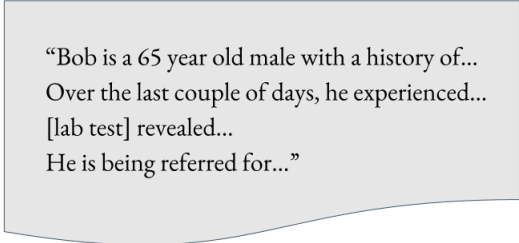
medications
medication name
coding system & code
medication date
quantity prescribed
provider, clinic

Table 3.7: Medications.

3.1.2 Unstructured Data

In addition to structured data, there is unstructured data present in the form of natural language clinicians' notes. Clinical notes are incredibly rich with relevant information about the presenting patient, and also provide a window into the clinician's thought process. Across 2000-2016, we have approximately 19.2 million outpatient notes (116 GB), and 3.2 million inpatient notes (42 GB).

Elements of a Note



“Bob is a 65 year old male with a history of...
Over the last couple of days, he experienced...
[lab test] revealed...
He is being referred for...”

Figure 3-2: An example history of present illness, common in clinical notes.

Clinicians' notes piece together many different types of information into a cohesive patient story. Below are some common elements of clinicians' notes:

- history of present illness (Figure 3-2)
- relevant/recent structured data (e.g: labs, diagnoses, medications, procedures)
- treatment recommendations/directions
- dialogue with the patient (could include events in life, timing of symptoms, ...)
- clinician's impression and observations of the patient
- description of the patient's physical, mental, and/or emotional state
- allergies, vaccines
- social history and family history

Note that much of this information is not readily extracted from structured data, and sometimes not even possible to extract. In the next section, preprocessing and feature extraction are discussed.

3.2 Preprocessing

Dataset construction involved careful consideration of the following:

1. Precisely what we **want to predict**
2. What information is available **at the time of empiric treatment**
3. What data in EMRs might be **informative** of antibiotic resistance

Based on these considerations, various data-driven design decisions were made along the preprocessing pipeline. This section discusses label (target) extraction, data filtering, and feature extraction.

3.2.1 Extracting Labels from the Microbiology Data

The ultimate goal of the project is to provide clinicians with a prediction of antibiotic resistance for use in the empiric treatment setting. However, there are multiple ways one could define “resistance.”

Phenotype vs. Raw Value

As previously mentioned, resistance is typically measured using techniques such as minimum inhibitory concentration (MIC) or disk diameter (DD), which yield raw numerical values that are then converted into phenotypes (S/I/R) using published cutoffs. While these raw numbers preserve information about *how* resistant a patient is, they are not directly comparable between testing methods (which differ in units and measurement scales). Both for simplicity and in order to have a common language between different testing methods, we decide to *predict phenotypes* instead. To further simplify the problem, the intermediate (I) phenotype is considered resistant (R), as this is what is done in practice.

Extracting Phenotypes

The raw microbiology data already contains the phenotype *at the time of prescription*. However, over time the standards for S/I/R have changed to become more

conservative, and so the same sample which is currently considered ‘resistant’ might have been labeled as ‘susceptible’ in the past (i.e. diameter or concentration thresholds used to be different). This is a problem with label quality, since we want to predict some ‘true’ resistance probability rather than what the published standards considered resistant at the time. Luckily, the microbiology data also contains the numerical test values (MIC/DD), and so we *retrospectively apply the published 2017 CLSI breakpoints* [15]. These breakpoints are applied across all antibiotics, with the exception of specific bug-drug combinations for which Dr. Sanjat Kanjilal applied additional clinical guidelines that take into account other clinical context.

Quality-Checking the Extracted Phenotypes

As shown in Figure 3-3, application of breakpoints led to unexpectedly high levels of resistance from 2000-2006. Through examination of the computed phenotypes, it became apparent that *shifts in testing methodology that made phenotypes from prior to 2007 incomparable to those afterwards*. More specifically, unexpected resistance levels were caused by: (1) less conservative/ higher breakpoints in the past, and (2) upgrades in 2006 which allowed MIC-measuring machines to measure lower concentrations.

Figure 3-4 shows an example of how the measured MICs for the cefepime (FEP) antibiotic changed across the years. As demonstrated by Figure 3-3, which shows that the phenotypes logged in the raw data (based on old breakpoints) were mostly susceptible (S), the breakpoints at the time were more conservative (higher threshold for marking a sample as resistant).

Figure 3-5 shows an antimicrobial susceptibility testing card. Different rows correspond to different concentrations of antibiotics being tested, and we hypothesize (and checked against our data) that in 2006 an additional row was added to allow testing of lower concentrations. For certain antibiotics, these lower concentrations were the only concentrations that fell below 2017 breakpoints.

For simplicity’s sake, we decide to *exclude samples from before 2007*. While special handling of data from 2000-2006 could still allow us to use these samples, there is still a significant volume of samples from 2007 and after.

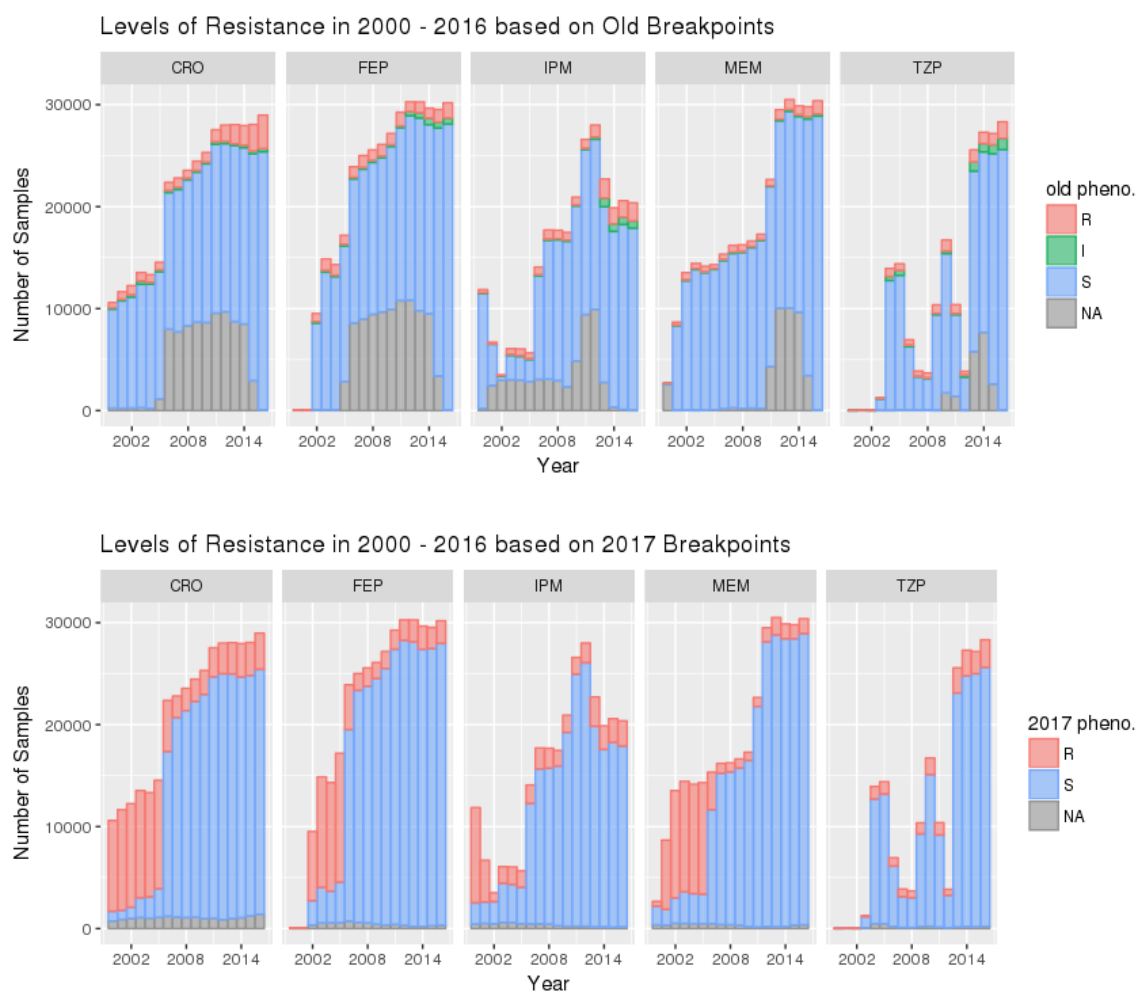


Figure 3-3: Levels of resistance based on different breakpoints. Note that based on 2017 breakpoints, the levels of resistance from 2000-2006 are much higher than reported based on the old breakpoints. Also note that depending on the year, certain drugs (e.g. TZP testing dips in 2007-2008, as well as 2012) are tested for in different quantities.

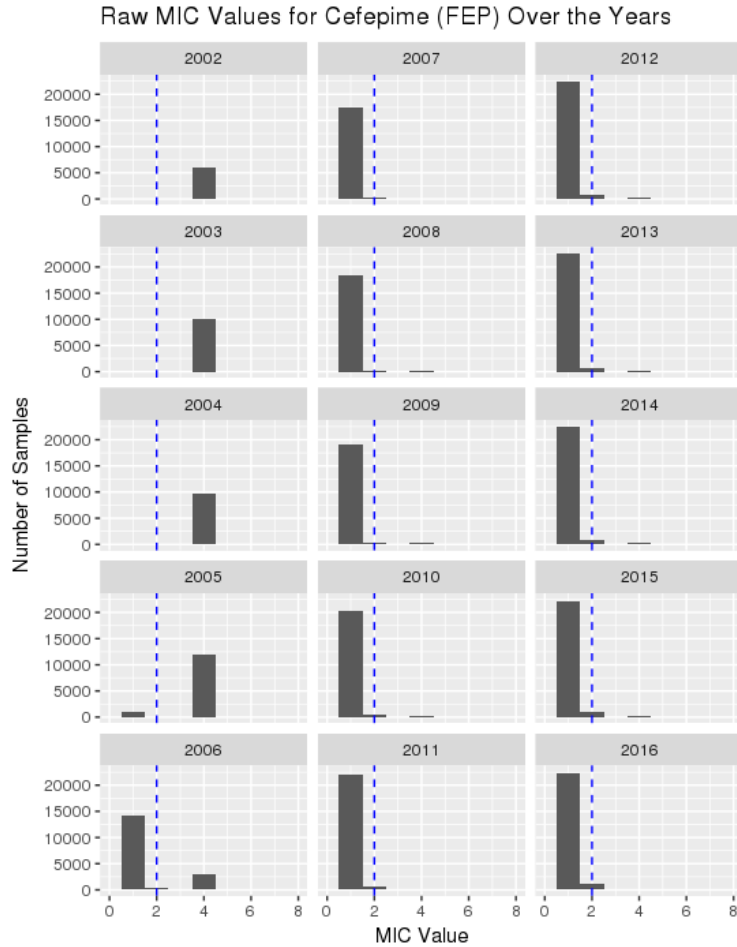


Figure 3-4: Raw minimum inhibitory concentration (MIC) values for cefepime (FEP). Over the years, the raw values have shifted due to upgraded testing equipment. Blue dotted lines mark the 2017 breakpoint, where values to the left are susceptible and values to the right are resistant. Due to upgrades in testing equipment, values prior to 2007 are mostly considered resistant by today’s standards.

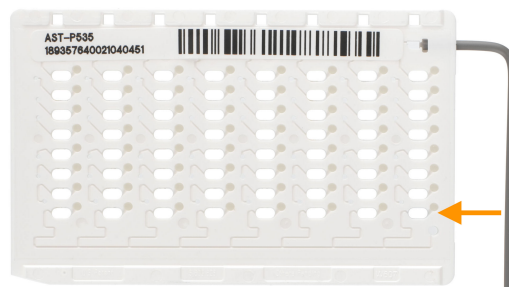


Figure 3-5: VITEK[®] antimicrobial susceptibility testing (AST) card. An upgrade in AST cards involved adding an additional row to the card, which allowed for testing of lower concentrations than before.

3.2.2 Filtering the Microbiology Data

Now that we have discussed *what* our labels are, we discuss *which* labels to keep. More microbiology data is available retrospectively than at the time of empiric treatment, so this data must be filtered in a way that is realistic to the deployment setting.

As discussed in the previous section, 2000 - 2006 had different testing equipment. Thus, we *filter to samples from 2007 - 2016*, and then *apply 2017 breakpoints*.

Especially for inpatients, multiple micro samples can be taken for a single infection. To mimic the empiric treatment setting, we would like to make predictions for the *first sample from an infection*. While a single infection can span across multiple encounters, clinicians generally consider separate infections as those with at least 14 days between them. Furthermore, the “same” infection should have the same recorded site of infection. Thus, we sort each patient’s samples chronologically and only take samples at least 14 days after any previous sample with the same site of infection. Same-day samples are merged or kept separate, depending on same or different sites of infection (respectively). This process is illustrated in Figure 3-6.

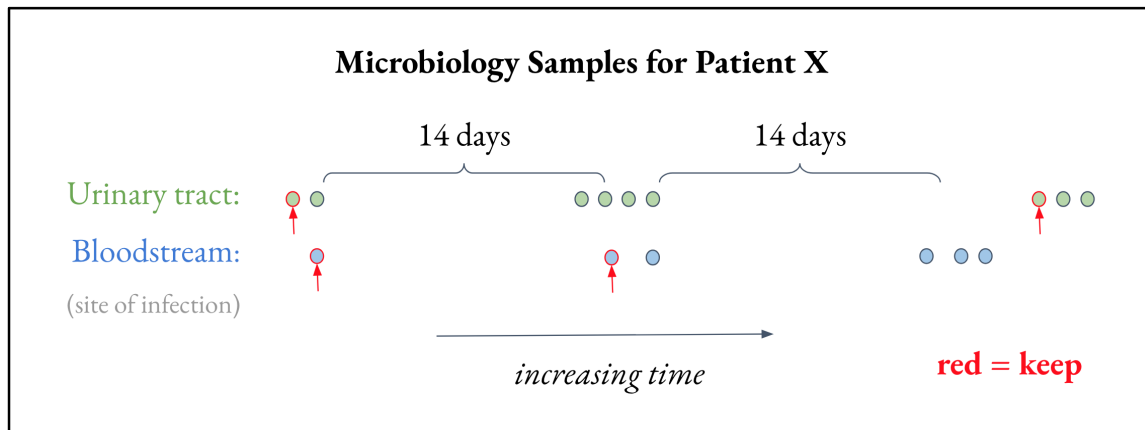


Figure 3-6: Illustration of filtering process. Keeps first sample from each of a patient’s separate infections (different site of infection or ≥ 14 days from the previous sample).

In summary:

1. Samples are filtered to those from 2007 - 2016.
2. 2017 CLSI microbiology breakpoints are applied to get new S/I/R phenotypes.
3. Samples are filtered such that each one is the first in a patient’s infection.

3.2.3 Feature Extraction

While there are many types of data available, the learning algorithms used in this work are ultimately fed a single consolidated design matrix (feature matrix). This section describes how features are extracted from their raw sources.

Types of Features

Broadly speaking, there are binary features, which indicate the presence or absence of something, and numerical features, which are normalized scalar values. Features can either be extracted from the visit at the time of sample collection, or computed over some time range in the patient's the medical history (these are called *windowed features*). Except for previous pathogen, previous antibiotic resistance, and colonization pressure, windowed features are computed over the last 7, 14, 30, 90, and 180 days, up until the day that the specimen/culture was taken. Previous pathogen and resistance are computed over the last 14, 30, 90, and 180 days, up until 7 days before the specimen date. Colonization pressure is computed over the last 90 days, up until 7 days before the specimen date. Figure 3-7 gives an example of how windowed features are computed.

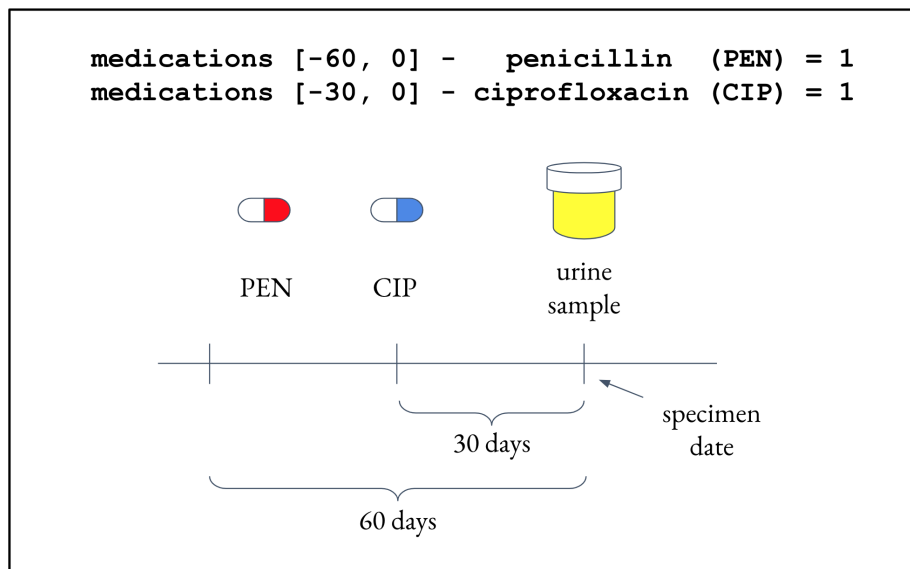


Figure 3-7: Illustration of windowed medication features. Binary windowed features are binary indicators over some window relative to the culture date.

Features from Structured Data

As described earlier in the chapter, our structured data sources include both past and present microbiology, encounters, demographics, diagnoses, procedures, lab values, and medications. Based on previous literature and Dr. Sanjat Kanjilal’s clinical intuition, the following features are extracted (B = binary, N = numerical, W = windowed):

- **Microbiology:** previous resistance (BW), previous pathogens (BW), year (N), hospital ward (B), hospital floor (B), colonization pressure (N, *see below*)
- **Encounters:** number of inpatient master encounters (BW)
- **Demographics:** gender (B), age (N), race (B), is veteran (B)
- **Diagnoses:** Elixhauser comorbidities¹ extracted from diagnosis codes (BW), and whether the patient was pregnant (BW)
- **Procedures:** specific types of procedures extracted from procedure description and procedure codes: mechanical ventilation (BW), parenteral nutrition (BW), central venous catheter (BW), surgery (BW), hemodialysis (BW)
- **Labs:** neutrophils (NW), white blood count (NW), and lymphocytes (NW), averaged over each window, imputed with the patient’s average value (if no labs at all, imputed with 0).
- **Combination of sources:** nursing home was extracted from the `admit_from` field in encounters, as well as procedures relevant only to those in nursing homes

Colonization pressure is defined as the proportion of resistance among previous samples taken from the same location. Location is defined at a high level by hospital ward (inpatient, outpatient, emergency room, or ICU), and at a more granular level by hospital floor (specific clinic, East Wing of a hospital, urology floor, etc.).

¹The Elixhauser Comorbidity Index categorizes patients’ comorbidities into 31 categories based on ICD9 and ICD10 codes. Categories include: paralysis, alcohol abuse, HIV/AIDS, lymphoma, weight loss, drug abuse, congestive heart failure, etc. [35]

Features from Unstructured Data

Unigrams and bigrams from both inpatient and outpatient notes are extracted, using a bag of words representation (vector of 1's and 0's, with 1's for words that show up in a note) to represent each note. Due to the enormous size of the unigrams and bigrams vocabularies, we also filter out stop words, and threshold for unigrams/bigrams that show up in at least 1% of notes and at most 95% of notes. The intuition behind this is that words that are too common (e.g: “infection”, “hospital”, etc.) or too rare (e.g: patient names, addresses, unique identification codes) might not provide as much signal as those that show up a moderate amount of the time, yet they would dramatically increase the dimensionality of our feature vectors.

Combining Structured and Unstructured Features

Structured and unstructured features are concatenated into one large vector for each patient, features with zero variance are filtered out, and numerical features with range outside of 0-1 are transformed to have zero mean and unit variance.

Summary: This chapter explained the available data sources, data-driven design choices made when constructing the dataset, and finally the features that were extracted. The next chapter digs into the microbiology data and attempts to characterize the nature of our prediction target.

Chapter 4

Exploratory Analysis of Micro Data

This chapter overviews exploratory analyses of the microbiology data. First, we quantify how much data is available, and where it comes from. Then we examine both resistance levels and the causative pathogens, which (as explained in Section 2.1.1) are a key consideration in the empiric antibiotic treatment setting. Finally, we summarize the main findings. All of the following analyses are reported on the dataset filtered according to the procedure described in Chapter 3 (samples from 2007 to 2016, from separate infections).

4.1 Sample Volume

In this section, yearly sample volume is broken down by which ward it came from, as well as where in the human body it came from.

4.1.1 Volume by Hospital Ward

From 2007 to 2016, the number of distinct patients with samples in our dataset has risen by approximately 8%, from about 23.5K to 25.5K distinct patient IDs per year. Over the same period of time, the number of distinct microbiology samples per year has risen by about 0.8%, from 36.4K to 36.7K. On average, each patient has approximately 2 samples, with the vast majority of patients (about 115K patients)

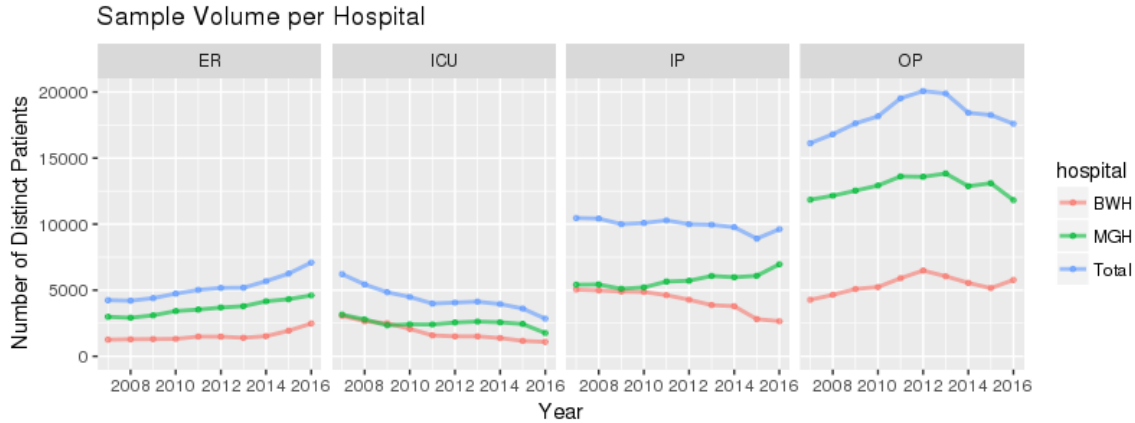


Figure 4-1: Yearly sample volume, by hospital ward.

having one sample from a distinct infection. The maximum number of samples from ‘separate infections’ (defined in Chapter 3) a patient has had per year is 31. Figure 4-1 breaks down the yearly trends in volume of microbiology samples for each hospital and hospital ward. The yearly trends for patient volume look similar.

Note that the majority of samples come from outpatient (OP) wards. Additionally, the volume of emergency room (ER) samples has been increasing over the years, and while the volume of inpatient (IP) samples as been increasing for MGH, for BWH it has been decreasing in recent years.

4.1.2 Volume by Site of Infection

Different hospital wards see different subpopulations of patients, which often have different classes of infections. The type of infection is a key determinant in the considerations that inform a doctor’s suggested treatment regimen. As shown in Figure 4-2, the most dominant infection in the ER, inpatient wards (IP), and outpatient wards (OP) is urinary tract infections. In intensive care units (ICU), respiratory tract infections are more common. Skin and soft tissue infections are more commonly observed in outpatient locations, whereas bloodstream infections are more common in the emergency room. In the ER, it appears that there has been a dramatic increase in urinary tract infections, along with an increase in bloodstream infections in recent years. Overall, about 48% of samples come from urinary tract infections, 12% from

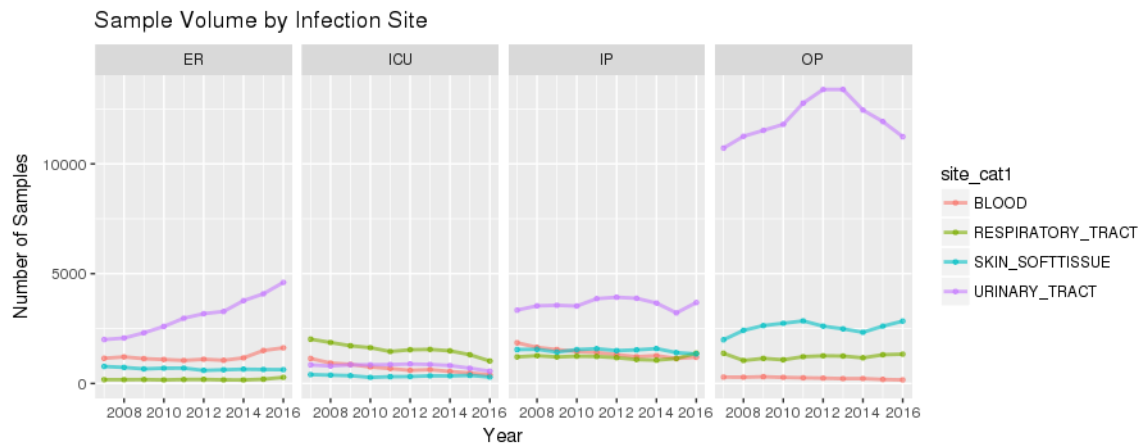


Figure 4-2: Yearly volume of samples for each infection site, by hospital ward.

skin and soft tissue infections, 10% from respiratory tract infections, and 8% from bloodstream infections.

4.2 Levels of Resistance

Levels of resistance vary across many different clinics, antibiotics, types of patients, and periods of time. Over the years, testing methodology has also changed, shifting from DD-based methods to predominantly MIC (Figure 4-3). In addition to upgrades in testing equipment and annually updated breakpoints, changing protocols have also moved different sets of antibiotics in and out of the standard sets of drugs to test for resistance. With these limitations in mind, we explore the levels of resistance that appear in our data.

Depending on the antibiotic, type of infection, and causative pathogen, clinicians can expect very different levels of resistance. As shown in Figure 4-4, levels of resistance in the MGH emergency room (ER) have changed over time, and different drugs have various base levels of resistance. For instance, among the samples tested from the MGH ER, resistance to penicillin and oxacillin, while still high, has decreased over the years, while resistance to nitrofurantoin (NIT) and ceftriaxone (CRO), while low, has increased.

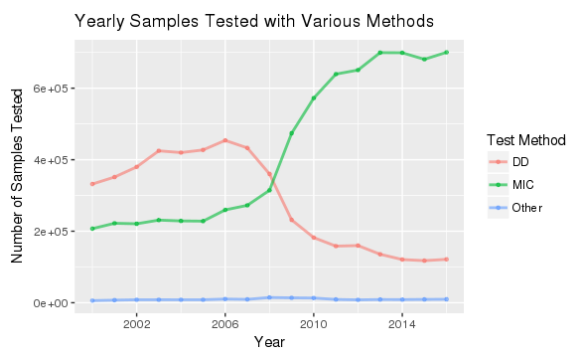


Figure 4-3: Test methods over the years. Minimum inhibitory concentration (MIC) is now the dominant method.

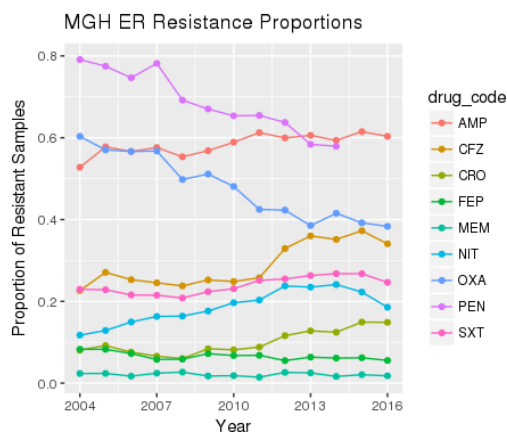


Figure 4-4: Resistance in MGH ER.

4.2.1 Pathogens' Resistance

At a fundamental level, the underlying phenomenon being measured is how resistant some pathogen (organism) is to various drugs. Some pathogens are more commonly associated with certain sites or types of infections. Figure 4-5 plots the presence of various pathogens at different sites of infection over the years. It shows that the dominant type of pathogen in bloodstream infections is *Coagulase-negative Staphylococcus* species, in respiratory and skin and soft tissue infections is *Staphylococcus*, and in urinary tract infections is *Escherichia*.

Different drug classes are active against different pathogens. Figure 4-6 shows levels of resistance in samples from urinary tract infections (UTIs), for three major pathogen genres, to various drugs of interest. NIT consistently has one of the lowest proportions of *Escherichia* resistance, and therefore would be quite effective against *Escherichia* (which are the most common pathogens in UTIs). However, other organisms in the *Enterobacter* and *Klebsiella* genres are highly resistant to NIT. Thus, one would expect that information about the organism, although unavailable at the time of empiric treatment, would be quite helpful in predicting resistance.

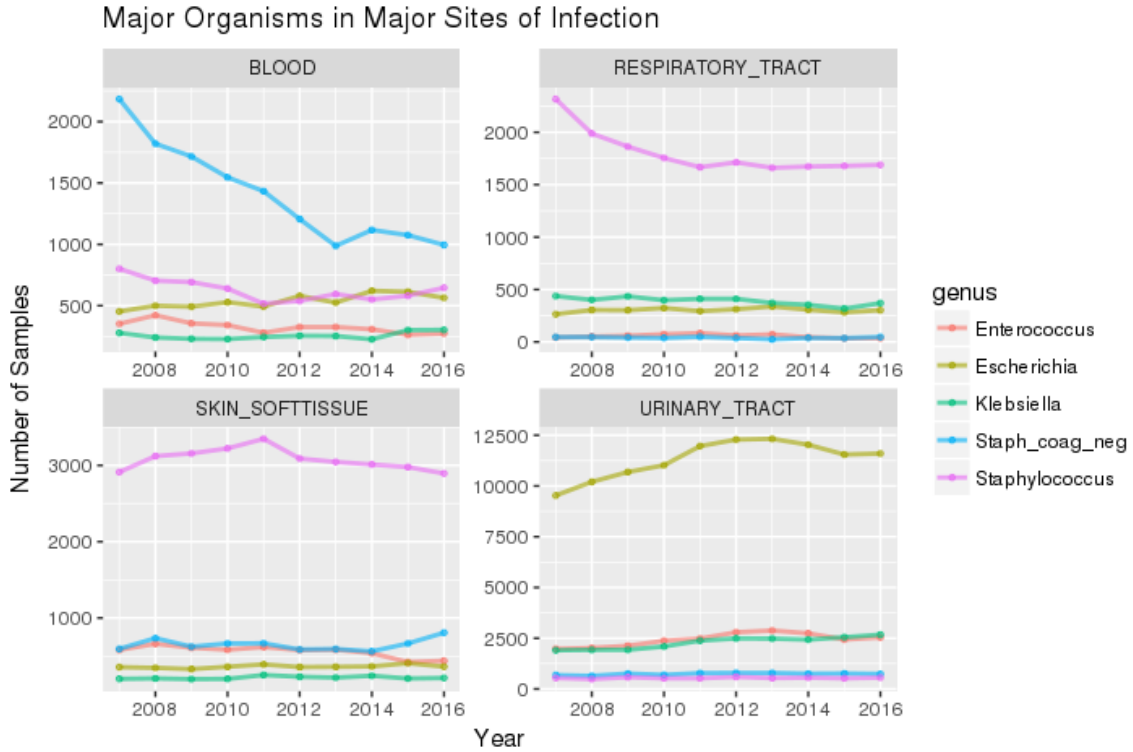


Figure 4-5: Organisms present in various sites of infection.

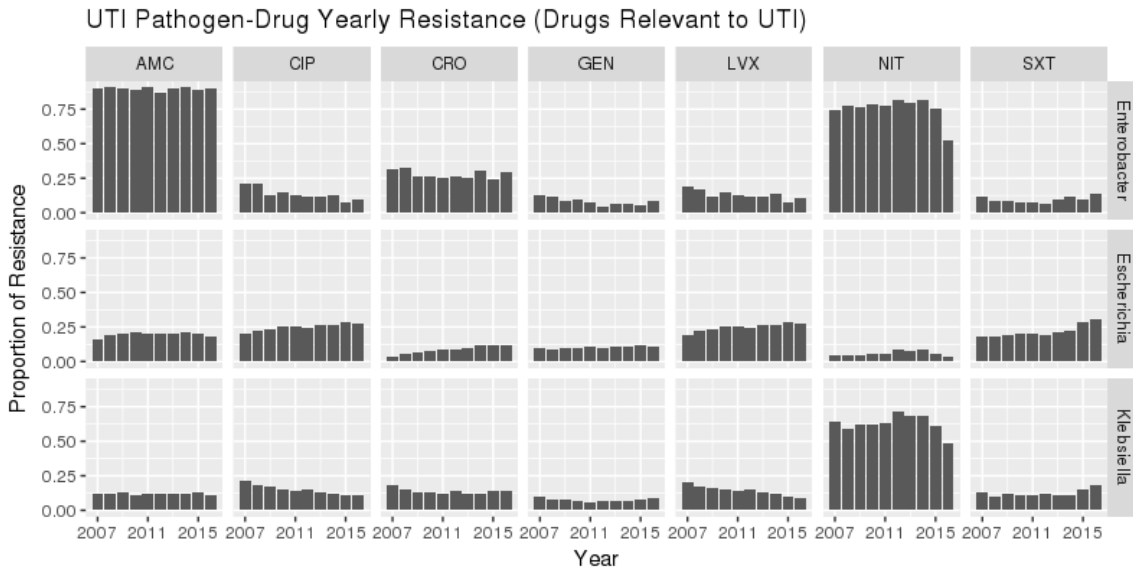


Figure 4-6: Levels of resistance of *Enterobacter*, *Escherichia*, and *Klebsiella* to various drugs relevant to treatment of urinary tract infections.

4.2.2 Correlations in Drug Resistance

Since drugs sometimes operate on the same biochemical pathway, certain pathogen mutations are known to convey resistance to multiple drugs. As shown in Figure 4-7, which plots how resistance to various drugs are correlated with each other, the strongest correlations are between ciprofloxacin (CIP) and levofloxacin (LVX), ceftipime (FEP) and meropenem (MEM), and imipenem (IPM) and meropenem (MEM). CIP and LVX are both members of the fluoroquinolone class, and FEP, MEM, and IPM are all beta-lactams.

The correlation between CIP and LVX is entirely expected. All members of the fluoroquinolone class act similarly, inhibiting certain bacterial enzymes to inhibit DNA replication and eventually lead to cell death. Resistance to fluoroquinolones arises from point mutations in regions encoding these target proteins, which confer resistance to all members of that antibiotic class. Therefore, selection of resistance to ciprofloxacin will always be expected to confer cross-resistance to levofloxacin.

Mechanisms explaining the correlation of resistance across different classes of antibiotics are less well understood, but are likely a combination of pleiotropic effects of antibiotic resistance mechanisms (when one gene influences multiple seemingly unrelated phenotypic traits), as well as co-localization of resistance genes on common mobile genetic vectors such as plasmids and integrons.

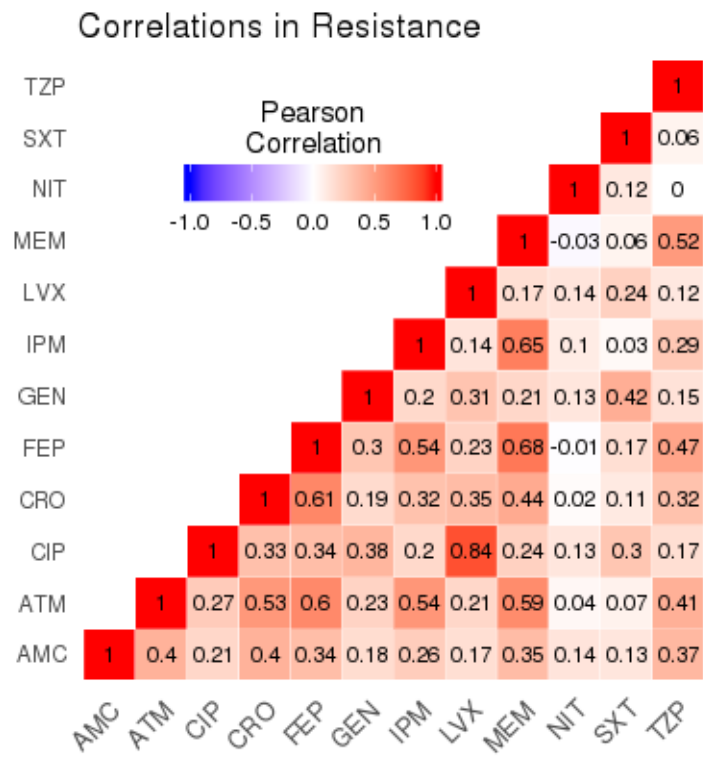


Figure 4-7: Pearson correlations of resistance among major drugs used to treat urinary tract infections. The strongest correlations are observed between CIP and LVX, FEP and MEM, and IPM and MEM.

4.3 Summary

Below are the key observations to take away from this chapter's exploratory analyses:

- Most of the patients in our data are outpatients.
- Over the years, samples collected at MGH and BWH have grown in volume and also changed in composition.
- The most common infection site is the urinary tract, followed by skin and soft tissue, respiratory tract, and bloodstream.
- Depending on the hospital ward, certain patient populations and infection sites are more common.
- Different antibiotics have varying levels of resistance. As observed in the collected microbiology samples, some have risen in proportion of resistance while others have decreased in resistance.
- Resistance to a particular drug is highly dependent on the underlying pathogen.
- Some drugs are correlated in their resistance patterns.

Over the years, the underlying populations of patients, samples, and pathogens have evolved. While some of these trends could be incorporated into modeling considerations (e.g: growing levels of resistance to penicillin, gradual changes in pathogenic composition of blood samples, etc.), other changes may be temporary fluctuations that may not generalize well without additional modeling considerations (e.g: when drugs are taken in and out of testing panels, or when testing standards change). Many of the aspects investigated in this chapter are sources of non-stationarity, and it is important to keep these in mind when conducting our experiments.

Chapter 5

General Experiment Setup

Most of our experiments fall within a single unified framework for making predictions. This chapter describes: (1) the setup of the prediction problem, (2) how data is partitioned into train and test sets, (3) how empiric prescriptions are extracted from raw medications data, and (4) which model classes are used for prediction. The next chapter walks through a more detailed account of experiments done within this framework using a specific cohort.

5.1 The Prediction Task

After constructing feature matrices and binary labels of S/R as described in Chapter 3, we are left with a multi-task binary prediction problem. That is, for each sample which is represented by a feature vector, the goal is to *predict probabilities of resistance to each drug of interest*. For simplicity, in this project separate models are trained to predict resistance to each antibiotic separately, but a logical extension of the work could predict resistance to multiple antibiotics jointly.

Additionally, patients are sometimes sent home without prescriptions because the infection is expected to resolve on its own. Since predictions of resistance are irrelevant in these situations, and such patients are identifiable by the clinician at the point of care, we *filter out patients without any antibiotics empirically prescribed* (this is operationally defined in Section 5.3).

5.2 Train and Test Splits

Data is partitioned into four disjoint sets of microbiology samples. As shown in Figure 5-1, distinct patient IDs are first split into disjoint sets, with 80% of patient IDs for training/development (train/dev), and 20% for testing. Then, to evaluate how well our models might generalize to the future, the data is split according to a train/dev time range of 2007-2013, and a test time range of 2014-2016.

All model tuning and design decisions are made based on dataset 1 in Figure 5-1, which comprises approximately 56% of all filtered microbiological samples from 2007-2016. Dataset 1 corresponds to the train/dev set for both patient IDs and time ranges, while dataset 2 corresponds to test patient IDs and train/dev time ranges, dataset 3 corresponds to train/dev patient IDs and test time ranges, and dataset 4 corresponds to test patient IDs and test time ranges. This thesis only reports test values for dataset 3, but future work will report values on the remaining test datasets as well.

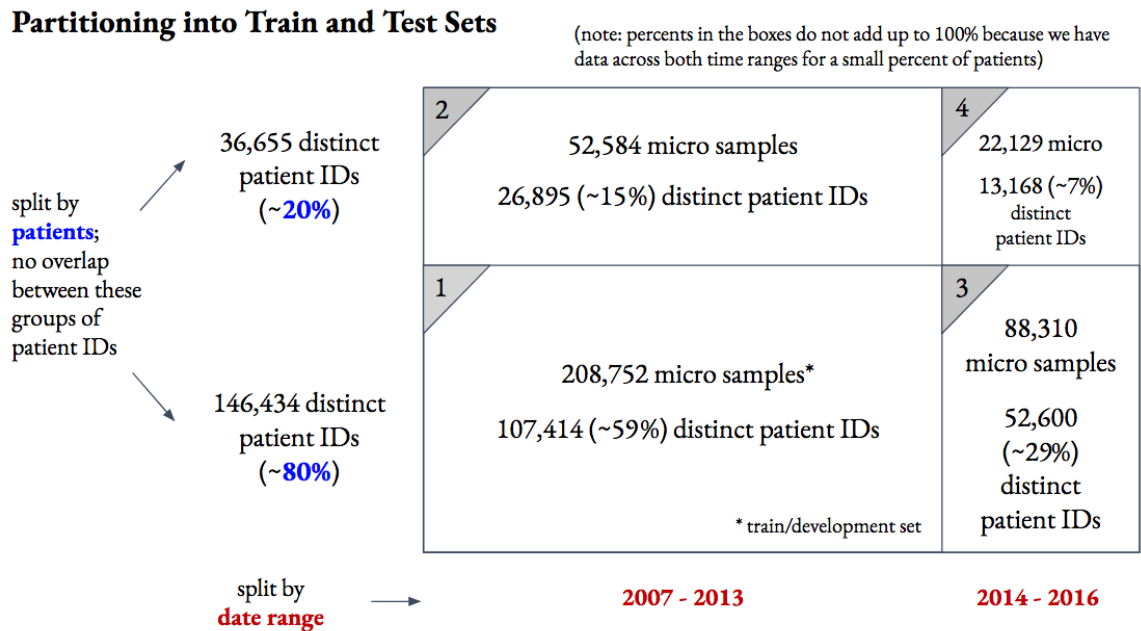


Figure 5-1: Partitioning of data into training/development and test sets, based on an 80-20 split of patient IDs and time intervals of 2007-2013 and 2014-2016.

5.3 Defining Empiric Prescriptions

In order to improve antibiotic prescription practices in the empiric treatment setting, we must first characterize current clinical practices. Based on exploratory analyses and knowledge of testing procedures at MGH and BWH, **we define empiric antibiotic prescriptions as any antibiotic medications from two days before to one day after the microbiologic specimen was collected.**

Figure 5-2 plots the number of antibiotics prescribed in the temporal vicinity of any specimen from a urinary tract infection (UTI). Before the date of specimen collection, there are relatively few prescriptions of antibiotics. On the same day as specimen collection, there is a large jump in prescriptions, followed by an immediate drop in prescriptions on the next day. The remaining days have further decreasing numbers of prescribed antibiotics.

To sanity-check our definition of empiric prescriptions, Figure 5-3 plots the distribution of drug classes for antibiotics prescribed in various time windows near the specimen collection date. A homogeneous group of patients (uncomplicated UTI patients) is used to narrow to a well-defined clinical scope. As shown in Figure 5-3, empiric prescriptions (in the time window of -2 days before collection to 1 day after) are primarily fluoroquinolones, which in this context are typically broad spectrum antibiotics such as ciprofloxacin and levofloxacin. The second-most prescribed class is folate inhibitors, which includes a relatively common narrow-spectrum antibiotic called trimethoprim-sulfamethoxazole. Comparing time windows $[-2, 1]$ and $[4, 14]$, the relative number of folate-inhibitor and nitrofurantoin prescriptions (both narrow-spectrum) increases greatly after microbiology results have come back. This makes intuitive sense because healthcare providers might be more willing to prescribe narrow-spectrum antibiotics (which generally have higher levels of resistance) after confirming that a patient is not resistant to them.

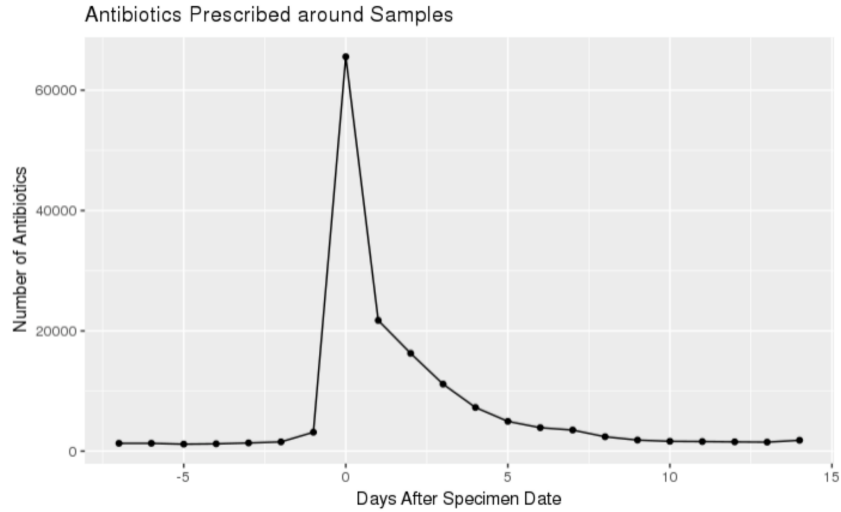


Figure 5-2: Antibiotics prescribed within a -7 day to +14 day time frame relative to the date of specimen collection.

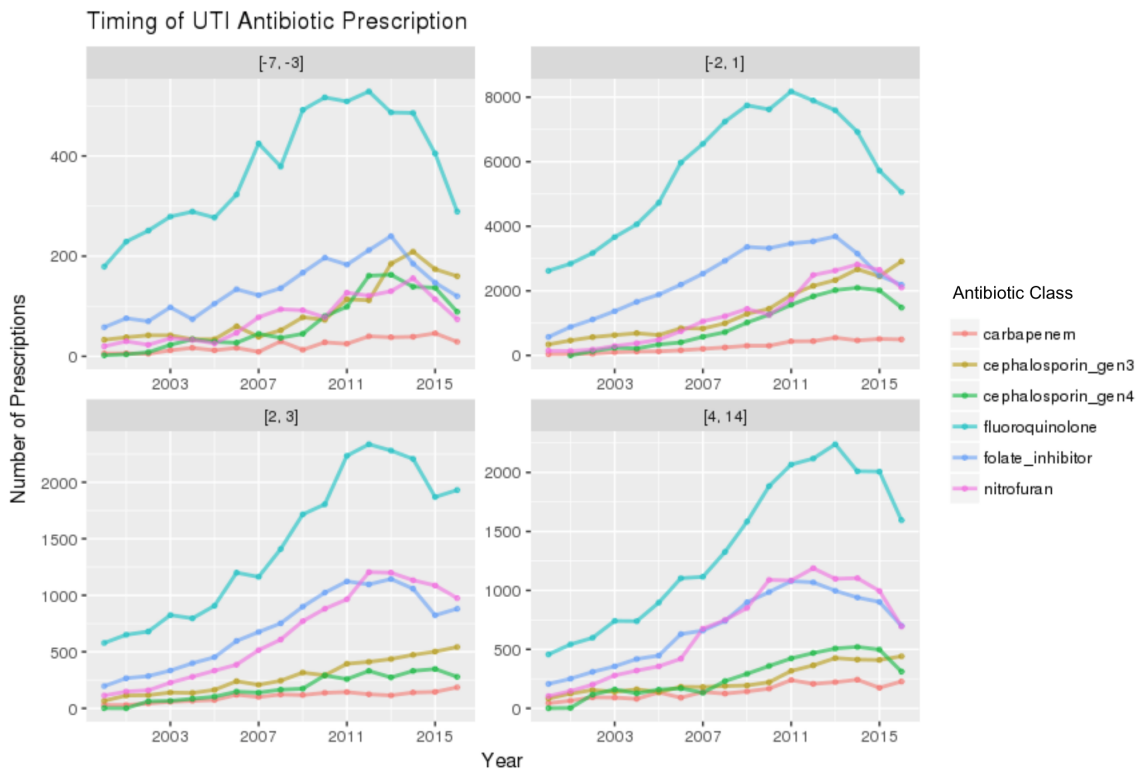


Figure 5-3: Prescription of various antibiotic classes to patients with urinary tract infections, in different time windows relative to specimen culture collection date. From 2000 to 2016, fluoroquinolones have been the most commonly prescribed, followed by folate inhibitors and nitrofurans. Also note that the proportion of folate inhibitor and nitrofurantoin prescriptions (narrow-spectrum antibiotics) increases in the [4, 14] window, when compared versus the [-2, 1] window.

5.4 Models

To predict whether a sample is resistant to a given drug (binary classification), logistic regression, decision trees, and random forests are used. These standard model classes are often more easily interpreted than more complicated models, while still achieving good performance and allowing for reasonable flexibility. We also experimented with shallow feed-forward neural networks initially, but they usually did not perform as well as logistic regression, were less interpretable, and took too much time to tune.

5.4.1 Logistic Regression

Logistic regression attempts to model the posterior probabilities of K classes. It is a generalized linear model which comes from the desire to have linear functions in some explanatory variable(s) x , while still ensuring that the probabilities remain in the range $[0, 1]$ and sum to one. In our problem setup, we would like to perform binary classification (where 1 is ‘resistant’ and 0 is ‘susceptible’), so $K = 2$.

To convert some real input $t \in \mathbb{R}$ to a probability value between 0 and 1, the logistic function $\sigma(t)$ is used:

$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

If we are modeling some probability $p(x)$ as a function of explanatory variables x and assume that t is a linear function of x , then we have

$$p(x) = \sigma(t(x)) = \sigma(\beta^T x) = \frac{1}{1 + e^{-\beta^T x}}$$

for some constants β , where x_0 is assumed to be a constant, such that β_0 is an intercept term. Rearranging, we get:

$$\ln \left(\frac{p(x)}{1 - p(x)} \right) = \beta^T x$$

This is the log-odds, also known as the logit. To fit the parameters β , maximum likelihood estimation is usually used. To interpret the coefficients β , note that a

positive β_i implies that an increase in x_i will increase the log-odds (and vice versa).

5.4.2 Decision Trees

The main idea behind decision trees is to partition an input space by fitting simple models, and predicting the output in each piece. After establishing a criteria to split the data according to (usually measures of impurity such as entropy or Gini index), a best single split is found for the given data. Next, each partition is further split according to the selected criteria. The tree-building process continues to recurse until some stopping condition is met, such as a minimum number of samples in each leaf. To regularize, the tree is usually built much too large, and then pruned back through "weakest-link" pruning (successively removing bottom-level splits that minimize the increase in overall error).

A major benefit of decision trees is how interpretable they are, especially if there are only a few levels. The next chapter visualizes and attempts to interpret some of the learned decision trees.

5.4.3 Random Forests

As the name implies, random forests are collections of trees. They are constructed by drawing bootstrap samples from the training data, growing a tree on each bootstrap sample, and for each tree selecting some random subset of variables to pick the best from and then split with. Finally, given an ensemble of trees trained on each of the bootstrap samples, a prediction is made based on voting or averaging.

While (depending on their complexity) random forests can be more difficult to interpret, this model class is more flexible than decision trees.

Chapter 6

Predicting Resistance in Urinary Tract Infections

This chapter focuses on predicting antibiotic resistance in urinary tract infections (UTIs). UTIs are very common, affecting 150 million people annually worldwide. Every year, UTIs cost the United States approximately \$3.5 billion [36]. In our dataset, UTIs account for approximately 48% of the microbiology samples.

6.1 Cohort Definitions

We analyze a **general UTI cohort**, as well as an **uncomplicated UTI cohort** that allows us to do a more clinically interpretable evaluation (both are defined below).

6.1.1 General UTI Cohort

As described in Chapter 5, we only consider patients who were prescribed antibiotics *empirically*, meaning their healthcare provider prescribed treatment before receiving results from microbiologic testing. To determine whether a sample is from a urinary tract infection, we use the `site_cat1` field in the Partners microbiology data, labeling the body site from which the culture was collected. After filtering to samples with the `URINARY_TRACT` infection site, there are approximately 116,900 samples remaining.

6.1.2 Uncomplicated UTI Cohort

In the clinical setting, urinary tract infections are categorized as either uncomplicated or complicated. Uncomplicated UTIs typically impact otherwise healthy patients, with no congenital, acquired anatomic, or neurological urinary tract abnormalities. Complicated UTIs are associated with host factors that may necessitate broader or more prolonged antibiotic treatment; these factors include host immune competence, comorbidities, male gender, pregnancy, and instrumentation of the genitourinary tract [36][37]. We restrict our analysis to patients with uncomplicated UTIs, as they are relatively straightforward to define.

Figure 6-1 outlines the criteria used to construct the uncomplicated UTI cohort. These criteria were based on definitions of uncomplicated UTI found in literature [38][39], and iteratively refined through chart review:

Out of samples in the general UTI cohort, we first filter to *non-pregnant female* patients with ages from *18 to 55*. In favor of a cleaner, more restrictive cohort, we *exclude patients with pyelonephritis* (infection of the upper genitourinary tract), and all cases with *any previous surgical procedure code in the past 90 days*. By utilizing procedure codes within the past 90 days, we also explicitly detect and *filter out patients with central venous catheters, mechanical ventilation, and parenteral nutrition*.

The standard treatment protocol for uncomplicated UTIs is limited to a few antibiotics (described further in Section 6.2.2). Chart review revealed that prescription of antibiotics other than these uncomplicated UTI drugs would indicate that the physician suspected a condition other than an uncomplicated UTI. For example, some patients filtered according to the previous criteria were only prescribed fluconazole, which is used against yeast infections rather than uncomplicated UTIs. Furthermore, as noted by procedure and diagnosis codes in our dataset, prescription of more than one of the uncomplicated UTI drugs usually indicates that a patient did not actually have an uncomplicated UTI. Thus, we further limit the cohort to patients *empirically prescribed exactly one of the uncomplicated UTI drugs on their last day of empiric therapy*.

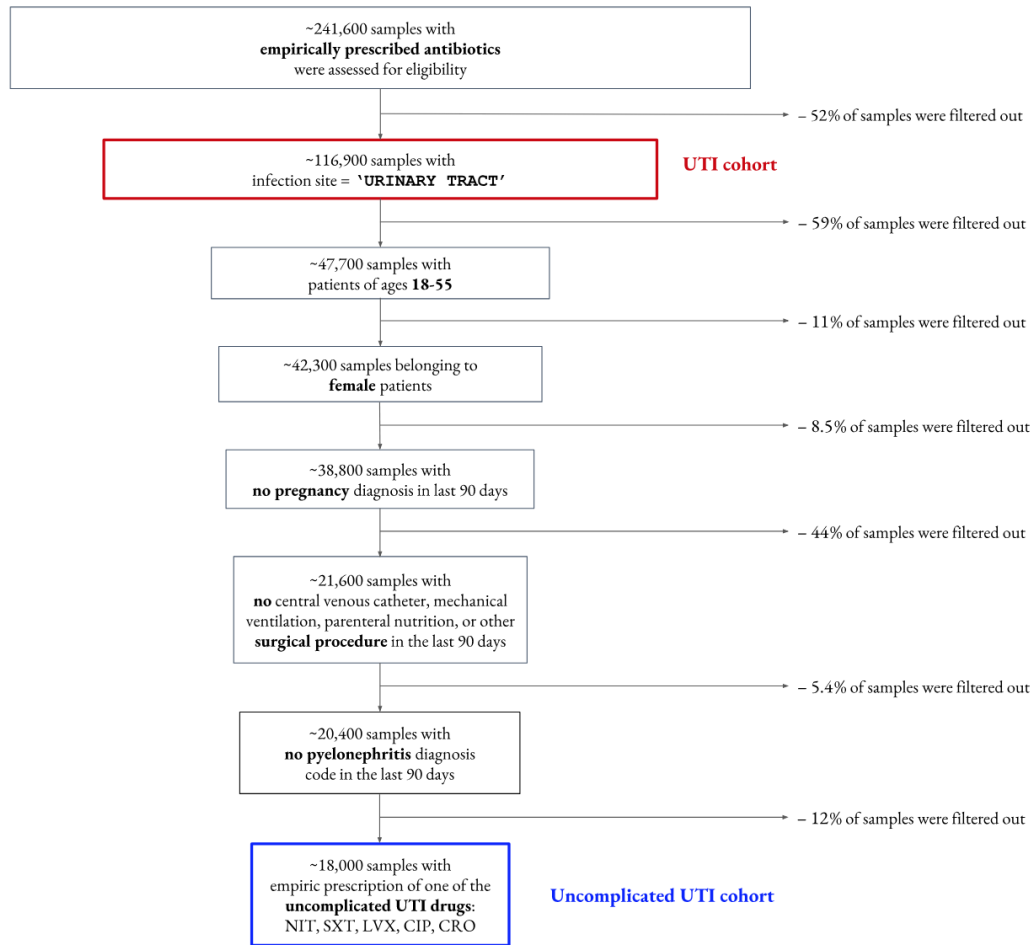


Figure 6-1: Inclusion/exclusion criteria for the general UTI cohort (red) and uncomplicated UTI cohort (blue). Quantities are across the train and test sets combined.

6.1.3 General vs. Uncomplicated UTI Cohort

Table 6.1 compares the general and uncomplicated UTI cohorts. The vast majority of uncomplicated UTI patients are outpatients with infections caused by *E. coli*, whereas general UTI patients have organisms more commonly found to reside in the emergency room, intensive care unit, and inpatient wards. One possible contributor to the difference is that approximately 62% of samples in the general UTI cohort had an associated prior surgical procedure within the past 90 days, whereas in the uncomplicated UTI cohort these samples were filtered out. Overall, the general UTI cohort contains a greater diversity of offending pathogens, as well as a greater proportion of resistant samples and samples from patients with a history of resistance.

Information	General UTI	Uncomplicated UTI
number of samples	116936	17991
age (std. dev.)	55.5 (22.9)	34.3 (10.9)
male (%)	24465 (20.9)	0 (0.0)
white (%)	83793 (71.7)	11284 (62.7)
from nursing home (%)	1925 (1.6)	5 (0.0)
central venous catheter 90D (%)	9122 (7.8)	0 (0.0)
surgery 90D (%)	72601 (62.1)	0 (0.0)
pregnant 90D (%)	3750 (3.2)	0 (0.0)
prev resist CIP 180D (%)	11951 (10.2)	307 (1.7)
prev resist LVX 180D (%)	13436 (11.5)	315 (1.8)
prev resist NIT 180D (%)	10462 (8.9)	316 (1.8)
prev resist SXT 180D (%)	8493 (7.3)	369 (2.1)
prev resist CRO 180D (%)	3756 (3.2)	83 (0.5)
<i>Pseudomonas</i> species (%)	4428 (3.8)	31 (0.2)
<i>E. coli</i> (%)	73675 (63.0)	14929 (83.0)
<i>Klebsiella</i> species (%)	13149 (11.2)	841 (4.7)
<i>Enterobacter</i> species (%)	3119 (2.7)	242 (1.3)
<i>Enterococcus</i> species (%)	11953 (10.2)	373 (2.1)
<i>Coagulase-neg. Staph.</i> species (%)	4257 (3.6)	910 (5.1)
<i>Staphylococcus aureus</i> (%)	2918 (2.5)	99 (0.6)
hospital ward - ER (%)	24124 (20.6)	1898 (10.5)
hospital ward - OP (%)	63770 (54.5)	15242 (84.7)
hospital ward - IP (%)	25179 (21.5)	811 (4.5)
hospital ward - ICU (%)	5134 (4.4)	46 (0.3)
resistant to CIP (%)	23513/110796 (21.2)	1178/17656 (6.7)
resistant to LVX (%)	25611/115959 (22.1)	1183/17957 (6.6)
resistant to NIT (%)	24343/111895 (21.8)	2025/17885 (11.3)
resistant to SXT (%)	19704/103686 (19.0)	2700/17651 (15.3)
resistant to CRO (%)	7159/97003 (7.4)	385/16674 (2.3)

Table 6.1: Comparison between the general and uncomplicated UTI cohorts. Numbers are for the full dataset (train and test combined).

Abbreviations: CIP, ciprofloxacin; LVX, levofloxacin; NIT, nitrofurantoin; SXT, trimethoprim/sulfamethoxazole; CRO, ceftriaxone; ER, emergency room; OP, outpatient; IP, inpatient; ICU, intensive care unit; ‘90D’ and ‘180D’ refer to the windows over which the features were computed (e.g: 90D = last 90 days).

6.2 Antibiotics of Interest

In total, the microbiology samples contain results for 54 distinct antibiotics. To better focus the analyses, this section defines drugs which are relevant to both cohorts.

6.2.1 Drugs Relevant to UTIs

Urinary tract infections can vary greatly, but the following 12 drugs are of particular interest: ceftriaxone (CRO), amoxicillin-clavulanate (AMC), aztreonam (ATM), ciprofloxacin (CIP), levofloxacin (LVX), gentamicin (GEN), trimethoprim-sulfamethoxazole (SXT), nitrofurantoin (NIT), cefepime (FEP), piperacillin-tazobactam (TZP), imipenem (IPM), and meropenem (MEM). These antibiotics treat the vast majority of pathogens responsible for UTI, and are the most commonly used by providers for this syndrome.

6.2.2 Drugs Relevant to Uncomplicated UTIs

In the uncomplicated UTI setting, standard treatment protocol is limited to a few first- and second-line antibiotics: nitrofurantoin (NIT), trimethoprim/sulfamethoxazole (SXT), ciprofloxacin (CIP), levofloxacin (LVX), and fosfomycin (FOF). In our dataset, FOF is rarely prescribed and almost never tested for, so it is excluded from the drugs of interest. In case the patient is resistant to all four drugs, the usual course of action is to consider CRO, which is a 3rd-line intravenous (IV) antibiotic that often requires admitting the patient to the hospital. Thus, CRO is also included with the drugs of interest. Table 6.2 summarizes basic information about each drug of interest:

Name	Abbreviation	Spectrum
ciprofloxacin	CIP	broad
levofloxacin	LVX	broad
nitrofurantoin	NIT	narrow
trimethoprim/sulfamethoxazole	SXT	narrow
ceftriaxone	CRO	broad

Table 6.2: Basic information about uncomplicated UTI drugs of interest.

Figure 6-2 shows the international practice guidelines for treatment of uncomplicated UTIs. The guidelines distinguish between cystitis (lower UTIs, included in the cohort), and pyelonephritis (upper UTIs, which are excluded) [36]. In Section 6.6, these guidelines are used to evaluate predicted probabilities within a clinical context.

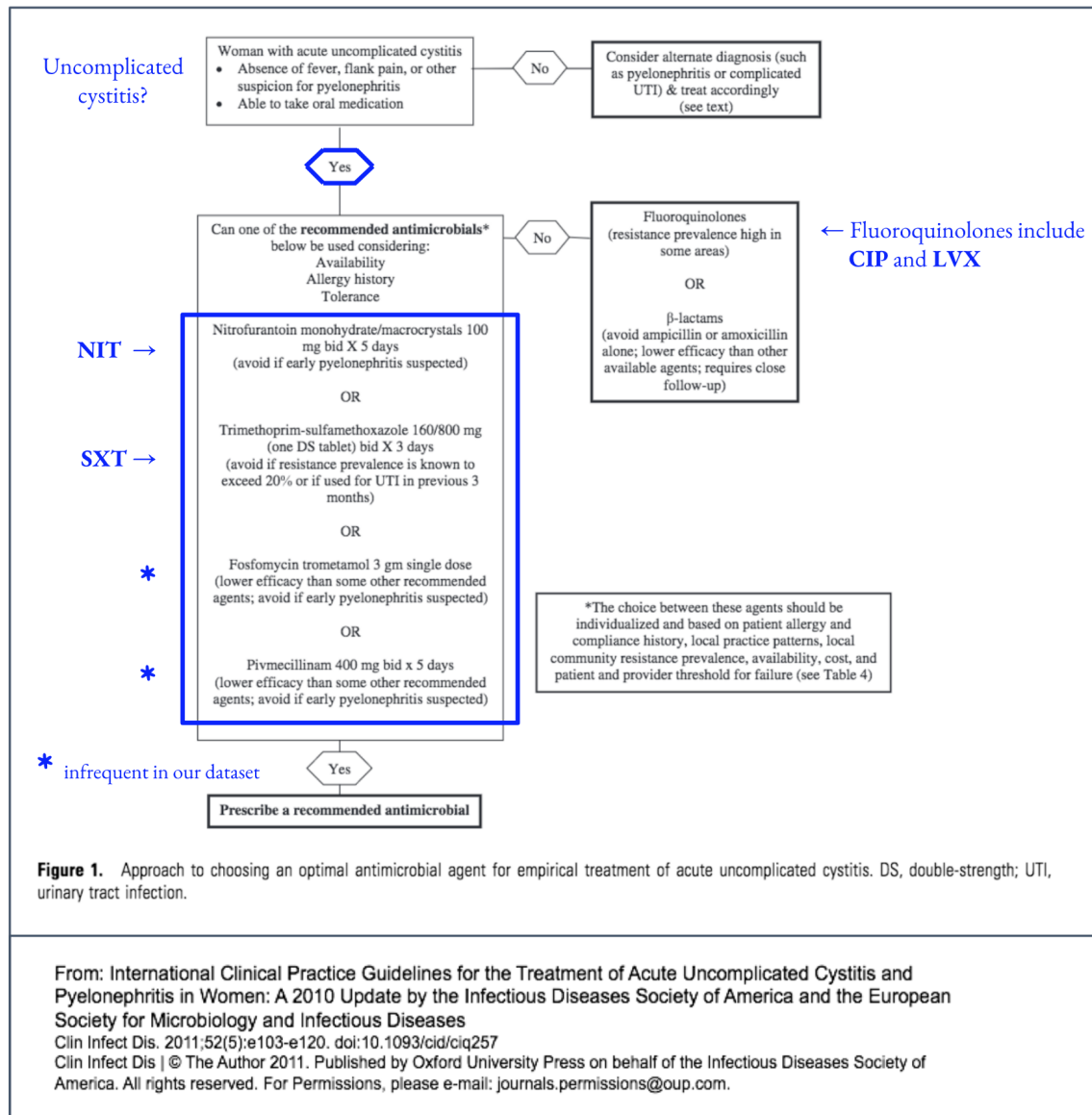


Figure 6-2: 2010 International Clinical Practice Guidelines for the Treatment of Uncomplicated Acute Uncomplicated Cystitis and Pyelonephritis in Women. If nitrofurantoin (NIT) or trimethoprim-sulfamethoxazole (SXT) can be used, then they should be used over fluoroquinolones such as ciprofloxacin (CIP) and levofloxacin (LVX). While fosfomycin and pivmecillinam are also recommended, tests for FOF are far less common in our dataset, and pivmecillinam is not used in the United States.

6.3 Experiment Setup

As described in Chapter 5, the data was split by time range and patient ID into one training/development set (which we call the train/dev set) and three testing sets. Figure 6-3 contains the breakdown of samples after filtering according to the general UTI cohort. Any model-tuning decisions were made entirely using the train/dev set, and various cohorts were used to answer different clinically relevant questions.

Partitioning UTI Samples into Train and Test Sets

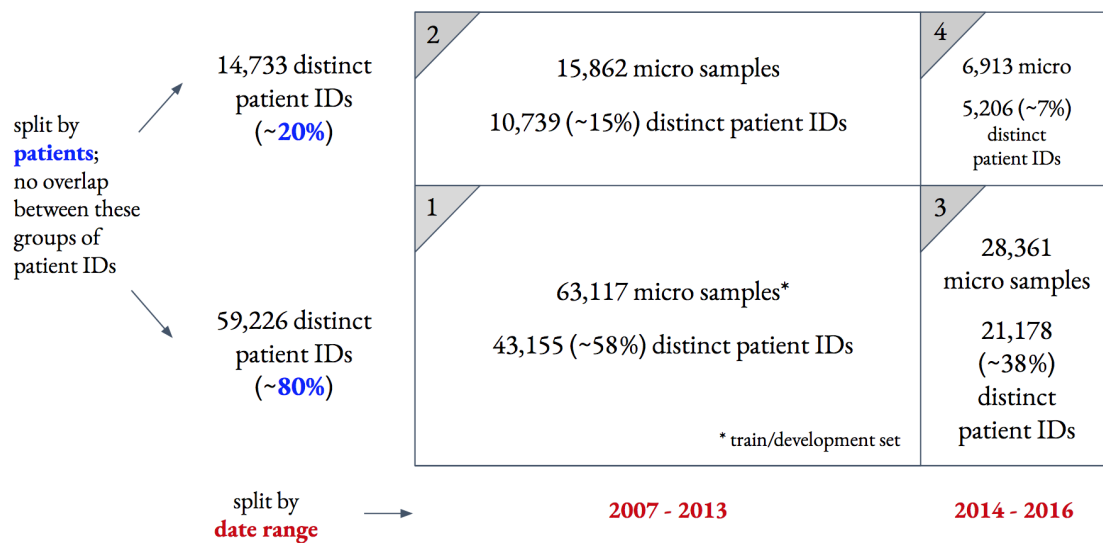


Figure 6-3: Breakdown of UTI samples into training/development and test sets. Samples were split 80-20 by patient ID, and then according to time range (2007-2013 for train/dev, and 2014-2016 for test).

For each antibiotic, we train a separate model. Each trained model has an associated training set, feature set, model class, and model hyperparameters. Any decisions on each of these aspects were made using the training/development set.

6.3.1 Training Set

There are two main cohorts of interest: (1) general UTIs and (2) uncomplicated UTIs. For both cohorts, we randomly split (70/30) by distinct patient ID's into training and validation sets, repeating this five times with different random seeds (since this is equivalent to sampling with replacement, we call this bootstrapping).

When evaluating performance on a given cohort, we also experiment with training on the broader population that it is a part of. For example, we train on the full UTI cohort and evaluate on the uncomplicated UTI cohort. Although uncomplicated UTIs might be more challenging to predict resistance in (patients are healthier and have more homogenous electronic medical records), there is approximately 6.5 times as much data in the general UTI cohort, and we hope that transfer learning might offer a boost in performance on uncomplicated UTIs.

6.3.2 Feature Set

Features can come from microbiology, encounters, demographics, diagnoses, procedures, labs, and clinicians' notes. To evaluate usefulness, we experimented with adding/removing in groups of features at a time. Chapter 3 describes these features in greater detail, and we report results on the full set of features.

6.3.3 Model Classes

As described in Chapter 5, we use logistic regression (lr), decision tree (dt), and random forest (rf) models.

6.3.4 Hyperparameter Tuning

For each label and training set, we search over the following grids of hyperparameters (defined in python's `scikit-learn` library) to find the configurations with the best dev AUCs (averaged over 5 independent 70/30 train/dev splits). For anything not specified below, `scikit-learn` defaults are used.

1. Logistic regression:

(a) `C` (inverse regularization): 0.001, 0.005, 0.01, 0.1, 1

(b) `penalty`: 'l1', 'l2'

(c) `solver`: 'liblinear'

(d) `intercept_scaling`: 1, 1000

(e) `max_iter`: 1000

2. Decision tree:

(a) `criterion`: 'gini', 'entropy'

(b) `max_depth` (or expansion until all leaves are pure): 3, 5, 10, None

(c) `min_samples_leaf`: 0.01, 0.02, 0.05

3. Random forest:

(a) `n_estimators`: 5, 10, 20

(b) `criterion`: 'gini', 'entropy'

(c) `max_depth` (or expansion until all leaves are pure): 3, 5, 10, None

(d) `min_samples_leaf`: 0.01, 0.02, 0.05

6.4 Model Performance

Figures 6-4 and 6-5 visualize the difference in performance between model classes by plotting heat maps of the average development set AUCs, with hyperparameters for each (model class, drug) pair tuned using grid search. Taking the highest average development set AUC scores for each label, Tables 6.3 and 6.4 contain information about model performance on the general UTI cohort and uncomplicated UTI cohort. Figures 6-6a and 6-6b plot the corresponding ROC curves.

Overall, we are better able to predict resistance for the general UTI cohort. This is likely because it is an easier prediction problem: compared to the uncomplicated UTI cohort, which is restricted to a homogeneous group of patients, there are more distinguishing factors that could separate resistant patients from susceptible patients. We also observe that for most antibiotics (with the exception of NIT), dev AUCs are similar across model classes.

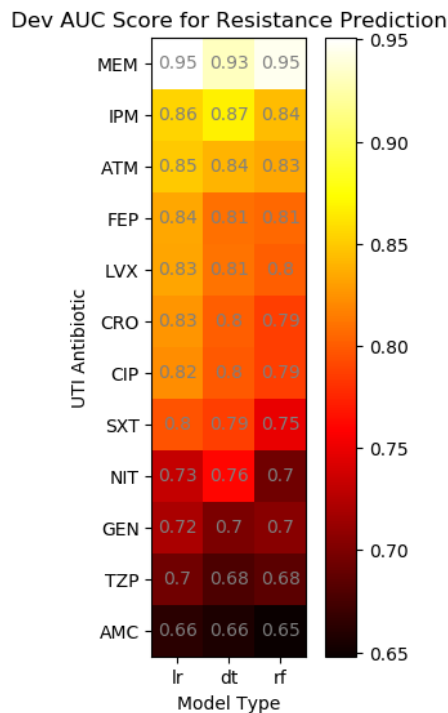


Figure 6-4: Average dev set AUCs for the general UTI cohort.

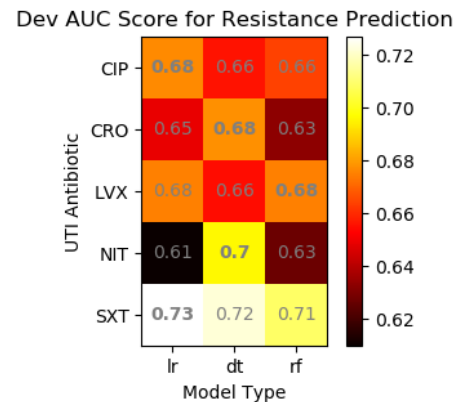


Figure 6-5: Average dev set AUCs for the uncomplicated UTI cohort.

Drug	Resistance	Train AUC	Dev AUC	Test AUC	Model
AMC	0.200	0.750 \pm 0.002	0.663 \pm 0.007	0.662 \pm 0.003	lr, C = 0.1, L1, intercept scl 1
CRO	0.064	0.858 \pm 0.003	0.826 \pm 0.007	0.788 \pm 0.005	lr, C = 0.1, L1, intercept scl 1
FEP	0.023	0.865 \pm 0.004	0.835 \pm 0.006	0.804 \pm 0.009	lr, C = 0.1, L1, intercept scl 1000
ATM	0.062	0.880 \pm 0.004	0.850 \pm 0.011	0.804 \pm 0.006	lr, C = 0.1, L1, intercept scl 1000
TZP	0.063	0.762 \pm 0.005	0.695 \pm 0.020	0.707 \pm 0.011	lr, C = 0.1, L1, intercept scl 1000
IPM	0.028	0.918 \pm 0.002	0.868 \pm 0.009	0.620 \pm 0.008	dt, gini, max depth None, min leaf 0.01
MEM	0.008	0.959 \pm 0.003	0.951 \pm 0.007	0.955 \pm 0.007	lr, C = 0.1, L1, intercept scl 1
CIP	0.205	0.857 \pm 0.002	0.823 \pm 0.004	0.799 \pm 0.003	lr, C = 0.1, L1, intercept scl 1000
LVX	0.216	0.861 \pm 0.001	0.833 \pm 0.004	0.801 \pm 0.002	lr, C = 0.1, L1, intercept scl 1000
NIT	0.217	0.777 \pm 0.001	0.761 \pm 0.003	0.665 \pm 0.003	dt, gini, max depth None, min leaf 0.01
SXT	0.175	0.838 \pm 0.001	0.796 \pm 0.004	0.718 \pm 0.004	lr, C = 0.01, L1, intercept scl 1000
GEN	0.082	0.783 \pm 0.001	0.721 \pm 0.009	0.693 \pm 0.008	lr, C = 0.1, L1, intercept scl 1000

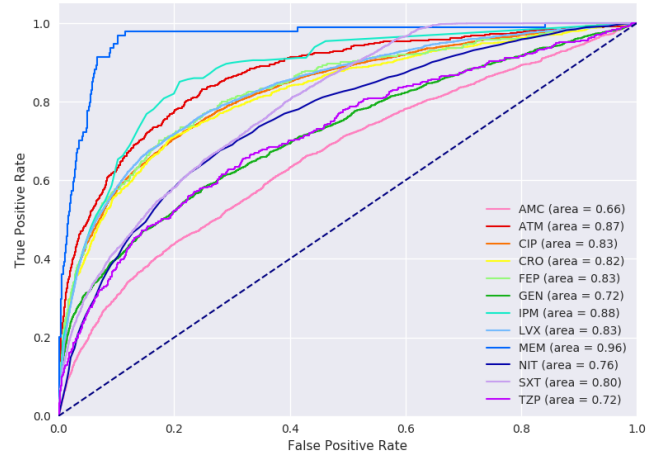
Table 6.3: Performance on general UTIs. Information is listed only for the models with the highest average dev set AUCs, and AUCs are reported \pm the standard deviation across 5 trials. Test AUCs are for the 2014-2016 time range, with overlapping patients from the train set. Drugs are grouped/ordered as follows by drug class: beta-lactams, fluoroquinolones, nitrofurantoin, folate-inhibitor, and aminoglycoside.

Drug	Resistance	Train AUC	Dev AUC	Test AUC	Model
CIP	0.061	0.699 \pm 0.014	0.677 \pm 0.025	0.640 \pm 0.014	lr, C = 0.1, L1, intercept scl 1000
LVX	0.060	0.702 \pm 0.009	0.675 \pm 0.030	0.664 \pm 0.025	lr, C = 0.1, L1, intercept scl 1000
NIT	0.114	0.806 \pm 0.010	0.696 \pm 0.027	0.585 \pm 0.016	dt, entropy, max depth 10, min leaf 0.01
SXT	0.142	0.758 \pm 0.005	0.727 \pm 0.005	0.683 \pm 0.007	lr, C = 0.1, L1, intercept scl 1000
CRO	0.019	0.774 \pm 0.012	0.678 \pm 0.065	0.545 \pm 0.014	dt, entropy, max depth 5, min leaf 0.01

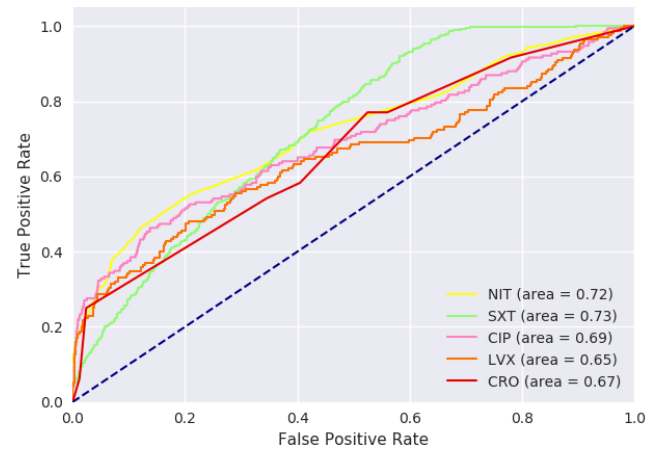
Table 6.4: Performance on uncomplicated UTIs (columns as described in Table 6.3).

As seen in Tables 6.3 and 6.4, the ‘best’ (highest dev AUC) model is usually a decision tree or logistic regression model with high regularization. AUCs are fairly consistent across multiple trials, and degrade from the train set, to the dev set, to the test set. The most dramatic decreases in AUC are seen in the narrow-spectrum antibiotic NIT, as well as the third-line antibiotic CRO. We also notice that logistic regression models tend to generalize better to future years, even achieving higher test AUCs than decision trees which achieved higher in-sample dev AUCs for the same label. In future work, this could be explicitly tuned for by subsampling train and development sets from non-overlapping time ranges.

Tables 6.3 and 6.4 also highlight differences between the general and uncomplicated UTI populations. Not only is there a decrease in dev AUCs, there is also a lower level of resistance to all antibiotics, with fluoroquinolone resistance decreasing from more than 20% to approximately 6%.



(a) General UTI cohort ROCs.



(b) Uncomplicated UTI cohort ROCs.

Figure 6-6: Development set ROC curves for (one split of) each cohort.

In Figures 6-6a and 6-6b, model performances are displayed across many different thresholds. For a low very major error, or false negative rate, one would expect a high true positive rate. Thus, the right side of the ROC curve is of particular interest. Note that in uncomplicated UTIs (Figure 6-6b), our performance on SXT is particularly good in this respect.

6.4.1 Additional Experiments

In order to characterize and further explore the strongest signals for antibiotic resistance, there were various additional experiments conducted. Without going into exhaustive detail, below are the main takeaways:

- **Bag-of-words text features:** Initially, the addition of text features boosted performance significantly. After carefully inspecting the most important text features, however, we gained inspiration for new features such as specimen date, previous resistance, previous organism, and hospital location. With the addition of these new features, bag-of-words features became far less important in our models. In fact, if we now remove bag-of-words features altogether, there is little to no difference in development set AUC.
- **Colonization pressure:** An individual’s previous antibiotic resistance has consistently been a strong predictor for current resistance. By creating the colonization pressure feature, we had a measure of previous population-level resistance, in various levels of locality to the patient we were making predictions for. During training, colonization pressure features were often selected as important. Compared to models without colonization pressure features, the addition of colonization pressure significantly improved development set performance on SXT, NIT, and CRO. On the test set, however, the only major improvement was for SXT.
- **Cohort of samples with previous resistance:** When doctors notice that a patient has previously been resistant to an antibiotic, they may want to avoid prescribing that antibiotic. However, it is often unclear how to choose among the remaining first- and second-line therapies. Using the uncomplicated UTI cohort, we trained models to predict resistance to CIP, LVX, NIT, and SXT, conditioned on previous resistance (in the past 180 days) to each of them. While we were able to achieve higher AUCs than on the regular uncomplicated UTI cohort, these cohorts were very small (around 150-200 samples each), and we decided that in this setting these models would not be clinically useful.

- **Pathogen as a feature:** In previous work predicting antibiotic resistance, the offending pathogen has sometimes been included as a feature. We exclude this information from our models because it is unavailable at the time of prescription, but when experimenting with inclusion of pathogen as a feature, it did indeed boost our development set performance significantly. Future work could attempt to predict the offending pathogen, or incorporate meta information about the pathogen that might be available at the time of empiric prescription (e.g: Gram stain results, etc.).

6.5 Model Interpretation

Many of the top-performing models were logistic regression or decision trees. This section interprets these models by discussing the top logistic regression coefficients, or the first five levels of the tree.

6.5.1 General UTIs

Tables 6.5 through 6.16 contain the top ten (greatest-magnitude) coefficients for tuned logistic regression models predicting resistance to each of the twelve drugs of interest. The best-performing models on the dev set were all logistic regression models with fairly strong L1 regularization, with the exception of IPM and NIT, for which decision trees had a *slightly* higher performance (≤ 0.03 AUC difference).

Across the board, previous exposure to the antibiotic (e.g: colonization pressure, previous resistance, previous medication) is highly indicative of current resistance. A variety of other features bubble up as well, including specific words in clinicians' notes, indications of exposure to other antibiotics, and gender.

As an illustrative example, let us consider ceftriaxone (CRO). Three of the top ten features indicate previous exposure/ resistance to ceftriaxone, two of the top ten features correspond to 'esbl' showing up in recent discharge summaries, and the

remaining features are ward-level colonization pressure of other antibiotics. Intuitively, previous exposure or resistance to ceftriaxone would predispose a patient to ceftriaxone resistance due to the mechanisms discussed in Chapter 2. ‘esbl’ stands for extended-spectrum beta-lactamases, which are enzymes that cleave beta-lactams such as ceftriaxone (and therefore render ceftriaxone ineffective). Finally, while these specific interactions with colonization pressure are less well-studied, they are clinically plausible. Given resistance to IPM, there are mechanisms which confer cross-resistance to other cephalosporins such as CRO. AMC inhibits esbls, and so if one is resistant to AMC, or if there are cross-resistance mechanisms at play, these could contribute to a greater likelihood of resistance to CRO.

Feature	Coeff
colp AMC 90 - higher level	1.55
colp ATM 90 - higher level	1.4
prev resistance AMC 180	0.94
prev resistance AMC 90	0.55
prev resistance SAM 180	0.32
demographics - is_male	0.32
prev organism Proteus 90	-0.31
DS_7_90 - ivf	-0.32
colp SAM 90 - higher level	-0.46
colp SXT 90 - higher level	-0.62

Table 6.5: AMC Features

Feature	Coeff
colp IPM 90 - higher level	2.71
prev resistance CRO 180	1.94
colp AMC 90 - higher level	1.76
DS_30_365 - esbl	1.04
DS_7_90 - esbl	0.69
colp TET 90 - higher level	0.69
prev resistance CRO 30	0.5
medication 7 - ceftriaxone	-0.52
colp AMP 90 - higher level	-0.64
colp SXT 90 - higher level	-0.97

Table 6.6: CRO Features

Feature	Coeff
colp AMC 90 - higher level	1.39
prev resistance FEP 180	1.28
colp TET 90 - higher level	0.72
DS_30_365 - esbl	0.68
prev resistance LVX 180	0.42
prev resistance CRO 180	0.4
ab type 180 - carbapenem	0.37
demographics - is_male	0.33
medication 180 - cefpodoxime	-0.33
demographics - is_white	-0.43

Table 6.7: FEP Features

Feature	Coeff
prev resistance ATM 180	1.19
colp ATM 90 - higher level	0.94
DS_30_365 - esbl	0.57
prev resistance CRO 180	0.51
ab type 180 - carbapenem	0.43
prev resistance ATM 30	0.41
hosp floor - UROLOGY	0.36
colp AMP 90 - overall	-0.4
hosp ward - OP	-0.5
colp SXT 90 - higher level	-2.6

Table 6.8: ATM Features

Feature	Coeff
prev resistance TZP 180	1.39
prev resistance CRO 180	0.39
colp AMP 90 - granular level	0.31
OP_7_90 - look	0.29
colp TET 90 - overall	0.21
ab type 90 - beta_lactams	0.21
OP_7_90 - mm	0.2
OP_7_90 - alert	-0.2
OP_30_365 - diameter	-0.21
OP_30_365 - order	-0.26

Table 6.9: TZP Features

Feature	Coeff
colp SAM 90 - overall	2.66
colp ATM 90 - higher level	1.75
prev resistance IPM 180	1.27
colp TET 90 - higher level	1.12
colp ERY 90 - overall	0.62
colp IPM 90 - granular level	0.58
demographics - is_male	0.53
colp SXT 90 - overall	-0.93
colp FOX 90 - higher level	-2.01
colp AMP 90 - overall	-2.69

Table 6.10: IPM Features

Feature	Coeff
prev resistance MEM 180	1.63
ab class 180 - aminoglycoside	0.66
demographics - is_male	0.58
comorbidity 90 - Paralysis	0.57
prev resistance ATM 180	0.45
ab type 180 - carbapenem	0.44
DS_30_365 - tigecycline	0.43
colp ERY 90 - higher level	0.39
hosp ward - OP	-0.64
colp AMP 90 - overall	-3.57

Table 6.11: MEM Features

Feature	Coeff
colp MXF 90 - higher level	1.1
prev resistance CIP 180	0.89
colp TET 90 - higher level	0.87
prev resistance LVX 180	0.78
OP_7_90 - resistant	0.76
medication 90 - norfloxacin	0.71
colp PEN 90 - overall	0.58
prev resistance MXF 180	-0.61
ab class 7 - fluoroquinolone	-0.7
colp SXT 90 - higher level	-0.85

Table 6.12: CIP Features

Feature	Coeff
colp CLI 90 - overall	1.11
prev resistance LVX 180	0.91
colp SAM 90 - overall	0.78
OP_7_90 - resistant	0.75
prev resistance CIP 180	0.71
colp TET 90 - higher level	0.62
colp OXA 90 - overall	0.59
infection_sites - BLOOD	-0.64
ab class 7 - fluoroquinolone	-0.69
colp CRO 90 - higher level	-1.11

Table 6.13: LVX Features

Feature	Coeff
colp CRO 90 - overall	9.5
colp TZP 90 - overall	5.67
colp ATM 90 - overall	1.71
colp GENS 90 - higher level	0.94
colp ATM 90 - higher level	0.82
prev resistance NIT 180	0.78
colp OXA 90 - overall	-0.74
colp AMP 90 - overall	-1.05
colp MXF 90 - overall	-1.21
colp IPM 90 - overall	-3.61

Table 6.14: NIT Features

Feature	Coeff
colp TET 90 - overall	6.23
colp SXT 90 - higher level	3.1
colp AMC 90 - higher level	1.64
prev resistance SXT 180	1.34
OP_7_90 - resistant	0.54
comorbidity 14 - Lymphoma	0.51
colp NIT 90 - higher level	-0.78
colp CIP 90 - overall	-2.4
colp NIT 90 - overall	-3.15
colp FOX 90 - overall	-8.09

Table 6.15: SXT Features

Feature	Coeff
prev resistance GEN 180	1.93
colp AMP 90 - overall	0.69
DS_30_365 - esbl	0.44
prev resistance GEN 90	0.42
colp SAM 90 - overall	0.38
procedure 7 - hemodialysis	0.34
OP_7_90 - qam	0.31
ab class 7 - fluoroquinolone	-0.33
OP_7_90 - tests	-0.34
prev organism Staph_coag_neg 180	-0.47

Table 6.16: GEN Features

Tables 6.5 - 6.16: Top 10 coefficients (greatest-magnitude) for predicting resistance to each drug in the General UTI cohort.

Abbreviations for Tables 6.5 - 6.16: colp, colonization pressure; ab, antibiotic; DS_x_y or OP_x_y, word from discharge summaries or outpatient notes (respectively), from x days back to y days back; higher level, computed over the ward (e.g: IP, OP, ER, ICU); granular level, computed over the hospital floor; number (e.g: 90, 180) refers to the window the feature was computed over; drug abbreviations are as described in Section 6.2.1.

6.5.2 Uncomplicated UTIs

Next, we inspect the models which achieved the greatest average dev AUCs on the uncomplicated UTI cohorts. Tables 6.17 through 6.21 contain the top ten coefficients of the tuned logistic regression models for each label. Again, logistic regression with strong L1 regularization outperformed other model variations, with the exception of NIT and CRO, whose ‘best performing’ models were decision trees (first five layers are visualized in Figures 6-7 and 6-8).

As part of model tuning, training on the general UTI cohort and directly evaluating on the uncomplicated UTI cohort (with hyperparameters tuned according to performance on the uncomplicated UTI cohort) was also experimented with. On average there was very little difference in performance on all drugs of interest. While future work could involve experimenting with other variations of transfer learning, due to similar performance we opt for models trained on the uncomplicated UTI cohort, a less computationally intensive option. Thus, only the values from training and evaluating on the uncomplicated UTI cohort are reported.

Based on the top coefficients for predicting resistance to each antibiotic, indicators of previous exposure to antibiotics are strong predictors of resistance. In the model for LVX, the top two factors are previous resistance to LVX and CIP in the past 180 days, which makes sense because CIP is in the same drug class as LVX, and it is known that resistance to these drugs is highly correlated.

Interestingly, environmental resistance to antibiotics belonging to other drug classes such as TZP and CRO, can predict resistance to NIT. While this requires further investigation, it could suggest a linkage phenomenon where exposure or acquisition of resistance to one antibiotic can lead to multi-drug resistance. The strongest negative predictor for LVX and SXT, `is_white`, might be informative of the patient’s economic, sociological, and epidemiological environment. Certain hospital floors also appear to be predictive of resistance in SXT and CRO, possibly indicative of colonization pressures or other epidemiological factors.

Feature	Coeff	Feature	Coeff	Feature	Coeff
prev resistance CIP 180	2.02	prev resistance LVX 180	1.37	colp TZP 90 - overall	7.39
prev resistance CIP 90	0.52	prev resistance CIP 180	0.94	colp CRO 90 - overall	5.64
OP_30_365 - nitrofurantoin	0.51	OP_30_365 - nitrofurantoin	0.7	prev resistance NIT 180	2.12
DS_7_90 - orders	0.35	comorbidity 90 - HTN	0.25	colp NIT 90 - overall	1.83
encounters - num inpatient 90	0.33	encounters - num inpatient 180	0.24	colp ATM 90 - higher level	1.72
DS_30_365 - seen	0.21	prev resistance CIP 90	0.23	OP_7_90 - switched	1.64
ab class 180 - fluoroquinolone	0.2	procedure 30 - had surgery	0.22	colp OXA 90 - overall	-1.95
OP_30_365 - home	0.2	encounters - num inpatient 30	0.22	colp PEN 90 - overall	-2.2
comorbidity 180 - HTN	0.16	ab class 90 - fluoroquinolone	0.19	colp AMP 90 - overall	-3.0
OP_7_90 - culture	0.15	demographics - is_white	-0.2	colp IPM 90 - overall	-4.37

Table 6.17: CIP Features

Table 6.18: LVX Features

Table 6.19: NIT Features

Feature	Coeff
colp TET 90 - overall	3.23
colp AMC 90 - overall	2.44
colp SXT 90 - higher level	2.4
prev resistance SXT 180	1.11
hosp floor - CHLSA WALK-IN CLN	0.38
hosp floor - CHLSA EVERETT	0.37
colp TET 90 - granular level	0.35
OP_30_365 - areas	0.23
OP_30_365 - cholesterol	-0.25
demographics - is_white	-0.29

Table 6.20: SXT Features

Feature	Coeff
colp AMC 90 - higher level	6.03
prev resistance CRO 180	4.59
hosp floor - BIMA, 3F	3.06
comorbidity 180 - WeightLoss	1.38
OP_30_365 - enlargement	1.32
DS_7_90 - remove	1.31
OP_30_365 - ab	1.19
OP_7_90 - primary	1.16
DS_7_90 - difficulty	-1.47
comorbidity 30 - HTN	-1.79

Table 6.21: CRO Features

Tables 6.17 - 6.21: Top 10 coefficients (greatest-magnitude) for predicting resistance to each drug in the uncomplicated UTI cohort.

Nitrofurantoin (NIT)

Figure 6-7 displays the first five levels of the 10-level decision tree model predicting resistance to NIT. As mentioned in Table 6.4, `min_samples_leaf = 0.01` and `criteria = 'entropy'`. In the uncomplicated UTI cohort, approximately 11.4% of samples are resistant.

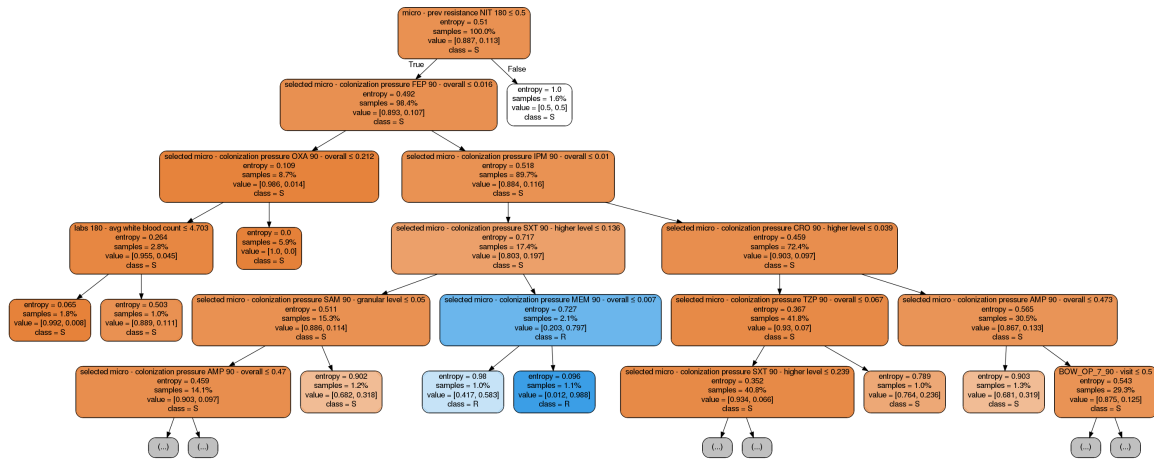


Figure 6-7: Decision tree for predicting resistance to NIT.

The root of the decision tree splits based on previous resistance to NIT. From there, samples are split mostly based on colonization pressure features, with the exception of one use of average white blood count labs, and one use of an outpatient bag of words token. We also note that different granularities of colonization pressures are used, where ‘overall’ colonization pressure refers to colonization pressure computed over the entire patient population, ‘higher level’ colonization pressure is computed over hospital wards (IP, OP, ER, or ICU), and ‘granular level’ colonization pressure is computed over hospital floor (e.g: specific outpatient clinics, parts of a hospital, etc.).

Based on the factors of resistance important to this decision tree, before prescribing NIT, a clinician might pay special attention to whether a patient has previously been resistant to NIT. While the combination of colonization pressure values are less easily interpreted, future studies may want to use these selected drugs as starting points for investigation into linked resistance.

Ceftriaxone (CRO)

Figure 6-8 displays the five-level decision tree model predicting resistance to CRO. As mentioned in Table 6.4, `criteria = 'entropy'` and `min_samples_leaf = 0.01`. In the uncomplicated UTI cohort, approximately 1.9% of samples have the resistant phenotype.

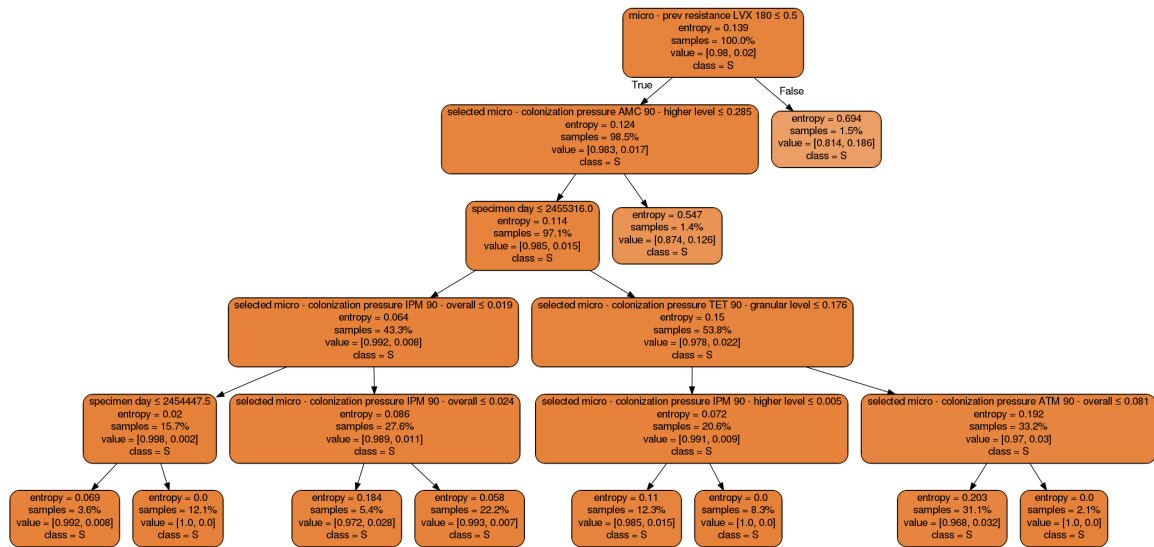


Figure 6-8: Decision tree for predicting resistance to CRO. Specimen day is reported in Julian days, so a specimen day of 2455316 refers to April 29th, 2010.

Rather than previous resistance to CRO, the root node splits based on previous resistance to LVX. While it is not immediately obvious why previous resistance to LVX is selected over CRO, resistance to the two drugs could be linked.

Also, except for previous resistance and specimen day, the vast majority of nodes use colonization pressure as a feature. One possible reason why colonization pressure and specimen day are picked up in the decision tree is that they are the most fine-grained numerical features. As a result, the decision tree may find it easier to split between samples using these features. Another hypothesis is that colonization pressure and specimen year are serving as proxies for nearest-neighbor matching of patients (e.g: coming from the same ward, around the same time), a point that should be further investigated. Either way, local resistance levels are a non-trivial factor in predicting individual resistance.

6.6 Evaluation in Clinical Context

To evaluate the practical utility of this work, the models' predictions are converted into treatment recommendations, and compared to the actual treatment decisions made by physicians in the dataset. In order to fairly evaluate these treatment decisions in retrospect, the analysis is restricted to a homogeneous cohort with a clear clinical guideline (Figure 6-2), where the only missing piece of information for determining the "correct" decision is the unknown resistance profile.

The metrics used to evaluate our recommendations against physicians' treatment decisions are: (1) the rate at which patients are prescribed something they are resistant to (**inappropriate antibiotic therapy**, or IAT), and (2) the number of **broad spectrum** antibiotics prescribed instead of narrow spectrum antibiotics. Note that the goal is to minimize both of these metrics. Using a decision algorithm to make treatment recommendations based on the models' predictions, and then experimenting with acceptable error rates, it is possible to reduce the spectrum of antibiotics prescribed by about 6.6% while achieving similar rates of IAT as physicians.

6.6.1 Treatment Decision Algorithm

Given a set of predicted probabilities of resistance across several drugs, the following decision algorithm is used to recommend which drug to prescribe:

1. From the predicted probabilities of resistance to NIT, SXT, LVX, CIP, obtain binary predictions of 'S' by thresholding.
2. If there are no drugs for which the model predicts 'S', fall back on the doctor's prescription.
3. Otherwise, recommend the first drug in the ranking: NIT, SXT, LVX, CIP.

Many variations of this algorithm are equally justifiable, but as a demonstration of how the model predictions could be used to improve clinical practice, the algorithm is kept relatively simple. Following is a discussion of the details and rationale for each step of the algorithm.

Step 1: Setting Thresholds

To translate predicted probabilities of resistance into clinical decisions, one must first set thresholds for declaring a sample ‘resistant’ (R) or ‘susceptible’ (S). The algorithm is also allowed to say it is ‘unsure.’ (Note: while the following specific decision algorithm will only use the ‘S’ binary predictions, one could easily imagine more complex algorithms that utilize predictions of ‘R.’)

Thresholds are set by varying desired false negative rates and false positive rates. For antibiotic resistance, false negatives are particularly harmful since they could lead to prescribing something that the patient is resistant to (inappropriate antibiotic therapy, or IAT), whereas a false positive would correspond to excluding a drug which the patient could have been susceptible to (suboptimal treatment). Additionally, not all antibiotics are equally harmful if IAT occurs (side-effect severity varies).

To convert desired false negative rates into thresholds for binary prediction, each model’s ROC curve is used. A low false negative rate implies confidence in predicting susceptibility, and so the threshold derived from a false negative rate is used to predict which samples are susceptible. On the flip side, desired false positive rates are converted into thresholds for predicting resistant samples. Remaining samples are declared ‘unsure.’ This procedure is illustrated in Figure 6-9.

Setting Thresholds from a Model’s ROC Curve

(‘R’ = 1 = positive, ‘S’ = 0 = negative)

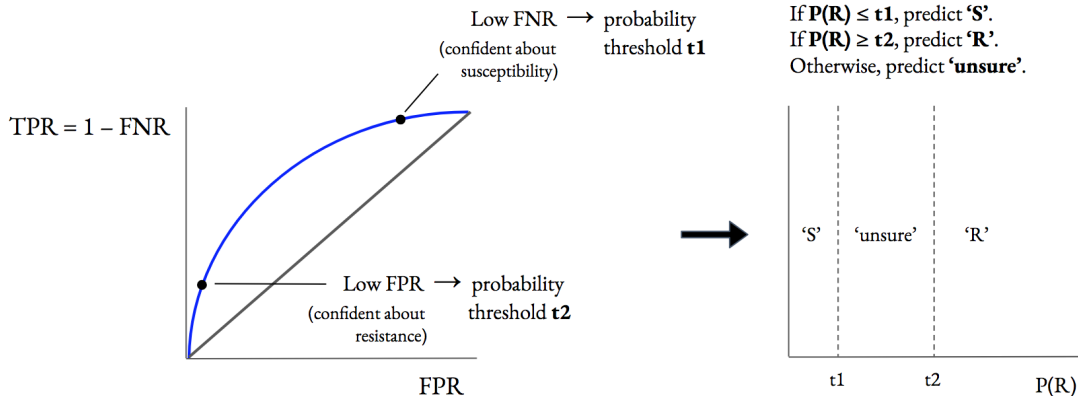


Figure 6-9: Using a model’s ROC curve to attain predictions of ‘R’, ‘S’, and ‘unsure’.

Steps 2 & 3: Making Recommendations

Predictions of ‘S’, ‘R’, and ‘unsure’ for each sample are converted into a recommendation of what to prescribe. For the sake of simplicity, only predictions of susceptibility are used in the recommendation. Also, although CRO is of interest, it is not a standard drug given to all uncomplicated UTI patients and so it is excluded from this analysis. Treatment recommendations are made as follows:

- out of the following rank ordering of drugs, recommend the first drug that the model predicts ‘S’ for: (1) NIT, (2) SXT, (3) LVX, (4) CIP
- if ‘S’ is not predicted for any of them, fall back on the doctor’s prescription

The ranking NIT, SXT, LVX, CIP is based on clinical guidelines for the cohort (Figure 6-2), which recommend narrow-spectrum antibiotics NIT and SXT over broad-spectrum quinolone antibiotics LVX and CIP (Figure 6-2). Within the broad/narrow categories, the antibiotics with lower overall resistance were ranked first. Moreover, other orderings were tried within the broad/narrow groups, but this made little difference in performance.

6.6.2 Tuning Thresholds and Model Classes

The metrics used in this evaluation are the relative rates of IAT and broad-spectrum prescriptions, as compared to the physicians in our dataset. Rather than setting a single constant value across all of the models for each drug, false negative rate (FNR) is treated as a hyperparameter that can be tuned to target different levels of IAT and improvements in spectrum. The model class is also varied between the logistic regression (lr) and decision tree (dt) model classes, which consistently achieved the best development (dev) set AUCs (and often had similar dev AUCs).

For each drug, predictions are made using both ‘dt’ and ‘lr’ models, in combination with FNR of 0.1%, 1.5%, and 10%. Across all four drugs of interest (CIP, LVX, NIT, SXT), all combinations of these model class/FNR settings are tried, except for CIP and LVX, which are set to the same FNR threshold due to computational constraints and virtually identical biochemistry.

6.6.3 Experiment Setup

The improvements in IAT and spectrum made relative to physicians in our dataset are computed for each combination of thresholds and model classes (let these be called ‘hyperparameters’), over 20 train/dev splits with a 70/30 proportion. Then, **doing no worse than physicians in IAT, hyperparameters are chosen such that improvement in spectrum is maximized**. Oppositely, doing no worse than physicians in spectrum, hyperparameters which maximally improve IAT are chosen.

More specifically, consider a single hyperparameter setting h . For each train/dev split $s \in [1, 2, \dots, 20]$, we compute the differences between our recommendations $r_{s,h}$ and physician treatment decisions $p_{s,h}$, for the IAT and spectrum metrics:

$$\Delta IAT_{h,s} = IAT(r_{h,s}) - IAT(p_{h,s})$$

$$\Delta spectrum_{h,s} = spectrum(r_{h,s}) - spectrum(p_{h,s})$$

For each hyperparameter set h , we then compute the empirical average and standard deviation of improvements in the IAT and spectrum metrics over the 20 splits:

$$\mu_{IAT,h} = E_{s \in [1,2,\dots,20]}[\Delta IAT_{h,s}] = \frac{\sum_{s=1}^{20} \Delta IAT_{h,s}}{20}$$

$$\sigma_{IAT,h} = \sigma_{s \in [1,2,\dots,20]}(\Delta IAT_{h,s}) = \sqrt{\frac{1}{20} \sum_{s=1}^{20} (\Delta IAT_{h,s} - \mu_{IAT,h})^2}$$

(similar equations for $\mu_{spectrum,h}$ and $\sigma_{spectrum,h}$.)

Inherently, there is a tradeoff between IAT and spectrum. To get a sense for the utility of our predictions for each metric, we choose hyperparameter sets that ‘do no worse’ on one criteria, while ‘maximally improving’ the other criteria. Note that since we would like to **reduce** IAT and spectrum, we would like to **minimize** ΔIAT and $\Delta spectrum$ (i.e.: we would like the average difference between our IAT and physicians’ IAT to be negative).

We define ‘doing no worse’ as the average reduction (e.g: how much IAT is reduced relative to physicians) being at most one standard deviation above zero:

$$\mu_{metric,h} - \sigma_{iat,h} \leq 0$$

We define ‘maximally improving’ as the average reduction that is the most standard deviations below 0:

$$\max_h \frac{-\mu_{metric,h}}{\sigma_{metric,h}}$$

6.6.4 Experiment Results

Under these definitions, the following reductions in development set IAT and spectrum are achieved, while doing no worse on the other metric:

- Doing no worse in spectrum, maximizing the decrease in IAT:
 - change in IAT: -26.1 ± 19.5 (-0.3% average reduction)
 - change in spectrum: 131.3 ± 145.4 (1.7% average increase)
- Doing no worse in IAT, maximizing the decrease in spectrum:
 - change in IAT (mean \pm std dev): 13.0 ± 13.8 (0.02% average increase)
 - change in spectrum: -246.7 ± 69.0 (-3.3% average reduction)

It is also possible to be more or less flexible in the definitions of ‘doing no worse.’ Instead of setting an average reduction at most *one* standard deviation *above zero* ($z' = 1$), we can set it to $z' = -2, -1, 0, 1$ or 2 standard deviations above zero. Note that **higher z' correspond to more relaxed definitions of ‘doing no worse.’**¹ Table 6.22 contains the reductions in development set IAT and spectrum for different settings of z' . Table 6.23 contains the reductions in test set IAT and spectrum.

¹Note: Within the empirical distribution of reduction in IAT or spectrum relative to clinical practice, z' is actually the negative of the lower bound on the Z-score of the data point 0. Said differently, if we observe that our IAT = physicians’ IAT, a z' of 0 would place us at the mean of the empirical distribution, a z' of 1 would place us one standard deviation below the mean of the empirical distribution, etc. If ‘no difference,’ i.e. a data point of 0 is one standard deviation above the mean of the empirical distribution, this implies that the empirical distribution had on average a more negative difference (a greater decrease) in IAT or spectrum.

	Holding IAT at various z'				Holding spectrum at various z'			
	IAT		spectrum		IAT		spectrum	
z'	mean	std	mean	std	mean	std	mean	std
2	32.3	19.0	-695.4	103.7	-56.1	26.1	563.1	282.9
1	13.0	13.8	-246.7	69.0	-26.1	19.5	131.3	145.4
0	-2.8	20.0	-189.4	130.4	-8.8	18.6	-23.3	138.5
-1	-18.1	17.7	111.2	222.8	-5.6	20.6	-153.7	140.4
-2	-56.1	26.1	563.1	282.9	1.6	21.3	-259.2	127.4

Table 6.22: Trade-off between development set IAT and spectrum for various settings of z' . Mean and standard deviations are taken across 20 train/dev bootstrap splits. On average, there were approximately 7,518 samples in each dev set. The bolded values correspond to $z' = 1$, our original definition of ‘do no worse’; under this definition of ‘doing no worse’ on IAT, our algorithm can on average prescribe $246.7/7518 \approx 3.3\%$ fewer broad-spectrum antibiotics. The highlighted row, which corresponds to $z' = 2$, appears to provide a significantly greater improvement in spectrum, while controlling IAT fairly well. In this setting of hyperparameters, our algorithm can prescribe $695.4/7518 \approx 9.2\%$ fewer broad-spectrum antibiotics.

	Holding IAT at various z'				Holding spectrum at various z'			
	IAT		spectrum		IAT		spectrum	
z'	mean	std	mean	std	mean	std	mean	std
2.0	0.9	9.7	-230.3	12.7	-39.8	10.4	478.1	25.5
1.0	9.3	9.9	-54.1	27.7	-32.7	12.7	342.5	39.3
0.0	-9.3	10.1	-51.4	12.7	-22.7	14.8	256.1	67.3
-1.0	-12.3	16.9	62.3	111.9	-10.4	9.8	-26.6	20.2
-2.0	-39.8	10.4	478.1	25.5	-15.6	11.8	71.0	33.2

Table 6.23: Trade-off between test set IAT and spectrum for hyperparameters from various development set settings of z' . On average, there were approximately 3,510 samples in each bootstrapped development set. The highlighted values correspond to $z' = 2$; under this definition of ‘doing no worse’ on IAT, our algorithm can prescribe $-230.3/3510 \approx 6.6\%$ fewer broad-spectrum antibiotics.

Table 6.22 shows that by setting $z' = 2$ in the development set, better improvements in spectrum can be made than by setting $z' = 1$. Specifically, the increase in inappropriate antibiotic therapy can be limited to $0.43\% \pm 0.25\%$ (mean \pm standard deviation over 20 trials), while reducing the proportion of broad-spectrum antibiotics prescribed by $9.2\% \pm 1.4\%$. Under this setting, the algorithm fell back on the doctor's decision for $5742.6 \pm 232.0 = 76.4\% \pm 3.1\%$ of samples on average. Doing no worse in spectrum while maximizing improvement in IAT does not yield as strong of an improvement, with a decrease in inappropriate antibiotic therapy of $0.7\% \pm 0.3\%$ accompanied by an increase in broad-spectrum prescriptions by $7.5\% \pm 3.8\%$.

Table 6.23 contains the test set IAT and spectrum when thresholds and model classes are decided according to the various development set z' s. Maximizing the improvement in spectrum for $z' = 2$, the increase in inappropriate antibiotic therapy is kept around $0.03\% \pm 0.3\%$, while the prescription of broad-spectrum antibiotics is reduced by $6.6\% \pm 0.4\%$. In making this improvement, the algorithm fell back on the doctor's decision for $2752.4 \pm 34.5 = 78.4\% \pm 1.0\%$ of samples on average.

In summary, using hyperparameters selected from the development set, the decision algorithm based on model predictions is able to reduce (relative to doctors) the proportion of broad-spectrum antibiotics prescribed in the test set by **$6.6\% \pm 0.4\%$** , at the cost of a small increase in test set inappropriate antibiotic therapy by **$0.03\% \pm 0.3\%$** .

6.7 Summary

Urinary tract infections are very common, and comprise approximately 48% of samples in our dataset. This chapter defined, analyzed, and discussed the models predicting resistance for a general UTI cohort and a more homogeneous uncomplicated UTI cohort. For both cohorts, the predictive models yielded clinical insights and confirmed previous knowledge of the strongest factors driving resistance. Models predicting resistance to the general UTI cohort achieved higher AUCs than the uncomplicated UTI cohort, likely due to the relative homogeneity of the uncomplicated UTI cohort.

The uncomplicated UTI cohort was defined such that established clinical guidelines were applicable. When a doctor sees an uncomplicated UTI patient, guidelines recommend four primary antibiotics to choose from in our dataset. By combining a guideline-based decision-making algorithm with the models' predicted probabilities of resistance, this chapter showed that it is possible to maintain similar levels of inappropriate antibiotic therapy to clinicians in our test dataset (2014-2016), while reducing the prescription of broad-spectrum antibiotics by approximately 6.6%.

Chapter 7

Discussion

This chapter describes some of the high-level themes and challenges faced throughout the project, highlighting possible improvements or directions for research in each theme. To conclude, we discuss the limitations and implications of this work.

7.1 Challenge 1: High-Dimensional, Sparse Data

This work draws from many data sources (diagnoses, procedures, medications, lab values, clinicians' notes, etc.), each of which is sparse and high-dimensional in its raw form (e.g: 45,500 distinct diagnoses codes, 25,700 distinct procedure codes, 1,000 distinct lab codes). In addition to the software engineering challenge that this creates (which is not discussed in this thesis), it is often difficult to extract clear signals from high-dimensional data.

To efficiently reduce the dimensionality of structured features, we started by relying on features previously used in literature, as well as domain knowledge from our clinical collaborator. Since these features were hand-picked (see Chapter 3), they were straightforward to interpret.

For clinicians' notes, a bag-of-words representation was used which helped motivate new features and guide our exploration of this high-dimensional space. When these text features were first added, AUCs jumped significantly across all of our models. Thus, we sought to characterize this signal and perhaps extract it in a cleaner

fashion. Since L1 regularization adds the L1-norm of the feature weights into the minimization objective, learned feature weights were pushed towards zero and sparser solutions were encouraged. In earlier experiments, the best models were consistently L1 logistic regression models with high regularization, thereby providing automatic variable selection and a way to focus our attention on the strongest signals.

Initially, logistic regression models with strong L1 regularization would select specimen year as one of the strongest coefficients. This led us to discover a shift in the underlying raw microbiology values, induced by a change in testing equipment (see Chapter 3). Initial models also selected words from discharge summaries indicative of previous microbiology results, and so we decided to add windowed features for previous organism and previous resistance. We also noticed mentions of specific hospital locations, and so we increased the granularity of our location feature from the hospital ward (emergency room, inpatient, outpatient, or intensive care unit) to the hospital floor (specific clinics, wings of the a hospital, etc.). To give our investigations more context, we also experimented with bigrams (which were too computationally intensive to compute for larger cohorts), and created a chart review tool to explore the words surrounding a token of interest, as well as other records belonging to that patient.

Overall, a greedy approach was taken towards extracting features and capturing signal from our high-dimensional dataset. Starting with a model which we found ‘interpretable enough,’ we repeatedly investigated unexpected features or boosts in performance until we were confident in the signal or had updated our models with a new feature explicitly meant to represent that signal.

There are several possible extensions to dealing with high-dimensional, sparse data. Currently, the signal in the bag-of-words representation is quite diluted, and individual tokens can be difficult to interpret without context. One might consider creating topics from the text instead, as described in Chapter 2. Seeking topics that discriminate between resistant and susceptible populations, one might also consider modeling/ extracting topics for the difference in word distributions between texts from the resistant and susceptible populations for each drug-pathogen combination.

Perhaps a more straightforward approach could be manual creation of various regular expressions in order to extract specific concepts. Features extracted from structured data could also be improved, since a selected subset of the information available is being used.

7.2 Challenge 2: Non-Stationarity

As microbiologic testing technologies, practices, and incentives have evolved, so have the distributions seen within our data. One instance of this was after breakpoints were applied to the raw microbiology values, and noticed unexpectedly high levels of resistance from 2000 to 2006. Plotting the raw minimum inhibitory concentrations, we noticed a two-year shift from almost all of the samples having an MIC of 4 to most of the samples having an MIC of 2. Since this was right at the breakpoint cutoff, the samples switched from being mostly susceptible to mostly resistant. This seemed highly unlikely, and upon further investigation, we realized that the testing equipment used to obtain these values had changed. This problem was dealt with simply by excluding the 2000 to 2006 data, but later work could certainly attempt to detect other types of change points more finely and utilize this additional data.

Since patients also change over their lifespans, we wanted to incorporate features that would capture information at different time scales. This was done by computing features over various backwards windows. Even with L1 regularization, different windows for the same type of feature were sometimes selected by the models. Since L1 regularization favors sparser solutions, this implied that a change in the patient's state across different time windows could be useful to the model.

Due to this issue of non-stationarity, we also evaluated how performance might degrade when training on one time window and evaluating on a later non-overlapping time window. This more closely mirrors the deployment setting, when previous information is available and we care only about performance in the future. In our initial experiments with the train/dev set, training on data from two years and evaluating on the following year mostly did not degrade performance, but on the test set of

overlapping patients and future time ranges performance degraded significantly for certain antibiotics. Often, these were models in which decision trees were selected on the development set.

On this dataset (and has been shown with other datasets [40]), we have seen evidence of non-stationarity issues resulting in degraded model performance. Thus, this is a promising dataset to work on improvements in models which may better account for non-stationarity. Already, we have seen changepoints which future work could try to detect. One could also further explore the role of colonization pressure, a feature which was added to attempt to capture the evolving population of pathogens in various locations.

7.3 Challenge 3: Interpretability

In order to gain trust in our models, we wanted to them to be relatively straightforward to interpret. For this reason, results from logistic regression and decision trees to guide most of the analyses. In logistic regression models for each drug we reviewed the coefficients with the greatest magnitude, and in decision trees we visualized the first five layers. Initially shallow feed-forward neural networks were also experimented with, but these took longer to tune, were more difficult to interpret, and performed similarly or worse than logistic regression, decision trees, or random forests. However, logistic regression, decision trees, and random forests are probably not the final story in terms of modeling. As an extension, one could attempt to jointly predict resistance to multiple drugs. There has also been significant work in interpreting more complex models [41][42][43], and there are certainly extensions to be made which could better account for the non-stationary, different types of patients, and underlying biological mechanisms present in our data.

7.4 Challenge 4: Retrospective Analysis

Since a randomized controlled trial was outside the scope of this work, it was difficult to reason retrospectively about how AUCs might translate to clinical practice. While the models output probabilities of resistance, there is no clear record of the doctor’s estimated probabilities at the time of empiric prescription. Additionally, although previous work also reports AUCs and positive predictive values, these values are not directly comparable because different cohorts, antibiotics, and pathogens were selected.

Instead, after extracting doctors’ empiric prescriptions, the predicted probabilities were converted into antibiotic recommendations which would be directly comparable. While this is difficult to do in general, by selecting a specific cohort with clear clinical guidelines, we were able to design and implement a simple decision algorithm that output treatment recommendations based on predicted probabilities. Initially, we experimented with additional options not mentioned in the guidelines (e.g: if a patient is resistant to all four oral antibiotics, consider giving an IV antibiotic), but this led our algorithm to recommend much more drastic measures than doctors would be comfortable with. As a result, the algorithm’s options were restricted just to the antibiotics mentioned in the guidelines. Since the recommendations were then directly comparable to doctors’ recommendations, we were able to evaluate how the models might improve clinical practice.

This challenge of formalizing treatment options and allowable policies is a challenge that has recently been discussed in the context of reinforcement learning for observational health settings [44][45][46]. While we were able to constrain the evaluation setting enough to create a simple algorithm using predicted probabilities to produce comparable treatment recommendations, this is just one facet of model performance. Targeted evaluation of additional cohorts, or more automated methods of finding cohorts for which standard guidelines are applicable, could be future extensions for evaluation. Another interesting research direction is examining the impact of actually applying our treatment recommendations. This moves us into the realm

of off-policy reinforcement learning, which has many interesting theoretical challenges that could be tackled within the context of antibiotic resistance.

7.5 Limitations and Implications

A common limitation seen in analyses using electronic medical records is that they capture infrequent and incomplete information about both the patient and the doctor. Patients may be allergic to certain antibiotics, have personal preferences, or have another confounding reason for which they should not be prescribed an antibiotic. While our clinical collaborator was able to chart review some patients in our cohort to verify that this was not too common, a comprehensive review of all samples would have been impractical. Another common limitation in observational health data is that the comorbidities and procedures are coded for the purpose of billing, which may provide a financially-motivated picture of the patient. Additionally, our dataset may be biased by which samples were sent in for testing in the first place.

One assumption in our uncomplicated UTI evaluation is that patients are only empirically prescribed one antibiotic, for the infection that appears in the microbiology data. While this is usually true for uncomplicated UTIs, in the general case it does not take into account the complexities associated with multiple infections or combination therapies. Thus, our evaluation relies on a clean cohort definition with clear clinical guidelines, both of which could be difficult to find in other cases.

In summary, we synthesize many modalities of data, and extract relevant signal to predict antibiotic resistance with good performance. We evaluate performance from multiple views, ranging from AUCs at a high level, to qualitative evaluation of the learned models, to evaluation on a specific cohort in the context of clinical practice. Our evaluation makes explicit a trade-off which doctors may choose to operate at different levels of depending on the severity of illness. In the future, we plan to extend this analysis to additional cohorts of interest, and hope that our models will allow doctors to more precisely determine the best course of action for a patient.

Appendix A

Dataset Construction

A.1 Feature Extraction

A.1.1 Procedures

There were five main concepts we wished to derive from procedures: presence or history of a central venous catheter, surgery, mechanical ventilation, hemodialysis, and parenteral nutrition. These features were extracted through a combination of string matching and CPT code lookup (note: code is in SQL):

- Central venous catheter:

```
lower(procedure_name) LIKE '%central venous catheter%'
OR (code_type='ICD9'
AND (code='38.97' OR (code>='999.31' AND code<='999.33'))))
OR (code_type='CPT' AND code>='36555' AND code<='36598')
```

- Surgery:

```
lower(procedure_name) LIKE '%surgery%'
OR lower(procedure_name) LIKE '%surgical%'
OR (code_type='CPT' AND code >= '10021' AND code <= '69990')
```

- Mechanical ventilation:

```
lower(procedure_name) LIKE '%ventilation%'
```

- Hemodialysis:

```
(lower(procedure_name) LIKE '%hemodialysis%'
```

```
AND lower(procedure_name) NOT LIKE '%than hemodialysis%')
```

```
OR (code_type='CPT' AND code >= '90935' AND code <= '90940')
```

```
OR (code_type='ICD9' AND code = '39.95')
```

- Parenteral nutrition:

```
lower(procedure_name) LIKE '%parenteral%nutrition%'
```

Bibliography

- [1] Centers for Disease Control and Prevention (CDC). Antibiotic Resistance Threats in the United States. Technical report, 2013.
- [2] Jim O’Neill. Tackling Drug-Resistant Infections Globally: Final Report and Recommendations. Technical report, 2016.
- [3] Centers for Disease Control and Prevention (CDC). Core Elements of Hospital Antibiotic Stewardship Programs. Technical report, 2014.
- [4] Sarah Cobey, Edward B Baskerville, Caroline Colijn, William Hanage, Christophe Fraser, and Marc Lipsitch. Host population structure and treatment frequency maintain balancing selection on drug resistance. *Journal of the Royal Society, Interface*, 14(133), 2017.
- [5] Thamir M Alshammari, E Paul Larrat, Haley J Morrill, Aisling R Caffrey, Brian J Quilliam, and Kerry L Laplante. Risk of hepatotoxicity associated with fluoroquinolones: A national case-control safety study. *American Journal of Health-System Pharmacy*, 71(1):37–43, 2014.
- [6] Sarah R Boggs, Kenji M Cunnion, and Reem H Raafat. Ceftriaxone-Induced Hemolysis in a Child With Lyme Arthritis: A Case for Antimicrobial Stewardship. *Pediatrics*, 128(5):e1289 LP – e1292, 2011.
- [7] Francesco Lapi, MacHelle Wilchesky, Abbas Kezouh, Jacques I. Benisty, Pierre Ernst, and Samy Suissa. Fluoroquinolones and the risk of serious arrhythmia: A population-based study. *Clinical Infectious Diseases*, 55(11):1457–1465, 2012.
- [8] Marjolein P M Hensgens, Abraham Goorhuis, Olaf M. Dekkers, and Ed J. Kuijper. Time interval of increased risk for *Clostridium difficile* infection after exposure to antibiotics. *Journal of Antimicrobial Chemotherapy*, 67(3):742–748, 2012.
- [9] Antimicrobial resistance. Global report on surveillance. *World Health Organization*, 61(3):383–394, 2014.
- [10] C Lee Ventola. The Antibiotic Resistance Crisis: Part 1: Causes and Threats. *P & T : A peer-reviewed journal for formulary management*, 40(4):277–83, 2015.

- [11] World Health Organization. Global action plan on antimicrobial resistance. page 28, 2015.
- [12] World Health Organization. Antibacterial Agents in Clinical Development. *World Health Organization*, page 48, 2017.
- [13] E Chang, J Daly, and D Elliott. *Pathophysiology Applied to Nursing Practice*. Elsevier Australia, 2006.
- [14] Hsi Liu, Thomas H. Taylor, Kevin Pettus, Steve Johnson, John R. Papp, and David Trees. Comparing the disk-diffusion and agar dilution tests for *Neisseria gonorrhoeae* antimicrobial susceptibility testing. *Antimicrobial Resistance and Infection Control*, 5(1):1–6, 2016.
- [15] CLSI. *M100 Performance standards for antimicrobial susceptibility testing. 27th ed.* 2017.
- [16] Jessica M A Blair, Mark A. Webber, Alison J. Baylay, David O. Ogbolu, and Laura J V Piddock. Molecular mechanisms of antibiotic resistance. *Nature Reviews Microbiology*, 13(1):42–51, 2015.
- [17] Dan I. Andersson and Diarmaid Hughes. Antibiotic resistance and its cost: Is it possible to reverse resistance? *Nature Reviews Microbiology*, 8(4):260–271, 2010.
- [18] C. K. Stover, X. Q. Pham, A. L. Erwin, S. D. Mizoguchi, P. Warrener, M. J. Hickey, F. S.L. Brinkman, W. O. Hufnagle, D. J. Kowallk, M. Lagrou, R. L. Garber, L. Goltry, E. Tolentino, S. Westbrook-Wadman, Y. Yuan, L. L. Brody, S. N. Coulter, K. R. Folger, A. Kas, K. Larbig, R. Lim, K. Smith, D. Spencer, G. K.S. Wong, Z. Wu, I. T. Paulsen, J. Relzer, M. H. Saler, R. E.W. Hancock, S. Lory, and M. V. Olson. Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. *Nature*, 406(6799):959–964, 2000.
- [19] T-F Mah, Betsey Pitts, Brett Pellock, Graham C Walker, Philip S Stewart, and George a O Toole. A genetic basis for *Pseudomonas aeruginosa* biofilm antibiotic resistance. *Letters to Nature*, 426(November):1–5, 2003.
- [20] J Davies. Inactivation of antibiotics and the dissemination of resistance genes. *Science*, 264(5157):375 LP – 382, 1994.
- [21] Timothy Sullivan, Osamu Ichikawa, Joel Dudley, Li Li, and Judith Aberg. The Rapid Prediction of Carbapenem Resistance in Patients With *Klebsiella pneumoniae* Bacteremia Using Electronic Medical Record Data. (September 2012):1–6, 2016.
- [22] M. Cristina Vazquez-Guillamet, Rodrigo Vazquez, Scott T. Micek, and Marin H. Kollef. Predicting Resistance to Piperacillin-Tazobactam, Cefepime and Meropenem in Septic Patients with Bloodstream Infection Due to Gram-Negative Bacteria. *Clinical Infectious Diseases*, 65(10):1607–1614, 2017.

- [23] Derek R. MacFadden, Jessica P. Ridgway, Ari Robicsek, Marion Elligsen, and Nick Daneman. Predictive utility of prior positive urine cultures. *Clinical Infectious Diseases*, 59(9):1265–1271, 2014.
- [24] Michael I Jordan David M Blei, Andrew Y Ng. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
- [25] Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Clinical Intervention Prediction and Understanding using Deep Networks.
- [26] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Supplement 1):5228–5235, 2004.
- [27] Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.
- [28] Genevieve Melton and George Hripcsak. Automated Detection of Adverse Events Using Natural Language Processing of Discharge Summaries. *Journal of the American Medical Informatics Association*, 12(4):448–458, 2005.
- [29] O. Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(90001):267D–270, 2004.
- [30] Alan R Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings of the AMIA Symposium*, page 17, 2001.
- [31] C Friedman, L Shagina, Y Lussier, and G Hripcsak. Automated encoding of clinical documents based on natural language processing. *Journal of the American Medical Informatics Association*, 11(5):392–402, 2004.
- [32] Jonathan A C Sterne, Ian R White, John B Carlin, Michael Spratt, Patrick Royston, Michael G Kenward, Angela M Wood, and James R Carpenter. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, 338, 2009.
- [33] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Scientific Reports*, 8(1):1–12, 2018.
- [34] Zachary C. Lipton. The Mythos of Model Interpretability. (Whi), 2016.
- [35] H Quan, V Sundararajan, P Halfon, and A Fong. Coding algorithms for defining comorbidities in. *Medical Care*, 43(11):1130–1139, 2005.

- [36] Ana Flores-Meireles, Jennifer Walker, Michael Caparon, and Scott Hultgren. Urinary tract infections: epidemiology, mechanisms of infection and treatment options. *Nature reviews Microbiology*, 13(5):269–284, 2015.
- [37] Thomas M. Hooton. Uncomplicated Urinary Tract Infection. *New England Journal of Medicine*, 366(11):1028–1037, 2012.
- [38] Tony Mazzulli. Diagnosis and management of simple and complicated urinary tract infections (UTIs). *Clinical infectious diseases*, 19(Suppl. 1):42–48, 2012.
- [39] Rupak Datta and Manisha Juthani-Mehta. Nitrofurantoin vs Fosfomycin: Rendering a verdict in a trial of acute uncomplicated cystitis. *JAMA - Journal of the American Medical Association*, 319(17):1771–1772, 2018.
- [40] Kenneth Jung and Nigam H. Shah. Implications of non-stationarity on predictive modeling using EHRs. *Journal of Biomedical Informatics*, 58:168–174, 2015.
- [41] Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems 30*, (Section 2):4765–4774, 2017.
- [42] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. 2016.
- [43] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-Precision Model-Agnostic Explanations. *Proc. of 32nd Conference on Artificial Intelligence (AAAI)*, pages 1527–1535, 2018.
- [44] Omer Gottesman, Fredrik Johansson, Joshua Meier, Jack Dent, Donghun Lee, Srivatsan Srinivasan, Linying Zhang, Yi Ding, David Wihl, Xuefeng Peng, Jiayu Yao, Isaac Lage, Christopher Mosch, Li-wei H. Lehman, Matthieu Komorowski, Matthieu Komorowski, Aldo Faisal, Leo Anthony Celi, David Sontag, and Finale Doshi-Velez. Evaluating Reinforcement Learning Algorithms in Observational Health Settings. pages 1–16, 2018.
- [45] Aniruddh Raghu, Matthieu Komorowski, Imran Ahmed, Leo A. Celi, Peter Szolovits, and Marzyeh Ghassemi. Deep reinforcement learning for sepsis treatment. *CoRR*, abs/1711.09602, 2017.
- [46] Niranjani Prasad, Li-Fang Cheng, Corey Chivers, Michael Draugelis, and Barbara E. Engelhardt. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. *CoRR*, abs/1704.06300, 2017.