# LEARNING THE STATE OF PATIENT CARE AND OPPORTUNITIES FOR IMPROVEMENT FROM ELECTRONIC HEALTH RECORD DATA

## WITH APPLICATIONS IN BREAST CANCER PATIENTS

By

Morgan R. Harrell

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

BIOMEDICAL INFORMATICS

May, 2017

Nashville, TN

Approved:

Daniel Fabbri, PhD

Mia Levy, MD, PhD

Mark Frisse, MD, MS, MBA

Thomas Lasko, MD, PhD

Robert Johnson, PhD

# ACKNOWLEDGMENTS

This work was financially supported by the National Library of Medicine (NLM) Training Grant T15 LM007450. I am grateful for the NLM's investment and proud to yield this work in return. This work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University, Nashville, TN. I am thankful for the technological access from Vanderbilt which enhanced this work. I am forever indebted to the faculty and staff at Vanderbilt's Department of Biomedical Informatics (DBMI) for building such an amazing training program. I may never fully appreciate the amount of work put into designing coursework, arranging seminars, planning retreats, and setting the curriculum for DBMI students. Nevertheless, I've benefitted from every experience and take incredible pride in the DBMI program.

I would like to thank my committee members: Daniel Fabbri, Mia Levy, Thomas Lasko, Mark Frisse, and Robert Johnson. This group was incredibly supportive and met all of my needs throughout my training. These members were not only invested in development and improvement in this work, but also in my personal progress and growth. Daniel Fabbri, my committee chair, encouraged extracurricular academic and industrial endeavors including an innovation course and internship. These enriching experiences helped me understand my research in a larger body of work and demand. Mia Levy allowed me to shadow her in her busy oncology clinic. This experience helped me understand the end result of my research, where providers deliver healthcare and patients request information and guidance. Thomas Lasko tutored me in machine learning. He broke down and helped me understand the complex concepts applied in this work. Mark Frisse helped immensely in framing this work in a broader context. His coaching helped me clearly communicate the motivation and value in this dissertation. Robert Johnson offered strategies and suggestions that built on foundational statistical methods. His input strengthened the analyses and findings in this work.

I would like to thank the people that gave me support not only in this academic endeavor, but throughout my full life. My dad, Scott Harrell, taught me integrity through example. He takes pride in doing things well regardless of scale - from the way he built his business to the way he fries an egg. He is my role model and in his voice, ”do the right thing” will always echo in my mind. My mom, Maria Harrell, goes above and beyond to ensure my well-being. My Aunt Laura, a teacher and strong advocate for education, always motivated me and rallied for my success. I would also like to thank my step-mom Alicia, my siblings Lauren, Ashley, and Corey, and two fantastic friends, Toni and Nicole, for surrounding me with love.

I met many people during this journey whose relationships will stay with me through life. Abigail Lind has been my side-by-side companion each year in DBMI and I am so appreciative to have met such a wonderful friend. I know we will be celebrating milestones together well beyond our time at Vanderbilt. I always enjoyed connecting with my DBMI peers including Josh Smith, Jacob VanHouten, Lina Sulieman, Laura Wiley, Ravi Atreya, Sharon Davis, Alex Cheng and Linda Zhang. Conversations with these folks, no matter how brief, were always enlightening, encouraging, and fun. If anyone begins to feel that graduate school is an isolating experience, just point them toward the break room at DBMI.

I would last like to thank Naqi Khan, a very special person who has immensely brightened my life since since meeting at Vanderbilt. Naqi is the first person I turned to many times for support. He always met my needs even when I couldn’t vocalize them. He has lifted me up, brought my feet back to the ground, challenged my ideas, and interjected time to relax. Naqi is responsible for much of my confidence and resolve, and I will forever appreciate his help in completing this dissertation.

Thank you all.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

# Chapter 1

# Introduction

## 1.1 Overview

Patient care is a multifaceted, complex process. Clinical guidelines outline preferred care practices, but because patients have unique needs, real-world practice can often differ from guidelines. Patient care is also imperfect. Missed opportunities and errors occur when the number of decision factors exceeds what is manageable for the care provider [1]. Nevertheless, computational tools including database systems and machine learning can help make complex problems more manageable. To meet the demand for improved patient care with technological solutions, it is necessary to design scientifically rigorous methods that characterize the **sufficiency** of patient data, **state** of patient care, and **opportunities** for improvement through the steps of data extraction, data analysis, and projections from data.

Characterizing the current state of patient care, and finding opportunities for improvement with computational tools requires sufficient patient data. Electronic health records (EHRs) are digital stores of patient health records, and capture data including patient appointment times, medication events, and billing codes [2]. EHRs have many clinical benefits including improved workflow and error reduction [3]. Additionally, the aggregated data from EHRs is beneficial in a secondary use - scientific research.

Applying the data stored in EHRs to scientific research is challenging. EHR data are often unstructured, and requires tools like Natural Language Processing (NLP) to transfer unstructured text into structured data fields [4]. EHR data is often inconsistent, incomplete and prone to errors [5]. For example, instances like death and treatment transfer are not always documented in the EHR. Therefore, terminal patient data does not inform

on the patient's current state. Furthermore, EHR data can follow different standards as EHRs evolve, and across EHR systems, requiring reconciliation for longitudinal or cross-institutional studies [6].

Although applying EHR data to research is challenging, it produces significant positive contributions and healthcare advancements. Contributions include improved communication between patients and care providers [7], identification of missed diagnoses [8], reduced errors [9] and improved patient care [10]. Advancements include genomic study enhancement [11], clinical decision support [12], and data mining techniques [13]. EHR data are an important factor for learning treatment practices and outcomes in the general patient population, and for identifying opportunities for improvement.

EHR data analysis can characterize the state of patient care and learn opportunities for improvement. State of patient care refers to the distribution of patients within a clinical workflow. For example, patients requiring long-term follow-up are distributed across 'adherent to follow-up' and 'loss to follow-up' states. These states are further divided into many other states, for example, 'receiving medication' and 'not receiving medication' states. Identifying an opportunity for improvement refers to 1) identifying a suboptimal patient state and 2) determining a means to drive patient distribution away from that state. If a patient's state is 'receiving medication' and 'reporting adverse symptoms,' switching their medication or treating their symptoms may move them to a 'no adverse symptom' state.

One patient cohort in which we can learn the state of patient care and opportunities for improvement is breast cancer patients undergoing long-term adjuvant endocrine therapy. Adjuvant endocrine therapy is prescribed to breast cancer patients post surgery, chemotherapy or radiation therapy to reduce risk of recurrence. Adjuvant endocrine therapy is recommended for at least a five year duration to minimize risk of recurrence [14]. The adjuvant endocrine therapy patient cohort is a model test-case to learn the state of patient care and opportunities for improvement for several reasons. 1) Although clinical trials report rates

of drug adherence and outcomes [15][16], there is little information on adjuvant endocrine therapy in practice in the general breast cancer patient population. 2) Adjuvant endocrine therapy has options for treatment, yielding many states in which patients are distributed. 3) Adjuvant endocrine therapy is a long-term treatment, during which patients move between states. Understanding the state of adjuvant endocrine therapy through EHR data, and finding opportunities for improvement, may drive improved patient care in the adjuvant endocrine therapy patient population.

This dissertation describes a novel set of methods to learn the state of patient care and opportunities for improvement from EHR data. Our approach is divided into three aims: 1) determine **sufficiency** of the data, 2) characterize the **state** of patient care, and 3) identify **opportunities** for improvement. Sufficient data is necessary to characterize the state of patient care, and characterizing the state of patient care is necessary to find opportunities for improvement. To meet standards for scientific discovery, these methods are designed to be reproducible and generalizable. These methods require common datatypes captured by EHRs, which facilitates application to many EHR systems. Additionally, these methods are founded at basic patient care and are extensible and customizable to other healthcare domains. Generalizable and reproducible methods for learning the state of care and opportunities for improvement from EHR data will yield valuable information for many patient cohorts, and drive better patient care in different settings.

## 1.2 Background

### 1.2.1 Electronic Health Record Systems

Electronic health records (EHRs) are digital stores of patient charts and healthcare information. Early EHR systems were introduced in the 1960s to allow clinicians quick access to the most current versions of patient records [17]. As EHRs developed, they began to provide clinical benefits that improve efficiency and quality of care such as increased

guideline adherence, enhanced monitoring, and decreased medication errors [18]. Clinics report fewer errors [19] and fewer costs [20] after transitioning from paper charts to an EHR system. The well documented benefits helped incite Meaningful Use, a government program that incentivizes EHR adoption and use [21] [22]. Through Meaningful Use, the percent of physicians using EHRs rose from 44% in 2009 [23] to 83% in 2015 [24].

The widespread adoption of EHRs opens doors for a beneficial secondary use of patient data: scientific research. Aggregated patient data allows for scientific analyses that give insight into patient care and outcomes. Patterns learned from aggregated EHR data facilitate clinical modeling [25] , data mining [26], and clinical decision support [27]. However, there are challenges in drawing knowledge from EHR data. Data incompleteness, inaccuracies, and inconsistencies are issues limiting secondary use studies [28] [29]. EHR data contains patient health information on care received within the system that the EHR spans. Care received in alternative EHR systems is not always shared, making patient health records fragmented across systems. Additionally, EHR data may contain errors, non-standardized data, and other issues which require data-cleaning and processing [5].

Vanderbilt University Medical Center (VUMC) is a group of hospitals and clinics that has served over 3 million patients. VUMC's EHR system has been evolving for close to two decades. Table 1.1 lists VUMC's EHR functionalities and the years they were implemented. The implementation year denotes the year in which VUMC's EHR began capturing respective datatypes. The system captures a wide range of patient data including billing codes, medications, and clinical communications. VUMC's EHR data are extracted, de-identified, and structured as part of the Synthetic Derivative (SD) for secondary use.

### 1.2.2 Adjuvant Endocrine Therapy

Hormone receptor-positive (HR+) breast cancers are uncontrolled growths dependent on hormone intake (mainly estrogen intake) for proliferation [30]. Approximately 70% of breast cancers are HR+ [31] and can be treated by preventing cells from binding estro-

Table 1.1: Vanderbilt University Medical Center's Electronic Health Record functionalities and the years they were implemented

| Function | Implementation Year |
|---|---|
| Inpatient electronic clinical documentation | 1997 |
| Outpatient electronic clinical documentation | 2001 |
| Patient summary service (PSS) for medications | 2003 |
| Outpatient e-prescribing | 2003 (Variable provider adoption over time) |
| Inpatient order entry system | 2004 |
| Nursing Bar Code Medication Administration (AdminRx) | 2007 |

gen. Endocrine therapy interferes with estrogen intake, and is prescribed as a treatment to shrink HR+ tumors prior to surgery. More commonly, endocrine therapy is prescribed as an adjuvant treatment to prevent cancer recurrence post treatment [32]. There are several classes of adjuvant endocrine therapy drugs, and they prevent estrogen intake by different mechanisms.

Selective estrogen-receptive modulators (SERMs) bind to estrogen receptors on cells rendering them unreceptive to extracellular estrogen. SERMs do not lower the overall levels of estrogen in the body, therefore are prescribed to pre- and peri- menopausal women. Potential side effects of SERMs include but are not limited to hot flashes, and increased risk of blood clots, cataracts, and uterine cancer. In the ATAC adjuvant trial, approximately 40% of patients on SERMs reported hot flashes [15].

Aromatase inhibitors (AIs) prevent aromatase enzymes from converting androgen into estrogen, which lowers the amount of estrogen in the body. AIs are effective when a patient's main source of estrogen is androgen conversion, a characteristic of post-menopausal women. Pre- and peri-menopausal woman must undergo natural or artificial menopause in order for AIs to be effective.

There are two types of AIs: steroidal and non-steroidal. Steroidal AIs form non-reversible bonds with aromatase enzymes, while non-steroidal AIs form reversible bonds and actively compete with androgen at binding sites. There is a lack of clinical trials focused on efficacy differences between steroidal and non-steroidal AIs, but because steroidal AIs are purported to have androgenic effects, non-steroidal AIs are the recommended first line AI treatment [33]. A common side effect of AIs is arthritis. A study by Henry et al reported that out of a 97 breast cancer patients taking AIs, 44 experienced musculoskeletal side effects, and 13 cases were severe enough for the patient to discontinue AI use [34]. Hot flashes, another common side effect of AIs, appeared in approximately 35% of patients on AIs in the ATAC adjuvant trial [15].

In 1977, the Food and Drug Administration (FDA) approved the first endocrine therapy, tamoxifen (a SERM), for use in metastatic hormone receptor positive breast cancer. In 1990, tamoxifen was approved for use in the treatment of early stage hormone receptor positive breast cancer [35]. It was subsequently approved for use in prevention of breast cancer in 1998 and for treatment of non-invasive ductal carcinoma in-situ (DCIS) in 2000. Clinical studies on tamoxifen show that it consistently reduces risk for death and tumor recurrence in HR+ breast cancer patients [36].

In 2002, the first aromatase inhibitor, anastrozole, was approved and demonstrated superiority over tamoxifen in the adjuvant treatment of post-menopausal woman with breast cancer [15]. Subsequently, in 2005, letrozole and exemestane, two alternative aromatase inhibitors, were also approved as adjuvant endocrine therapies in this setting.

In 1998, clinical guidelines recommended adjuvant endocrine therapy for a five year duration post initial treatment[16]. Recent guidelines extended treatment duration up to ten years in at-risk populations [14]. With extended timeframes for duration of treatment and the growing number of available drug options, adjuvant endocrine therapy treatment paths can vary across patients. Furthermore, clinical trials report varying rates of drug termination (31-73%) [15][37], adverse events [34][38], and drug switches [39]. There is

limited information on adjuvant endocrine treatment, including outcomes and follow-up, in real-world settings.

Varying drug use and rates of adverse symptoms create complex states of patient care for adjuvant endocrine therapy patients. Furthermore, since treatment is long-term, patients can change states over the course of care. EHR data, including medications and billing codes, from an adjuvant endocrine therapy patient can identify the patient's state to the extent of the EHR system. Sufficient data from a cohort of adjuvant endocrine therapy patients can characterize the collective state of patient care, which in turn aids in identifying opportunities for improvement.

### 1.2.3 Cohort Selection

Cohort selection for retrospective studies is mostly performed on an ad hoc basis. Cohort and data selection directly drives results and limitations. For example, building a cohort consisting of patients with perfect and complete EHR data biases results to possibly 1) local patients and 2) patients diligent towards maintaining their health. Excluding patients that are not local to the clinic's location and are at higher risk for non-adherence fails to represent real patient populations. Many retrospective studies require cohort selection, and many studies depend on inter-study comparison of results when there is no gold-standard. Methodology for cohort and data selection is useful in standardizing the cohort selection process. Studies that use standard and consistent methodology for cohort selection reduce variability in cohort selection that affects results. Comparing study results that follow the same methodology for cohort selection has fewer limitations than comparing study results that follow different cohort selection processes. Building a logical methodology for cohort selection in retrospective EHR studies may benefit many secondary use studies.

### 1.2.4 Machine Learning

Machine learning is a computational method to learn patterns from datasets. Learned patterns take the form of a function that generalizes to new data points. Machine learning is used to solve prediction problems where a vector maps to a specific outcome and there is a desire to compute the correct outcome.

Machine learning is divided into supervised and unsupervised methods. Unsupervised machine learning identifies structure or similarities among data [40]. Unsupervised machine learning can be abstract. There is no right answer to the underlying relationship of the data, and there is no direct evaluation of the model's performance. For example, given a dataset of patient medications from both male and female patients, unsupervised machine learning may predict the females as one patient group and males as a different patient group.

Supervised machine learning requires labeled data to train models [40]. Models should generalize to correctly predict labels for new labeled data points. For example, given a training dataset of patient medications mapped to a diagnosis that they treat, supervised machine learning will build a function to best determine the diagnosis from the medications. Then, given a set of medications withheld from the training dataset, the function will predict a diagnosis that the medications treat. Metrics like accuracy or area under the receiver operating characteristic curve (ROC AUC) can evaluate model performance [40]. Accuracy is the percent of correct predictions out of all predictions. The AUC is the probability that a model will assign a higher probability prediction to a random positive class vector than a random negative class vector.

Matching machine learning methods with clinical datasets yields predictive models for clinical outcomes. Electronic health record systems are valuable data sources for feature vector construction for many prediction problems. For example, unsupervised learning methods can predict phenotypic patterns from EHR data that facilitate personalized medicine [41]. Supervised machine learning methods can predict re-admittance rates from

congestive heart failure patient EHR data [42]. Predicting outcomes in patient care is one step toward patient care improvement. For example, predicting follow-up in the case of long-term patients may identify features for patients that don't follow-up. Informing clinicians that a patient may not follow-up allows them to coach the patient, or adjust treatment, to improve follow-up rates.

### 1.2.5 Pre-Existing Efforts to Learn the State of Patient Care and Opportunities for Improvement

There are previous efforts to learn the state of patient care and opportunities for improvement in healthcare domains through data collection and statistical analysis. Two of those efforts include The Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute, and The Patient-Centered Outcomes Research Institute (PCORI).

The SEER Program compiles and distributes information on cancer statistics for the purpose of informing the U.S. population on cancer burdens [43]. SEER began collecting data in tumor registries in 1973. Since then, SEER expanded it's tumor registry number to cover a wide range of demographics. SEER currently holds data on approximately 28 percent of the US population, including 26 percent of African Americans, 38 percent of Hispanics, 44 percent of American Indians and Alaska Natives, 50 percent of Asians, and 67 percent of Pacific Islanders. Data that SEER collects includes primary tumor sites, stage at diagnosis, first course of treatment, vital status, etc. From the SEER registries, overall cancer incidence, survival, and the first course of treatment can be summarized. Although SEER data can characterize the state of care through survival and first course of treatment, SEER data lacks other data that characterizes the complete state of care. For example, SEER data does not include changes in treatment and patient side effects. In order to characterize the complete state of care, additional data such as those in electronic health records is necessary. Nevertheless, SEER data is a valuable resource for cancer research

and are used by many researchers, clinicians, public health officials, and patients.

The PCORI organization is an independent non-profit group with the aim to deliver information that drives desired health outcomes. PCORI grants fund patient-centered comparative clinical effectiveness research (CER). Their research mission is to help patients and providers make informed healthcare decisions, improve healthcare delivery and outcomes, by creating empirical research guided by patients, providers, and the broader healthcare community [44]. PCORI breaks down their initiative into five priorities: 1) Assessment of prevention, diagnosis, and treatment options, 2) Improving healthcare systems, 3) Communication and dissemination research, 4) Addressing disparities, and 5) Accelerating patient-centered outcomes research and methodological research. PCORI's interests align with learning the state of patient care and opportunities for improvement, and PCORI grants fund research germane to the methods in this dissertation.

# Chapter 2

# Research Statement

This dissertation constructs methodology that determines the **sufficiency** of patient data, characterizes the **state** of patient care, and identifies **opportunities** for improvement from electronic health record (EHR) data. The methods are applied to an adjuvant endocrine therapy patient cohort treated at Vanderbilt University Medical Center (VUMC) using VUMC's EHR system. Chapter 3 describes steps to determine the sufficiency of EHR data availability, and the role of data availability in data selection and interpreting results (i.e. inferring failure to follow-up vs treatment transfer). Chapter 4 describes statistical and visualization methods to learn the state of patient care. We measure drug use, frequencies of adverse symptoms, and disease recurrence. Last, chapter 5 describes steps to explore opportunities for improvement in patient care with machine learning. We look for signals to indicate failure to follow-up at VUMC for five-years (the minimum recommended treatment duration). The complete workflow is shown in Figure 2.1.

## 2.1 Aims

### 2.1.1 Aim 1: Determine sufficiency of the data

Data sufficiency is the availability and limitations of data for research. For data to be sufficient for a long-term patient cohort analysis, data must span the timeframe for which treatment occurs. Data should begin at or before a minimum time prior to data collection, and data should persist throughout the timeframe. However, when data are inconsistent or missing, they may be sufficient *with limitations*. Rather than including or excluding all imperfect data, we can weight data importance based on availability of alternative data types. Determining data sufficiency is challenging due to characteristics of the EHR. EHR

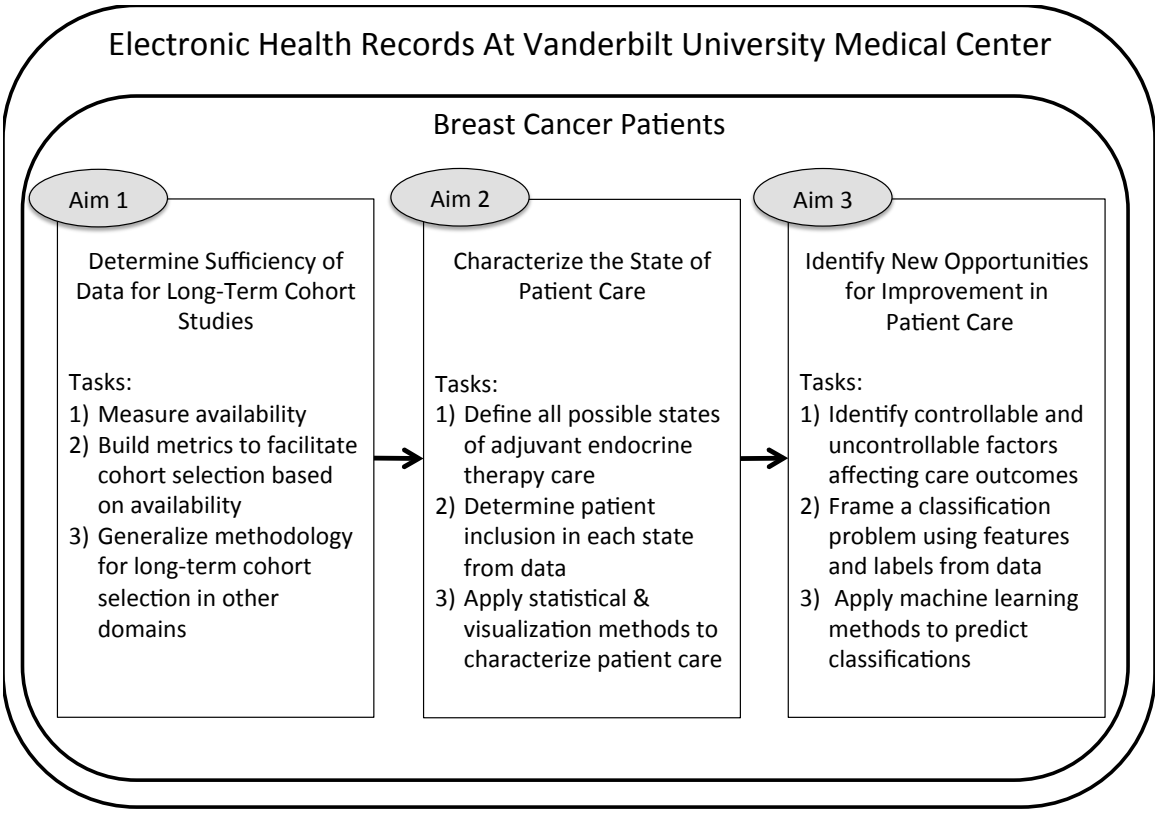



Figure 2.1: Complete workflow of the dissertation for learning the state of patient care and opportunities for improvement from EHR data. All work is completed from electronic health record data from breast cancer patients treated at Vanderbilt University Medical Center.

adoption and functionality implementation occurs at different rates, which leads to differing data availability. Patient data spans the timeframe of the patient's return visits, which may not always be the optimal timeframe. Determining data sufficiency allows for more reliable data, and accurate interpretation of results. Tasks to determine data sufficiency include 1) measuring data availability, 2) building metrics based on data availability to facilitate data selection, and 3) generalizing the methodology for application to other health care domains.

### 2.1.2 Aim 2: Characterize the state of patient care

The state of patient care is the distribution of patients across a clinical workflow. Characterizing the state of patient care from EHR data allows for insight on treatment patterns in the real-world setting. Information on the state of patient care in real-world setting can give clinicians and patients realistic expectations about treatment, and help them make decisions regarding care. For example, in an adjuvant endocrine therapy patient population, we understand that patients often discontinue treatment, or change their treatment plan, but the rate of those occurrences and the timeframe in which they happen are unmeasured. If and patient and provider learn that patients using a specific drug change their treatment plan at a certain time, the patient and provider can expect and plan for that event. The current state of patient care is challenging to characterize because it requires many aggregated structured statistics and data types like appointment times, medication events and billing codes. Analyzing the state of patient care is particularly challenging in long-term cohorts due to changes in data availability and data types over time. Tasks to characterize the state of patient care include: 1) mapping the possible states of care, and 2) determining distribution of patients across states from EHR data.

### 2.1.3 Aim 3: Identify new opportunities to improve patient care

Opportunities for patient care improvement are signals in the EHR data that indicate a patient is in a suboptimal patient state. Signals can facilitate patient care improvement

suggest a means of shifting patient distribution away from the suboptimal state into a more favorable state. For example, patients in a state with adverse side effects may be moved into a state without adverse side effects by 1) treating their side effects or 2) altering their medications. Finding signals for patient care improvement is challenging because it requires large amounts of structured data and many different data types. Tasks to identify new opportunities to improve patient care include 1) identifying controllable and uncontrollable factors affecting care outcomes, 2) framing a classification problem using features and labels from EHR data, and 3) applying machine learning methods to identify signals for patient care improvement.

## 2.2 Value of the Work

### 2.2.1 Biomedical Informatics Techniques

Technical achievements of this work include measuring the completeness and limitations of VUMC's EHR data for characterizing and improving breast cancer treatment. Our methods and metrics facilitate secondary use of VUMC's EHR data in breast cancer studies. Additionally, our exploration of machine learning applications to EHR data support machine learning as a useful tool to improve breast cancer treatment.

This work enhances biomedical informatics techniques by defining tasks necessary to learn the state of patient care and opportunities for improvement. We address limitations and steps to improve methodology. For example, we promote health information exchange and the need for additional data sources to enhance the understanding of healthcare problems. Additionally we took steps to understand the generalizability of our methods to other healthcare domains.

### 2.2.2 Clinical Knowledge

This work benefits the breast cancer clinicians and patients by providing an empirical summary of adjuvant endocrine therapy treatment at VUMC. Summary statistics on adjuvant endocrine therapy hallmarks, including drug switches, drug discontinuation, and recurrences, provide realistic expectations for treatment at VUMC. Realistic expectations translate into better planning for care. Furthermore, predicting follow-up informs clinicians of the significant features that indicate whether a patient will complete adjuvant endocrine therapy at VUMC, which facilitates planning for continuing care.

The broader value of this work benefits the biomedical informatics research community. Our methods to determine data sufficiency, characterize the state of patient care, and identify opportunities for improvement among adjuvant endocrine therapy patients act as a scaffold to generalize and extend to other healthcare domains. This work defines necessary tasks to learn the state of patient care and opportunities from improvement.

# Chapter 3

# Aim 1: Determine Sufficiency of the Data

## 3.1 Overview

This chapter aims to determine the **sufficiency** of Electronic Health Record data to learn the state of patient care and opportunities for improvement. The text is an extended version of an academic journal article titled *Evaluating EHR Data Availability for Cohort Selection in Retrospective Studies* published at the *IEEE International Conference on Health Informatics*. The article describes metrics for data availability among adjuvant endocrine therapy patients treated at Vanderbilt University Medical Center. We address the rise of data population as VUMC's EHR evolves, data persistence over time, and strategies to handle missing data. Additionally, we generalize our methods for applications of EHR data extraction in other healthcare domains. Once data completeness and limitations were properly defined, we tested our methods to select a cohort of adjuvant endocrine therapy patients for a longitudinal study on five-year follow-up.

The methods in this chapter are related to measurements of exposure and outcomes in the field of statistics. Exposure and outcome are observable traits with a relationship such that exposure affects the outcome [45]. In this chapter, data availability is the exposure variable, and data persistence over some time is the outcome. As time periods for data availability increases, the probability of data persistence decreases.

This chapter shows the impact of data sufficiency on secondary use of EHR data and contributes generalized methods for data selection based of EHR data sufficiency. The main finding of this study are 1) data sufficiency can drive data selection for secondary use studies, 2) data sufficiency metrics can serve as weights for missing data points, and 3) and data sufficiency affects secondary use study results.

## 3.2 Introduction

Retrospective studies with longitudinal medical data can answer questions about medication compliance, adverse drug events, and therapeutic efficacy in a population. Sources for robust longitudinal medical data include Electronic Health Record (EHR) systems, billing systems, and registries. EHRs collect various medical data-types for patients over time including billing codes, medication events, and diagnoses. However, there are several challenges to using longitudinal EHR data in retrospective studies.

Retrospective studies require a patient cohort with reliable longitudinal data, including accurate start and end points. Due to characteristics of the EHR, it is often unclear when reliable data begins and ends. Evolving EHR functionality and disparate provider adoption leads to discrepancies in EHR data population. For example, Vanderbilt University Medical Center (VUMC) initially implemented its EHR in the inpatient setting prior to the outpatient setting, resulting in temporally more complete data for patients admitted as inpatients. Additionally, EHRs infrequently document patients exiting the healthcare system, leaving an indeterminate end date for patient EHR data. Applying longitudinal EHR data in retrospective studies requires 1) determining when past data becomes consistently available and 2) determining when patient data ends.

One strategy for using longitudinal EHR data in a retrospective study is selecting a patient cohort with data that 1) begins after a threshold for consistent data availability and 2) persists for a specified timeframe measured through records and clinical activity in the EHR. This can be achieved by defining data availability metrics, and applying the metrics to maximize the amount of reliable data and minimize the amount of ambiguous data included in the study. Currently, patient cohort selection for retrospective studies is done on an ad hoc basis. While there have been many studies using billing codes and medication events extracted from the EHR [46] [47] [48], and many studies using weighted EHR data [49] [50], to our knowledge there have been no studies applying EHR data availability to retrospective study cohort selection. We sought to improve cohort selection by creating

a generalizable heuristic based on EHR data availability. We 1) created data-driven metrics for longitudinal EHR data availability, 2) utilized the metrics for data restriction and weighting in a cohort selection heuristic, and 3) tested our heuristic on a cohort of stage I-III breast cancer patients for a retrospective study on adjuvant endocrine therapy adherence.

## 3.3 Background

There are many real-world scenarios that make identifying longitudinal cohorts from EHR data challenging. Differing timeframes for EHR adoption and added functionality results in staggered data availability. VUMC's EHR system has been evolving for close to two decades. Inpatient electronic clinical documentation began in 1997, and outpatient electronic clinical documentation began in 2001. The patient summary service (PSS) for medications and outpatient e-prescribing began in 2003, although with variable provider adoption over time. The inpatient order entry system began in 2004, and Nursing Bar Code Medication Administration (AdminRx) began in 2007. As a result, the available data for patients changed over time in direct relation to added EHR functionality. Furthermore, lack of documentation on patients leaving the EHR system result in ambiguous data termination. Reasons that a patient may leave the healthcare system include 1) death, 2) completion of care, 3) transfer of care to another institution and 4) discontinuation of care. Although death events can be clearly documented in most EHRs, many patients die outside of the hospital. As a result, death events are not consistently communicated back to the managing healthcare system and thus not documented in the EHR. Alternative resources for ascertaining death for a large population are typically required [51]. Likewise, when a patient completes their treatment and is discharged from active care, there is limited structured documentation in the EHR to indicate this plan. Finally, when patients transfer their care to a different facility, the managing provider may never be notified. A request for outside medical records may be the only indication of such a transition, but there is no place to document this in a structured way in current EHR systems.

Despite data inconsistencies, EHRs are a valuable data source for longitudinal retrospective studies. VUMC's Synthetic Derivative is a data source of de-identified health records from over 2 million patients [52]. The Synthetic Derivative includes over 200 million billing codes in the form of ICDs and CPTs, and over 400 million medication events derived from medication lists and clinical notes with the natural language processing method, MedEx [4]. Additionally, the Synthetic Derivative also includes a tumor registry that contains detailed diagnosis, staging and treatment information for over 90,000 patients with cancer linked to their electronic health record.

One retrospective study requiring longitudinal medical data is an analysis on adjuvant endocrine therapy adherence. Adjuvant endocrine therapy is prescribed to hormone receptor positive breast cancer patients to prevent tumor recurrence. Women who complete five years of the adjuvant endocrine therapy drug, Tamoxifen, have a significantly lower risk of breast cancer recurrence and mortality than women who only complete 1-2 years [53]. Unfortunately, drug side effects like hot flashes, arthritis pains, mood disturbances, and bone loss often make it difficult for patients to complete five years of treatment [54]. In the ATAC clinical trial, approximately 85% of tamoxifen-treated patients adhered to five years of treatment [15]. Studies on adherence in the general patient population report lower and varying rates. One study reports 69% five-year adherence through patient interviews during treatment [55]. This study does not depend on EHR data availability, but patient interviews are not always viable for retrospective studies. An alternative study reports 49% five-year adherence from automated pharmacy records of hormonal therapy prescriptions and refills [56]. Patients in the study were censored at date of dis-enrollment in the system (among other reasons), which may be underreported. The study does not seek alternative data availability in the EHR to confirm patient enrollment, leaving an opportunity to improve results.

Applying EHR data availability metrics to cohort selection for a retrospective adjuvant endocrine therapy adherence study can optimize data utility and lead to more accurate

Figure 3.1: Generalized cohort selection heuristic employing EHR data availability metrics

**Generalized Heuristic for Patient Cohort Selection in Retrospective EHR Studies**

- Select patient population with desired medical event

  *(i.e. adjuvant endocrine therapy patients)*

- Set desired length of longitudinal data and data intervals

  *(i.e. 5 years, 12 month intervals)*

- Set date restrictions to data based on the selected medical event

  *(i.e. clinical guidelines began recommending adjuvant endocrine therapy for at least 5 years in 1998)*

- Set restrictions based on data availability metrics

  Set threshold for data availability

  Set weighting parameter (include/exclude)

- Select cohort meeting all rescrictions

- Perform analysis (if necessary, recalibrate and repeat)

results. Furthermore, a generalized heuristic employing EHR data availability metrics can facilitate cohort selection for a multitude of longitudinal retrospective studies.

## 3.4 Methods

Our objective is to create a heuristic for selecting cohorts with sufficient longitudinal EHR data. The heuristic should consider when data becomes consistently available, and adjust for the rate at which patients leave the EHR system. To construct this heuristic, we define data availability metrics that determine the start and end of EHR data per patient, and then use these metrics in rules for cohort selection. We applied this heuristic to a breast cancer cohort for an adjuvant endocrine therapy adherence analysis.

Table 3.1: Selected Adjuvant Endocrine Therapy Drugs

| Generic Name | Other Names | Drug Class | Year of FDA approval for metastatic therapy | Year of FDA approval for adjuvant therapy |
|---|---|---|---|---|
| Anastrozole | Arimidex | Non-steroidal AI | 2000 | 2002 |
| Exemestane | Aromasin | Steroidal AI | 2005 | 2005 |
| Letrozole | Femara | Non-steroidal AI | 2005 | 2005 |
| Tamoxifen | Nolvadex | SERM | 1977 | 1990 |

### 3.4.1 Data extraction

Data was collected from the VUMC Synthetic Derivative and includes 1) billing codes (ICDs, CPTs), which depict any billable event recorded in the EHR system, 2) medication event data, which depict drug information documented in the EHR for the patient, and 3) tumor registry data, which is populated for patients diagnosed at VUMC or receiving the majority of their treatment at VUMC. Tumor registry data includes the cancer diagnosis site (e.g. breast) and histology (e.g invasive ductal carcinoma), date of diagnosis, cancer stage, and vital status including date of last known contact and date of death. The adjuvant endocrine therapy drugs identified for our adherence analysis are listed in Table 3.1.

### 3.4.2 Metrics

To determine when data are consistently available, we first summarized the growth of billing code, medication event, and tumor diagnosis data points in the EHR for all patients and for the subset of patients diagnosed with stage I-III breast cancer.

Second, we calculated the percentage of patients who had a particular data type (billing code, medication event, or adjuvant endocrine therapy medication event) available in a particular twelve-month interval before or after their diagnosis year. Specifically, given a

year $y$ in which a clinical event begins, and the $i^{th}$ 12-month interval post clinical event, let cohort $C_{yi}$ be the $n$ patients with the clinical event in year $y$ with no recorded death before interval $i$. Then, the number of patients, in $C_{yi}$ with EHR data in interval $i$ is:

$$D(C_{yi}) = \sum_{c=0}^{c=n} \begin{cases} 0 & \text{if } \emptyset \text{ datapoints in } i \\ 1 & \text{if } \geq 1 \text{ datapoints in } i \end{cases} \tag{3.1}$$

Finally, the percentage of data available is calculated as the number of patients with greater than zero data points during the 12 month interval divided by all patients in the given diagnosis-year group.

$$\textit{Data Availability}(C_{yi}) = \frac{D(C_{yi})}{n} \tag{3.2}$$

Patients were censored from subsequent intervals if they had a reported death in the tumor registry.

The change in data availability from a given year over subsequent 12 month intervals measures how long data are consistently available. A rise in data availability denotes that previous data was not substantially populated. A drop in data availability denotes patients leaving the EHR system. For example, if we measure 95% data availability for breast cancer patients diagnosed in 2006, and 73% data availability in that population post five years, than we know 22% of those patients left the EHR system in that timeframe, whether it be from transferred treatment or discontinued treatment etc. We use these data availability measures as weights for patients in the cohort (described in subsection 3.4.4).

### 3.4.3 Visualization

To visualize our data availability measures, we built a series of heat maps. Availability metrics are calculated for patients grouped by initial diagnosis year and normalized on the initial diagnosis year. Each measure in the heatmap is the percentage of data available

calculated by equation 3.2. Patients were censored from subsequent intervals if they had a reported death in the tumor registry.

### 3.4.4 Cohort Creation Heuristic

We defined a heuristic that employs cohort restrictions specific to adjuvant endocrine therapy and EHR data availability metrics. For patients to be included in the cohort, their date of diagnosis must be equal or later than 1998, when clinical guidelines began to recommended five years of adjuvant endocrine therapy [16], and the date of diagnosis must be equal to or later than dates where 90% of a given EHR data-type is populated. Our desired study timeframe is five years, and we chose 12-month intervals for which to extract data points since patients on adjuvant endocrine therapy typically have follow-up appointments at least every 6 months for the first five years. If a patient does not have an instance of a given data type for a 12-month interval, we infer that their longitudinal data has ended.

Given our available data-types (billing codes and medication events) and data availability metrics, we created six separate cohorts for adjuvant endocrine therapy adherence analyses Table 3.2. The cohorts are restricted by availability of data types or combination of data types, and weighted by data persistence. Data type availability restrictions require that patients have greater than zero data points of a specific data type per 12-month interval. Data persistence, defined as the data availability in the $i^{th}$ year, can be used as a weight to estimate a patient's persistence in treatment, and allows for an upper and lower bound on patients included in the cohort.

We can propose two strict assumptions for adjuvant endocrine therapy patients with discontinued data prior to five years: 1) all are non-adherent to adjuvant endocrine therapy, or 2) all are continuing treatment elsewhere. We know that the true explanation lies between these extremes. So, we create an upper bound cohort, where we exclude patients with a missing data point for a 12-month interval, and a lower bound cohort, where we assume all patients with a missing data point for a 12 month interval are non-adherent at the rate

Table 3.2: Cohort Restrictions for Adjuvant Endocrine Therapy Analysis

| | No Weighting (Exclude Patients with Missing Data) | Weight Patients with Missing Data |
|---|---|---|
| Restrict on Billing Code Availability | Cohort 1 Upper Bound | Cohort 1 Lower Bound |
| Restrict on Medication Event Availability | Cohort 2 Upper Bound | Cohort 2 Lower Bound |
| Restrict on Both Billing Code and Medication Event Availability | Cohort 3 Upper Bound | Cohort 3 Lower Bound |

of data availability in that interval. We believe this is a reasonable lower bound because this represents a natural rate of patients leaving the system. Determining adherence within these different cohorts allows for a range of adherence rates and a margin of error.

We generalized our steps for adjuvant endocrine therapy cohort selection so they may be employed toward cohort selection in a multitude of retrospective studies requiring longitudinal EHR data (Figure 3.1).

### 3.4.5 Adjuvant Endocrine Therapy Adherence Analysis

We determined the completion rate each year for five years of adjuvant endocrine therapy in each of the six patient cohorts. The rate of completion is the number of patients with at least one adjuvant endocrine therapy medication event documented in the EHR per year divided by the total number of patients in the cohort.

Table 3.3: Summary of data availability from VUMC EHR.

| Data source | Tumor Registry | Billing data | Any medication event | Endocrine therapy medication event |
|---|---|---|---|---|
| Dates data available in source | 1960-2014 | 1987-2014 | 1984-2014 | 1991-2014 |
| Count of patients with data available in source | 84857 | 2.2 million | 777000 | 8534 |
| Count of data elements available in source | 90000 | 167 million | 363 million | 690000 |
| Year data first available for breast cancer cohort | 1964 | 1987 | 1988 | 1991 |
| Count of stage I-III breast cancer patients with data in source | 5824 | 5240 | 3492 | 1945 |
| Count of data elements available for cohort in source | 5824 | 2 million | 8 million | 401000 |

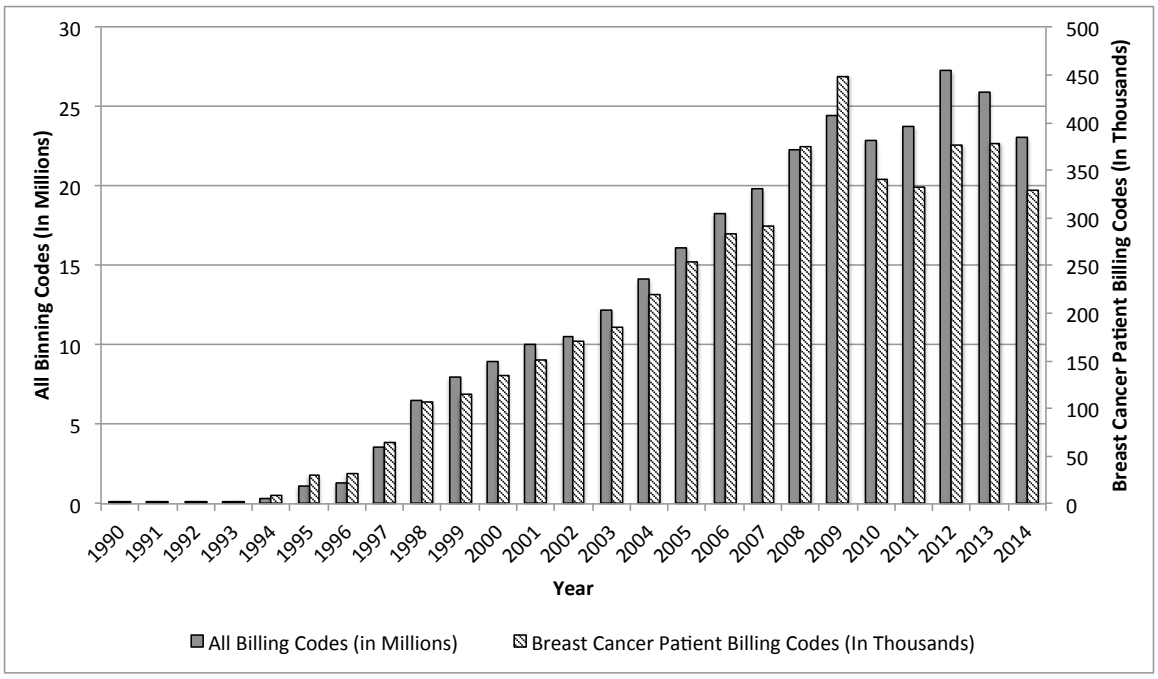
## 3.5 Results

### 3.5.1 Metrics and Visualization

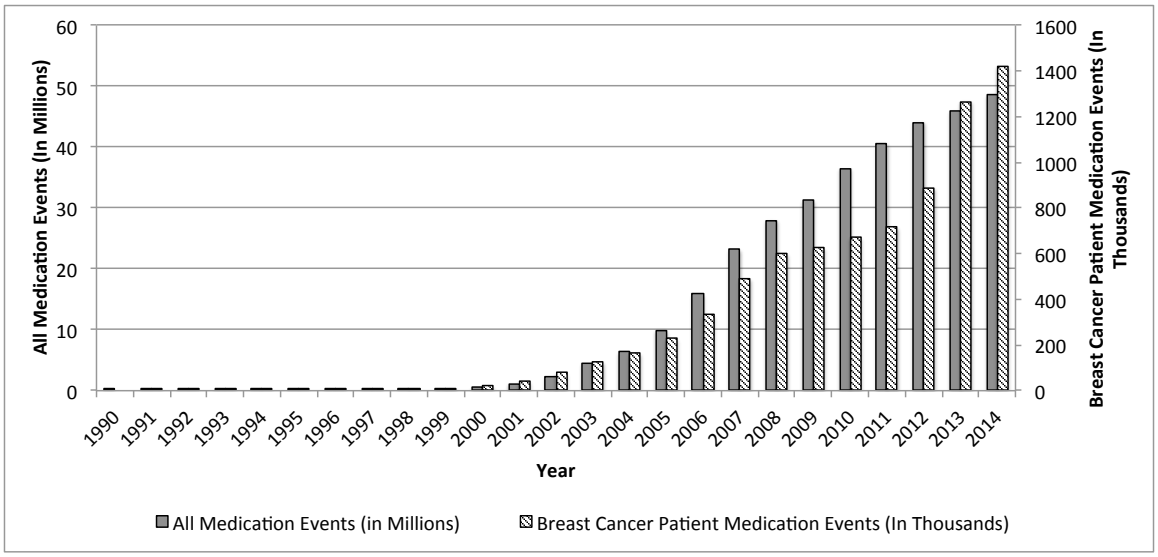
The rise in data availability of billing codes, medication events, and tumor registry diagnoses within VUMC's EHR for both the general population and the subset of stage I-III breast cancer patients are shown in Figure 3.2. Both EHR utilization and VUMC patient volumes increased during the forty-year period. Summary statistics for data availability are

in in Figure 3.3. We identified 5824 patients with stage I-III breast cancer in the VUMC Tumor Registry beginning in 1964. The first billing data became available for this patient cohort in 1987, the first medication events in 1988, and the first endocrine therapy events in 1991. Although tumor registry data predates billing and medication event data by over 20 years, 59% percent of all patients in in the tumor registry and 60% of stage I-III breast cancer patients in the tumor registry had all three data types available.

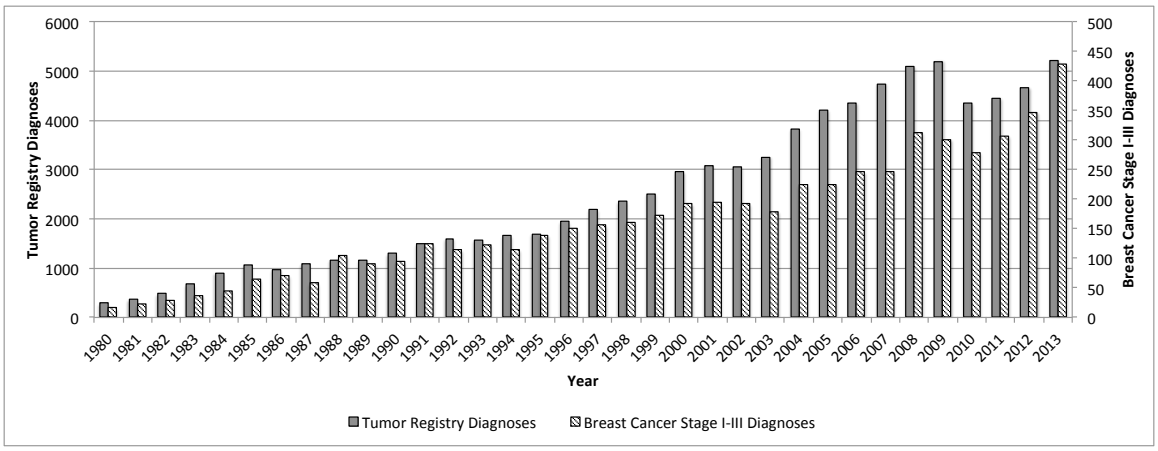
The percent of data population in the EHR for billing codes, medication events, and adjuvant endocrine therapy medication events for stage I-III breast cancer patients are shown as heatmaps in Figures 3.3, 3.4, and 3.5. In each heatmap, the diagnosis year is represented along the y-axis, and 12-month intervals are represented along the x-axis. Y-1 is 12 months prior to the diagnosis date, Y0 is 12 months after the diagnosis date, and so on. The values in the heatmap are the percentages of patients in a given group with greater than zero data points for that term. Ideally, all patients would have billing codes and medication events during their year of diagnosis, resulting in 100% data population for all diagnosis years in the Figure 3.3 and Figure 3.4 Y0 columns. Lesser percentages denote unpopulated data in the EHR. Patients who receive adjuvant endocrine therapy are those with hormone receptive positive breast cancers, which make up approximately 70-80% of breast cancers [31]. Therefore, data population in Figure 3.5 Y0 column is ideally 70-80%, and extend through Y4 column for five years of therapy adherence.




(a) Billing Codes




(b) Medication Events




(c) Stage I-III Breast Cancer Diagnoses

Figure 3.2: Counts of billing codes, medications and tumor registry patients by year in the VUMC EHR for all patients and for the subset of stage I-III breast cancer patients.

| | Y-1 | Y0 | Y1 | Y2 | Y3 | Y4 | Y5 | Y6 | Y7 | Y8 | Y9 | Y10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1990** | 0.00 | 0.01 | 0.00 | 0.00 | 0.10 | 0.49 | 0.57 | 0.54 | 0.66 | 0.65 | 0.60 | 0.50 |
| **1991** | 0.01 | 0.05 | 0.01 | 0.05 | 0.54 | 0.51 | 0.52 | 0.55 | 0.50 | 0.56 | 0.45 | 0.40 |
| **1992** | 0.00 | 0.06 | 0.11 | 0.47 | 0.66 | 0.65 | 0.58 | 0.55 | 0.50 | 0.45 | 0.41 | 0.46 |
| **1993** | 0.00 | 0.15 | 0.63 | 0.64 | 0.63 | 0.63 | 0.60 | 0.61 | 0.54 | 0.55 | 0.55 | 0.52 |
| **1994** | 0.16 | 0.71 | 0.65 | 0.65 | 0.63 | 0.57 | 0.56 | 0.52 | 0.45 | 0.45 | 0.51 | 0.51 |
| **1995** | 0.46 | 0.86 | 0.72 | 0.73 | 0.68 | 0.66 | 0.67 | 0.58 | 0.56 | 0.43 | 0.43 | 0.44 |
| **1996** | 0.38 | 0.82 | 0.78 | 0.74 | 0.67 | 0.61 | 0.61 | 0.55 | 0.51 | 0.49 | 0.47 | 0.46 |
| **1997** | 0.31 | 0.87 | 0.73 | 0.65 | 0.60 | 0.58 | 0.52 | 0.48 | 0.44 | 0.40 | 0.43 | 0.46 |
| **1998** | 0.45 | 0.90 | 0.85 | 0.77 | 0.74 | 0.69 | 0.66 | 0.60 | 0.64 | 0.64 | 0.65 | 0.62 |
| **1999** | 0.39 | 0.93 | 0.88 | 0.80 | 0.75 | 0.69 | 0.70 | 0.65 | 0.62 | 0.58 | 0.55 | 0.52 |
| **2000** | 0.39 | 0.95 | 0.82 | 0.78 | 0.76 | 0.72 | 0.70 | 0.66 | 0.62 | 0.57 | 0.54 | 0.50 |
| **2001** | 0.45 | 0.94 | 0.81 | 0.76 | 0.67 | 0.66 | 0.63 | 0.62 | 0.62 | 0.52 | 0.50 | 0.47 |
| **2002** | 0.44 | 0.94 | 0.87 | 0.74 | 0.72 | 0.67 | 0.68 | 0.61 | 0.64 | 0.56 | 0.56 | 0.57 |
| **2003** | 0.46 | 0.93 | 0.86 | 0.86 | 0.79 | 0.71 | 0.71 | 0.72 | 0.67 | 0.68 | 0.67 | 0.63 |
| **2004** | 0.44 | 0.94 | 0.87 | 0.82 | 0.78 | 0.75 | 0.72 | 0.69 | 0.68 | 0.67 | 0.65 | 0.44 |
| **2005** | 0.42 | 0.94 | 0.85 | 0.77 | 0.75 | 0.73 | 0.68 | 0.66 | 0.64 | 0.60 | 0.45 | |
| **2006** | 0.41 | 0.95 | 0.87 | 0.84 | 0.77 | 0.73 | 0.75 | 0.74 | 0.72 | 0.46 | | |
| **2007** | 0.45 | 0.97 | 0.91 | 0.83 | 0.74 | 0.77 | 0.78 | 0.74 | 0.54 | | | |
| **2008** | 0.51 | 0.97 | 0.87 | 0.81 | 0.80 | 0.79 | 0.73 | 0.55 | | | | |
| **2009** | 0.50 | 0.96 | 0.88 | 0.86 | 0.81 | 0.78 | 0.60 | | | | | |
| **2010** | 0.44 | 0.94 | 0.88 | 0.86 | 0.86 | 0.66 | | | | | | |
| **2011** | 0.34 | 0.95 | 0.90 | 0.86 | 0.57 | | | | | | | |
| **2012** | 0.48 | 0.96 | 0.92 | 0.68 | | | | | | | | |
| **2013** | 0.45 | 1.00 | 0.74 | | | | | | | | | |
| **2014** | 0.59 | 0.97 | | | | | | | | | | |

Figure 3.3: BILLING CODES: Percentage of breast cancer stage I-III patients with greater than zero billing codes within 12-month intervals surrounding their diagnosis date. Diagnosis date is represented along the y-axis and 12-month intervals are represented along the x-axis. Y-1 is the 12-month period prior diagnosis date, Y0 is the 12-month period after diagnosis, and so on.

| | Y-1 | Y0 | Y1 | Y2 | Y3 | Y4 | Y5 | Y6 | Y7 | Y8 | Y9 | Y10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1990 | 0.00 | 0.04 | 0.20 | 0.09 | 0.16 | 0.19 | 0.14 | 0.34 | 0.39 | 0.42 | 0.36 | 0.44 |
| 1991 | 0.01 | 0.19 | 0.18 | 0.14 | 0.17 | 0.18 | 0.29 | 0.32 | 0.30 | 0.29 | 0.34 | 0.32 |
| 1992 | 0.05 | 0.17 | 0.13 | 0.16 | 0.17 | 0.24 | 0.27 | 0.28 | 0.26 | 0.31 | 0.30 | 0.43 |
| 1993 | 0.05 | 0.30 | 0.17 | 0.17 | 0.30 | 0.39 | 0.40 | 0.40 | 0.44 | 0.40 | 0.51 | 0.51 |
| 1994 | 0.06 | 0.50 | 0.30 | 0.48 | 0.48 | 0.47 | 0.39 | 0.40 | 0.43 | 0.43 | 0.44 | 0.48 |
| 1995 | 0.05 | 0.50 | 0.43 | 0.53 | 0.50 | 0.50 | 0.51 | 0.46 | 0.48 | 0.44 | 0.44 | 0.43 |
| 1996 | 0.09 | 0.69 | 0.62 | 0.53 | 0.53 | 0.48 | 0.48 | 0.43 | 0.42 | 0.44 | 0.43 | 0.40 |
| 1997 | 0.17 | 0.84 | 0.61 | 0.52 | 0.55 | 0.53 | 0.48 | 0.43 | 0.41 | 0.37 | 0.39 | 0.42 |
| 1998 | 0.24 | 0.86 | 0.72 | 0.71 | 0.65 | 0.61 | 0.61 | 0.57 | 0.57 | 0.60 | 0.64 | 0.64 |
| 1999 | 0.24 | 0.89 | 0.79 | 0.67 | 0.68 | 0.67 | 0.69 | 0.65 | 0.61 | 0.56 | 0.54 | 0.55 |
| 2000 | 0.30 | 0.90 | 0.73 | 0.73 | 0.73 | 0.67 | 0.67 | 0.65 | 0.63 | 0.59 | 0.54 | 0.54 |
| 2001 | 0.29 | 0.89 | 0.75 | 0.74 | 0.61 | 0.60 | 0.60 | 0.60 | 0.59 | 0.53 | 0.52 | 0.48 |
| 2002 | 0.38 | 0.93 | 0.85 | 0.74 | 0.71 | 0.68 | 0.68 | 0.63 | 0.65 | 0.64 | 0.59 | 0.60 |
| 2003 | 0.40 | 0.91 | 0.82 | 0.84 | 0.79 | 0.72 | 0.72 | 0.72 | 0.71 | 0.68 | 0.67 | 0.64 |
| 2004 | 0.40 | 0.94 | 0.86 | 0.82 | 0.77 | 0.73 | 0.72 | 0.70 | 0.68 | 0.66 | 0.67 | 0.53 |
| 2005 | 0.38 | 0.93 | 0.82 | 0.77 | 0.73 | 0.75 | 0.70 | 0.66 | 0.66 | 0.61 | 0.51 | |
| 2006 | 0.37 | 0.93 | 0.86 | 0.85 | 0.78 | 0.77 | 0.76 | 0.73 | 0.75 | 0.64 | | |
| 2007 | 0.41 | 0.97 | 0.92 | 0.85 | 0.82 | 0.79 | 0.79 | 0.75 | 0.66 | | | |
| 2008 | 0.48 | 0.96 | 0.89 | 0.83 | 0.81 | 0.82 | 0.77 | 0.66 | | | | |
| 2009 | 0.48 | 0.97 | 0.90 | 0.86 | 0.84 | 0.80 | 0.72 | | | | | |
| 2010 | 0.44 | 0.94 | 0.90 | 0.88 | 0.87 | 0.78 | | | | | | |
| 2011 | 0.37 | 0.96 | 0.91 | 0.89 | 0.74 | | | | | | | |
| 2012 | 0.52 | 0.99 | 0.93 | 0.78 | | | | | | | | |
| 2013 | 0.52 | 1.00 | 0.85 | | | | | | | | | |
| 2014 | 0.61 | 0.98 | | | | | | | | | | |

Figure 3.4: MEDICATION EVENTS: Percentage of breast cancer stage I-III patients with greater than zero medication events within 12-month intervals surrounding their diagnosis date. Diagnosis date is represented along the y-axis and 12-month intervals are represented along the x-axis. Y-1 is the 12-month period prior to diagnosis date, Y0 is the 12 month period after diagnosis, and so on.

| | Y-1 | Y0 | Y1 | Y2 | Y3 | Y4 | Y5 | Y6 | Y7 | Y8 | Y9 | Y10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1990 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 | 0.07 | 0.03 | 0.14 | 0.18 | 0.18 | 0.15 | 0.08 |
| 1991 | 0.00 | 0.00 | 0.00 | 0.02 | 0.03 | 0.05 | 0.15 | 0.22 | 0.20 | 0.26 | 0.15 | 0.16 |
| 1992 | 0.00 | 0.01 | 0.00 | 0.04 | 0.06 | 0.19 | 0.23 | 0.17 | 0.11 | 0.09 | 0.07 | 0.15 |
| 1993 | 0.00 | 0.03 | 0.04 | 0.11 | 0.28 | 0.34 | 0.32 | 0.28 | 0.29 | 0.29 | 0.26 | 0.25 |
| 1994 | 0.00 | 0.03 | 0.09 | 0.32 | 0.40 | 0.38 | 0.36 | 0.35 | 0.31 | 0.27 | 0.26 | 0.26 |
| 1995 | 0.01 | 0.08 | 0.20 | 0.33 | 0.31 | 0.30 | 0.31 | 0.25 | 0.25 | 0.26 | 0.23 | 0.19 |
| 1996 | 0.00 | 0.35 | 0.44 | 0.40 | 0.36 | 0.33 | 0.29 | 0.26 | 0.24 | 0.21 | 0.24 | 0.24 |
| 1997 | 0.02 | 0.61 | 0.50 | 0.44 | 0.42 | 0.44 | 0.38 | 0.32 | 0.28 | 0.28 | 0.28 | 0.26 |
| 1998 | 0.02 | 0.71 | 0.60 | 0.52 | 0.47 | 0.44 | 0.43 | 0.39 | 0.38 | 0.38 | 0.35 | 0.33 |
| 1999 | 0.01 | 0.72 | 0.57 | 0.46 | 0.47 | 0.43 | 0.42 | 0.35 | 0.35 | 0.30 | 0.29 | 0.30 |
| 2000 | 0.00 | 0.70 | 0.56 | 0.47 | 0.45 | 0.44 | 0.42 | 0.35 | 0.37 | 0.33 | 0.30 | 0.28 |
| 2001 | 0.01 | 0.65 | 0.48 | 0.45 | 0.37 | 0.37 | 0.35 | 0.31 | 0.31 | 0.27 | 0.22 | 0.24 |
| 2002 | 0.02 | 0.71 | 0.54 | 0.47 | 0.47 | 0.46 | 0.46 | 0.38 | 0.42 | 0.41 | 0.39 | 0.38 |
| 2003 | 0.00 | 0.68 | 0.52 | 0.51 | 0.50 | 0.48 | 0.47 | 0.44 | 0.41 | 0.38 | 0.39 | 0.36 |
| 2004 | 0.02 | 0.67 | 0.57 | 0.55 | 0.53 | 0.51 | 0.49 | 0.47 | 0.47 | 0.42 | 0.42 | 0.35 |
| 2005 | 0.00 | 0.59 | 0.52 | 0.49 | 0.47 | 0.49 | 0.49 | 0.43 | 0.42 | 0.39 | 0.34 | |
| 2006 | 0.01 | 0.59 | 0.53 | 0.52 | 0.48 | 0.51 | 0.51 | 0.47 | 0.47 | 0.42 | | |
| 2007 | 0.02 | 0.66 | 0.63 | 0.58 | 0.57 | 0.59 | 0.57 | 0.54 | 0.49 | | | |
| 2008 | 0.02 | 0.68 | 0.61 | 0.55 | 0.56 | 0.57 | 0.53 | 0.46 | | | | |
| 2009 | 0.00 | 0.63 | 0.58 | 0.56 | 0.55 | 0.54 | 0.48 | | | | | |
| 2010 | 0.02 | 0.66 | 0.62 | 0.61 | 0.59 | 0.54 | | | | | | |
| 2011 | 0.02 | 0.67 | 0.62 | 0.62 | 0.58 | | | | | | | |
| 2012 | 0.01 | 0.71 | 0.67 | 0.62 | | | | | | | | |
| 2013 | 0.01 | 0.76 | 0.71 | | | | | | | | | |
| 2014 | 0.04 | 0.74 | | | | | | | | | | |

Figure 3.5: ADJUVANT ENDOCRINE THERAPY: Percentage of breast cancer stage I-III patients with greater than zero adjuvant endocrine therapy medication events (medications listed in Table 1) within 12 month intervals surrounding their diagnosis date. Diagnosis date is represented along the y-axis and 12-month intervals are represented along the x-axis. Y-1 is the 12-month period prior to diagnosis date, Y0 is the 12 month period after diagnosis, and so on.

### 3.5.2 Adjuvant Endocrine Therapy Cohort Selection

The cohort restrictions imposed on stage I-III breast cancer patients in the VUMC tumor registry are described in Figure 3.6. We required that 1) patients be diagnosed between 1998 and 2010, and 2) patients must have at least one adjuvant endocrine therapy medication event. Restrictions imposed though our heuristic required that 1) patient data begins at or after 90% availability for a given data-type and 2) data persists each year up to five years measured by existence of greater than 0 data points per year. We found 1710 patients with appropriate billing code availability. 1711 patients with appropriate medication event availability, and 1627 patients with both appropriate billing code and medication event availability.

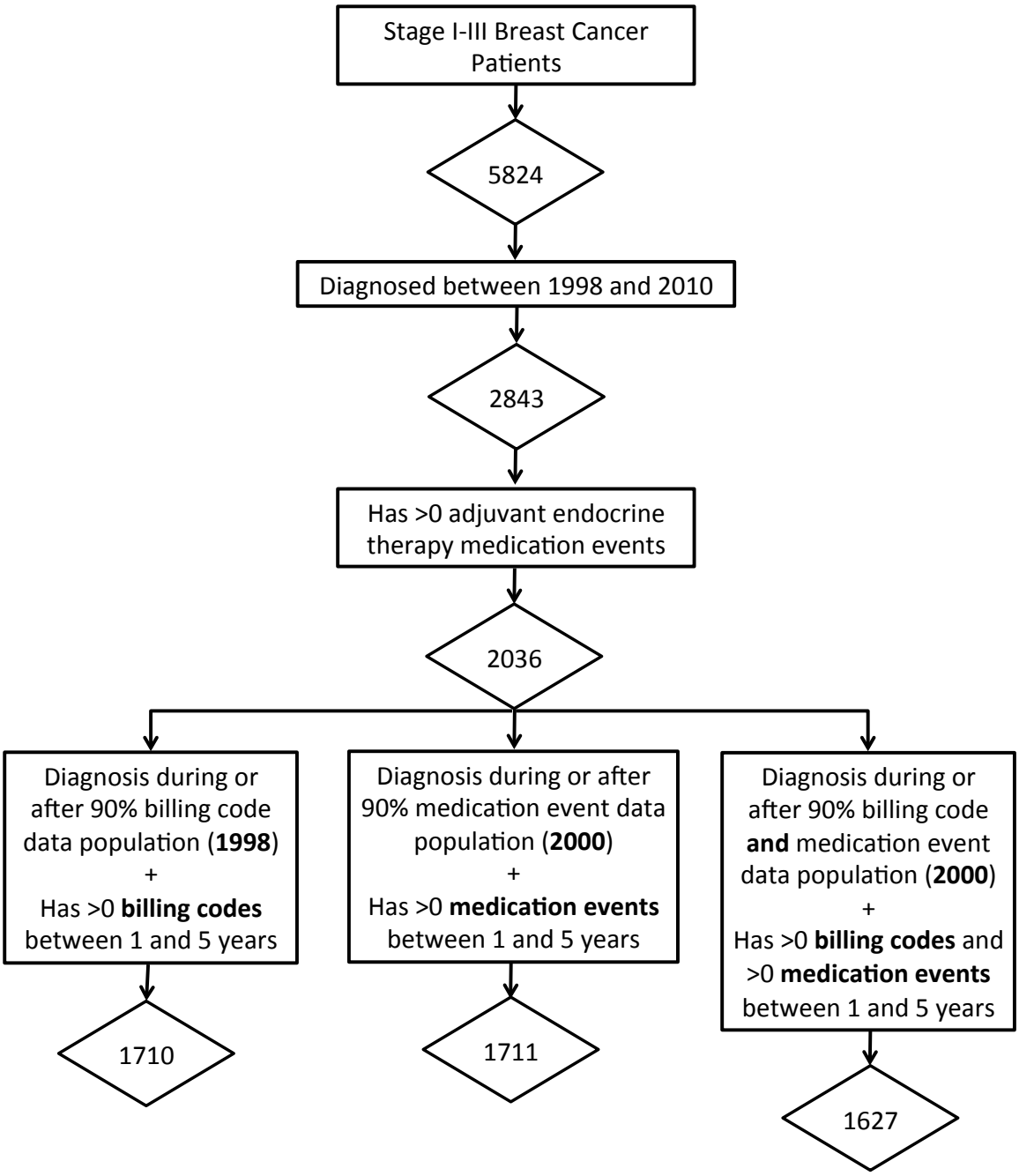



Figure 3.6: Patient cohort selection flowchart.

### 3.5.3 Adjuvant Endocrine Therapy Adherence

Our adjuvant endocrine therapy adherence analysis determines the number of patients with at least one adjuvant endocrine therapy medication event per year, beginning from

their first adjuvant endocrine therapy drug up to five years, divided by the total number of patients in the given cohort. Adherence rates per cohort are graphed in Figure 3.7. The five-year upper bound completion rate is between 74 and 78% and the five-year lower bound completion rate is between 55 and 57% across the respective cohorts.

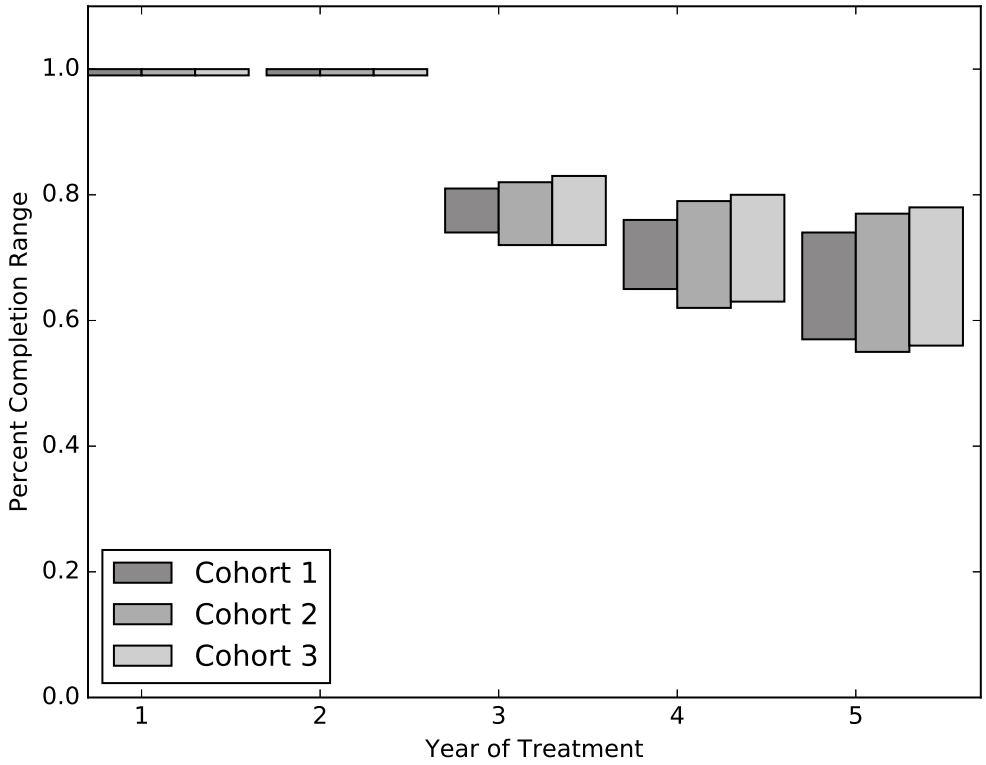



Figure 3.7: Upper and lower bound adjuvant endocrine therapy adherence through five years in cohorts selected with the EHR data availability heuristic (Cohorts described in Table II).

## 3.6 Discussion

Electronic health record (EHR) systems are a valuable source for longitudinal medical data, but applying that data to retrospective studies requires determining accurate data start and endpoints. To determine data start and endpoints, we constructed data availability metrics using data from VUMC's EHR, a system that has been evolving for nearly two

decades. We applied those metrics in a heuristic for cohort selection in a retrospective, adjuvant endocrine therapy adherence study. Then, we generalized our heuristic so it may be applied to a multitude of retrospective studies requiring longitudinal medical data. From our study, we found several opportunities and limitations in applying longitudinal EHR data to retrospective studies.

Data population in the EHR at Vanderbilt University Medical Center has been growing since implementation due to increased functionality of the EHR, improved documentation, and growing number of patients seen at the medical center. Differences in EHR implementation yield disparate availability for data-types over time. For instance, in breast cancer patients, billing code data reaches 90% population at an earlier year than medication data (year 1998, column Y0 for billing code data (Figure 3.3) and year 2000, column Y0 for medication event data (Figure 3.4)). Learning when EHR data becomes consistently available allows cohort restriction for reliable data.

We found that data availability prior to breast cancer diagnosis (column Y-1, in Figure 3.3 and Figure 3.4) increases over time, but does not reach 90% population and may not be sufficient for inclusion in longitudinal studies. Few data points for patients prior to a breast cancer diagnosis is expected due to VUMC's referral patterns.

We found that in some cases, a smaller percentage of patients had adjuvant endocrine therapy medication events prior to their breast cancer diagnosis (column Y-1, Figure 3.5). There are a few possible explanations for this: 1) They had DCIS (stage 0) non-invasive breast cancer prior to their Stage I-III invasive breast cancer diagnosis are were receiving risk reducing endocrine therapy, 2) The patient never had breast cancer but was at high risk for getting breast cancer and was on risk reducing endocrine therapy, or 3) the NLP algorithm extracted the information in error. We also found that many patients had more than five years of adjuvant endocrine therapy medication events. This could be due to 1) clinical trials open to VUMC patients that extend adjuvant endocrine therapy, or 2) patients with locally recurrent or metastatic recurrent disease receiving endocrine therapy

for a different indication.

Through applying our cohort selection heuristic to adjuvant endocrine therapy patients, we determined upper and lower bound adherence rates at each year for five years. At the end of five years, adjuvant endocrine therapy adherence falls between 55% and 78% for patients treated at VUMC. Our heuristic yields higher higher rates of adherence than a previous retrospective study using EHR data without data availability metrics (49%) [55] and lower rates of adherence than recent clinical trials (85%) [15].

The major contribution of this work is that it shows the impact of data availability on secondary use of EHR data in retrospective studies. Missing data is the main contributing factor to the wide range of adherence rates. VUMC's EHR lacks follow up information for 23% (78% adherence rate - 55% adherence rate) of stage I-III breast cancer patients over a five-year time frame. While missing data may always be a challenge in retrospective studies, improved communication to managing healthcare systems and improved documentation in the EHR can reduce missing data in future time points.

Our generalized heuristic extracted from the methods and findings in this study can facilitate cohort creation for retrospective studies using longitudinal EHR data. Longitudinal EHR data is a valuable basis for retrospective studies provided that the limits of the data are investigated.

## 3.7 Limitations

This study is limited by the data captured in the Electronic Health Record system and VUMC. Patient loss in the system is not reported in the EHR. When a patient no longer has data in the EHR, it can be due to 1) an unreported death, or 2) discontinued care at VUMC. In the second event, it is unclear whether the patient is receiving care in a non-VUMC system or whether they have ended treatment. Additional datasets can help enrich this study and better understand the patient's state. Health plan datasets and additional health information exchange can clarify patient loss in the VUMC system and allow for improved

measures of adjuvant endocrine therapy adherence.

This work could further determine availability of patient death records by comparison to the National Death Registry. Bridging to an external dataset in this case is challenging because our data source was de-identified health records, which do not map to external datasets. In different cases with identified data, there is potential to map to other datasets and improve these methods.

## 3.8 Aim 1 Conclusions

This study describes metrics for data availability among adjuvant endocrine therapy patients treated at Vanderbilt University Medical Center. We measure the rise of data population as VUMC's EHR evolves, data persistence over time, and strategies to handle missing data. Additionally, we generalize our methods for applications of EHR data extraction in other healthcare domains. Application of these metrics facilitated cohort selection for a longitudinal study on adjuvant endocrine therapy follow-up. We show the impact of data sufficiency on secondary use of EHR data in retrospective studies and contribute generalized methods for cohort selection based of EHR data sufficiency.

The main findings of the study include:

- Data sufficiency can drive data selection for secondary use studies
- Data sufficiency metrics can serve as weights for missing data points
- Data sufficiency affects secondary use study results

# Chapter 4

# Aim 2: Characterize the State of Patient Care

## 4.1 Overview

This chapter aims to characterize the **state** of patient care using Electronic Health Record data. The text is an extended version of an academic article titled *Analysis of Adjuvant Endocrine Therapy from Breast Cancer Patient EHR Data* and published by *JCO Clinical Cancer Informatics*. The article defines the states of adjuvant endocrine therapy, which include receiving patient care at VUMC, receiving care outside of VUMC, discontinuing care, or death. We measure patient inclusion in states within the VUMC system, and discusses the limitations of knowledge on patient states outside of the VUMC system. We measure hallmarks of adjuvant endocrine therapy including drug switches, drug discontinuation, recurrence and death - occurrences reported in clinical trials, but previously unmeasured in a real-world patient population.

This chapter shows the potential of electronic health record data to characterize the state of patient care for adjuvant endocrine therapy patients. Additionally, this chapter describes the limitations of electronic health records in characterizing the entirety of care for a condition, and discusses generalizability for characterizing the state of patient care in other healthcare domains. the main conclusions of the paper are 1) 49% of VUMC adjuvant endocrine therapy patients were lost to follow-up or did not complete adjuvant treatment throughout five years, 2) 52% percent of VUMC patients switched to a different endocrine therapy drug during their treatment, and 3) VUMC's EHR is a valuable resource for characterizing the state of adjuvant endocrine therapy for VUMC patients.

## 4.2 Introduction

Adjuvant endocrine therapy is prescribed to hormone receptor positive breast cancer patients for recurrence prevention post surgery. Clinical guidelines recommend patients use adjuvant endocrine drugs for a five-year duration, with recent guidelines extending the recommendation to ten years for some high-risk patient populations [14]. In this timeframe, patients may deviate from their original treatment plan in the form of a drug switch or termination [54]. Although clinical trials report rates of drug switches and termination [15], the prevalence and motivations for these events in the general breast cancer patient population are unmeasured.

There are several reasons a patient may change their adjuvant endocrine therapy treatment. A drug switch may result from a change in menopausal status, intolerable side effects, tumor recurrence, generic drug alternatives, or physician preference. Drug discontinuation may result from intolerable side effects, financial factors, or death. Measuring rates of drug switches, termination, and their possible cause in patients in the general breast cancer population yields an empirical projection for treatment strategies and outcomes.

Electronic health record systems (EHRs) store patient medical data including medications, billing codes, and diagnoses. EHRs allow for mining large quantities of adjuvant endocrine treatment data to determine treatment trends over time. To understand treatment patterns for adjuvant endocrine therapy in the general patient population, we analyzed Vanderbilt University Medical Center's (VUMC) EHR data for 1,587 stage I-III breast cancer patients. Treatment data includes adjuvant endocrine therapy drugs taken by the patient, timestamps for each drug, and ICD9 codes. Our goals are to 1) determine the frequencies of drug switches and discontinuation, and 2) determine the potential cause for drug switches and discontinuation. This study describes long-term adjuvant endocrine treatment in real-world settings, and demonstrates the ability to use electronic health record data to characterize oral medication treatment patterns in patients with cancer.

## 4.3 Background

Hormone receptor-positive (HR+) breast cancers make up 70% of breast cancers [31] and can be treated by preventing cells from taking in estrogen. Endocrine therapy interferes with estrogen intake, and is prescribed as a neoadjuvant treatment to shrink HR+ tumors prior to surgery. More commonly, endocrine therapy is prescribed as an adjuvant treatment to prevent cancer recurrence post-surgery. There are several classes of adjuvant endocrine therapy drugs, and they prevent estrogen intake by different mechanisms.

Selective estrogen-receptive modulators (SERMs) bind to estrogen receptors on cells and make them unreceptive to extracellular estrogen. SERMs do not lower the overall levels of estrogen in the body, therefore are prescribed to pre- and peri- menopausal women. Potential side effects of SERMs include but are not limited to hot flashes, and increased risk of blood clots, cataracts, and uterine cancer. In the ATAC adjuvant trial, approximately 40% of patients on SERMs reported hot flashes [15].

Aromatase inhibitors (AIs) prevent aromatase enzymes from converting androgen into estrogen, which lowers the amount of estrogen in the body. AIs are effective when a patient's main source of estrogen is androgen conversion, a characteristic of post-menopausal women. Pre- and peri-menopausal woman must undergo natural or artificial menopause in order for AIs to be effective.

There are two types of AIs: steroidal and non-steroidal. Steroidal AIs form non-reversible bonds with aromatase enzymes, while non-steroidal AIs form reversible bonds and actively compete with androgen at binding sites. There is a lack of clinical trials focused on efficacy differences between steroidal and non-steroidal AIs, but because steroidal AIs are purported to have androgenic effects, non-steroidal AIs are the recommended first line AI treatment [33]. A common side effect of AIs is arthritis. A study by Henry et al reported that out of a 97 breast cancer patients taking AIs, 44 experienced musculoskeletal side effects, and 13 cases were severe enough for the patient to discontinue AI use [34]. Hot flashes, another common side effect of AIs, appeared in approximately 35% of patients

on AIs in the ATAC adjuvant trial [15].

In 1977, the FDA approved the first endocrine therapy, tamoxifen (a SERM), for use in metastatic hormone receptor positive breast cancer. In 1990, tamoxifen was approved for use in the treatment of early stage hormone receptor positive breast cancer [35]. It was subsequently approved for use in prevention of breast cancer in 1998 and for treatment of non-invasive ductal carcinoma in-situ (DCIS) in 2000. Clinical studies on tamoxifen show that it consistently reduces risk for death and tumor recurrence in HR+ breast cancer patients [36].

In 2002, the first aromatase inhibitor, anastrozole, was approved and demonstrated superiority over tamoxifen in the adjuvant treatment of post-menopausal woman with breast cancer [15]. Subsequently, in 2005, letrozole and exemestane, two alternative aromatase inhibitors, were also approved as adjuvant endocrine therapies in this setting.

In 1998, clinical guidelines recommended adjuvant endocrine therapy for a five year duration [16]. Recent guidelines extended treatment duration up to ten years in at-risk populations [14]. With extended timeframes for duration of treatment and the growing number of available drug options, adjuvant endocrine therapy treatment paths can vary across patients. Clinical trials report varying rates of drug termination (31-73%) [15][37], rates of adverse events [34][38], and drug switches [39], but there is limited information on adjuvant endocrine treatment in the general patient population.

Electronic health records (EHRs) contain patient medical information including medication events and diagnoses. As of 2014, over 75% of hospitals in the United States use at least a basic EHR [57], and that percentage is growing. EHR use benefits patients and providers by facilitating clinical workflow and improving healthcare quality. Additionally, EHRs benefit research in that the data stored in the EHR allows for empirical analyses of practice patterns and clinical outcomes [58]. EHRs for breast cancer patients receiving adjuvant endocrine therapy provide data to characterize longitudinal treatment patterns and changes.

Vanderbilt University's Synthetic Derivative is a data source that contains deidentified health records from over 2 million patients [52]. MedEx, a natural language processing method, identified over 400 million medication events from medication lists and clinical notes in the Synthetic Derivative [4]. From these resources, patient diagnoses, medication events, and International Classification of Disease codes (ICDs), along with their respective timestamps, form a basis for analyzing adjuvant endocrine therapy treatment in breast cancer patients.

## 4.4 Methods

### 4.4.1 Data Collection

Our data source is de-identified electronic health records collected at VUMC as part of the Synthetic Derivative. The patient cohort met the following criteria: 1) Patients were diagnosed with stage I-III breast cancer, determined from the Vanderbilt tumor registry, 2) Patients received one or more of the following adjuvant endocrine therapy drugs: Anastrozole/Arimidex, Exemestane/Aromasin, Letrozole/Femara, Tamoxifen/Nolvadex (described in Table 4.1, and 3) Patient's adjuvant endocrine therapy began between 1998 and 2011. These date restrictions enforce that patients were 1) recommended five years of treatment and 2) can have at least five years of follow-up data. This study was done with the approval of Vanderbilt's IRB 140691, type exempt.

Data used in the study are adjuvant endocrine therapy medication events, patient ICD9 codes, and their respective timestamps. Death and recurrence data were collected from Vanderbilt's Tumor Registry. We used minimum medication event dates for 'start' times (i.e. the earliest medication start time). 'Stop' times are estimated at 6 months post the maximum medication event date. These estimates are based on expected patient follow-up every six months in the first five years after diagnosis. Patients do not take more than one adjuvant endocrine therapy drug simultaneously, so we interpreted the presence of a new

Table 4.1: List of the endocrine therapy drugs used by patients in the study.

| Generic Name | Other Names | Drug Class | Year of FDA approval for metastatic therapy | Year of FDA approval for adjuvant therapy |
|---|---|---|---|---|
| Anastrozole | Arimidex | Non-steroidal AI | 2000 | 2002 |
| Exemestane | Aromasin | Steroidal AI | 2005 | 2005 |
| Letrozole | Femara | Non-steroidal AI | 2005 | 2005 |

endocrine therapy drug in a record as a switch to the new drug.

### 4.4.2 Population Overview

To review our patient population, we began by defining all-inclusive states of adjuvant endocrine therapy treatment at VUMC including unknowns due to limitations on the data. We calculated statistics for patient inclusion in each group, and on patients grouped by five-year treatment completion status and switching status. We calculated outcomes of completion, death, and recurrence to determine differences among the patient groups.

### 4.4.3 Treatment Trends

Changes in the adjuvant endocrine therapy drug market, and new adjuvant endocrine therapy knowledge, leads to changes in treatment switch and stop frequencies over time. To determine drug prescription trends, we extracted prescriptions for the selected adjuvant endocrine therapy drugs and calculated prescription frequencies over time. We graphed the results along with the percent of patients who switched or stopped drugs by year. We hypothesized that as drug options rise, drug switch frequency increases and drug stop frequency decreases.

### 4.4.4 Identifying a Cause for Treatment Changes

We hypothesize that the two adjuvant endocrine drug classes, SERMs and AIs, incite different stop and switch patterns due to properties of the drugs. We hypothesize that switches from a SERM to an AI are most likely concomitant with changes in pre- or peri-menopause to post-menopause status, since AIs are only effective in post-menopausal women. Alternatively, we hypothesize that a switch from an AI is most likely concomitant with drug toxicity, since patients beginning on AIs are post-menopausal and their menopausal status will not change.

We explored causes for drug switches or discontinuation with computational chart reviews. We performed a text search for the stem words 'stop,' 'switch,' 'complain,' and 'discontinue' in Health Plan clinical notes written within a month (before or after) of a patient's treatment change. Next, we manually reviewed a random sample of notes containing a search word for direct references to changes in adjuvant endocrine therapy.

To explore the likelihood that switches from SERMs are correlated with a change in menopause status, we calculated the age distribution and p-values of patients on SERMs and patients switching from SERMs. For comparison, we calculated the age distribution and p-values for patients on AIs, and patients switching from AIs. We hypothesized that the age of patients switching from a SERM is greater than the age of patients taking SERMs, indicating that their age and menopause progression is correlated with switching. We also checked for evidence of artificial postmenopause progression through 1) CPT codes for oophorectomies prior to a switch from a SERM and 2) medication events for estrogen-suppressing drugs Goserelin and Leuprorelin prior to a switch from a SERM. Patients undergoing oophrectomies or taking estrogen-suppressing drugs become postmenopausal, and will likely switch medications.

To explore the likelihood that switches or discontinuation from AIs are correlated with toxicity, we probed for a correlation between switching/discontinuation and ICD codes for hot flashes and arthritis pain. The targeted ICD9 codes are 714.*, 715.*, 716.*, 729.*,

627.2, and 782.62. We compared the rate of ICD codes in patients that switch/stop during treatment to the rate of ICD codes in patients that complete treatment with no change. ICD codes must have occurred within a year before treatment change or completion to ensure 1) we compare equal timeframe across groups and 2) ICD codes are relevant to the current treatment plan. We hypothesized that ICD prevalence for hot flashes and arthritis would be higher in patients who switch/stop treatment, indicating adverse events as a reason for change. Furthermore, we hypothesized that patients who stop/switch from AIs have higher ICD rates than patients who stop/switch from SERMs since stop/switch from a SERM may result from a different cause - menopause status. To determine significant differences between ICD prevalence, we calculated Fisher's Exact test P-values.

SERMs and AIs have different mechanisms of action, resulting in disparate rates of toxicity, and disparate efficacy among age groups. Consequently, age-related and toxicity-related reasons for change exhibit different distributions across the two drug classes. To explore if changes in SERM and AI treatment occur at different times, we examined the time until treatment change in both drug classes. Switch/stop from an AI due to adverse events may be localized to a certain time of toxicity onset. Conversely, changes in treatment due to menopause progression are not localized to a certain time since the change is dependent on a patient's age and estrogen levels. We tested these hypotheses by graphing time until drug switch or discontinuation for patients on SERMs and AIs.

## 4.5 Results

### 4.5.1 Population Overview

The patient states for adjuvant endocrine therapy treatment at VUMC are illustrated in Figure 4.1 and define knowns and unknowns due to nature of the data. Our study includes 1,587 stage I-III breast cancer patients taking adjuvant endocrine therapy drugs. The average age of patients at treatment start is 56.9 (12.3 standard deviation). The average number

of unique endocrine therapy drugs per patient is 1.5 (0.7 standard deviation). Approximately 64% of patients continued care at Vanderbilt post five years of adjuvant endocrine therapy treatment.

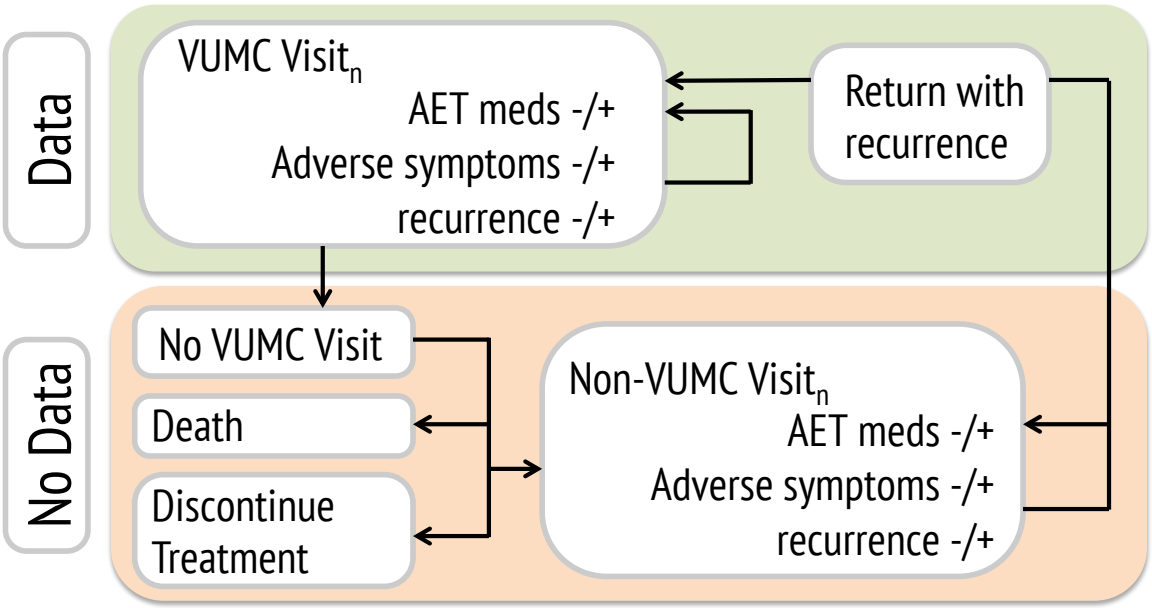



Figure 4.1: Adjuvant endocrine therapy treatment states at Vanderbilt University Medical Center. Highlights differentiate between the data that is included in the VUMC EHR, and data that is excluded from the VUMC EHR but exists in health plan information and claims data.

Table 4.3 illustrates our patient population broken down by 5-year completion status and switch status and Table **??** includes p-values for inter-group comparisons of recurrence and death rates. Data comparisons of patient grouped by switch-status and completion-status are illustrated in Figure 4.2. We show that approximately 48% of our patient population does not complete at least 5 years of endocrine therapy for a reason other than death, and therefore may not achieve the lowest possible risk of recurrence. The probability of a patient completing five years of therapy given that they switched drugs is 60%, while the probability for a patient to complete five years given that they did not make a drug switch is 37%. Patients in our population who completed at least five years of treatment recurred at a rate of 3.4% and patients who did not complete five years of treatment recurred at a

Table 4.3: Adjuvant endocrine therapy statistics grouped by completing five years of treatment, death, switching drugs, report of recurrence, and maintained visits to a Vanderbilt facility.

| Patient Category | | Number of Patients | Percentage of Patients | Recurrences before five years | Percentage of recurrences | Deaths before five years | Percentage of deaths | Number of patients with Vanderbilt visits post five years | Percentage of patients with Vanderbilt visits post five years |
|---|---|---|---|---|---|---|---|---|---|
| Switch before 5 years | Stop before 5 years | | | | | | | | |
| Yes | Yes | 334 | 21.0% | 35 | 10.4% | 10 | 2.9% | 163 | 48.8% |
| Yes | No | 500 | 31.5% | 19 | 3.8% | 0 | 0.0% | 383 | 76.6% |
| No | Yes | 469 | 29.5% | 30 | 6.4% | 26 | 5.5% | 229 | 48.8% |
| No | No | 284 | 18.0% | 8 | 2.8% | 0 | 0.0% | 232 | 81.7% |
| Total | | 1587 | 100.00% | 92 | 5.8% | 36 | 2.3% | 1007 | 63.5% |

rate of 8.0%.

subsectionTreatment Trends

Figure 4.3 shows the endocrine therapy prescription frequencies at VUMC with the percentage of patients stopping or switching their drugs each year. The graphs reflect a rise in aromatase inhibitors after 2004. The percent of patients stopping treatment decreases over time, and the percent of patients switching drugs increases.

subsectionExploring Causes for Treatment Change

Of the 1,303 patients who switched or stopped adjuvant endocrine therapy prior to five years, 383 (29%) had an instance of the selected stem words in their clinical notes near the time of change. 120 (9%) possessed 'stop,' 73 (6%) possessed 'switch,' 298 (23%) possessed 'complain,' and 59 (5%) possessed 'discontinue.' In a random selection of 100 patients with stem-word-positive notes, 16% possessed a documented cause for changes to adjuvant endocrine therapy, 20% possessed documented stopping or switching without an explicit cause, and in the remaining 64%, stem words did not reference adjuvant endocrine therapy.

Table 4.4 shows the average age of switch from a SERM is higher than the average age of patients on SERMs. In contrast, the average age of patients switching from AIs is less than the average age of patients on AIs. The distributions support that patients switching from SERMs to AIs may be attributed to change in postmenopausal status.

Out of the patients who switched from a SERM to an AI, 16.5% had a reported oophorectomy through either a CPT code 58940 or 58720, or an ICD9 parent code of 65 before their switch date. Additionally, 11.5% of patients switching from a SERM to and AI received either Goserelin or Leuprorelin (estrogen-suppressing drugs that place a patient in a post-menopausal state). All together, 28% of the patients who switched from a SERM to AI underwent an artificial change to post-menopausal status.

Arthritis ICD rates in patients who stopped or switched drugs were 42% and 72% greater than arthritis ICD rates for patients who completed treatment without a switch for SERMs and AIs, respectively (14.2% and 30.6% compared to 10.0% and 22.0%). Hot

flash ICD rates in patients who stopped or switched drugs were 46% and 39% greater than hot flash ICD rates for patients who completed treatment without a switch for SERMs and AIs, respectively (24.8% and 25.0% compared to 17.0% and 18.0%). Furthermore, patients who stop or switch from an AI have approximately double the rate of arthritis ICD codes of patients who stop or switch from a SERM (Table 4.5).

Figure 4.3a shows the time distributions for stop and switch from AIs and SERMs. All drug change distributions peak within the first year of treatment. Switching and stopping from and AI occurs within a localized time. Switching or stopping from a SERM has a broad distribution.

Table 4.4: Age distributions for patient at the time of AI use, SERM use, a switch from an AI, and a switch from a SERM. P-values measure differences between the distributions. Significant differences are denoted with an asterisk.

| Drug Event | Average | Standard Deviation | Comparison Group | Fisher's Exact Test |
|---|---|---|---|---|
| | | | | P-value |
| AI Use | 64.1 | 10.1 | AI Use vs SERM Use | <0.001* |
| SERM Use | 55.4 | 12.8 | AI Use vs AI Switch | <0.001* |
| Switch from AI | 57.4 | 12.2 | SERM Use vs SERM Switch | <0.001* |
| Switch from SERM | 60 | 10.7 | AI Switch vs SERM Switch | <0.001* |

## 4.6 Discussion

With electronic health record data from a cohort of 1,587 stage I-III breast cancer patients receiving adjuvant endocrine therapy, we found that approximately 48% of patients did not complete the recommended minimum of five years treatment, and 52% of patients

Table 4.5: Prevalence of billing codes for adverse events in patients, and Fisher's exact test p-values to determine significant differences between prevalence. Patients are grouped by drug class and change in original treatment plan. Significant differences are denoted with an asterisk.

| Patient Group | Arthritis Positive (Percent) | Hot Flash Positive (Percent) | Comparison Group | Fisher's Exact Test | |
|---|---|---|---|---|---|
| | | | | P-value Arthritis | P-value Hot Flashes |
| Completed SERM without switch | 10 | 17 | Complete AI vs SS from AI | 0.040* | .062 |
| Stop/Switch (SS) from SERM | 14.2 | 24.8 | Complete SERM vs SS from SERM | 0.694 | 0.037* |
| Completed AI without switch | 22 | 18 | SS from AI vs SS from SERM | $<$0.001* | 0.947 |
| Stop/Switch (SS) from AI | 30.6 | 25 | Complete AI vs Complete SERM | 0.005* | 0.874 |

switched to a different endocrine therapy drug during their treatment. Using ICD codes, we found that patients who changed their adjuvant endocrine treatment experienced higher rates of arthritis and hot flashes than other patients. Changes in treatment in patients on SERMs follow menopause progression inferred through age, administration of estrogen-suppressing drugs or surgeries. Additionally, switching treatment from an AI is likely to occur at the beginning of treatment, while switching from a SERM, is not localized to a treatment time.

Patients who switched drugs at some point during their treatment are more likely to

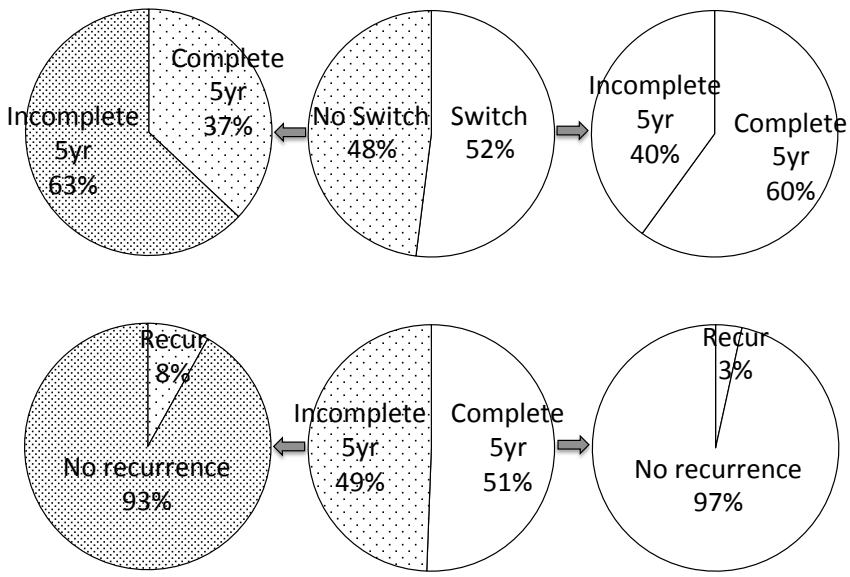
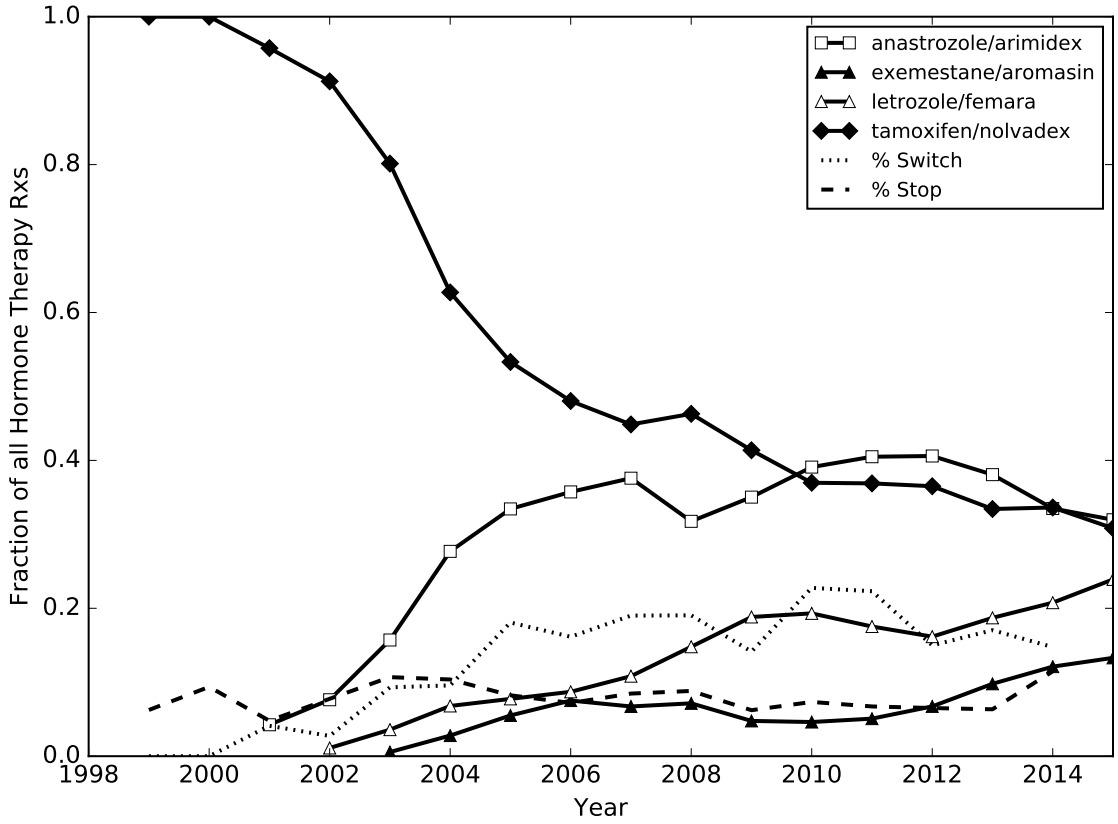



Figure 4.2: Rates of completion from patients grouped by presence of a drug switch, and rates of recurrence from patients grouped by five-year treatment completion. The center circle represents the full patient cohort, and the marginal circles.

complete at least five years of adjuvant endocrine therapy (Figure 4.2). Switching drugs may be an expedient that encourages patients to continue treatment, and the additional treatment time due to the switch may benefit the patients. We support that patients who complete at least five years of adjuvant endocrine therapy gain a lower rate of recurrence (3.4% compared to 8.0%) (Figure 4.2). The recurrence rates fall within the confidence interval of population recurrence rates reported for stage 1 breast cancer (95% CI = 3% to 15%) [59]



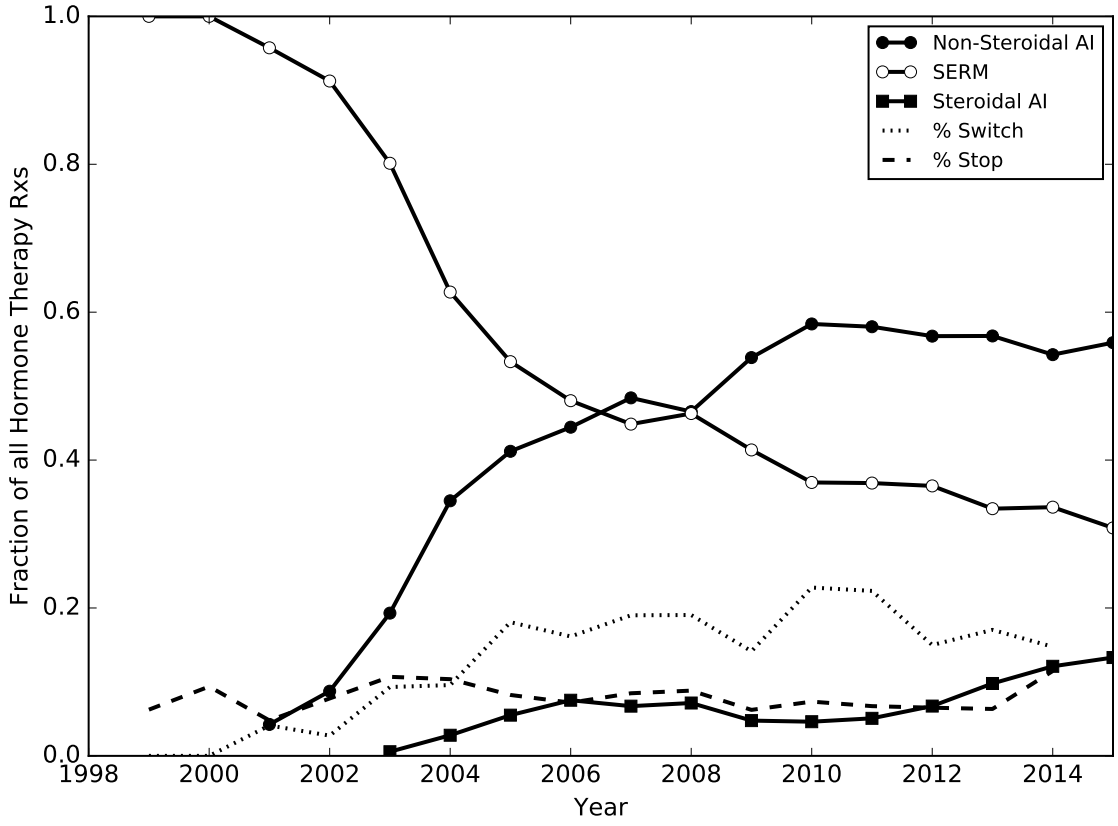

(a) Frequencies by individual adjuvant endocrine therapy drug.

### 4.6.1 Treatment Trends

Endocrine therapy prescription changes at VUMC reflect a rise in options for patients, and as a result, the percent of patients switching adjuvant endocrine drugs exceeds the percent of patients stopping adjuvant endocrine therapy in 2004 (shown in Figure 4.3). The option to switch between endocrine therapy drugs may encourage adherence to the recommended treatment duration.

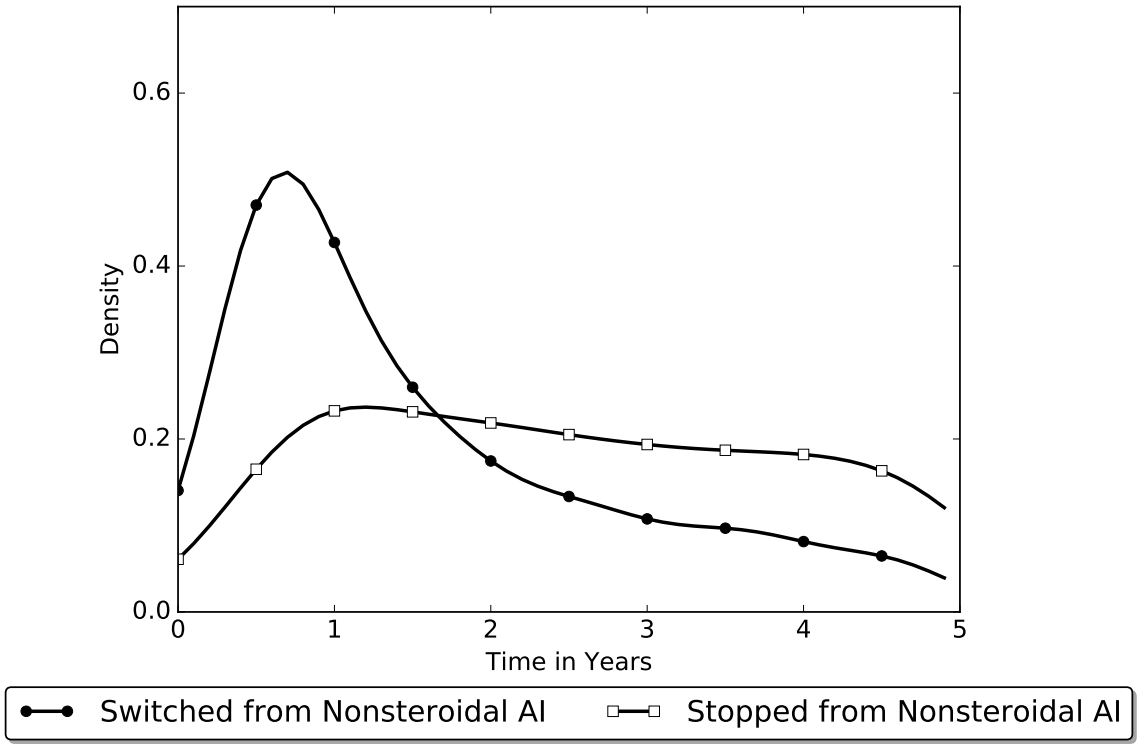
### 4.6.2 Identifying a Cause for Treatment Changes

From our computational and manual chart review, we estimate that 5% of patients have a clearly documented cause for adjuvant endocrine therapy treatment change. When documentation on the cause of treatment change is sparse, inferring potential reasons for change




(b) Frequencies by adjuvant endocrine therapy drug class.

Figure 4.3: Changes in adjuvant endocrine therapy prescription frequencies and percentage of patients stopping or switching at Vanderbilt University Medical Center over time. AI prescriptions frequencies increase and SERM prescriptions frequencies decrease.

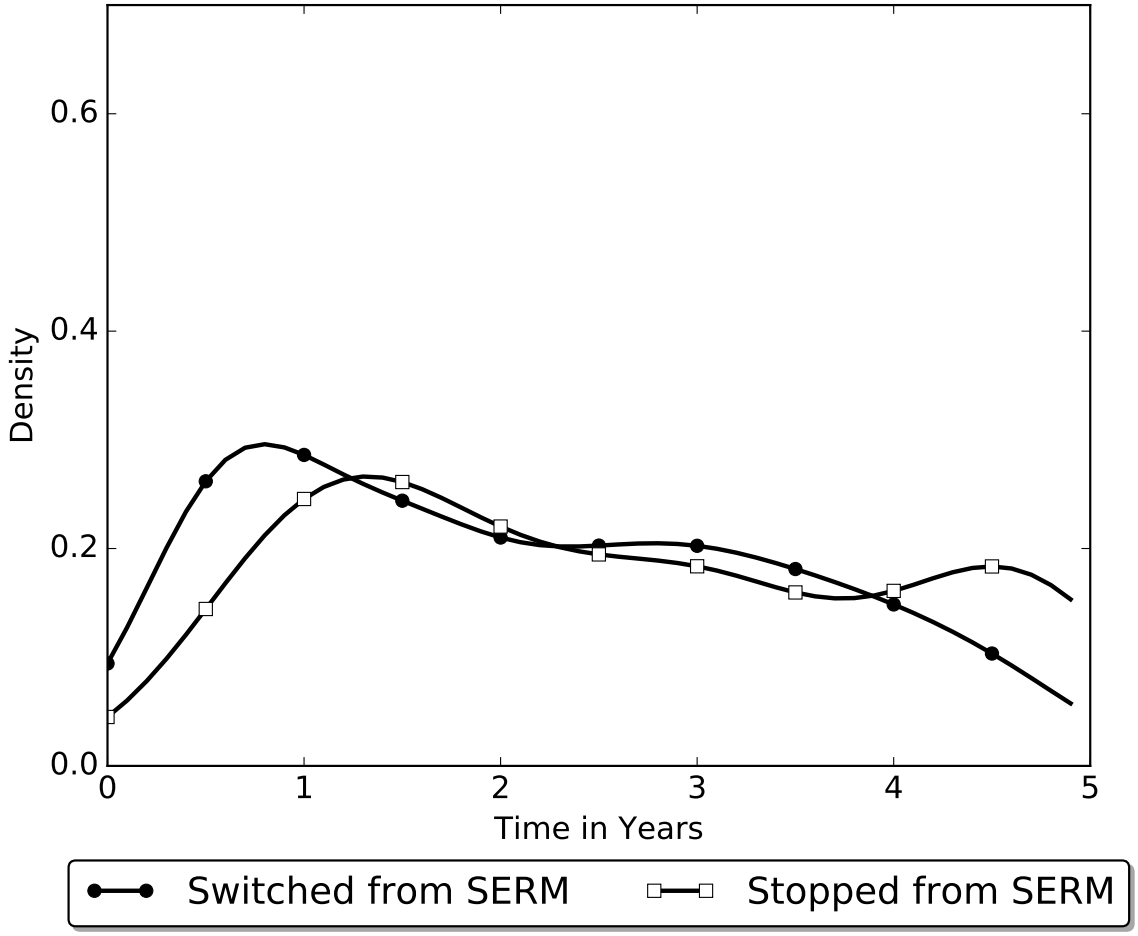



(a) Density plot of time patient's spent on AIs prior to stopping or switching.

with EHR data-applied informatics methods is an alternative way to retrieve such information, but presents many challenges.

The age distributions for patients at the time of a drug switch (Table 4.5) support that switching from a SERM is related to age. The average age during a switch from a SERM is higher than the average age of patients treated with SERMs. In contrast, patient age at a switch from AIs is less than patient age during AI use. A switch from a SERM is more likely related to age than a switch from an AI. Additionally, 4.4 shows that switching from a SERM is not localized to a specific duration of treatment, while switching from an AI is. Switches from an AI are more likely due to adverse events that manifest within a year of treatment. For example, Henry et al measured that musculoskeletal symptoms from AI use peaked within 6 months 9. Switches from a SERM are likely dependent on a patient's individual menopause progression and appear at any time during treatment.

Patients who stop or switch their treatment experience higher rates of arthritis and hot




(b) Density plot of time patient's spent on SERMs prior to stopping or switching.

Figure 4.4: Density plots showing temporal trends in adjuvant endocrine therapy drug use, switches, and discontinuations. AI switches are localized to the first year while SERM switches and not localized to a specific time point

flashes than patients who complete their treatment without changes. The highest rates of arthritis occurred in patients who stopped or switched AI treatment. Patients who stopped or switched either AI treatment or SERM treatment experience similar rates of hot flashes but more than those who did not switch. Pervious studies report musculoskeletal side effects at a rate of 44% [34], and 44-47% [60] and 23% [61], and hot flashes from AI use occur at a rate of 30% [61] and 35% [15]. Our lower rates are due to the source of our data - the ICD codes. ICD codes appear in the patient record if the provider considers the symptom severe enough to bill for it or document it. A symptom may not be documented, but that does not guarantee the symptom's absence.

### 4.6.3 Generalizability to Other Healthcare Domains

The methodology used in this study comprised of building a state diagram, statistical summarizations, t-tests, density measures, and visualization. Provided a similarly sufficient dataset to the one extracted in this study, these methods could be useful in characterizing the state of patient care in other healthcare domains beyond adjuvant endocrine therapy. The methodology at hand applies to long-term patient care, due to measures over time. Characterizing the state of patient care in a short-term setting, such as outpatient emergency room visits, is limited by this methodology because of the lack of longitudinal data on which to measure patient states. Characterizing the state of adjuvant endocrine therapy at VUMC supports that our methods have potential use in alternative long-term healthcare domains.

## 4.7 Limitations

This study is first limited by the completeness of EHR data. We show in Figure 4.1 the extent to which VUMC EHR data captures data on adjuvant endocrine therapy patients. Data outside the realm of VUMC is missing, limiting the extent of characterizing complete adjuvant endocrine therapy care. This study could be enriched with additional datasets,

mainly health plan data, that follows patients across all facilities in which they receive care. These datasets would clarify the unknowns in adjuvant endocrine therapy treatment outside of the VUMC system.

This study is also limited by the medication event extraction tool used to extract medications and timestamps collected in the EHR. We depend on timestamps for adjuvant endocrine therapy 'start,', 'stop,' and 'switch' dates. Timestamps may be incorrect when the medication event extraction tool identifies medications in the patient note that the patient is not currently using. Also, clinical notes may be erroneously copied forward leading to over-projected dates for treatment stop. Further limitations in the data include under-documented recurrence, death, and adverse symptoms. Recurrence and death may be under-reported due to lack of follow-up with patients, and ICD codes for adverse symptoms may be under-documented since they are captured by billing codes and providers do not always bill for them.

Last, although we can make inferences on a reason to stop or switch adjuvant endocrine therapy treatment, cause of treatment change is infrequently documented in physician notes, making validation difficult. In addition to change in menopause status, adverse events, and recurrence, changes in treatment can be due to financial factors [62] or physician preference, neither of which are documented in the EHR.

## 4.8 Conclusions

This study demonstrates the ability to leverage longitudinal electronic health records to characterize treatment trends and the state of patient care in a cohort of adjuvant endocrine therapy patients. We support that EHR data are a source for real-world frequencies of adjuvant endocrine therapy patient tumor recurrence, drug-switch, and drug-stop, as well as a source for exploratory analyses on causes for treatment change. We defined the states of patient care for adjuvant endocrine therapy at VUMC and estimated patient inclusion in states using data recorded in the EHR.

The main findings of the study include:

- 49% of VUMC adjuvant endocrine therapy patients were lost to follow-up or did not complete adjuvant treatment throughout five years.
- 52% percent of VUMC patients switched to a different endocrine therapy drug during their treatment.
- Age and adverse events are correlated with changes in adjuvant endocrine therapy.
- VUMC's EHR is a valuable resource for characterizing the state of adjuvant endocrine therapy for VUMC patients.

# Chapter 5

# Aim 3: Identify New Opportunities to Improve Patient Care

## 5.1 Overview

This chapter aims to identify new **opportunities** to improve patient care. In the previous chapters, we find that adjuvant endocrine therapy patients often fail to follow-up at VUMC for treatment for the recommended five-year duration. Reasons for this include transfer of care to a non-VUMC provider, discontinuation of care, or unreported death in the VUMC EHR. Although it is not always clear why a patient fails to follow-up, there are predictors in the EHR data that assists in identifying follow-up status. Supervised machine learning is a computational solution to identify predictors from large, labeled datasets. The following text is an extended version of an intended academic article titled *Supervised Machine Learning to Predict Follow-Up Among Adjuvant Endocrine Therapy Patients*, which aims to identify predictors for follow-up in EHR data using machine learning methods. We build machine learning classifiers to predict follow-up with appointments and medication events, and construct predictors for follow-up using EHR derived, appointment, and demographic features.

This chapter demonstrates the potential of machine learning techniques to identify new opportunities for patient care improvement, specifically improving patient follow-up, using electronic health record data. We find that random forests are useful models for predicting follow-up, and identify features that differentiate patients that follow-up. The main conclusions from this chapter are 1) VUMC adjuvant endocrine therapy follow-up can be predicted with and AUC of 0.74 using supervised machine learning methods, 2) EHR data from adjuvant endocrine therapy patients holds predictors for follow-up, and 3) supervised machine learning is a useful method for learning new opportunities for improvement in

patient care through EHR data

## 5.2 Introduction

Adjuvant endocrine therapy is prescribed to hormone receptor positive breast cancer patients to prevent tumor recurrence [16] [14]. Clinical guidelines recommend patients use adjuvant endocrine therapy drugs for at least five years to minimize recurrence risk. However, long-term care can be onerous for patients. Adjuvant endocrine therapy patients must follow-up with their physician every six months, and may suffer side effects from their adjuvant endocrine drugs [54] [15]. As a result, patients may fail to follow-up with their physician for the recommended time, which results in suboptimal care and higher risk for tumor recurrences [**?** ].

There are several reasons a patient may fail to follow-up with their care provider: poor communication between patient and provider, burdensome distance between the patient and care facility, negative side effects that discourage continued treatment, etc. Identifying controllable reasons for failure to follow-up can promote interventions that improve follow-up rates. However, reasons for failing to follow-up are not reported by patients, making it challenging to find controllable factors that can improve follow-up rates.

One predictor may be distance to the VUMC facility. Breast cancer patients are often referred to VUMC for treatment and subsequently travel for care. VUMC is centrally located and around 200 miles from the next nearest cancer center (Figure 5.1). Once a patient begins adjuvant endocrine therapy, they may chose to transfer care to a primary care provider within closer proximity to their home.

Although reasons for failure to follow-up are not always reported back to the care provider, information in Electronic Health Records (EHRs) may act as predictors for failure to follow-up. This information includes demographics, appointment patterns, and other treatment information. One method that can predict follow-up in EHR data is supervised machine learning. Supervised machine learning builds a function from labeled data that can


Figure 5.1: Map of Vanderbilt University Medical Center location and locations of nearest cancer centers.

be generalized to additional data points. There are several methods of supervised machine learning including random forest and neural networks.

Random forests are ensemble learning methods that classify data and identify predictors for classes. Random forests achieve accurate prediction for tasks involving electronic health record data including predicting risk in hypertension control [63] and predicting adverse drug events [64]. Features of random forests include an ability to capture non-linear relationships, interpretable models, and included feature selection, where features are selected based on their ability to split data [40].

Neural networks are machine learning methods that combine logistic regressions to classify data. Neural networks achieved accurate prediction from electronic health records data on tasks including heart failure detection [65]. Features of neural networks include the ability to capture complex, non-linear relationships. However, the complexity of neural networks makes models difficult to describe.

Supervised machine learning classifiers built with adjuvant endocrine therapy patient EHR data to predict follow-up uncovers predictors that may facilitate interventions and improve follow-up rates. Improving follow-up rates in the adjuvant endocrine therapy patient population will reduce recurrence rates and ultimately improve patient care in the cohort. This study demonstrates the ability to use EHR data to find opportunities for improvement in patient care and guide clinical decision-making.

## 5.3 Background

Adjuvant endocrine therapy is prescribed to reduce risk of tumor recurrence in hormone receptor positive breast cancer patients. In 1998, clinical guidelines recommended adjuvant endocrine therapy for a five year duration 1, and recent guidelines extended treatment for up to ten years in at-risk populations [14]. These guidelines require patients to follow-up with their care providers long-term. Unfortunately, follow-up times often fall short of the recommended duration [**?** ]. There are many reasons for failure to follow-up, and a reason is not always reported to the care provider when a patient fails to follow-up. However, predictors for failing to follow-up may be found in the medical documentation leading up to the follow-up failure.

Electronic health records (EHRs) contain patient medical information including appointment times, clinical communications, medication events and diagnoses. Vanderbilt University Medical Center (VUMC) holds EHRs on over 3 million patients, and has a tumor registry with records on over 90,000 patient tumors linked to EHR records. VUMC holds health records on 2900 stage I-III breast cancer patients with at least one adjuvant endocrine therapy medication event. Furthermore, VUMCs appointment audit logs contain appointment scheduling patterns, including appointment frequency and cancellation frequency, on over 2 million patients, approximately 20,000 of which are prescribed adjuvant endocrine therapy drugs. These resources form a basis for predicting follow-up among adjuvant endocrine therapy breast cancer patients.

Random forests are ensemble learning methods that can accomplish prediction tasks, including those based on EHR data [64] [63] [66]. Benefits of random forests include high accuracy [40], resistance to over-fitting [67], ability to capture non-linear relationships, and estimates of feature importance in the classification. Limitations of random forests include over-fitting when underlying decision trees are too large, and misleading feature importance estimates if features are correlated [40]. Random forests can find predictors for follow-up within electronic health record data from adjuvant endocrine therapy patients. Learning predictors for follow-up can lead to interventions that improve follow-up rates, lower recurrence rates, and ultimately improve patient care.

Neural networks are combinations of activation functions that yield a classification prediction [68]. Neural networks have accomplished complex prediction tasks from handwriting and facial recognition [69] [70] to clinical predictions for mortality risk [71] and decision support [72]. Benefits of neural networks include the ability to capture signals from large noisy datasets. Neural networks are limited by the large amount of data points necessary to train the models, and by the challenges for interpretation due to dependence on complex combinations of features.

## 5.4 Methods

Our methods begin with a basic measure for follow-up prediction, and continues to complex measures for follow-up with supervised machine learning. We measure the average distance between patient home address and VUMC for patients that continue follow-up at VUMC, measured through consistent appointments with any VUMC provider, over five years from adjuvant endocrine therapy start. Then, we plot the odds ratio for follow-up against failure to follow-up against distance. Our supervised machine learning methods include data collection, feature matrix construction, classifier construction, evaluation and optimization. A workflow for our methods is illustrated is Figure 5.2.

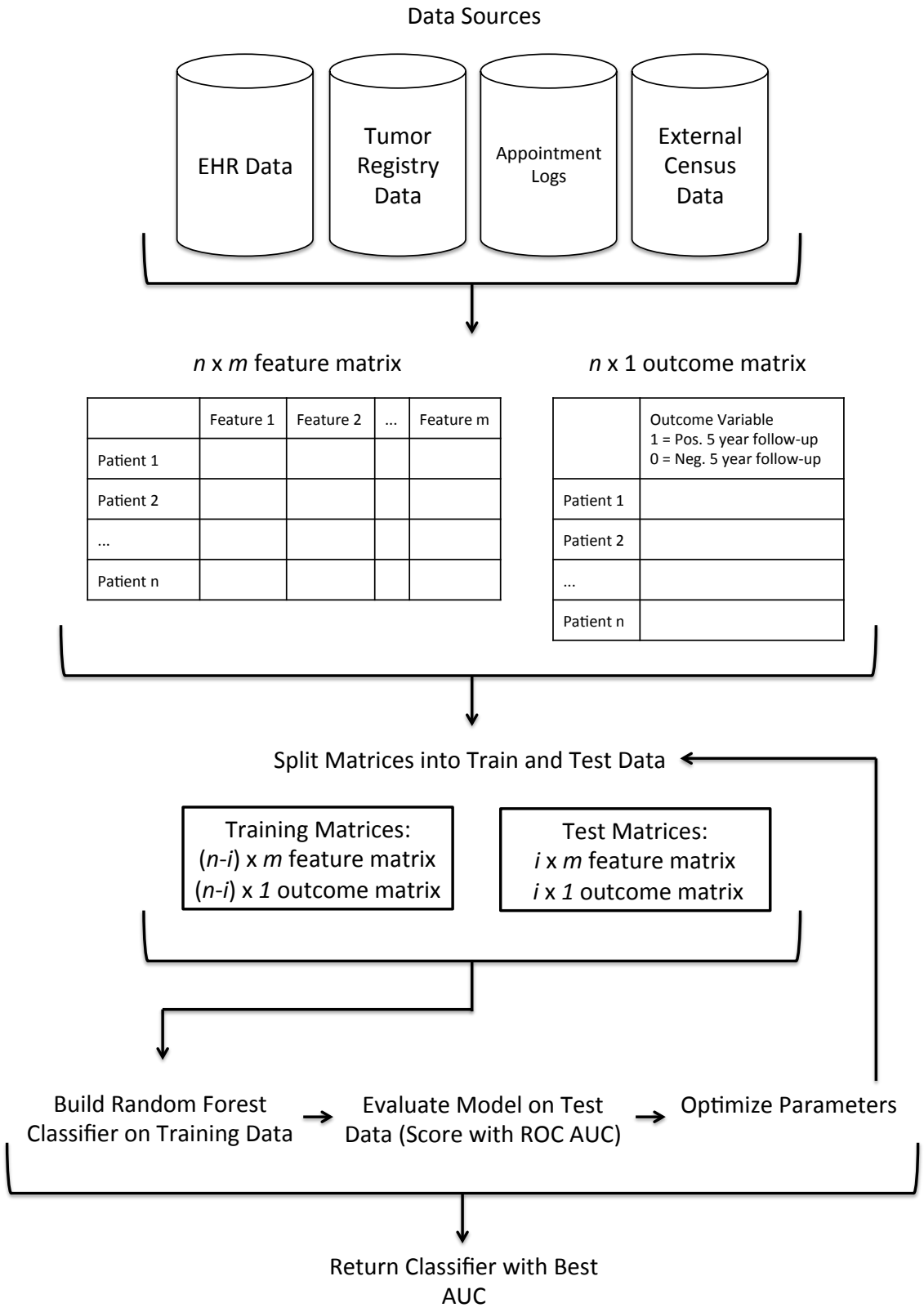



Figure 5.2: A workflow for our supervised machine learning classification methods. We begin with data collection from electronic health records, tumor registry entries, and appointment logs. We build treatment, appointment, and demographic features from the data and model them as a matrix with the outcome variable '0' for patients who fail to follow-up for five years and '1' for patients who follow-up with their care providers for five years. We train and test a random forest classifier, evaluate performance and optimize parameters to achieve the highest prediction AUC.

### 5.4.1 Data Collection

Our data source is electronic health records, tumor registry data, and appointment records collected at VUMC. Additionally, we obtained zip code census data from a public source, and medication classes from Unified Medical Language System's RxNorm. We selected a patient cohort based on the following restrictions: 1) stage I-III breast cancer diagnosis, determined from the Vanderbilt tumor registry, 2) medication event for at least one of the following adjuvant endocrine therapy drugs: Anastrozole/Arimidex, Exemestane/Aromasin, Letrozole/Femara, Tamoxifen/Nolvadex, and 3) began adjuvant endocrine therapy between 1998 and 2011. These date restrictions ensure that patients were recommended five years of treatment and could have at least five years of follow-up data. This study was done with the approval of Vanderbilt's IRB 160839, type exempt.

### 5.4.2 Feature Matrix Construction

We extracted data from electronic health records, tumor registry entries, and appointment logs for each patient in the cohort to build features for a classification matrix. This data includes all interactions with the VUMC system with any provider and any department. Features are listed in Table 5.1. N-categorical features were split into n-1 binary features. The matrix has both binary and normalized continuous feature types.

We built three different binary outcome variables to measure: 1) consistent, yearly appointments with any VUMC provider, 2) consistent, yearly appointments with a VUMC oncologist, and 3) consistent yearly VUMC medication events for adjuvant endocrine therapy drugs. Outcome variables two and three are subsets of outcome variable one. Outcome one captures patients who may transfer follow-up appointments with a primary care physician. Outcome two captures patients who strictly see oncologists. Outcome three captures patients who not only follow-up with their clinician, but complete their adjuvant endocrine treatment. Figure 5.3 illustrates the follow-up measures. We used years +/- a 3 month buffer

as appropriate intervals to confirm follow-up. Patients are recommended to follow-up every six months, therefore patients appropriately following-up have, at minimum, data once per year. Patients who had a recorded death in the tumor registry before five years from their initial adjuvant endocrine therapy medication event were censored from the matrix.

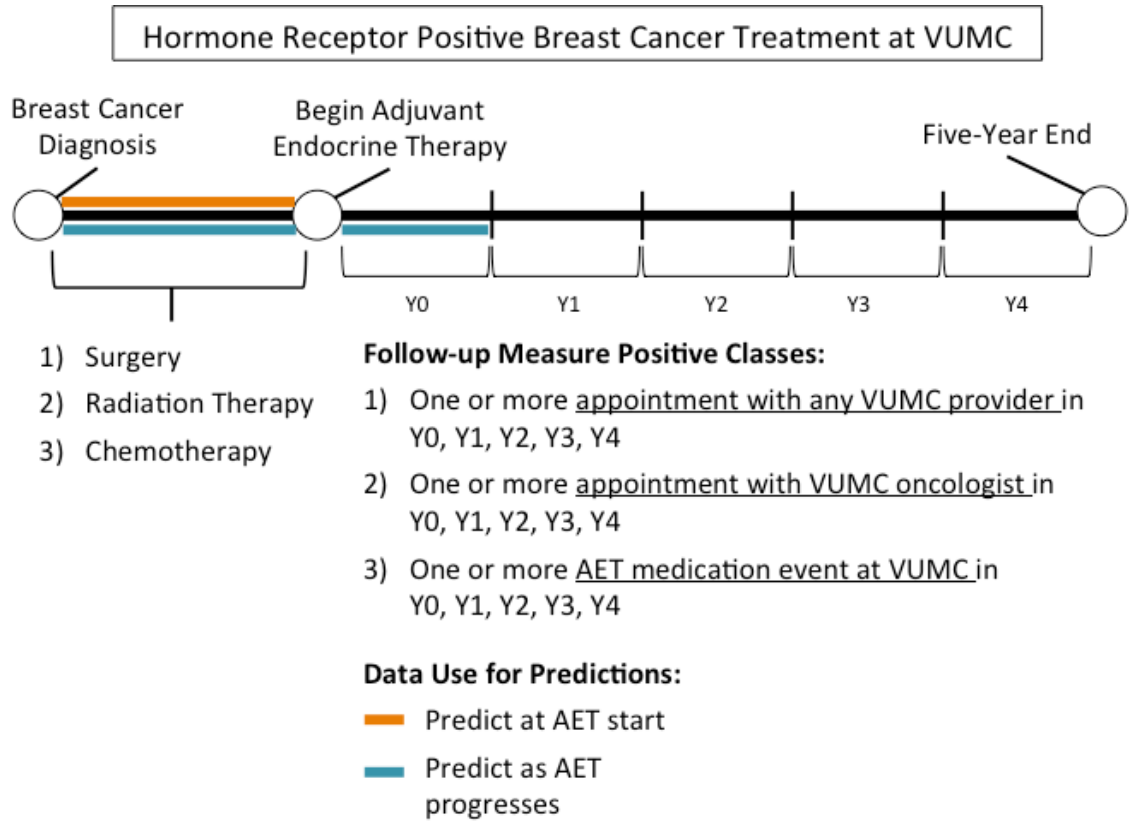



Figure 5.3: Projected patient treatment at VUMC with a guide for follow-up measures and data used for predictions.

### 5.4.3 Classifier Construction

We used the Python machine learning packages SciKitLearn [73] to build random forest classifiers, and Keras [74] with Theano [75] to build neural networks. We depended on the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University, Nashville, TN for building models with GPUs. We built classifiers from EHR data, appointment data, demographic data, and a combination of all three data types. Clas-

sifiers make predictions from patient data at both the start of adjuvant endocrine therapy and as treatment data accrues. For each classifier, we divided data into train and test data sets, trained a classifier on the training set, and used the sequestered test data to test the classifier's performance.

### 5.4.4 Classifier Evaluation and Optimization

We evaluated model performance by measuring the area under the Receiver Operator Characteristic curves (AUC). The area under the ROC curve measures the probability at which the classifier, given a positive and negative class, will assign a higher value to the positive class than the negative class. We used five-fold cross validation to train and test five classifiers per set of parameters, then averaged the AUCs together as a final performance measure.

To optimize our classifiers, we adjusted underlying parameters and reevaluated ROC AUCs. The parameters we adjusted for random forest classifiers include the maximum number of features (m) to search for the best split (m, log m, or square root of m), the max depth of the underlying decision trees (1 to m), the minimum number of samples required to split a node (1 to 10), and the split criteria for features (entropy/information gain or gini/impurity). The parameters we adjusted for neural networks include the number of hidden layers and number of nodes per layer. We built classifiers for each combination of parameters and returned the model with the best AUC.

### 5.4.5 Temporal Classifier Construction

After building base models on data collected between breast cancer diagnosis and adjuvant endocrine therapy start, we introduced data from the first year of treatment to our training data. We hypothesize that prediction power improves as treatment data accrues. We built classifiers with additional training data derived from quarter-year increments of treatment up to one year. We measured the change in AUC and changes in significant

features over time.

## 5.5 Results

### 5.5.1 Data Collection

1455 VUMC patients met our cohort criteria for adjuvant endocrine therapy follow-up prediction. 838, or 57% of patients followed-up for the recommended five-year duration with any provider in the VUMC system. 748, or 51% followed-up with an oncologist, and 576 or 39% followed-up on adjuvant endocrine therapy medications. The data we extracted for patients in our cohort include over 2 million medication events, 293,000 ICD codes, 193,000 appointment logs, and 3,257 clinical communications. Furthermore, patients had home addresses in 700 different zip codes, from which we collected distance and census information.

### 5.5.2 Distance Measures

Over time the average distance of patients continuing to follow-up at VUMC shrinks while the average distance of patients discontinuing follow-up at VUMC grows 5.4. The ratio of density curves for distance among patients that follow-up and fail to follow-up at VUMC over five years. This curve shows the ratio of follow-up to no follow-up for a given distance from VUMC 5.5. When a patient is located less than 50 miles away from VUMC, the odds that they will complete five years of follow-up is 3:2. For patients living at least 200 miles from VUMC, the odds that they will complete five-years of follow-up is less than 2:3.

### 5.5.3 Feature Matrix Construction

Overall, we constructed 321 features for 1455 adjuvant endocrine therapy patients at VUMC. 307 features were built from EHR data, 7 features were built on appointment

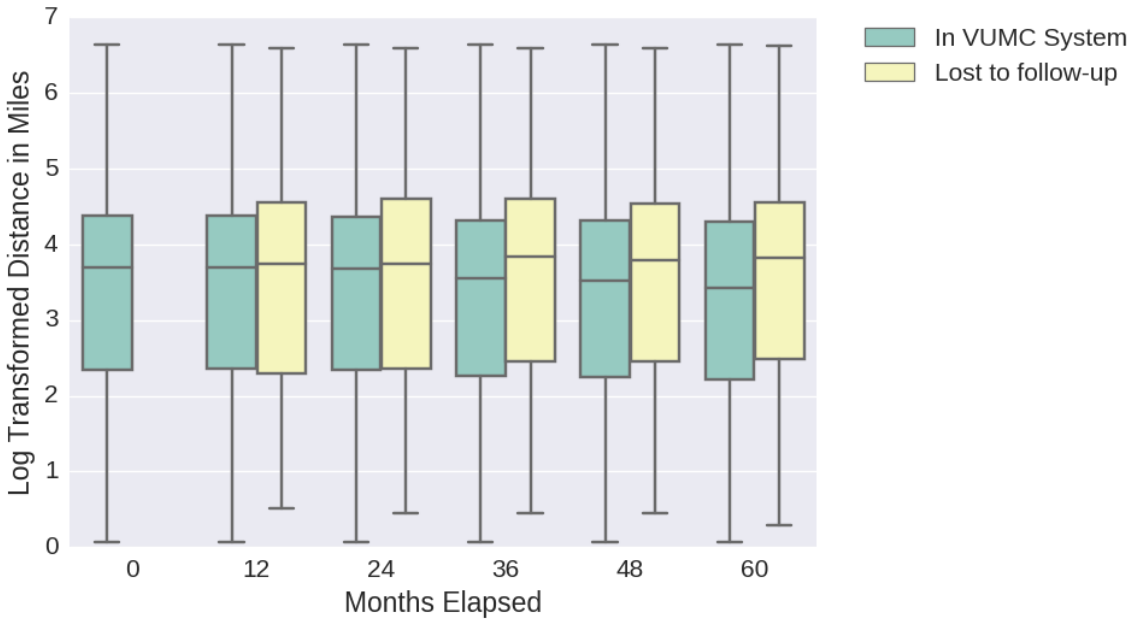



Figure 5.4: Box Plot of the log-transformed miles to VUMC center in adjuvant endocrine therapy patients as they follow-up, or do not follow-up with a VUMC provider across five years from adjuvant endocrine therapy start.

log data, and 7 features were built on demographic data. The majority of features were counts for treatment with specific providers and departments, and presence of CPT codes in electronic health record data. Consequently, our feature matrix was sparse.

### 5.5.4 Supervised Machine Learning

Our best classifier predicted follow-up with any VUMC provider and used combined EHR, appointment, and demographic features. This classifier yields an AUC of 0.74 (Figure 5.6). The top five significant features are 1) total medications count, 2) age at diagnosis, 3) median zip code income, 4) distance in miles, and 5) counts of ICD9 parent code 719 for unspecified joint disorders.

Classifiers built solely on demographic, appointment, and medical features had less prediction power than combined features, but medical features yielded the most prediction power. Significant features for each of the models are listed in Table 5.2. Furthermore, pre-

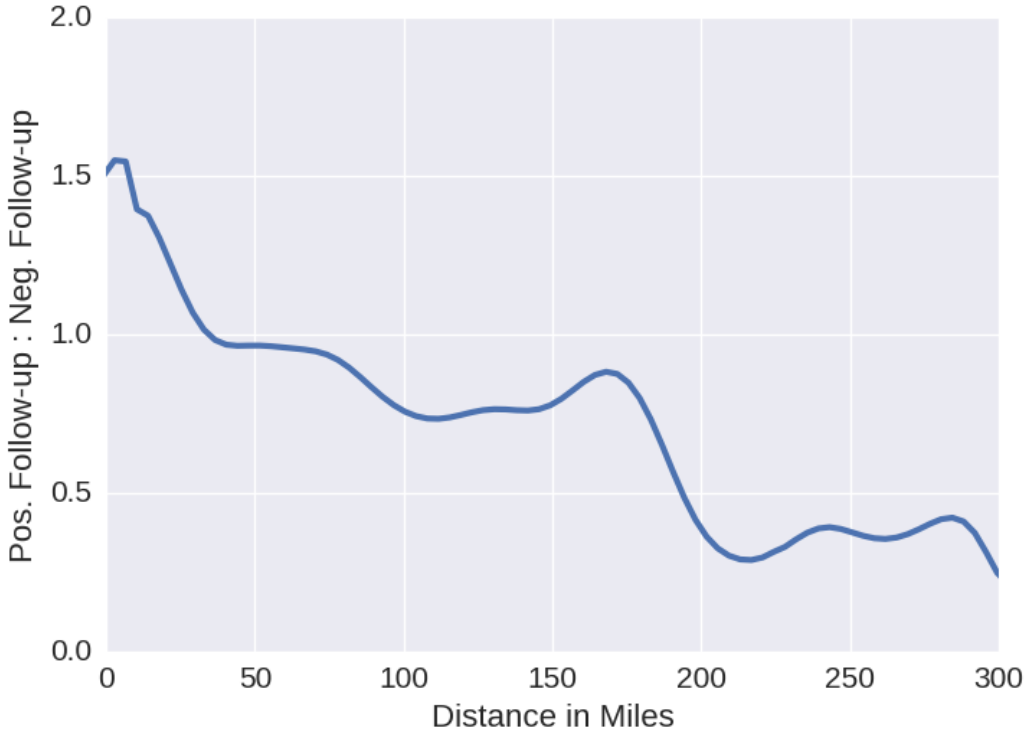



Figure 5.5: Odds ratio for follow-up status (pos five-year follow-up : neg. five-year follow-up) against distance to VUMC.

dictions for follow-up when follow-up was measured through oncologist appointments and adjuvant endocrine therapy medications yielded less prediction power than when follow-up was measured through appointments with any provider at VUMC (Figure 5.6).

When optimizing our random forest model, we found that the AUC increased as max depth of underlying decision trees increased until stabilizing at max depth of 15. Therefore, we may prune the model to include our 15 most significant features, i.e eliminate 306 out of the 321 features, and maintain our predictive power. Altering the minimum number of samples required to split a node did not appear to have a significant impact on AUCs. We found that splitting nodes based on gini index rather than entropy, resulted in the higher AUCs.

We found that our neural network models achieved stable AUCs with two hidden layers,

each with a number of nodes half the number of features. However, the neural networks under-performed compared to the random forest models. A meta analysis on the size of our data set suggests that our cohort is not large enough to achieve the maximum predictive power of a neural network. Neural network AUCs do not stabilize as our dataset size reaches its maximum (Figure 5.7).

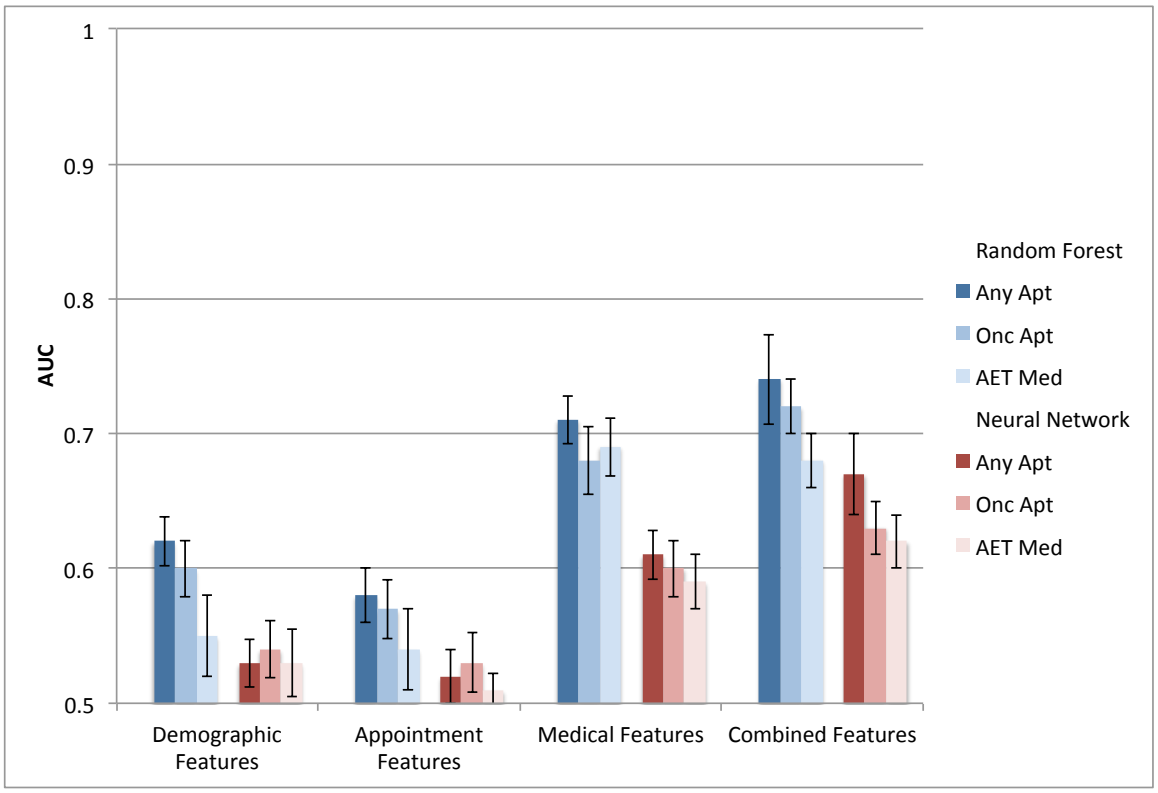



Figure 5.6: AUC measures for random forest and neural network classifiers built with demographic, appointment, medical features and all features combined. All classifiers are built with data collected from the time of breast cancer diagnosis to the start of adjuvant endocrine therapy.

Table 5.2: Top three features from random forest classifiers built from EHR, appointment log, and demographic data to predict five-year follow-up with any VUMC provider among adjuvant endocrine therapy patients

| | Primary Feature | Secondary Feature | Tertiary Feature |
|---|---|---|---|
| EHR Derived Features | Total medication count | ICD9 719 - Unspecified joint disorders | ICD9 715 - Osteoarthritis and allied disorders |
| Appointment Features | Average copay | Average appointment duration | Percent no-shows |
| Demographic Features | Age at diagnosis | Median zip income | Diagnosis year |

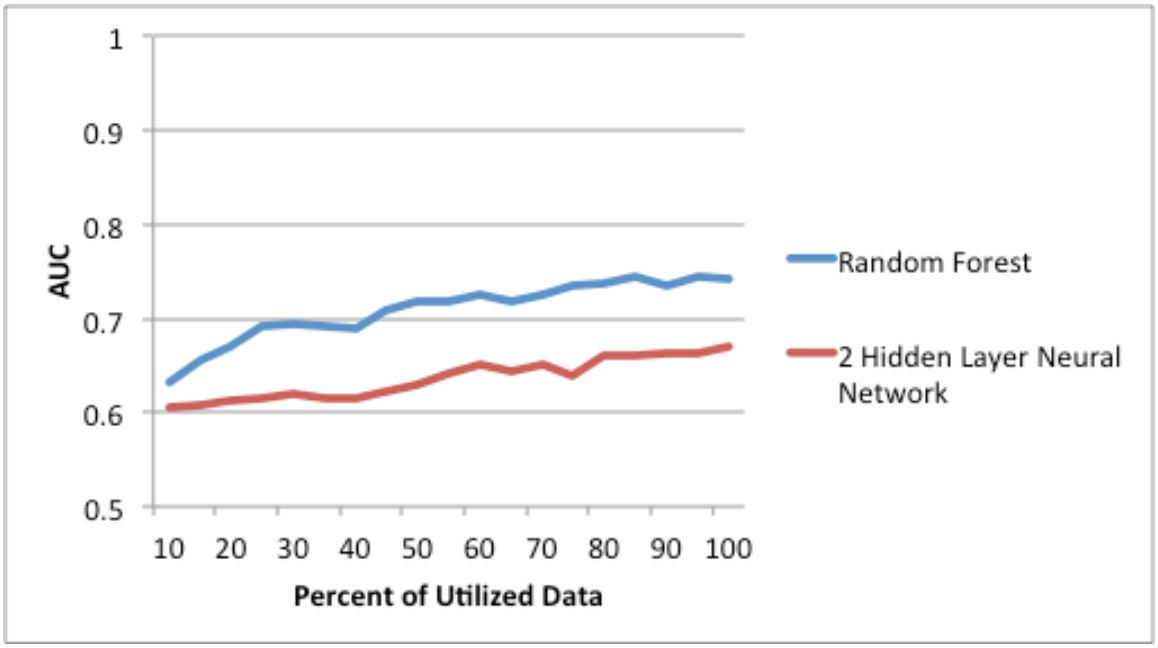



Figure 5.7: Change in AUC as size of dataset increases. The random forest AUC converges, indicating saturated data for prediction. The 2-layer neural network AUC does not converge, indicating additional data is required to achieve the best possible AUC. The data used in this classifier was collected between the date of diagnosis and adjuvant endocrine therapy start. Follow-up was measures through consistent appointments with ant provider.

### 5.5.5 Random Forest Temporal Classifiers

Predicting failure to follow-up in adjuvant endocrine therapy patients improves slightly as subsequent data is collected from the patients. At the time of adjuvant endocrine therapy start, the data available to predict failure to follow-up results in 0.74 AUC for random forest classification. The AUC increases with additional data collected over time, and rises 11% for follow-up predictions built from one year of treatment data (Figure 5.8). Additionally, the top significant features for random forest classifiers change over time, indicating a rise in important predictors as treatment data is collected (Table 5.3).

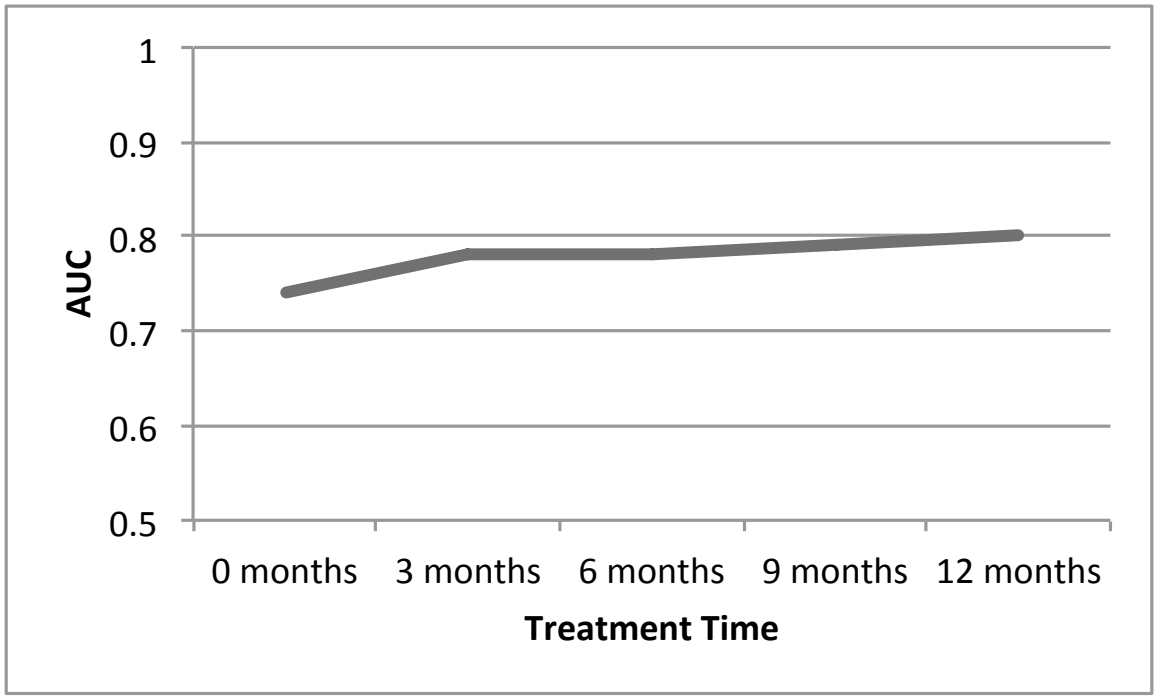



Figure 5.8: AUC for classifiers predicting follow-up in adjuvant endocrine therapy patients built with quarter year increments of data collected from treatment start. Each prediction is for five-year follow-up and the prediction window decreases as data is gathered. The outcome variable is consistent follow-up with any VUMC provider.

Table 5.3: Top three significant features from classifiers built on data collected in treatment timeframes with follow-up outcome as consistent appointments with any VUMC provider.

| Treatment Time | 0 months | 3 months | 6 months | 9 months | 12 months |
|---|---|---|---|---|---|
| 1st Feature | Total medications count | Total medications count | Total medications count | Total medications count | Total medications count |
| 2nd Feature | Age at diagnosis | ICD9 709 - Other disorders of skin and subcutaneous tissue | ICD9 709 - Other disorders of skin and subcutaneous tissue | ICD9 611 - Other disorders of the breast | ICD9 611 - Other disorders of the breast |
| 3rd Feature | Median zip income | ICD9 611 - Other disorders of the breast | ICD9 611 - Other disorders of the breast | ICD9 627 - Menopausal and postmenopausal disorders | ICD9 627 - Menopausal and postmenopausal disorders |

## 5.6 Discussion

Adjuvant endocrine therapy patients often fail to follow-up with their care providers for the recommended five-year time frame. Although reasons for failing to follow-up are not always reported to the care provider, medical-related, appointment-related, and demographic-related data stored in EHRs may hold predictors for follow-up. Learning predictors for follow-up may allow for interventions that improve follow-up rates, reduce recurrence rates, and ultimately improve patient care in the adjuvant endocrine therapy population.

Distance to VUMC may predict follow-up among adjuvant endocrine therapy patients

at VUMC, but there are many more factors that affect patient follow-up. Electronic Medical Records collect a wide variety of patient data that hold additional predictors for follow-up. Finding signals among the noise requires more complex methods than single measures.

We found that supervised machine learning is capable of predicting failure to follow-up using data from an adjuvant endocrine therapy patient cohort. The best classifier is a random forest, and incorporates medical-related, appointment-related, and demographic related features. Out of the three feature-types, medical-related features derived from the EHR are the most significant predictors for follow-up among adjuvant endocrine therapy patients.

The top three significant features in a combined random forest classifier include total number of medications, age at diagnosis, and median income in zip code. A high total number of medications for a patient suggest that 1) the patient has multiple conditions and 2) the patient receives multi-faceted care at VUMC. Whether a patient is treated at VUMC for other conditions may indicate that they are likely to follow-up at VUMC for their adjuvant endocrine therapy. Patient age may affect decision making for adjuvant endocrine therapy treatment. Older patients may be less inclined to follow-up, while younger patients may be more motivated to complete treatment. Last, median income for zip code suggests that financial burdens impact patient's follow-up rates.

Out of three measures for follow-up we found that consistent appointments with any provider at VUMC yields the best predictive power. This follow-up measure identifies patients that are more likely to have a primary care physician at VUMC or are being treated for other conditions at VUMC. Measuring follow-up through VUMC oncologist appointments excludes patients that transfer their adjuvant endocrine therapy to their primary care physician. Predictions for follow-up measured by adjuvant endocrine therapy medications were most difficult to predict, likely because of multiple reasons for follow-up failure by this measure (i.e. transferring care outside of VUMC, discontinuing drugs due to side effects, or discontinuing drugs due to metastatic disease).

Through training, testing and optimizing our classifiers, we found optimal conditions to predict follow-up given our feature set. Meta analyses measuring AUCs as a function of model complexity (tree depth in random forest and hidden layers in neural networks) identify minimally complex models that best describe the data without over-fitting. Furthermore, a meta analysis on dataset size identified a limitation in neural networks that did not affect random forests. We support that our random forest model is a valid model for predicting follow-up among adjuvant endocrine therapy patients.

Predicting follow-up in adjuvant endocrine therapy patients can be achieved with an as early as the time of treatment start with an AUC of 0.74. Predictive power increases as health data is collected from the patient throughout the first year of treatment. The changes in significant features predicting follow-up over time indicate a rise in predictors as treatment data as collected. At the time of treatment start, total number of medications is the most significant feature. This is likely correlated with the amount of treatment patients receive at VUMC. As treatment data is collected, we see that ICD9 codes for menopause and breast disorder become significant features predicting follow-up, indicating that the patient's health has impact on follow-up time.

## 5.7 Limitations

This study has limitations in both the dataset and the methods. The dataset used in this study is drawn from VUMC's EHR, which is not complete healthcare data for patients. Data applicable to the study, but unavailable for use, includes health plan data and claims data that describe patient's complete care rather than treatment received solely at VUMC. We measured adverse symptoms through billing codes, a strategy that captures adverse symptoms severe enough for a physician to bill. However, this approach can excludes many other patient complaints that are not billed for. The study is also limited by the medication even extraction tool used to identify patient current medications. This tool may erroneously identify drugs that a clinician records in patient notes for alternative reasons as

a drug the patient receives. Last, we assume that patients are adherent to the drugs reported as medication event despite evidence of non-adherence in other patient populations [76] [77].

We built a large feature matrix in which groups of features are likely to be correlated. For example, patients with an ICD code for cardiovascular disease likely have VUMC cardiologist appointments and likely have medication events for anticoagulants. Correlated features in random forests distort feature importances, as importance is spread out across the features. Determining groups of features within our feature matrix could alleviate this limitation. Our feature matrix did not include enough rows for a neural network to achieve the highest possible AUC. Supplementing out dataset with patient data from other healthcare centers would enrich our neural network classifier.

Last, there are more features that may predict follow-up that are not recorded as part of healthcare data therefore were inaccessible for this study. These features may include patient access to transportation and childcare, or patient flexibility in work schedule. Socioeconomic factors are shown to impact adherence to healthcare [78] [79], and would enrich models for follow-up predictions among breast cancer patients and patient in other healthcare domains.

## 5.8 Conclusions

We built a supervised machine learning classifier capable of predicting five-year follow-up measured three ways among adjuvant endocrine patients using medical, appointment, and demographic data recorded in EHRs. Our classifier supports that total medication count at VUMC, which is a measure of amount of care received at VUMC, is the most significant predictor for follow-up long-term. Furthermore, although follow-up may be predicted as early as the start of adjuvant endocrine therapy, predictive power increases as treatment data is collected throughout the first year of treatment. Learning predictors for follow-up can facilitate interventions that improve follow-up rates, guide clinical decision-making, and

ultimately improve patient care. This study shows that EHR data and supervised machine learning are valuable resources for finding opportunities for improvement in patient care.

The main findings of the study include:

- VUMC adjuvant endocrine therapy follow-up can be predicted with and AUC of 0.74 using supervised machine learning methods
- EHR data from adjuvant endocrine therapy patients holds predictors for follow-up
- One significant predictor for adjuvant endocrine therapy follow-up at VUMC is the total number of medications, a measure of illness and care received at VUMC
- Supervised machine learning is a useful method for learning new opportunities for improvement in patient care through EHR data

Table 5.1: Delineated list of EHR, appointment, and demographic features calculated for each patient. These features are used in a random forest classifier to predict failure to follow-up among adjuvant endocrine therapy patients.

| |
|---|
| EHR Derived Features<br>• Cancer stage<br>• Adjuvant endocrine therapy medication type (0/1 for each type)<br>• Non-adjuvant endocrine therapy medication classes (Derived from RxNorm, count for each class)<br>• ICD Parent Codes (count)<br>• CPT Codes (count)<br>• Providers seen (count)<br>• Departments visited (count)<br>• Number of clinical communications |
| Appointment Features<br>• Referral (0/1)<br>• Percent of appointments scheduled at time of previous appointment<br>• Percent of appointments cancelled<br>• Percent no-shows<br>• Number of appointments scheduled<br>• Average time between appointments<br>• Average appointment duration |
| Demographic Features<br>• Patient Age<br>• Diagnosis year<br>• Insurance type (public/private)<br>• Average copay<br>• Distance to care facility in miles<br>• Median Income for zip code<br>• Percent of population filing taxes for zip code |

# Chapter 6

## Overarching Conclusions

This dissertation describes methods to learn the state of patient care and opportunities for improvement from electronic health record (EHR) data. We approached this task through three aims: 1) measure the **sufficiency** of EHR patient data, 2) characterize the **state** of patient care, and 3) identify **opportunities** for improvement in patient care. For proof of concept, the methods were applied to a cohort of adjuvant endocrine therapy patients treated at VUMC. However, the methods were built to be generalizable to other EHRs and other healthcare domains. Overall, this dissertation matches clinical datasets with computational methods to derive new clinical knowledge.

The scope of this dissertation is breast cancer patients receiving adjuvant endocrine therapy at VUMC. Adjuvant endocrine therapy is a long term treatment that is challenging to characterize in real-world settings due to the need for consistent longitudinal data. Nevertheless, characterizing adjuvant endocrine therapy in practice benefits patients and care providers by providing realistic expectations for treatment and guiding in treatment planning. The scope of this work directly benefits the breast cancer clinicians and patients at VUMC describing providing previously unmeasured clinical workflows.

Work for Aim 1 measured the **sufficiency** of EHR patient data for study of adjuvant endocrine therapy at VUMC. Chapter 3 describes three tasks to satisfy the aim: 1) measure data availability, 2) build metrics based on data availability to facilitate data selection, and 3) generalize the methodology for application in other health care domains. We learned that 1) data sufficiency can drive data selection for secondary use studies. We depended on the availability of different EHR datatypes to create adjuvant endocrine therapy cohorts. 2) Data sufficiency metrics can serve as weights for missing data points. We weighted missing data from adjuvant endocrine therapy patients to balance loss from the EHR system with

discontinued treatment. 3) Data sufficiency affects secondary use study results. We measured adherence rates across adjuvant endocrine therapy patient cohorts built with different data availability metrics and found varying results.

Work for Aim 2 characterized the **state** of patient care among adjuvant endocrine therapy patients at VUMC. Chapter 4 describes three tasks to satisfy the aim: 1) define all possible states of adjuvant endocrine therapy care, 2) determine patient inclusion in each state from EHR data, and 3) build statistical and visualization methods to characterize patient care. We learned that 1) EHR data can determine the distribution of patient across clinical workflows. We were able to estimate rates of drug switches, drug discontinuations, recurrence and death among a adjuvant endocrine therapy patients at VUMC. 2) Despite the limitation of including only healthcare administered within and EHR system, EHR data are valuable resources for drawing information on the state of patient care. The information on adjuvant endocrine therapy patients at VUMC is valuable to patients and providers because it informs patients and facilitates treatment planning.

Work for Aim 3 identified **opportunities** for improvement in patient care among adjuvant endocrine therapy patients at VUMC. Chapter 5 describes three tasks to satisfy the aim: 1) extract features with the potential to affect care outcomes from EHR data, 2) frame a classification problem using features and labels from the data, and 3) apply machine learning methods to predict clinical outcomes. We learned that 1) EHR data contains features that can predict the clinical outcome of follow-up among adjuvant endocrine therapy patients at VUMC and 2) supervised machine learning is an appropriate method to predict follow-up with EHR data. A significant predictor for adjuvant endocrine therapy follow-up at VUMC is the total number of medications, a measure of illness and care received at VUMC. This predictor informs clinicians that patients that are less ill or receive the majority of their care outside of VUMC are less likely to follow-up. To translate this predictor into and opportunity to improve care, a provider can take steps to refer the patient to a non-VUMC provider to make it easier to them for continue their treatment. Patients may

be more likely to continue treatment when it is simple for them.

A consistent limitation in this work is the boundary of VUMC's EHR data, which captures patient care only administered in the VUMC system. Although VUMC data covers a large number of clinics, it is likely that many patients in our study sought health care outside of the VUMC system. Consequently, our dataset fails to capture complete health care data for our cohort, and has ambiguities in care. By supplementing our datasets with health plan data, or additional public and private clinical datasets, we can achieve a more complete trajectory of healthcare for patients. However, matching clinical datasets is challenging. With the necessity to protect patient privacy, patient identifiers are often unique to systems and slows patient identifying and tracking [80] [81]. Nevertheless, matching clinical datasets informs clinicians of complete patient care, and improves secondary use studies like those described in this dissertation. Therefore, there is a drive for data standards, interoperability, and health information exchange. This study would benefit from EHR data enriched with other public and private datasets like health plan data. Capturing complete health data from patients would minimizes ambiguity and inconsistencies while improving accuracy of these results.

## 6.1 Future Directions

In future work, this study would greatly benefit from 1) enrichment with additional datasets, 2) testing the generalizability to other healthcare domains, and 3) implementing an opportunity for improvement into practice.

Repeating these studies using a more enriched dataset would reduce ambiguities and inconsistencies in the current results, and yield a more comprehensive understanding of adjuvant endocrine therapy in and outside of VUMC. Additional datasets that would enrich this work include datasets from alternative EHRs and health plan data. To use these additional datasets, we need to extract the datatypes necessary for our methods: medications, billing codes, and appointment data. We would determine the sufficiency of the data with

Table 6.1: Summary of conclusions from the dissertations divided into knowledge from conclusions and beliefs about the broader applications.

| Aim | Knowledge | Beliefs |
|---|---|---|
| 1. Sufficiency of the data | EHR data sufficiency impacts studies on adjuvant endocrine therapy at VUMC by driving cohort selection and serving as weights for missing data. | Methodologies to determine data sufficiency among adjuvant endocrine therapy patients at VUMC are generalizable to secondary use studies in other healthcare domains. |
| 2. State of patient care | The EHR is valuable data source for characterizing VUMC adjuvant endocrine therapy. We were able to estimate rates of drug switches, drug discontinuation, adverse symptoms and outcomes, using patient data. | The EHR is a valuable resource for characterizing the state of patient care for healthcare domains beyond adjuvant endocrine therapy. |
| 3. Opportunities for improvement | EHR data from adjuvant endocrine therapy patients at VUMC holds predictors for follow-up, and supervised machine learning is a useful tool to learn follow-up probabilities. | EHR data matched with appropriate computational methods is a useful approach to identify opportunities for patient care improvement across many healthcare domains. |

our methodology, which includes determining the start and end of data on patients. We would then characterize the state of patient care. Health plan data is especially beneficial in this task. Health plan data includes care received at any institution regardless of the EHR system, and will include death. Complete healthcare data eliminates ambiguities that exist when a patient receives care outside of a single EHR system. Identifying opportunities for improvement would be simplified with enriched datasets. We predicted follow-up for adjuvant endocrine therapy patients, but the cohort of patients that fail to follow-up is diverse. Patients that fail to follow-up include patients that continue care outside of VUMC, patients that discontinue care, and patients that die. Distinguishing between patient groups allows for more focused predictions. For example, rather than predict follow-up at VUMC, we can predict discontinued care. Truly improved patient care extends beyond where the patient follows-up for care.

In additional future work, we will test the extent of the methodology's generalizability to other healthcare domains and other EHR systems. Other healthcare domains that are appropriate test cases for these methods are long-term treatments. Therefore, we could test the methods on other neoplasms for chronic conditions such as diabetes or chronic heart failure. We could also test the methods for short-term conditions to learn how the methods may be adapted to characterize the state of patient care and opportunities for improvement among short-term patients. Short-term care includes emergency room visits, flu or infections. To generalize to other EHR systems, we need access to other EHR datasets. We may learn additional steps for data clean-up and standardization necessary for our methodology. For example, EHRs implement functionality and update standards at different rates, so data availability changes across EHRs. We may also explore the potential to match patients across EHR datasets and gain a more complete outlook of care. This would be unnecessary with health plan data, but valuable if only EHR data is accessible.

Applying the methods in this dissertation to other healthcare domains and other EHR datasets would identify further opportunities and limitations, strengthen our findings, and

make the methods comprehensive. The demand to draw new knowledge from clinical data holds across all healthcare domains. Scientific methods to characterize the state of patient care and identify opportunities for improvement from EHR data are proportionately valuable to the number of healthcare domains in which they are applicable.

Last, opportunities for improvement identified through these methods can be implemented into healthcare practice to determine efficacy in shifting distribution of patients into optimal states. We found that patients that receive primary care at VUMC are more likely to follow-up for adjuvant endocrine therapy at VUMC. To shift the distribution of patients away from failure to follow-up at VUMC to follow-up at VUMC, we may alert providers when a patient is less likely to follow-up. This could occur as a notification in the EHR at the time of an appointment. Then, the provider may recommend a primary care physician at VUMC so complete treatment is received at VUMC. Once this process is implemented, we can measure the rate of follow-up over time. If the rate of follow-up for adjuvant endocrine therapy at VUMC increases in relation to the historical follow-up rate, than 1) we have successfully found an opportunity for improvement, and 2) acted on it leading to an improvement in healthcare for adjuvant endocrine therapy patients.

# BIBLIOGRAPHY


[1] Wayne Kondro. Medical errors increasing because of complexity of care and breakdown in doctor-patient relationship, physician consultant says. *Canadian Medical Association Journal*, 2010.

[2] M Eichelberg, T Aden, J Riesmeier, A Dogac, and G B Laleci. ELECTRONIC HEALTH RECORD STANDARDS A BRIEF OVERVIEW.

[3] Sharon Silow-Carroll, Jennifer N Edwards, and Diana Rodin. Using electronic health records to improve quality and efficiency: the experiences of leading hospitals. *Issue Brief (Commonw Fund)*, 17:1–40, 2012.

[4] Hua Xu, Shane P Stenner, Son Doan, Kevin B Johnson, Lemuel R Waitman, and Joshua C Denny. MedEx: a medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association : JAMIA*, 17(1):19–24, 2010.

[5] Nicole Gray Weiskopf and Chunhua Weng. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, 20(1):144–151, 2013.

[6] Susan Rea, Jyotishman Pathak, Guergana Savova, Thomas A Oniki, Les Westberg, Calvin E Beebe, Cui Tao, Craig G Parker, Peter J Haug, Stanley M Huff, et al. Building a robust, scalable and standards-driven infrastructure for secondary use of ehr data: the sharpn project. *Journal of biomedical informatics*, 45(4):763–771, 2012.

[7] Frank Ueckert, Michael Goerz, Maximilian Ataian, Sven Tessmann, and Hans-Ulrich Prokosch. Empowerment of patients and communication with health care professionals through an electronic health record. *International Journal of Medical Informatics*, 70(2):99–108, 2003.

[8] Hilary K Wall, Judy A Hannan, and Janet S Wright. Patients with undiagnosed hypertension: Hiding in plain sight. *JAMA*, 312(19):1973–1974, 2014.

[9] David C Radley, Melanie R Wasserman, Lauren EW Olsho, Sarah J Shoemaker, Mark D Spranca, and Bethany Bradshaw. Reduction in medication errors in hospitals due to adoption of computerized provider order entry systems. *Journal of the American Medical Informatics Association*, 20(3):470–476, 2013.

[10] P. J. O'Connor, J. M. Sperl-Hillen, W. A. Rush, P. E. Johnson, G. H. Amundson, S. E. Asche, H. L. Ekstrom, and T. P. Gilmer. Impact of Electronic Health Record Clinical Decision Support on Diabetes Care: A Randomized Trial. *The Annals of Family Medicine*, 9(1):12–21, jan 2011.

[11] Catherine A McCarty, Rex L Chisholm, Christopher G Chute, Iftikhar J Kullo, Gail P Jarvik, Eric B Larson, Rongling Li, Daniel R Masys, Marylyn D Ritchie, Dan M Roden, and W et al Burke. The eMERGE Network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Medical Genomics*, 4(1):13, dec 2011.

[12] Max J Romano and Randall S Stafford. Electronic health records and clinical decision support systems: impact on national ambulatory care quality. *Archives of internal medicine*, 171(10):897–903, 2011.

[13] I-N Lee, S-C Liao, and M Embrechts. Data mining techniques applied to medical information. *Medical informatics and the Internet in medicine*, 25(2):81–102, 2000.

[14] H. S. Rugo, R. B. Rumble, E. Macrae, D. L. Barton, H. K. Connolly, M. N. Dickler, L. Fallowfield, B. Fowble, J. N. Ingle, M. Jahanzeb, S. R. D. Johnston, L. A. Korde, J. L. Khatcheressian, R. S. Mehta, H. B. Muss, and H. J. Burstein. Endocrine Therapy for Hormone Receptor-Positive Metastatic Breast Cancer: American Society of Clinical Oncology Guideline. *Journal of Clinical Oncology*, page JCO671487, may 2016.

[15] A Howell, J Cuzick, M Baum, A Buzdar, M Dowsett, J F Forbes, G Hoctin-Boes, J Houghton, G Y Locker, and J S Tobias. Results of the ATAC (Arimidex, Tamoxifen, Alone or in Combination) trial after completion of 5 years' adjuvant treatment for breast cancer. *Lancet*, 365(9453):60–2, jan 2005.

[16] Early Breast Cancer Trialists' Collaborative Group. Tamoxifen for early breast cancer: an overview of the randomised trials. Early Breast Cancer Trialists' Collaborative Group. 351(9114):1451–67, 1998.

[17] R.A. Greenes, A.N. Pappalardo, C.W. Marble, and G.Octo Barnett. Design and implementation of a clinical data management system. *Computers and Biomedical Research*, 2(5):469 – 485, 1969.

[18] Basit Chaudhry, Jerome Wang, Shinyi Wu, Margaret Maglione, Walter Mojica, Elizabeth Roth, Sally C Morton, and Paul G Shekelle. Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. *Annals of internal medicine*, 144(10):742–52, may 2006.

[19] Jennifer King, Vaishali Patel, Eric W. Jamoom, and Michael F. Furukawa. Clinical benefits of electronic health record use: National findings. *Health Services Research*, 49(1pt2):392–404, 2014.

[20] Samuel J Wang, Blackford Middleton, Lisa A Prosser, Christiana G Bardon, Cynthia D Spurr, Patricia J Carchidi, Anne F Kittler, Robert C Goldszer, David G Fairchild, Andrew J Sussman, Gilad J Kuperman, and David W Bates. A cost-benefit analysis of electronic medical records in primary care. *The American journal of medicine*, 114(5):397–403, apr 2003.

[21] David Blumenthal. Launching hitech. *New England Journal of Medicine*, 362(5):382–385, 2010. PMID: 20042745.

[22] David Blumenthal and Marilyn Tavenner. The meaningful use regulation for electronic health records. *New England Journal of Medicine*, 363(6):501–504, 2010. PMID: 20647183.

[23] Furukawa King, Patel. Physician Adoption of Electronic Health Record Technology to Meet Meaningful Use Objectives: 2009-2012. (7):2009–2012, 2012.

[24] Dawn Heisey-Grove and Vaishali Patel. Any, Certified, and Basic: Quantifying Physician EHR Adoption through 2014. 2014.

[25] Jionglin Wu, Jason Roy, and Walter F Stewart. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. *Medical care*, 48(6 Suppl):S106–13, jun 2010.

[26] Peter B Jensen, Lars J Jensen, and Søren Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405, 2012.

[27] Hunt DL, Haynes R, Hanna SE, and Smith K. Effects of computer-based clinical decision support systems on physician performance and patient outcomes: A systematic review. *JAMA*, 280(15):1339–1346, 1998.

[28] Taxiarchis Botsis, Gunnar Hartvigsen, Fei Chen, and Chunhua Weng. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. *AMIA Joint Summits on Translational Science proceedings AMIA Summit on Translational Science*, 2010:1–5, jan 2010.

[29] Nicole G. Weiskopf, George Hripcsak, Sushmita Swaminathan, and Chunhua Weng. Defining and measuring completeness of electronic health records for secondary use. *Journal of Biomedical Informatics*, 46(5):830 – 836, 2013.

[30] James D Yager and Nancy E Davidson. Estrogen carcinogenesis in breast cancer. *New England Journal of Medicine*, 354(3):270–282, 2006.

[31] Shahla Masood. Estrogen and progesterone receptors in cytology: A comprehensive review. *Diagnostic Cytopathology*, 8(5):475–491, 1992.

[32] F Lumachi, A Brunello, M Maruzzo, U Basso, and S MM Basso. Treatment of estrogen receptor-positive breast cancer. *Current medicinal chemistry*, 20(5):596–604, 2013.

[33] W. R. Miller, J. Bartlett, A. M. H. Brodie, R. W. Brueggemeier, E. di Salle, P. E. Lonning, A. Llombart, N. Maass, T. Maudelonde, H. Sasano, and P. E. Goss. Aromatase Inhibitors: Are There Differences Between Steroidal and Nonsteroidal Aromatase Inhibitors and Do They Matter? *The Oncologist*, 13(8):829–837, aug 2008.

[34] Henry et al. Prospective characterization of musculoskeletal symptoms in early stage breast cancer patients treated with aromatase inhibitors. *Breast cancer research and treatment*, 111(2):365–72, sep 2008.

[35] B Fisher. A randomized clinical trial evaluating tamoxifen in the treatment of patients with node-negative breast cancer who have estrogen-receptor positive tumors. *The New England Journal of Medicine*, 320(8), 1989.

[36] ANTONIO C. WOLFF and NANCY E. DAVIDSON. Use of SERMs for the Adjuvant Therapy of Early-Stage Breast Cancer. *Annals of the New York Academy of Sciences*, 949(1):80–88, jan 2006.

[37] Caitlin C Murphy, L Kay Bartholomew, Melissa Y Carpentier, Shirley M Bluethmann, and Sally W Vernon. Adherence to adjuvant hormonal therapy among breast cancer survivors in clinical practice: a systematic review. *Breast cancer research and treatment*, 134(2):459–78, jul 2012.

[38] L Kligman and J Younus. Management of hot flashes in women with breast cancer. *Current oncology (Toronto, Ont.)*, 17(1):81–6, feb 2010.

[39] Uwe Güth, Mary Elizabeth Myrick, Andreas Schötzau, Nerbil Kilic, and Seraina Margaretha Schmid. Drug switch because of treatment-related adverse side effects in endocrine adjuvant breast cancer therapy: how often and how often does it work? *Breast cancer research and treatment*, 129(3):799–807, oct 2011.

[40] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[41] Thomas A Lasko, Joshua C Denny, and Mia A Levy. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PloS one*, 8(6):e66341, 2013.

[42] Kiyana Zolfaghar, Naren Meadem, Ankur Teredesai, Senjuti Basu Roy, Si-Chi Chin, and Brian Muckian. Big data solutions for predicting risk-of-readmission for congestive heart failure patients. In *Big Data, 2013 IEEE International Conference on*, pages 64–71. IEEE, 2013.

[43] Lynn A Gloeckler Ries, Marsha E Reichman, Denise Riedel Lewis, Benjamin F Hankey, and Brenda K Edwards. Cancer survival and incidence from the surveillance, epidemiology, and end results (seer) program. *The oncologist*, 8(6):541–552, 2003.

[44] Joe V Selby, Anne C Beal, and Lori Frank. The patient-centered outcomes research institute (pcori) national priorities for research and initial research agenda. *Jama*, 307(15):1583–1584, 2012.

[45] Silva I Santos et al. Cancer epidemiology, principles and methods. *Cancer epidemiology, principles and methods*, 1999.

[46] Prakash M Nadkarni. Drug safety surveillance using de-identified EMR and claims data: issues and challenges. *Journal of the American Medical Informatics Association : JAMIA*, 17(6):671–4, 2010.

[47] Brownstein John S, Sordo Margarita, Kohane Isaac S, and Mandl Kenneth D. The tell-tale heart: population-based surveillance reveals an association of rofecoxib and celecoxib with myocardial infarction. 2(9):e840, 2007.

[48] Jose-Franck Diaz-Garelli, Elmer V Bernstam, Mse, Mohammad H Rahbar, and Todd Johnson. Rediscovering drug side effects: the impact of analytical assumptions on the detection of associations in EHR data. *AMIA Joint Summits on Translational Science proceedings AMIA Summit on Translational Science*, 2015:51–5, 2015.

[49] Lingling Li, Changyu Shen, Xiaochun Li, and James M Robins. On weighting approaches for missing data. *Statistical methods in medical research*, 22(1):14–30, 2013.

[50] Jing Zhao. Temporal weighting of clinical events in electronic health records for pharmacovigilance. In *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 375–381. IEEE, 2015.

[51] Brian J Wells, Kevin M Chagin, Amy S Nowacki, and Michael W Kattan. Strategies for handling missing data in electronic health record derived data. *EGEMS (Washington, DC)*, 1(3):1035, 2013.

[52] D M Roden, J M Pulley, M A Basford, G R Bernard, E W Clayton, J R Balser, and D R Masys. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clinical pharmacology and therapeutics*, 84(3):362–9, 2008.

[53] No authors listed. Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. *Lancet*, 365(9472):1687–1717, 2005.

[54] Dawn L Hershman. Perfecting breast-cancer treatment–incremental gains and musculoskeletal pains. *The New England journal of medicine*, 372(5):477–8, 2015.

[55] Timothy L Lash, Matthew P Fox, Jennifer L Westrup, Aliza K Fink, and Rebecca A Silliman. Adherence to tamoxifen over the five-year course. *Breast cancer research and treatment*, 99(2):215–20, 2006.

[56] Dawn L Hershman, Theresa Shao, Lawrence H Kushi, Donna Buono, Wei Yann Tsai, Louis Fehrenbacher, Marilyn Kwan, Scarlett Lin Gomez, and Alfred I Neugut. Early discontinuation and non-adherence to adjuvant hormonal therapy are associated with increased mortality in women with breast cancer. *Breast cancer research and treatment*, 126(2):529–37, 2011.

[57] Dustin Charles, Meghan Gabriel, and Talicia Searcy. Adoption of Electronic Health Record Systems among U.S. NonFederal Acute Care Hospitals: 2008-2014, 2014.

[58] Dean F Sittig, Adam Wright, Jerome A Osheroff, Blackford Middleton, Jonathan M Teich, Joan S Ash, Emily Campbell, and David W Bates. Grand challenges in clinical decision support. *Journal of biomedical informatics*, 41(2):387–92, apr 2008.

[59] Abenaa M Brewster, Gabriel N Hortobagyi, Kristine R Broglio, Shu-Wan Kau, Cesar A Santa-Maria, Banu Arun, Aman U Buzdar, Daniel J Booser, Vincente Valero, Melissa Bondy, and Francisco J Esteva. Residual risk of breast cancer recurrence 5 years after adjuvant therapy. *Journal of the National Cancer Institute*, 100(16):1179–83, aug 2008.

[60] K. D. Crew, H. Greenlee, J. Capodice, G. Raptis, L. Brafman, D. Fuentes, A. Sierra, and D. L. Hershman. Prevalence of Joint Symptoms in Postmenopausal Women Taking Aromatase Inhibitors for Early-Stage Breast Cancer. *Journal of Clinical Oncology*, 25(25):3877–3883, sep 2007.

[61] Jennifer R. Garreau, Tammy DeLaMelena, Deb Walts, Kasra Karamlou, and Nathalie Johnson. Side effects of aromatase inhibitors versus tamoxifen: the patients' perspective. *The American Journal of Surgery*, 192(4):496–498, 2006.

[62] S Yousuf Zafar, Jeffrey M Peppercorn, Deborah Schrag, Donald H Taylor, Amy M Goetzinger, Xiaoyin Zhong, and Amy P Abernethy. The financial toxicity of cancer treatment: a pilot study assessing out-of-pocket expenses and the insured cancer patient's experience. *The oncologist*, 18(4):381–390, 2013.

[63] Sun Jimeng, Candace D McNaughton, Ping Zhang, Adam Perer, Aris Gkoulalas-Divanis, and Joshua C Denny. Predicting changes in hypertension control using electronic health records from a chronic disease management program. *Journal of the American Medical Informatics Association : JAMIA*, 21(2):337–44, 2014.

[64] Guan Wang, Kenneth Jung, Rainer Winnenburg, and Nigam H Shah. A method for systematic discovery of adverse drug events from clinical notes. *Journal of the American Medical Informatics Association*, 2015.

[65] Edward Choi, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association*, 2016.

[66] Fabian Güiza, Jan Ramon, and Maurice Bruynooghe. Machine learning techniques to examine large patient databases. *Best Practice & Research Clinical Anaesthesiology*, 23(1):127–143, 2009.

[67] E. M. Kleinberg. An overtraining-resistant stochastic modeling method for pattern recognition. *The Annals of Statistics*, 24(6):2319–2349, 1996.

[68] Stephen I. Gallant. *Neural Network Learning and Expert Systems.* A Bradford Book, 1993.

[69] Le Cun, Y Le Cun, B Boser, J S Denker, D Henderson, R E Howard, W Hubbard, and L D Jackel. Handwritten Digit Recognition with a Back-Propagation Network.

[70] S. Lawrence, C. L. Giles, Ah Chung Tsoi, and A. D. Back. Face recognition: a convolutional neural-network approach. *IEEE Transactions on Neural Networks*, 8(1):98–113, Jan 1997.

[71] M Aczon, D Ledbetter, L Ho, A Gunny, A Flynn, J Williams, and R Wetzel. Dynamic Mortality Risk Predictions in Pediatric Critical Care Using Recurrent Neural Networks. jan 2017.

[72] William G. Baxt. Use of an Artificial Neural Network for Data Analysis in Clinical Decision-Making: The Diagnosis of Acute Coronary Occlusion. *Neural Computation*, 2(4):480–489, dec 1990.

[73] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[74] François Chollet. Keras. https://github.com/fchollet/keras, 2015.

[75] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016.

[76] Michael A. Fischer, Margaret R. Stedman, Joyce Lii, Christine Vogeli, William H. Shrank, M. Alan Brookhart, and Joel S. Weissman. Primary Medication Non-Adherence: Analysis of 195,930 Electronic Prescriptions. *Journal of General Internal Medicine*, 25(4):284–290, apr 2010.

[77] Carmel M Hughes. Medication Non-Adherence in the Elderly. *Drugs & Aging*, 21(12):793–811, 2004.

[78] Ken S. Field and David J. Briggs. Socio-economic and locational determinants of accessibility and utilization of primary health-care. *Health and Social Care in the Community*, 9(5):294–308, 2001.

[79] E. Vermeire, H. Hearnshaw, P. Van Royen, and J. Denekens. Patient adherence to treatment: three decades of research. a comprehensive review. *Journal of Clinical Pharmacy and Therapeutics*, 26(5):331–342, 2001.

[80] Richard E Gliklich, Nancy A Dreyer, and Michelle B Leavy. Managing Patient Identity Across Data Sources. 2014.

[81] Beth Haenke Just, David Marc, Megan Munns, and Ryan Sandefer. Why Patient Matching Is a Challenge: Research on Master Patient Index (MPI) Data Discrepancies in Key Identifying Fields. *Perspectives in health information management*, 13(Spring):1e, 2016.