GENERATING EVIDENCE FOR COPD CLINICAL GUIDELINES USING EHRS

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Amber M. Johnson

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

August 2019

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF DISSERTATION APPROVAL

Dr. Bharat Bhargava, Chair

Department of Computer Science

Dr. Mohammad Adibuzzaman

Regenstrief Center for Healthcare Engineering

Dr. Christopher W. Clifton

Department of Computer Science

Dr. Eugene H. Spafford

Department of Computer Science

Dr. Samuel S. Wagstaff, Jr.

Department of Computer Science

Approved by:

Dr. Voicu S. Popescu

Head of Department Graduate Program

In

Loving Memory

of

Ethel Mae.

February 9, 1952 - July 24, 2014

ACKNOWLEDGMENTS

Proverbs 3:5-6 says, “Trust in the Lord with all your heart and lean not on your own understanding. In all your ways acknowledge Him, and He will direct your path.”

This work is evidence that God has a plan.

To God be the glory.

PREFACE

Advancements in technology have allowed us to collect and analyze data, capturing and exposing information that may have otherwise been buried within the large amounts of data we produce each day. Each day, billions of bytes of healthcare data are collected via smart devices and in care settings. Similarly to the data, billions of dollars in healthcare costs are collected annually for COPD care and treatments. According to the National Institutes of Health (NIH) [1], 12 million adults in the U.S. are diagnosed with COPD and, each year, 120,000 die from the incurable disease.

COPD is a slow-developing disease whose origins are not fully understood. The more information we know about COPD, the more we understand the impact that it has on our community and the opportunities we have to develop tools for informed decision-making and treatment selection using the technological advances and data made available to us. Thus, this dissertation leverages Computer Science concepts and applications to provide a framework for analyzing raw clinical data collected from ICU patients. The methods in this body of work can potentially enable medical researchers to study the history of millions of individual COPD patients to learn what they were treated for prior to being diagnosed with COPD.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

SYMBOLS

ARF_1	acute respiratory failure 1
ARF_2	acute respiratory failure 2
FiO_2	fraction of inspired oxygen
NARF	no acute respiratory failure
$PaCO_2$	partial pressure of carbon dioxide
PaO_2	partial pressure of oxygen

ABBREVIATIONS

ABG	Arterial Blood Gas
AECOPD	Acute Exacerbation of Chronic Obstructive Pulmonary Disease
API	Application Programming Interface
ARF	Acute Respiratory Failure
CDM	Common Data Model
CO2	Carbon Dioxide
COPD	Chronic Obstructive Pulmonary Disease
CPT	Current Procedural Terminology
CSV	Comma Separated Value
DB	Database
EAV	Entity-attribute Value
EHR	Electronic Health Record
ETL	Extract, Load, Transform
FEV_1	Forced Expiratory Volume
FHIR	Fast Healthcare Interoperability Resources
GOLD	Global Initiative for Chronic Obstructive Lung Disease
HIPPA	Health Insurance Portability and Accountability Act
HR	Heart Rate
I2B2	Informatics for Integrating Biology and the Bedside
ICD-9	International Classification of Diseases, Ninth Revision, Clinical Modification
ICU	Intensive Care Unit
IRB	Institutional Review Board
MCMC	Markov Chain Monte Carlo

MIMIC	Medical Information Mart for Intensive Care III
MIMIC-III	Medical Information Mart for Intensive Care III
NDC	National Drug Codes
NLP	Natural Language Processing
OOP	Object Oriented Programming
OHDSI	Observational Health Data Science and Informatics
OMOP	Observational Medical Outcomes Partnership
PCORNET	Patient-centered Clinical Research Network
PICO	Patient/population/problem, Intervention, Comparison, Outcome
RCT	Randomized Controlled Trial
RR	Respiratory Rate
SABD	Short Acting Bronchodilators
SBP	Systolic Blood Pressure
SQL	Structured Query Language
SSH	Secure Shell
TPM	Transition Probability Matrix

ABSTRACT

Johnson, Amber M. Ph.D., Purdue University, August 2019. Generating Evidence for COPD Clinical Guidelines Using EHRs. Major Professor: Bharat Bhargava.

The Global Initiative for Chronic Obstructive Lung Disease (GOLD) guidelines are used to guide clinical practices for treating Chronic Obstructive Pulmonary Disease (COPD). GOLD focuses heavily on stable COPD patients, limiting its use for non-stable COPD patients such as those with severe, acute exacerbations of COPD (AECOPD) that require hospitalization. Although AECOPD can be heterogeneous, it can lead to deterioration of health and early death. Electronic health records (EHRs) can be used to analyze patient data for understanding disease progression and generating guideline evidence for AECOPD patients. However, because of its structure and representation, retrieving, analyzing, and properly interpreting EHR data can be challenging, and existing tools do not provide granular analytic capabilities for this data.

This dissertation presents, develops, and implements a novel approach that systematically captures the effect of interventions during patient medical encounters, and hence may support evidence generation for clinical guidelines in a systematic and principled way. A conceptual framework that structures components, such as data storage, aggregation, extraction, and visualization, to support EHR data analytics for granular analysis is introduced. We develop a software framework in Python based on these components to create longitudinal representations of raw medical data extracted from the Medical Information Mart for Intensive Care (MIMIC-III) clinical database. The software framework consists of two tools: *Patient Aggregated Care Events (PACE)*, a novel tool for constructing and visualizing entire medical histories of both individual patients and patient cohorts, and *MarkSIM*, a Markov Chain

Monte Carlo modeling and simulation tool for predicting clinical outcomes through probabilistic analysis that captures granular temporal aspects of aggregated, clinical data.

We assess the efficacy of antibiotic treatment and the optimal time of initiation for in-hospitalized AECOPD patients as an application to probabilistic modeling. We identify 697 AECOPD patients of which 26.0% were administered antibiotics. Our model simulations show a 50% decrease in mortality rate as the number of patients administered antibiotics increase, and an estimated 5.5% mortality rate when antibiotics are initially administrated after 48 hours vs 1.8% when antibiotics are initially administrated between 24 and 48 hours. Our findings suggest that there may be a mortality benefit in initiation of antibiotics early in patients with acute respiratory failure in ICU patients with severe AECOPD.

Thus, we show that it is feasible to enhance representation of EHRs to aggregate patients' entire medical histories with temporal trends and support complex clinical questions to drive clinical guidelines for COPD.

1 INTRODUCTION

The Global Initiative for Chronic Obstructive Lung Disease (GOLD) national clinical practice guidelines [2] have been used worldwide by healthcare professionals for treating and managing chronic obstructive pulmonary disease (COPD). COPD refers to a group of diseases (e.g., chronic bronchitis and emphysema) that affects over 16 million people in the US and is the fourth leading cause of death in the world [2, 3]. GOLD recommends antibiotic therapy for patients with severe acute exacerbation of COPD (AECOPD), a sudden worsening of respiratory symptoms [2, 4], as they can shorten hospital length-of-stay (LOS) and decrease mortality [5]. Though several studies [6] have been conducted to assess the short-term and long-term efficacy of antibiotics for AECOPD, none have explored the effect of the timing of antibiotic administration on mortality for AECOPD patients in the intensive care unit (ICU). Hence, there are no guidelines or recommendations for the initial timing for administering antibiotics. This dissertation presents a framework for using electronic health records (EHRs) to systematically generate evidence for such guidelines.

Acute exacerbations can have a significant impact on health status, potentially leading to deterioration of health and early death [7–9]. AECOPD not only accounts for a majority proportion of the total cost that COPD inflicts on the healthcare system, but are associated with a 6% risk of inpatient mortality [10]. According to the American Lung Association, in 2010, nearly $50 billion was attributed to COPD costs, and of that, nearly $30 billion was spent on direct healthcare costs alone [11]. During AECOPD hospitalizations, physician care decisions are collected in a patients' EHRs and clinical notes. Each visit is a source of patient information (e.g., diagnosis, treatment, outcome) about clinical events that affect their health state during their hospital encounters. While this information is crucial for treating patients, there is often not a standardized method of processing and documenting the information [12].

The current state of COPD treatment generation is a largely manual process. "Currently, [emergency department] physicians must rely largely on their experience and the patient's personal criteria for gauging how an [exacerbation of COPD] will evolve." [13]. Creating treatments based on human knowledge without evidence can introduce bias and inaccuracy into this process. For instance, several investigators found that the benefits of some treatments administered to COPD patients were widely-adopted without evidence, which later resulted in safety concerns [14, 15]. While the highest quality of evidence for clinical guidelines comes from published systematic reviews and meta-analyses, the GOLD Science Committee members, a group of recognized leaders in COPD research and clinical practice, meet twice annually to discuss publications that potentially have an impact on COPD management [2]. During these two, yearly meetings, the committee reaches a consensus on whether to reference the information as support of the current recommendations or modify the guidelines to reflect new findings [2]. While this approach has been widely adopted and valuable, there are potential limitations surrounding the availability and access to new findings and relevant information that is yet to be published during the process of developing and updating the guidelines [16]. Clinical practice can become quickly outdated and it is important to have timely mechanisms for updating guidelines to incorporate new evidence [17]. Thus, there is "a significant gap between up-to-date clinical evidence for best practices, as reflected by the clinical guidelines and actual practice patterns." [18]

To address this issue, longitudinal, clinical data such as EHRs can be used for discovering trends as well as monitoring and tracking patients over time by analyzing treatments (e.g., antibiotics administrations) and outcomes (e.g., mortality) found in patient medical histories [19]. However, representing a patient's medical history coherently is challenging as this information is typically scattered across different clinical databases (DBs) such as pharmacy, ICU, and Emergency medicine, and can be challenging to retrieve, analyze, and properly interpret as a consequence of its high volume and unstructured nature [20,21]. Even more, existing software systems [22,23]

do not allow for granular analysis and longitudinal processes such as chronic disease progression to be observed directly, nor do they support analysis of how patients transition from one clinical condition (e.g., mild to severe exacerbation) to another as a direct cause of an intervention (e.g., drug administration, oxygen therapy, surgery).

While several previous studies have used probabilistic modeling approaches such as Markov chains (or models) to provide support for decision-making under uncertainty and identify medical trends for COPD, directly using EHR data as input for such models is challenging [24–27]. Typically models used for analyzing temporal data assume that data is time-invariant, collected with some fixed sampling frequency [28]. However, EHRs contain highly-dimensional, time-variant data that is observed at irregular time intervals. Hence, the nature of EHR data limits the ability to represent the granular timing of clinical events, which can lead to process misrepresentation in the model. Thus, in addition to transforming temporal aspects of EHR data into input parameters for such models, methods for estimating model parameters must be developed.

This dissertation presents a novel approach that curates clinical data, systematically captures the effect of interventions during medical encounters, and hence, may support evidence generation for clinical guidelines in a systematic and principled way.

An outline of the contributions of this work is as follows:

1. *Provide a framework for large integrated EHR data for access and extraction.* We introduce a conceptual framework that includes a series of components to structure EHR data, enhancing its representation to allow capture of underlying temporal characteristics and injection of clinical domain knowledge. We develop a unified extraction application programming interface (API) for a clinical database, providing flexibility in data extraction and a structured way to formulate and execute complex queries to curate patient histories.

2. *Generate caretrails with temporal trend from EHR data to aggregate patients' medical histories.* A ***care-trail*** is defined as: **a chronological collection of**

events, occurring during a patient's hospital encounter, integrated with clinical domain knowledge. We introduce a tool with a rule-based state coding engine that encodes patient histories using clinical domain knowledge to define health states based on clinical conditions. Our tool also has capabilities for generating multiple caretrails for different patients and organizing patient cohorts for studies and analysis. These caretrails not only introduce an enhanced structure of EHRs but provide a way to aggregate, visualize, and model patients' entire medical histories coupled with clinical domain knowledge.

3. *Develop computational methods to answer clinical questions using patient histories.* We introduce a Markov Chain Monte Carlo modeling and simulation tool that encodes clinical conditions as computable definitions of health states using raw EHR data. Our methods capture exact timing information from patient histories to estimate model parameters as a function of time, by calculating the time between changes in health states. We are the first to use this approach to estimate the efficacy of the initial timing of antibiotics treatment for in-hospitalized AECOPD patients.

1.1 Thesis Statement

Data found in EHRs is heterogeneous and complex, making it difficult to represent it in a way that captures data characteristics, such as temporality, necessary for reproducible granular analysis. Using prior EHRs from several thousands of patients to influence treatment generation could revolutionize many health regimens, including COPD [29]. However, current tools that provide methods for analyzing EHR data require processes that can cause data loss and only allow limited statistical analysis. We hypothesize that:

> *It is feasible to generate evidence for clinical guidelines by enhancing the representation of electronic health records to aggregate patients' entire medical histories and support complex clinical questions.*

According to the Centers for Disease Control and Prevention [30], six in ten adults in the U.S., have a chronic disease and four in ten have two or or more. These diseases are the leading causes of death and disability, driving 3.3 trillion in annual health care costs [30]. While this dissertation is presented as a case study in using computational analysis of EHRs to improve the GOLD best practices for COPD, we believe the approach taken, methodology, and tools are applicable to a wide range of chronic conditions such as heart disease, cancer, stroke, Alzheimer's diabetes, and chronic kidney disease [30].

1.2 Dissertation Order

The remainder of this dissertation is divided into six chapters as illustrated in Figure 1.1.

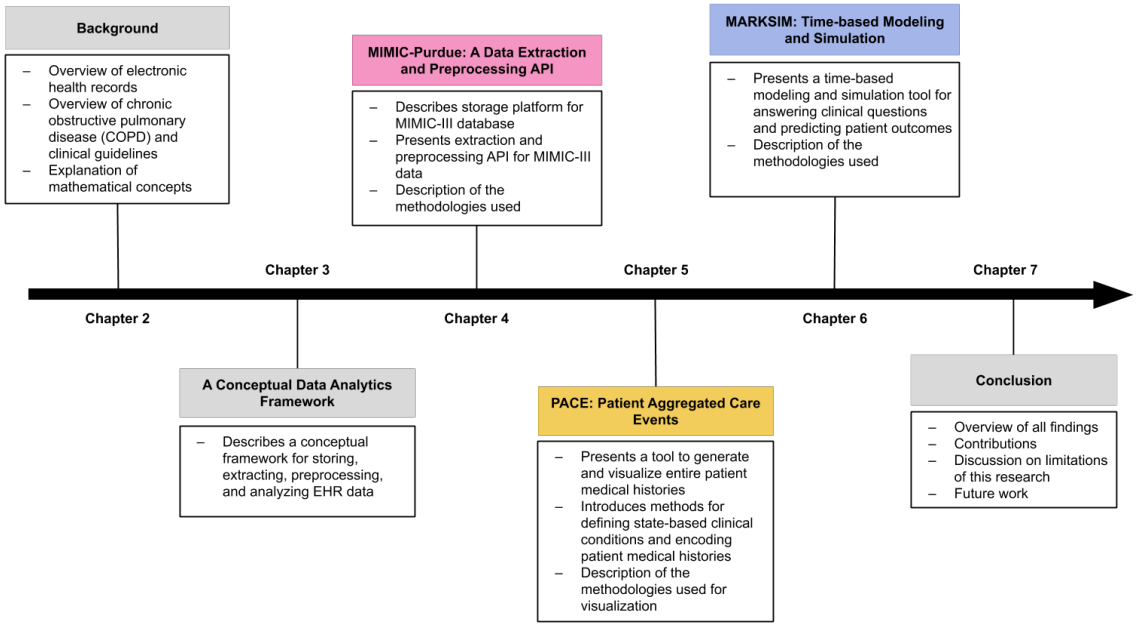


Figure 1.1.: Dissertation order.

2 BACKGROUND

In this chapter, we review the available literature regarding clinical guidelines for COPD, benefits and challenges of data analysis with electronic health records, and introduce some mathematical concepts for modeling.

2.1 Chronic Obstructive Pulmonary Disease (COPD)

2.1.1 Definition and Overview

Chronic Obstructive Pulmonary Disease (COPD), the fourth leading cause of death in the world, affects an estimated 65 million people worldwide and is projected to be the 3rd leading cause of death by 2020 [2, 31, 32]. The Global Initiative for Chronic Obstructive Lung Disease (GOLD) clinical guidelines [2] define COPD,

> "a common, preventable and treatable disease that is characterized by persistent respiratory symptoms and airflow limitation that is due to airway and/or alveolar abnormalities usually caused by significant exposure to noxious particles or gases."

There are two main forms of COPD: chronic bronchitis, a long-term cough with mucus, and emphysema, irreversible damage to the lungs over time [9]. The slow progression of the disease causes the airways of the lungs to be inflamed and become obstructed (i.e., blocked) [9]. All forms of COPD are caused by exposure to air pollution, cigarette smoking, or a rarely inherited, alpha 1-antitrypsin deficiency [2]. Spirometry, a test to assess lung function, is used to measure airflow limitation, which is required to make the diagnosis in this clinical context [2]. The severity of airflow limitation is based on forced expiratory volume (FEV_1), the amount of air a person can exhale during a forced breath. There are four stages, shown in Figure 2.1, that

classify the severity of COPD. Though there is no cure for the life-long, terminal

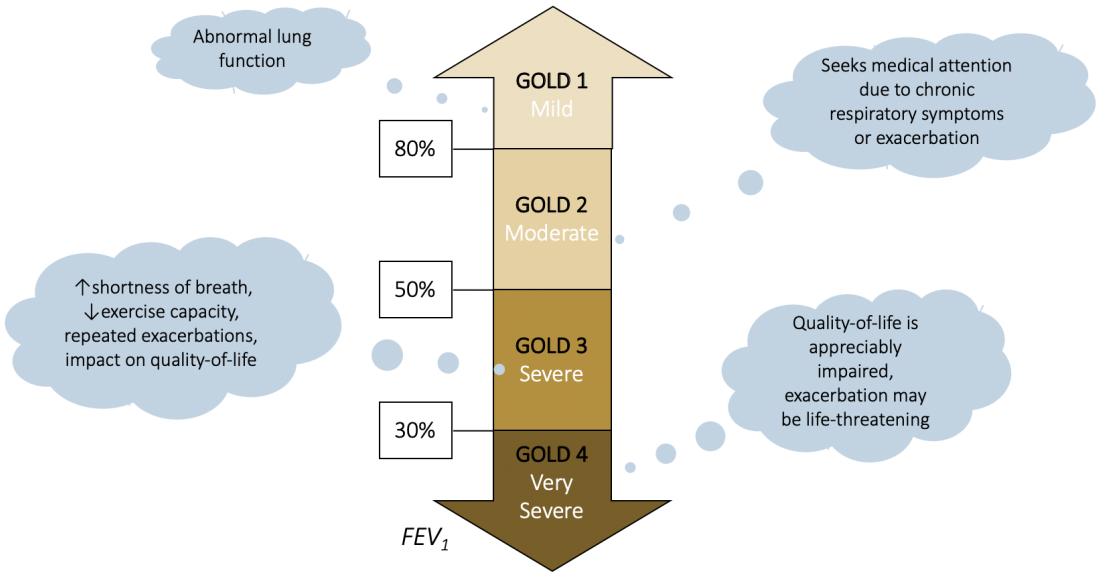


Figure 2.1.: GOLD classifications [2].

disease, healthcare costs continue to rise as a result of managing and treating the disease. According to the American Lung Association, in 2010, nearly 50 billion dollars was attributed to COPD costs, and of that, nearly 30 billion dollars was spent on direct healthcare costs alone [11]. Acute exacerbations of COPD (AECOPD), a sudden worsening of symptoms, account for a majority of the total cost that COPD inflicts on the healthcare system [7–9]. Even more, AECOPD has a significant impact on health status, potentially leading to deterioration of health and early death [7–9].

2.1.2 Clinical Guidelines

Clinical practice guidelines are systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances. [33] Clinical practice guidelines are one the foundations to improving care

for patients and providing a collective effort for improving healthcare [34]. They are typically developed through either evaluation of the "best" available evidence, which is measured by the level of evidence of the study [35]. That is, there is a hierarchy of evidence assigned to studies based on the method, design, validity, and applicability to patient care [36]. Evidence obtained from systematic reviews or meta-analyses of randomized controlled trials are considered the highest form of evidence [37, 38].

Development

Several guidelines, including North American and European [39, 40], have been developed and published regarding COPD management. Among these, GOLD, is the most widely distributed international COPD guidelines [2]. The GOLD program was formed in 1998 to increase awareness regarding the management and prevention of COPD as well as to produce recommendations for management based on scientific information available during that time. In 2001, the GOLD organization released its first consensus report (i.e., guidelines) entitled, *Global Strategy for the Diagnosis, Management, and Prevention of COPD* [2]. Since then, complete revisions, based on public research, have been published in 2006, 2011, and 2017. The GOLD Science Committee members, a group of recognized leaders in COPD research and clinical practice, review and evaluate published research on COPD management and prevention to produce the GOLD report. The published information is found by searching for keywords such as COPD and clinical trial via PubMed, a search engine maintained by The United States National Library of Medicine [2]. Such literature typically derives from clinical studies that are intended to answer clinical questions regarding risk factors diagnostics and prognostic criteria for conditions and treatments [17, 34]. These clinical questions are typically well-formulated, using methods such as PICO (patient/population/problem, intervention, comparison, and outcome), and target specific patient groups [41, 42]. Each publication that is considered to potentially have an impact on COPD management is discussed by the committee during two,

yearly meetings. The committee reaches a consensus on whether to reference the information as support of the current recommendations or modify the report to reflect new findings [2].

The GOLD document is a global document [43] that serves as a basis for expansion in both developed and under-developed countries and on a local scale within countries. Therefore, the recommendations, assessments, and principles provided by GOLD can be tailored to a particular setting. The University of Michigan Health System developed clinical care guidelines for hospitalized patients with AECOPD using material from the GOLD guidelines as a reference point for comparison and support throughout their publication [44]. When Wexner Medical Center at The Ohio State University [45] created their guidelines entitled, *Inpatient Management of Chronic Obstructive Pulmonary Disease (COPD) Exacerbations*, they also adopted recommendations from GOLD. In addition to referencing information provided by GOLD, these organizations conduct a similar process as GOLD by reviewing published literature and agreeing on a final report.

On a global and local scale, developers of clinical guidelines consist of clinical experts who form committees to review the most current evidence found in published literature, and make decisions regarding the best practices that should be included in the guidelines [17]. While this approach has been widely adopted and valuable, there are challenges. Specifically, there are limitations surrounding the availability and access to new findings and relevant information that is yet to be published during the process of developing and releasing the guidelines [16]. This presents motivation to develop and implement new methods for evaluating and discovering evidence during the guideline creation process. Clinical practice become quickly outdated, and there are no timely mechanisms to update guidelines and incorporate new evidence [17]. Thus, there is "a significant gap between up-to-date clinical evidence for best practices, as reflected by the clinical guidelines and actual practice patterns." [18]

Recommendations

Though not standardized, GOLD recommendations are used in clinical practice to help select treatments for COPD patients. The guidelines are updated from year-to-year to include or exclude these recommendations based on the best available evidence. For example, earlier versions of the GOLD guidelines used airflow limitation as a measure to assess the severity of COPD and further make treatment selections accordingly. Later, studies found that FEV_1 alone is not sufficient for determining therapeutic options on an individual level. This led to the development and enhancement of the *ABCD* assessment tool, which incorporates a comprehensive assessment of symptoms as well as other clinical parameters, risk of exacerbations, and lung function. GOLD states that a management approach should match assessment to treatment objectives such that it *"can be used in any clinical setting anywhere in the world and moves COPD treatment towards individualized medicine"* [2]. The assessment tool (see Appendix A.1), assigns a grade (i.e., A, B, C, D) to patients by separating airflow limitation from clinical parameters; thereby encouraging better treatment selection that reflects parameters that are influencing a patient's symptoms at any given time [2]. Based on information used in the ABCD assessment, the document includes pharmacological treatment algorithms and pathways (Figure A.2) for each GOLD Grade. However, the treatments associated with each assessment group, defined by the ABCD assessment tool, are for patients with stable COPD, implying that patients experiencing severe episodes (e.g., AECOPD) that lead to hospitalization are not considered. This may be attributed to the complexity of COPD and various impacting factors such as comorbidities, lifestyle, and environment.

While GOLD does not mention the use of the tool for non-stable COPD, utilizing the ABCD methodology and other GOLD recommendations can be a starting point for developing a methodology to identify treatment selection tools for AECOPD patients. AECOPD can negatively impact the overall health status of patients, the frequency of hospitalization and readmission, and disease progression [2]. Thus, the

main goal in treating these episodes is to prevent and minimize the negative impact that they introduce, as they can lead to health deterioration and early death. Corticosteroids, antibiotics, and oxygen therapy are all forms of treatment options used for exacerbations [13, 46]. COPD exacerbations are complex events that can be triggered and/or amplified by respiratory viral infections and environmental factors. GOLD recommends the following classifications exacerbations based on the additional therapy needed as follows [2]:

Mild – treated with short acting bronchodilators (SABDs) only

Moderate – treated with SABDs plus antibiotics and/or oral corticosteroids

Severe – patient requires hospitalization or visits to the emergency room

These classifications potentially lack sufficient information and recommendations regarding the course of action to be taken before or during the exacerbation. Exacerbations are heterogeneous in that determining cause is difficult to identify, and GOLD suggests that severe exacerbations be based on the patients clinical signs [2]. The guidelines state that severe exacerbations may be associated with acute respiratory failure (ARF), which is a buildup of fluid in the air sacs of the lungs that inhibits the release of oxygen into the blood [2,4,47]. Appendix A.1 contains the classifications of exacerbations for hospitalized AECOPD. These classifications provide more details about the condition of the patient and the clinical signs to measure for assessing the severity of the exacerbation.

2.2 Electronic Health Records (EHRs)

While medicine has been practiced for thousands of years, dating back at least to 2000BC [1], technology's application to medicine is relatively new. Despite its relatively short time in use, dating to the 1500s [2], technology has impacted medical practice in various ways. One area where technology has advanced clinical practice and research deals with the collection of data. Data collection healthcare is defined as:

[1] https://blogs.uoregon.edu/hgoldenw14gateway/timeline/

[2] https://www.infoplease.com/math-science/health/medical-advances-timeline

> "the on-going, systematic assembling and measuring of information, analysis and illustration of health data necessary for integration, implementing, designing, and evaluating public health prevention [programs], which then enables one to answer relevant questions and evaluate outcomes".

From data collection through wearables to data storage on servers to data analysis on state-of-the-art processing machines, the impact of technology has increased the diversity, availability, and volume of healthcare data such as genomic, sensor, public health, and electronic health record. Such data is made available through various collection mechanisms and for a range of purposes, creating a paradigm shift for data-driven, evidence-based analytics and discoveries in the healthcare industry. Genomic data, the genome and DNA sequences of an organism, is used to discover and analyze genome structures and other genomic parameters [48]. Public health data derives from monitoring population health via national surveys and reports from clinical studies, claims, and costs. Sensor and behavior data is collected via social media networks and wearable devices (e.g., fitness trackers, medical devices, and smart-watches) that have enabling technologies such as sensors that capture bodily or environmental impulses and transmitters that send data for analysis [49]. Clinical data found in electronic health records is a critical component of healthcare data and the building block for healthcare data digitization [50].

Dating back to the 1990s, EHRs, digital, comprehensive, and longitudinal collection of a patient's healthcare data [52], have evolved to become a valuable source of information to a variety of stakeholders (e.g., insurance companies, researchers, medical professionals). EHR data is collected directly from the patient at the time of care via medical monitoring devices and medical professionals (e.g., clinicians, nurses). They are comprised of patient information such as demographics, laboratory results, vital sign measurements, diagnosis and procedure codes.

EHR data was, which once consisted of handwritten and typed reports, is primarily designed for internal use in medical settings as well as medical billing purposes (e.g., insurance claims) [52]. However, the digitization of healthcare data has al-

Table 2.1.: Types of healthcare data and sources (Adapted from [51]).

Type	**Description**		**Source**
Clinical	EHRs	Detailed patient-related information (physician prescriptions, medications, medical history)	Hospitals & Clinics
	Diagnostic	Diagnostic Results (imaging results, laboratory reports)	Laboratories, Radiology Departments
	Biomarkers	Molecular data (genomic, proteomic, transcriptomic, metabolomic)	Diagnostic Companies
	Ancillary	Administrative data (admission, discharge, transfer) & financial data (claims)	Hospitals & Clinics
Claims	Medical Claims	Medical reimbursement data (procedures, hospital stay, insurance policy details)	Payers
	Prescription Claims	Prescription reimbursement data (drugs, dose, duration)	Payers
Clinical Research	Clinical Trails	Design parameters (compound, size, end points)	Pharma Companies, Medical Journals
Patient-generated Data	Social Media	Community discussions	Web Health Portals, Social Media Websites
	Wearables & Sensors	Wellness & lifestyle data (smartphones, fitness monitors)	Device Data Systems

lowed for secondary use of EHRs in clinical analysis and research [19]. The American Medical Informatics Association (AMIA) states,

> "Secondary use of health data can enhance healthcare experiences for individuals, expand knowledge about disease and appropriate treatments, strengthen understanding about the effectiveness and efficiency of our healthcare systems, support public health and security goals, and aid businesses in meeting the needs of their customers." [53].

On a population scale, EHRs assist in better understanding of disease progression and patient trajectories as well as informed decision-making at the point of care [51]. In recent years, several institutions such as the National Patient-Centered Clinical Research Network (PCORnet) [54], Informatics for Integrating Biology and the Bedside (i2b2) [55], and the Observational Health Data Science and Informatics (OHDSI) [22] have created clinical data repositories in conjunction with analytic, cohort identification, and data sharing tools for advancing clinical research as well. Additionally, clinical datasets such as Medical Information Mart for Intensive Care III (MIMIC-III) have been made available for clinical research. While these data sources and others include large amounts of granular clinical information that have be useful for pushing clinical research forward, researchers are tasked with overcoming challenges that effect the utilization of EHRs. [56] states, "Using [EHR] data for research is fundamentally different from using prospectively collected data, as has historically been done in randomized controlled clinical trials." In this section, we discuss access, storage and preprocessing, and data sources regarding analyzing EHR data.

2.2.1 Data Access

The number of healthcare data sources continues to increase as innovative technologies surface. This leads to the assumption that the availability of such data may be abundant, allowing for simpler access. While this assumption may hold for data obtained from sources such as public health initiatives and social network crowd-

sourcing, one of the largest barriers for utilizing EHR data is its inaccessibility to researchers [57]. Challenges regarding access to patient medical data are ongoing for clinical researchers despite efforts [58] that promote and address issues across the healthcare domain. We focus on three main areas of concern in relation to data access: patient privacy, data sharing, and interoperability.

Privacy Clinical data reveals unique, personal, sensitive, and critical information about an individual. While such information can be used to gain clinical insight regarding a patient, the misuse of such data, leading to violation of a patient's privacy, is an abuse from both an ethical perspective and a legal standpoint. The importance of patient privacy was realized in 1951 when Henrietta Lacks' medical information not only became public but was used for research unbeknownst to her and her family [59]. Though Ms. Lacks' data resulted in one of the most important discoveries in medical research history, the HeLa cell line [59], the misuse by researchers caused great controversy and inflicted devastation on the Lacks family. This later lead to the creation and modification of local and federal privacy laws such as the Health Insurance Portability and Accountability Act (HIPAA) [60]. HIPAA protects patients and their medical information by defining protocols for accessing and handling medical data, mainly a de-identification process that requires the removal and/or generalization of protected health information such as name, address, phone number, and dates. The de-identification process preserves the rights and privacy of patients [61] while providing datasets for analysis and evaluation. In addition to HIPAA requirements, clinical research involving human subjects may be subject to an approval from an institutional review board (IRB), a committee that ensures proposed research methods are ethical [62]. While the IRB process is designed to protect subjects as well as add structure to the clinical research pipeline, IRB review and approval is time-consuming and burdensome and can impede or even prevent research [62].

Data Sharing Violation of privacy laws (i.e., HIPAA) can result in fines, penalties, and felony offenses [63], which are consequences that could discourage clinical data

providers from sharing patient data out of fear of litigation [64]. While laws have been established to protect the privacy of patient data, data accessibility is impacted by other factors such as data governance: the process by which responsibilities of collecting and securing information while also getting value from that information [65]. EHR data is usually collected for a specific purpose and, traditionally, not intended nor expected to be used beyond the purposes in-which it is collected. EHRs can be viewed as professional medical opinions that reflect the clinicians and institutions that interact with patients. This information is collected and stored on systems that are in the possession of the care providers, who in essence, have access and control over patients' health data, making them the primary stewards of EHRs [66]. Data holders are more-likely to manage this data in ways that are most conducive to their needs, creating data silos. Furthermore, providers may not want to share the data or have the resources in place to do so. Even after patient data is de-identified and data owners are willing to give access to researchers, data sharing in the absence of interoperability is an issue.

Interoperability Lack of interoperability, caused by the lack of technological compatibility and lack of standardize coding, has plagued the healthcare system for years [67]. In 2009, President Barack Obama signed the American Recovery and Reinvestment Act, an initiative supporting the development of a national system of EHRs, digitizing all patient health records, in hopes to achieve interoperability across care settings and promote the meaningful use and value in EHR development [68]. Despite the development of this initiative, healthcare systems that are tailored to fit the needs of individual organizations continue to be developed, limiting the ability for EHR systems to exchange information among providers. Further, patients may have multiple providers, leaving their medical information scattered across organizations. Initiatives, such as the Standard Health Record [69], are intended to solve interoperability by standardizing EHR data, creating a unifying template across the medical field that contains all of a patient's health data from multiple clinicians [69].

While such a standardized template is befitting, it requires a restructuring of an entire domain of systems that have been adapted by medical institutions and created specifically to fit their needs.

Researchers have made several attempts to solve data sharing issues using more technical approaches such as blockchain. For instance, MedRec uses a blockchain solution to allow distributed data collection and access to address data fragmentation, slow access to medical data, and system interoperability [66]. Other researchers [50, 70] have also adopted blockchain technology to promote distributed management and access of patient data. Such solutions help to avoid data silos and data cemeteries [71] as well as target challenges relating to data access and availability. On a wider scale, data warehouses and clinical research networks have been developed to efficiently integrate data from various sources and provide a platform for sharing data as well as reproducible research. Data warehouses are recent trends in healthcare data analytics and gathering, which consist of repositories of information from clinical and research records, usually with integration to query de-identified data [72]. Platforms such as the i2b2 [55] facilitate data access by linking multiple data sources together in one place. The main goal is to have a common data representation that encompasses standards across the healthcare domain [72], strengthening the link between a network of providers, researchers, and other stakeholders.

2.2.2 Data Storage and Structure

EHR systems are primarily designed for "routine clinical care," and in-turn, the wealth of information that they store is not in "readily minable formats" [73]. Issues related to data structure and representation are common hurdles in healthcare research using EHRs. Many of these issues are attributed to lack of EHR data standardization. One attempt at standardization is the use of common coding terminologies within EHR for billing purposes. Various clinical and administrative terminologies such as International Classification of Diseases, Ninth Revision, Clinical Modification

(ICD-9), Current Procedural Terminology (CPT), RxNorm, and National Drug Codes (NDC) have been adopted across EHR systems for clinical diagnoses, procedures, and drugs representation. These different coding terminologies provide flexibility to practitioners as they generate EHR data for patients, but introduce challenges to analysts as they must identify and learn coding terminologies across datasets. Moreover, current medical systems differ in their acceptance and support of international standards, protocols, and formats and semantics [74]. Because there is no standard data representation of EHRs across medical systems [75], generating broad data models that can be widely applied across medical systems remains a challenge.

Traditionally, DB technologies [76] have been used to handle various forms of clinical data for storage, accessibility, and retrieval. DBs provide direct access to data and dynamic extraction of specific data entities. Some DB technologies are optimized to handle more structured data while others are built to handle both structured and unstructured data. Relational DBs such as MySQL, Oracle DB, and Microsoft SQL Server were developed with a corresponding query language model, structured query language (SQL), to store structured data, following a relational model. The data is organized as tables that include rows for representing records and columns for representing attributes of records [77, 78]. These relational databases are often structured and work with a well-defined schema, a physical implementation of a data model. DB schemas include details such as data types, constraints, foreign or primary keys [76].

As data continued to grow and change, a newer DB technology, called Non-relational or NoSQL, was developed to support both structured and unstructured data as well as scalability [79]. NoSQL DBs differ from relational DBs as they do not follow a relational model nor a fixed schema [80]. This is an important difference because NoSQL is a form of unstructured storage, allowing a simpler, more flexible structure [80]. NoSQL databases store types include document, column, key-value, and graph, where each value in the DB usually has a key [80]. Both schema or schema-less databases define how data is stored for specific database technologies.

Data representation and functionality of these database technologies are limited by the features available in specific DB platforms. Because of complexity issues with representing complex relationships, object relational databases such as MySQL and PostgreSQL (or Postgres) were developed [81].

Once data extracted from EHR systems is made available to researchers, they must perform costly and time-intensive processes such as restructuring of the data before it is analyzed [82]. Aside from current traditional database schemas, there exists data models. Current approaches used to rectify issues surrounding data structure include attempts to standardize data terminology through the development of data models [83,84]. Data models, similar to database schemas, describe how data is stored. However, unlike a DB schema, a data model is not specific to an implementation but rather is a data design [76]. Data models are general organization/architecture of data elements and their relationships and data schemas are the specific implementation of a data model in a particular database management system. These models are developed to accommodate healthcare data from disparate data sources such as administrative claims, EHRs, longitudinal surveys, and registries [85]. They determine the structure of data and describe data organization regardless of how the data is represented in the underlying system [76]. Organizations such as i2b2 [23] and OHDSI [22] attempt to standardize representation and define structure by developing common data models (CDMs), a way of organizing data into a standard structure [56].

i2b2 developed a core data model based on an Entity-Attribute Value (EAV) star schema to provide a common structure to all of the data sources it houses from heterogeneous sources (e.g., clinical trials, EHR systems, and other clinical data systems) and allow data to be aggregated and optimized efficiently. i2b2's star schema data model consists of five tables. At the center of this model is a fact table that represents a patient object. Each row of the fact table represents a single observation about a patient. Facts, which are quantitative or factual data, include diagnosis, procedure, lab data, demographics, health history, genetic data, and provider data. The remaining four tables, called dimension tables, represent attributes such as provider numbers,

concept codes, start and end dates, and other parameters, regarding a patient object. Dimensions are groups of hierarchies and descriptors that define facts [23]. The remaining tables are dimension tables, which contain descriptive information about facts. i2b2 requires standardization across its system. Hence, the star schema, which maps concepts and other clinical information. The extracted data must be transformed into an i2b2 compatible star schema [86,87]. i2b2's star schema data model is instantiated similar to a single relational model and is considered to be the simplest style of schema for a data warehouse [23].

OHDSI's Observational Medical Outcomes Partnership (OMOP) CDM was developed to provide consistency and standardization for healthcare data from heterogeneous sources. Like i2b2, OHDSI achieves this by requiring data to be transformed from varying sources into a database with a common format or model with common representations (e.g., terms, vocabulary, coding schemes) [88].

2.2.3 Data Preprocessing

EHR data is optimized to support activities related to billing and reviewing clinical observations as opposed to research [89]. Understanding the context in which data is collected is important for leveraging EHRs for data analysis. Data is routinely collected from various departments within clinical settings, categorizing data according to the clinical and administrative activities performed. [89] *et al.*, categorized common issues with raw clinical data (Table 2.2). Such issues affect the quality of clinical data used for secondary purposes and requires data to be preprocessed. Data preprocessing includes a series of steps that can be an iterative and repetitive process to properly organize the data for analysis. These steps include data cleaning, integration, transformation, and reduction. Specifically, data cleaning and transformation are important for data analysis using EHR data.

Data cleaning involves removing erroneous data, handling missing, noisy, and duplicate data [90]. These issues can be introduced at the point of data entry or the

point of conversion when prepared for secondary use [91]. Missing data is handled by ignoring the record, manual completion, and filling with predictive values [91]. Noisy data refers to random error in an observed variable (e.g., abnormal values differing from expected baseline), which is a common problem EHR data [91]. EHR data that contains duplicate, erroneous, or missing data can produce misleading results when used as input analysis [90]. While there are several tools to automate the data cleaning process, correcting all possible data inconsistencies and errors would have to occur during data entry or extraction [72, 91].

Data transformation is intended to represent the data in a format that is suitable for analysis [91]. Processes such as data aggregation involve combining values of the same data attribute or simply gathering all information pertaining to a patient. Data transformation is a common approach required by data warehouses. They require a time-consuming process called extraction, transform, and load (ETL) to convert data into a desired syntactic and semantic standard, usually defined as a platform-specific CDM [72, 90]. This process is often not automated and requires additional steps to assure the underlying data has been processed properly. While preprocessing steps such as data cleaning is considered to be a common problem in data warehousing, additional preprocessing for transforming the data to fit CDMs introduces other challenges such as data loss.

2.2.4 Data Analysis

To evaluate data analytics for longitudinal EHRs in clinical research, we define three types of healthcare data analytics models: (i) descriptive, (ii) predictive, (iii) prescriptive. Each analytic model is designed to address questions such as "What happened?", "What could happen?", or "What should be done?" Techniques such as computational modeling, text mining, natural language processing (NLP), and visual based processing of data are used to generate these three types of data analytic models and derive insights from data [92]. The data analytics pipeline can be viewed

Table 2.2.: Overview of common categories of hospital data and common issues to consider during analysis (Adapted from [89]).

Category	**Examples**	**Common issues to consider**
Demographics	Age, gender, ethnicity, height, weight	Highly sensitive data requiring careful de-identification. Data quality in fields such as ethnicity may be poor
Laboratory	Creatinine, lactate, white blood cell count, microbiology results	Often no measure of sample quality. Methods and reagents used in tests may vary between units and across time
Radiographic images and reports	X-rays, computed tomography scans, echocardiograms	Protected health information, such as names, may be written on slides. Templates used to generate reports may influence content
Physiologic data	Vital signs, electrocardiography waveforms, electroencephalography waveforms	Data may be pre-processed by proprietary algorithms. Labels may be inaccurate (for example, fingerstick glucose measurements may be made with venous blood)
Medication	Prescriptions, dose, timing	May list medications that were ordered but not given. Time stamps may describe point of order not administration
Diagnosis and procedural codes	ICD-9 codes, CPT codes, Diagnosis Related Groups codes	Often based on a retrospective review of notes and not intended to indicate a patients medical status. Subject to coder biases. Limited by suitability of codes
Caregiver and procedural notes	Admission notes, daily progress notes, discharge summaries, Operative reports	Typographical errors. Context is important (for example, diseases may appear in discussion of family history). Abbreviations and acronyms are common

as an iterative, not sequential process that encompasses a range of steps to answer clinical questions and delve deeper into informed clinical decision-making. For the remainder of this subsection, we discuss the aforementioned types of data analytics models further.

Descriptive Descriptive healthcare analytic models are considered the most commonly used type of analytics [93], providing insight into the past. Descriptive analytics helps answer questions such as: "How many patients were administered antibiotics? How many patients over the age of 65 were treated?" Descriptive analytics, which emphasizes the use of the underlying data, rather than information, was introduced to describe data without complex calculations, allowing reporting of simple statistics of data. Healthcare professionals and researchers use descriptive data analytics as a starting point to understand past and current healthcare decisions. Descriptive data models are structured by categorizing, classifying, characterizing, aggregating, and converting data to analyze what actually happened in the form of summaries [93]. They summarize raw data for human interpretability such as a count or aggregate that can be input to basic mathematical formulas (e.g., summations, averages). These summaries are usually presented as charts, reports, or visualizations that illustrate patient outcomes, characteristics of cohort study participants, healthcare costs, etc. Specifically, visualizations are used often, as it allows data to be analyzed in a graphical format by constructing graphs, histograms, etc. to identify and explore trends.

Predictive While predictive analytics is more advanced than descriptive analytics in that it emphasizes the use of information versus the data, descriptive analytics can be used as a building block for predictive analytics. Predictive analytics target "understanding the future," determining what might happen given summarized and historical data. Simulation and modeling techniques are the most common approaches in predictive data analytics, where predictions are made by estimating the likelihood of a future outcome with some certainty. [93]. For instance, questions such as "What

is the efficacy of oral antibiotics in patients with lung diseases?" and "Which patients are most likely to survive a heart transplant after receiving mechanical ventilation?" can be answered through predictive analytics. One benefit to developing predictive analytics models with EHRs is that predictive algorithms attempt to optimize the use of available data by filling in missing data. This is especially useful in EHRs, as they contain longitudinal time-series data points that are susceptible to error and missing information. The temporal dynamics of EHRs can be exploited using predictive models such as Markov Chains and Gaussian processes [94].

Prescriptive Prescriptive analytics uses medical knowledge and expertise in addition to data and information to determine what should be done [93]. For example, the choice of antibiotic versus steroids during a COPD exacerbation may be determined based on the severity of the exacerbation and the treatment that maximizes the best outcome. Given a situation, prescriptive analytics target optimizing possible courses of actions, going beyond descriptive and predictive analytics [93].

2.2.5 Data Sources

Clinical DBs contain routinely collected for administrative healthcare data or data collected specifically to assess particular clinical outcomes [95]. EHR, administrative, claims, clinical trial, and health surveys are type of data found in clinical DBs that are made available both privately and publicly [95]. Private DBs [96] are obtained through private clinical networks and medical institutions. Such DBs can be difficult to access, as they are legally bound to the entities which holds the data. Some DBs offer data to researchers through partnerships and collaborations that align with their institutional goals. There are a limited number of publicly available resources for EHR clinical DBs that contain information for diverse patient populations [97], and even less than with a range of granular clinical information. Cerner's APACHE Outcomes DB includes data from roughly 150,000 ICU stays, but lacks physician notes, waveform data, and complete physiological and lab measurements [97]. Phillips

eICU DB [98] is populated with data from over 160,000 patients who were admitted to critical care units in 2014 and 2015. eICU also does not contain complete physician notes, and is only made available to selected researchers by submitting a proposal to the eICU Research Institute [97]. The MIMIC-III database is the only publicly and freely available critical care database of its kind.

MIMIC-III is a nationally recognized relational DB curated by the MIT Laboratory for Computational Physiology from Beth Israel Deaconess Medical center [99]. The MIMIC-III (or MIMIC) database contains both high resolution waveform data as well as clinical information on Intensive Care Unit (ICU) patients, admitted between 2001 and 2012. MIMIC-III is publicly available and comprised of deidentified data for roughly 46,520 distinct patients and 58,976 hospital admissions for patients in critical care units [99]. The database includes a range of detailed, granular patient-level, medical information such as time-stamped, physiological measurements, laboratory tests, and demographics (See Figure 2.2).

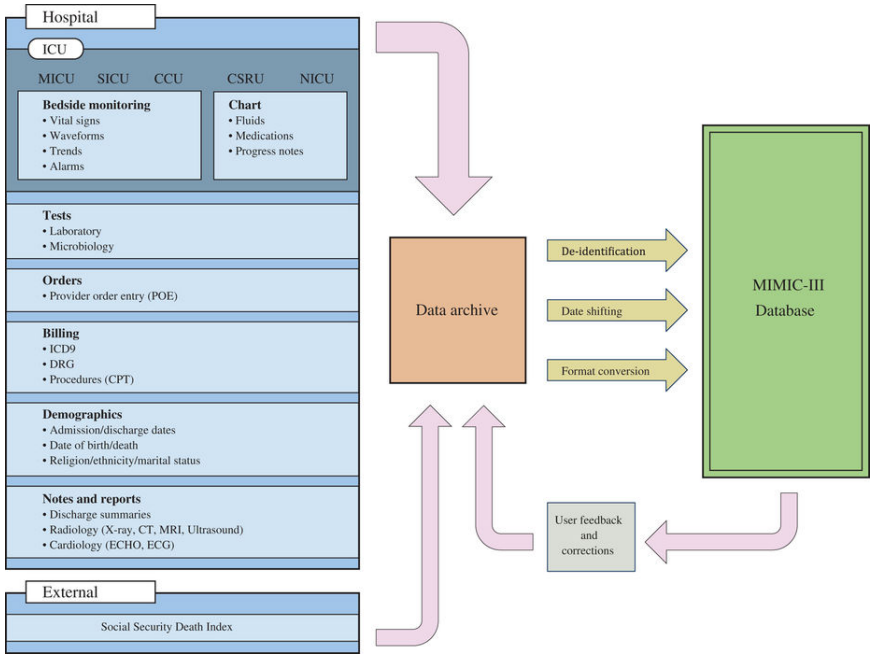


Figure 2.2.: Overview of the MIMIC-III clinical database [99].

MIMIC-III consists of 26 tables (Figure 2.3) that are linked by identifiers using a suffix 'ID' [99]. Specifically, *subject_id* and *hadm_id* refer to a unique patient and hospital admission, respectively. Charted events such as notes, lab results, and vital signs are stored in several 'events' tables that include measurement values as well as time-stamps of when the event or measurement was recorded [99]. Other tables include demographics information such as date of birth, date of death (if the patient died), admission and discharge time-stamps, prescriptions, and dictionaries to identify concepts. Tables for medical information such as ICD-9 and CPT are also included to identify diagnoses and procedures that were recorded during hospital stays.

The MIMIC-III database is HIPAA compliant, and IRB determined that individual patient consent be waved [57]. Prior to incorporating data into the MIMIC-III database, it was first deidentified in accordance with HIPAA standards through a process of structured data cleansing and date shifting, removing all identifying data elements such as patient name, address, and dates [99]. Time of day, day of the week, and dates related to clinical observations were shifted to preserve intervals regarding patient stays. MIMIC-III access is granted to researchers after signing a data user agreement and completing a course in protecting human research participants. This process, along with a relaxed IRB requirement allows researchers unrestricted analysis to use the MIMIC-III dataset, bridging the gap between research studies and access to real-world, clinical data. Access to MIMIC-III data is originally provided as a collection of comma separated value (CSV) files that can be used to import into databases systems such as PostgreSQL and BigQuery to create new instances of the MIMIC-III DB.

The MIMIC Code Repository is an online, open source repository comprised of standardized scripts in languages including Structured Query Language (SQL), Python, and R [101]. This repository is shared and actively contributed to by the research community (i.e. individuals who have been granted access to the MIMIC-III database).

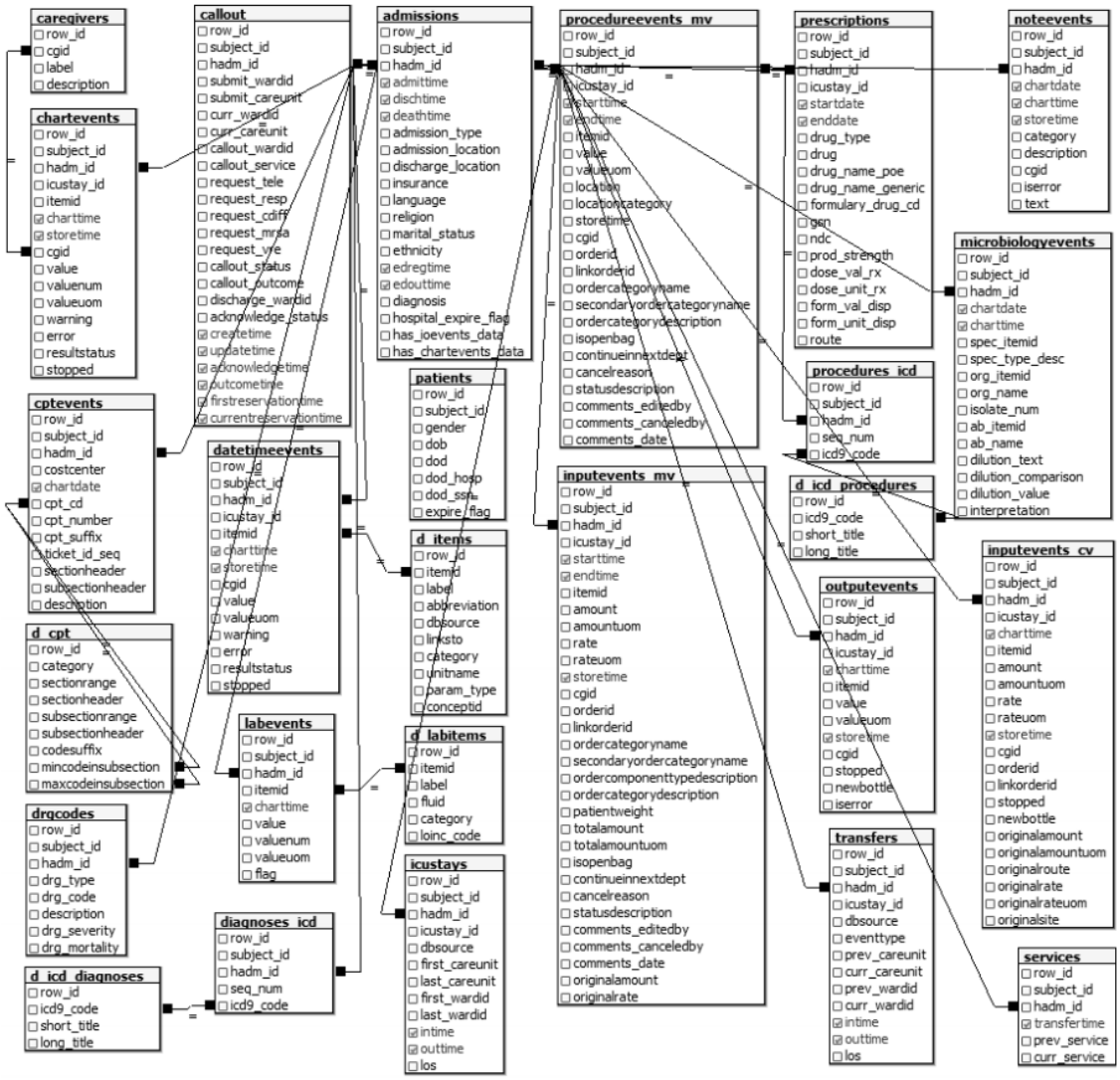


Figure 2.3.: The MIMIC-III reconstruction (PostgreSQL) [100].

Properties of Clinical Data Sources

EHRs also have data characteristics issues that make analyzing EHRs difficult. To show how these issues manifest in a collection of medical data, MIMIC-III is used [99] to illustrate challenges in analyzing the longitudinal representation of EHR data. The MIMIC-III database has limitations such as redundancy across tables, sparsity, heterogeneity, and temporality, making it useful for the context of this research.

Challenges effecting large-scale analysis of EHR data that are exemplified in the MIMIC-III database include:

1. **Heterogeneity** EHRs contain many types of data, having unique characteristics; thereby, making it difficult to analyze such heterogeneous data. They are multi-dimensional in that they contain transactions, time-series data, and unstructured text. These dimensions have data types such as date/time, categorical (e.g., demographic attributes), and units of measurements.

2. **Sparsity** Data sparsity is a common issue in longitudinal data analysis and representation. Data may be missing or erroneous as a result of human error or machine failure. For example, there are clinical variables in MIMIC-III that are measured continuously such as vital signs, while others are measure infrequently such as laboratory tests. A patient's blood pressure may be measured continuously for several hours during a hospital stay. This measurement could be recorded in their EHR both manually by a care provider or via a medical monitoring device. However, because of error or missing information, the patient's blood pressure measurement may be unsuitable for use. This can be caused by errors in the measuring process or the fact that the patient's blood pressure was no longer measured.

3. **Redundancy** EHR data, such as the information collected in the ICU, contains clinical data that can also introduce a considerable amount of redundancy [102]. The same observations with varying and/or duplicate values may be repeated during a patient's medical encounter. For instance, MIMIC-III contains information regarding drugs administered to a patient during their ICU visit in several different tables: `PRESCRIPTIONS`, `INPUTEVENTS_MV`, and `LABEVENTS`. Each of these tables contain time-stamps, drug names, dosage amounts, and other information. As shown in Figure 2.4, these drug observations are associated with multiple unique identifiers and value sets in each table. Also, to gather this data, several tables were accessed to obtain this information. This suggests

that EHR data can be duplicated across multiple tables, making it difficult to reconstruct a single medical encounter.

PRESCRIPTIONS					LABEVENTS						INPUTEVENTS					
drug	startdate	enddate	dose_val_rx	dose_unit_rx	itemid	label	charttime	value	valuenum	valueuom	itemid	label	starttime	endtime	amount	amountuom
Vancomycin	12/11/79 0:00	12/12/79 0:00	1000	mg												
Vancomycin	12/11/79 0:00	12/12/79 0:00	1000	mg							225798	Vancomycin	12/11/79 1:20	12/11/79 1:21	1	dose
					51009	VANCOMYCIN	12/11/79 3:44	43.9	43.9	ug/mL						
Vancomycin	12/12/79 0:00	12/17/79 0:00	1000	mg												
					51009	VANCOMYCIN	12/12/79 3:29	20.6	20.6	ug/mL						
					51009	VANCOMYCIN	12/13/79 6:00	14.3	14.3	ug/mL						
											225798	Vancomycin	12/13/79 8:00	12/13/79 8:01	1	dose
					51009	VANCOMYCIN	12/14/79 5:41	15.3	15.3	ug/mL						
											225798	Vancomycin	12/14/79 8:30	12/14/79 8:31	1	dose
					51009	VANCOMYCIN	12/15/79 5:54	15.7	15.7	ug/mL						
					51009	VANCOMYCIN	12/16/79 6:07	16.8	16.8	ug/mL						
											225798	Vancomycin	12/16/79 8:00	12/16/79 8:01	1	dose
Vancomycin	12/17/79 0:00	12/20/79 0:00	1000	mg												
					51009	VANCOMYCIN	12/17/79 6:01	24.7	24.7	ug/mL						
					51009	VANCOMYCIN	12/18/79 5:43	7.4	7.4	ug/mL						
					51009	VANCOMYCIN	12/20/79 11:00	27.3	27.3	ug/mL						

Figure 2.4.: Administration of *vancomycin*, an antibiotic drug, during one unique ICU stay via the MIMIC-III database.

4. **Temporality** Accurately capturing the temporal characteristics of medical observations is challenging as a result of poor documentation of temporal ordering. This makes it difficult when developing mechanisms to represent a patient's medical history using EHRs. A patient's EHR for a single hospital encounter may contain multiple clinical events, and each clinical event may be repeated. Many observations, such as lab results and vital signs, are continuously recorded during a patient's hospital encounter, resulting in a collection of time-series information [103]. This time-series is susceptible to irregularity as observations are recorded in various intervals. Also, several observations may be recorded simultaneously. For instance, a heart rate monitor and a device to measure blood oxygen may be connected to a patient to measure both heart rate and blood oxygen levels. The measurements are then sent to a EHR system and append to the patient's EHR. Though this information is captured simultaneously and they will have the same time-stamps in the EHR, they will be recorded in separate rows in the database table.

While these challenges can make data analysis difficult, the promises of leveraging EHR data for secondary purposes rely on several factors. Specifically, the ability to

understand the underlying data source and identifying the preprocessing steps needed to transform EHR data representations for analysis. Researcher and analyst must be aware of these challenges to reap the full value of EHR data and reduce or remove errors and biases that may impact the data.

2.3 Mathematical Concepts

2.3.1 Markov Chain

A discrete first-order Markov chain is a type of stochastic process that describes a sequence of possible events, where the probability of the next event depends only on the current event [104]. Such events are called *states*, s_i, which make up a finite *state space*, $S = \{s_0, s_1, s_2, ..., s_n\}$ used to model the changes in a system over time. The system moves randomly between states, where each move is considered a step. State changes are called *transitions*, and the corresponding probabilities for different state changes are called *transition probabilities*. Transition probabilities, p_{ij}, represent the likelihood (or probability) of changing to another state or remaining in the same state, p_{ii}, where i and j are states. Transition probabilities can be represented by a transition probability matrix (TPM), P, or a graph (Figure 2.5). A TPM consists of rows and columns with probabilities for each state.

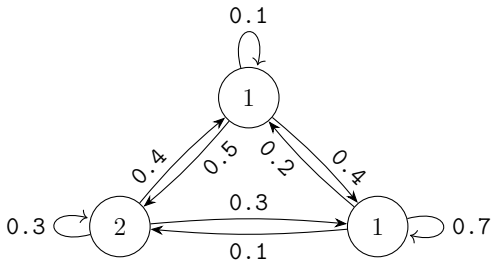
Absorbing states are states that are impossible to leave and terminate the Markov process once reached. Transition probabilities must follow the constraint that each row sums to 1.

2.3.2 Monte Carlo Simulations

Monte Carlo simulations are a technique for understanding the impact of uncertainty in a model [105] and are used to "assess the validity, reliability, and plausibility of inferential techniques" [106]. A main feature of Monte Carlo simulations is its ability to estimate the likelihood of outcomes. They can be used to estimate the

$$P = \begin{bmatrix} 0.1 & 0.5 & 0.4 \\ 0.4 & 0.3 & 0.3 \\ 0.2 & 0.1 & 0.7 \end{bmatrix}$$

(a) TPM



(b) Transition probability graph.

Figure 2.5.: A simple 3-state Markov chain with transition probabilities represented by (a) TPM. Rows represent the probability of moving from the corresponding state to the state corresponding to the column. (b) Transition probability graph. Nodes represent states. Edges represent probability of moving from one state to another or remaining in the same state.

probability of different outcomes in processes, such as Markov chains, that are difficult to predict or too complex to solve analytically [105, 106]. The simulation is an iterative process, which repeatedly generates random values and selects next states based on the range of the model parameter estimates (i.e., transition probabilities). The simulation ends once a particular state has been reached or a specified number of random draws has been completed. The results are based on repeated sampling from a probability distribution for a random process.

3 A CONCEPTUAL EHR DATA ANALYTICS FRAMEWORK

3.1 Introduction

A conceptual framework is a way to organize components and ideas needed to achieve a goal [107]. It provides a representation of an overall picture of what is needed. Understanding the overall picture is an important aspect of data analytics. Thousands of terabytes of data are generated each day from various sources (e.g., social networks, online shopping, medical sensors), focusing more attention onto big data analytics [108]. Big data analytics is the process of collecting, organizing and analyzing large amounts of data [109, 110]. The ability to leverage large amounts of data from various sources has presented opportunities for analytics in areas of the healthcare industry such as population health, business, and clinical research [111]. While these areas differ according to the data sources (e.g., EHR, claims, clinical trials) used for analysis, they all intend to extract meaningful insights from data.

Data analytics solutions can differ across areas in healthcare (e.g., business, clinical) as well as within the same area. Thus, depending on the underlying data, as well as the information and analytics needed, various tools and frameworks can be developed. Specifically, the EHR data analysis pipeline is comprised of several components that are useful and/or necessary for generating evidence and providing insight into healthcare processes. EHR analysis tools provide research and medical professionals the ability to analyze various forms of data (e.g., structured, unstructured). However, depending on the form of the data, tools equipped with specific capabilities may be required to perform analysis. These specialized tools can create data silos, which are isolated sources of data that are only accessible or usable by a small population of medical professionals or analysis [71].

Despite the potential for data silos and isolation in practice, a conceptual framework can be developed that encompasses the main components of an EHR analysis system. While the components represent the architectural structure of the system, the capabilities of those components provide the ability to manipulate (e.g., remove errors, label) the data for analysis. Once the components and corresponding capabilities are identified, tools and techniques that have such capabilities can be mapped to the components, creating a framework. This chapter presents a novel conceptual framework that identifies and explains five major components necessary for EHR data analysis. These components are: i) *data storage* ii) *data extraction and preprocessing* iii) *data aggregation* iv) *data analysis* and v) *data visualization.* We explain how longitudinal EHR data impacts the capabilities of each component to show how a conceptual framework can be developed to define the comprehensive requirements for generating clinical evidence.

3.2 Related Work

Designing analytics frameworks can be challenging, as developers must have knowledge of the analytics goals to capture data without loss of information and preserving data integrity. Previously developed frameworks [110, 112, 113] for healthcare data analytics are structured around the following architectural components: data storage, aggregation, analysis, and visualization. While these main components support analysis for various types of healthcare data (e.g., insurance claims, health surveys, pharmacological, EHRs), the capabilities of each component must be able to support the domain of the data used for input. For example, analyzing health survey data differs from analyzing longitudinal EHR data. Health surveys provide prevalence estimates to evaluate population health, while EHRs contain routinely collected, time-series clinical information regarding medical encounters for individual patients. Further, EHRs include various types of data, (i.e., structured and unstructured) which may require a variation of processing techniques for analysis. For example, SemEHR [114]

is an information extraction and retrieval framework. While the performance capabilities of SemEHR have been demonstrated through querying concepts in EHR DBs and lab tests measurements [113], the SemEHR framework provides two different components for handling structured and unstructured data. Structured data (e.g., ICD-9 codes, vital signs) are extracted and processed using SQL queries, while BioYODIE [115], a clinical NLP system, is used to extract concepts in the unstructured EHR data (i.e., "free-text" clinical notes).

Various analytics frameworks have been created to fit the underlying data, while others have been developed for a specific need, as detailed [108, 116]. Though these frameworks are developed for specific analytics needs, they are all based on similar components. For example, Wang *et al.*, [112] defined a framework based on data aggregation, data processing, and data visualization to evaluate the business value of big data in healthcare. Khazaei1 *et al.*, [117] describes a cloud-based reference framework for providing health-analytics-as-a-service for both real-time and retrospective analysis using componenets such as data acquisition, transformation, storage, analytics, knowledge extraction, and visualization. Chawla *et al.*, [116] describes a data-driven, personalized healthcare framework using a collaborative filtering approach, expressing the importance of data aggregation. Sarkar *et al.*, [108] introduced a framework for a secure healthcare system that includes data extraction and aggregations and components. Saggi *et al.*, [110] provides an overview of an architecture for a big data analytics framework for value-creation in business based on several components, including: data generation, data acquisition, data storage, advanced data analytics, and data visualization [110].While each of these frameworks were created for a specific need, they all conceptualized the solutions needed to achieve the corresponding analytic goal.

In addition to frameworks, several tools have been developed and integrated into existing healthcare analytics platforms and infrastructures. Specifically, several research networks [22, 54, 55] offer tools that can be utilized with their specific data models. For example, OHDSI [22] offers a suite of data analytic tools for explor-

ing data and generating evidence to improve health decisions. *HERMES* (Health Entity Relationship and Metadata Exploration System), a vocabulary browsing tool that allows searching and exploring of terms and concepts, *PLATO* (Patient-Level Assessment of Treatment Outcomes), a predictive modeling tool to assess patient outcome probability, ACHILLES (Automated Characterization of Health Information at Large-scale Longitudinal Exploration System), a visualization tool for clinical databases [85], and ATLAS [118], a tool for researchers to conduct scientific analyses, are among the tools were designed to facilitate data exploration, data analysis, and cohort definition. Each of these tools serve a specific purpose that can generate output that is used by a separate tool as input. For example, clinical concepts found using HERMES can be used to identify and explore patient cohorts with ATLAS. However, OHDSI, like other research networks, requires data to be transformed to their (i.e., OMOP) platform-specific CDM to use these tools.

Several frameworks have been developed for specific analytics needs, while others have developed software packages to accompany CDMs for existing platforms. Currently, there is no single tool that encompasses all of the necessary features to provide comprehensive and granular analysis for longitudinal data; however a combination of several tools with the capabilities necessary for meeting analytics needs can be identified with a conceptual framework. EHR data analytics frameworks must support key functions that are necessary for the anlytics goal [119]. While frameworks have been developed for specific analytics needs, they are structured by common architectural components (i.e., storage, aggregation, analysis, and visualization). Thus, the development of EHR frameworks that are equipped with capabilities to manipulate the data at various levels of granularity are conceivable through through these components.

3.3 Methodology

EHR data analytics is driven by the overall analytics goal. Some analytic goals target the identification of trends in the data (e.g., what has happened), while others are interested in what will potentially happen in the future. Clinical data collected over time requires specific techniques for understanding temporal characteristics, as time is the most important parameter when analyzing longitudinal data.

Time-series data is represented by timestamps and generates a temporal ordering of data points that are based on when an event occurs. Incorporating timing information such as the exact time when an event occurred as a key analysis input adds more specificity and granularity to the analysis. For the purposes of this study, we consider time to be comprised of two parts, continuous and discrete. Discrete time includes events (e.g., hospital discharge) or time measures at a specific time (e.g., drug administration). In the medical field, these events include X-Ray results, lab tests, interventions, etc. Discrete time events are used to identify change of measurement values using static data points. Continuous time includes data that is measured repeatedly. Examples of continuous time events include singular physiological waveform data such as heart rate, blood oxygen level and other data that can be monitored via medical monitoring devices [120]. Both discrete and continuous data capture the temporal nature of the time context that can be used for data analysis.

The components presented in Figure 3.1 represent a framework of an EHR analysis system. This framework is comprised of a group of sequential steps that transform raw EHR data into usable inputs for visualization and/or modeling and simulation algorithms. This framework is an exhaustive explanation of the major components of an EHR data analytics system.

3.3.1 Data Storage and Access Architecture

Data is at the core of the EHR data analytics framework. Identifying data storage formats that establishing the foundation for accessing and extracting the data with

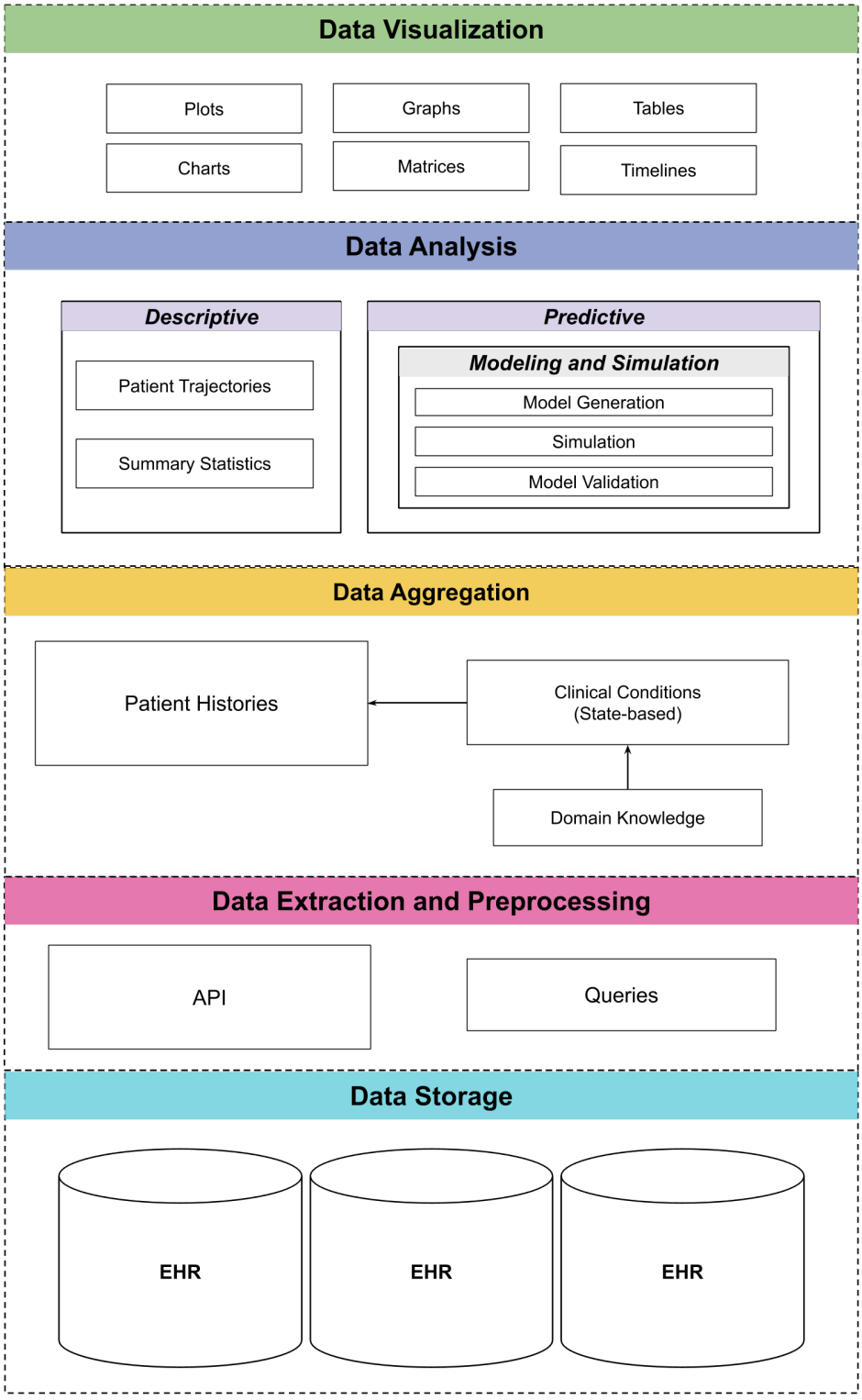


Figure 3.1.: Conceptual framework for EHR data analytics.

limited complexity in regards to our needs is important. EHR data is typically made available as a relational DBs, a collection of tables that are linked together by shared keys (Figure 3.2). This structure helps maintain data integrity and enable faster analysis and more efficient storage [89]. EHR is can also be exported from a DB as

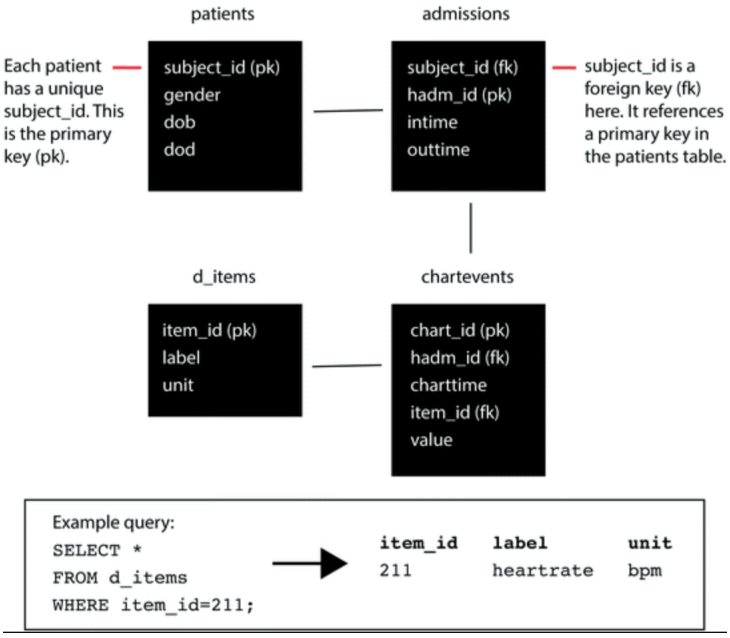


Figure 3.2.: Example of tables in the MIMIC-III relational database pk is primary key. fk is foreign key. [89]

a file, most common being a CSV file [89]. CSV files can be loaded as a spreadsheet in existing software packages such as Microsoft Excel, edited with text editor such as Microsoft Word, and imported and processed by most data analysis packages, making them an intermediate data format used to hold data [89]. Privacy laws such as HIPAA have storage and security requirements for data that contain personal and identifiable information. While access control measures and data policies are put in place for EHR data access, such methods are beyond the scope of this research.

3.3.2 Data Extraction and Preprocessing

Data Extraction Once data is collected and stored, outside researchers and other medical practitioners can extract the data to perform additional analysis [120]. Data acquisition resulting in useful information requires background knowledge of the dataset characteristics. For example, determining a patient's age requires knowledge of the structure of the data. Patients can have multiple hospital stays, and their age can change between stays, and their information can be spread across several tables. The `PATIENTS` table contains a unique identifier, `subject_id`, used to identify the patient and a field, `dob`, which represents the patient's date of birth. The `ADMISSIONS` table contains timestamps and unique identifiers for each hospital stay related to the patient. These two tables must be joined using the `subject_id` to retrieve this information. Once extracted, the `dob` timestamp and hospital admission timestamps are used to compute the patient's age for each hospital stay event for the patient. This extraction depends on the data storage technique as well as the type of data being extracted. Without this information, the extraction can not occur in way that is digestable by the tool; thus, the desired data will not be analyzed. This need forces data extraction techniques to be developed that align with the structure and semantics of the underlying data source. Data can be extracted from the DB by constructing SQL queries or via an application programming interface (API) that maps pre-constructed and customized queries to functions that are applied to the data. API methods with pre-constructed queries do not require comprehensive knowledge of the underlying storage and structure of data. For example, the API can include a function to return a patient's age based on all hospital admissions. Such a function does not require the user to know the necessary tables to access. API functions that allow customizable queries as input will require the user to have some knowledge of the underlying data and structure. These functions provide flexibility and allow dynamic data extraction for more complex or advanced data acquisition.

Data Preprocessing Once extracted, the EHR data may have to be preprocessed (or transformed) before it can be used as input to analysis algorithms. This transformation involves cleaning, sorting, and validating the data. Preprocessing steps for data cleaning includes removing duplicate, implausible, and incomplete data records. Data can be simulated to fill in the missing data; however, generating synthetic data is outside the scope of this work. Mathematical operations can be performed on the data to generate new summary data or change individual EHR values. For example, the data could contain multiple measurements for heart rate, all measured within a five minute period. These measurements combined by calculating the average of their values to create a single measurement for heart rate with a single timestamp. Data sorting adds order to the data based a numerical (e.g., `subject_id`) or timestamp (e.g., `charttime`) field.

Data preprocessing is an iterative task that is not only implemented for cleaning and sorting data. Preprocessing and transforming patient data into meaningful information that supports data analysis for generating evidence is necessary to use the data as input into analysis functions. Thus, additional preprocessing steps may include integrating clinical domain knowledge into the data.

3.3.3 Data Aggregation

Data preprocessing transforms the data so that it can be analyzed for discovering trends and underlying information. Data aggregation tools organize information to create a holistic summary of patient events. Such tools can integrate clinical domain knowledge into patient aggregated data to understand and analyze patient health changes. Belle *et al.*, [120] state

> "Understanding and [analyzing] clinical conditions and disease progression requires an aggregated approach where structured and unstructured data stemming from a myriad of clinical and nonclinical modalities are utilized for a more comprehensive perspective of [health and] disease states."

This will enable analysis that reflects the current state of clinical practice as well as care decisions captured in the EHR.

3.3.4 Data Analysis

Data analysis is characterized by the type of data as well as the purpose of the analysis [112]. There are three main kinds of data analytics: descriptive, predictive, and prescriptive [112]. To select the most appropriate kind of analytics approach, there must be an understanding of the type of question being asked and the level of measurement being used for input variables [121]. Prescriptive data analytics provide optimal solutions or potential courses of action to help understand what should be done in the future [112]. While this research intends to generate evidence through EHR data analytics to drive clinical guidelines, prescriptive analytics is beyond the scope of this research. Below, we describe the remaining two data analytic types, descriptive and predictive analysis.

Descriptive Analysis In healthcare, descriptive studies are considered the most commonly used type of analytics, providing insight into the past [93]. Data can be presented in the form of summary statistics or by creating patient trajectories to understand past medical events and how such events could potentially affect patient outcomes. Descriptive analyses using summary statistics can also be used to compare results between studies such as mean mortality rate among men and women with COPD. Chapter 5 discusses aggregating care events of patients' EHR data, which is a type of descriptive analysis.

Predictive Analysis Predictive analytics techniques predict outcomes based on probability estimation [112]. Predictive analytics approaches include probabilistic modeling and simulation such as Markov chains and Monte Carlo and machine learning algorithms such as Bayesian networks, can be used to capture granular temporal information regarding EHR events and patient behaviors. Specifically, probabilistic

modeling and simulation can be used to answer both temporal and atemporal clinical research questions using time-aware and time-agnostic techniques [122]. Probabilistic modeling algorithms are based probability distributions that are realized through knowledge of past values. While, these algorithms provide additional insight into temporal trends and can identify patterns and hidden relationships regarding clinical variables, models must be developed that support analysis of time-series data [28,112,123]. Analyzing longitudinal data requires more than what common modeling methods such as Markov chains and Bayesian networks alone can provide [124]. Such models assume fixed sampling frequencies, such as evenly spaced time events. However, such models cannot be directly applied to EHRs as they contain irregularly sampled data. Using data that does not meet these assumptions can fail to capture what is present in the data and can lead to spurious and inaccurate results [124]. Thus, analysis components with modeling capabilities must support the temporal dynamics of EHR data. Additionally, predictive analytics techniques such as probabilistic modeling should be accompanied by validation approaches to evaluate the robustness of the model predictions. Chapter 6 discusses time-based modeling and how complex queries can be generated and analyzed, which can be categorized as a prescriptive analysis.

3.3.5 Data Visualization

The data visualization component generates outputs derived from the analysis component [112]. Visualization encompasses any visual transformation of the raw data or mathematically transformed data. Visualizations can be presented as individual patient and cohort summaries using scatter graphs and charts, as well as timelines of time-series data to provide a comprehensive view of the evolution of patient conditions. Visualizations of longitudinal data should reveal any changes that exist within individual patient data as well as data from patient cohorts. Generating comprehensive views of longitudinal EHR data using graphical methods provide a sense of the time elapsed and the events that occurred during a hospital stay. Ad-

ditionally, visualizing the data can provide greater insight into overall trends in the data, expose anomalies, and identify subsets of data for additional analysis.

3.3.6 Framework Development

An EHR data analytics framework can be be developed using an object-oriented programming (OOP) language such as Python [125]. Python is particularly useful for creating visualizations, processing data, and performing granular analysis. Its modular programming structure streamlines maintenance and existing code modification, defines abstract data types, and creates objects that can be reused within and across applications [125]. Python is equipped with a variety of libraries that can be adapted and modified to develop components that, together, create a comprehensive analytics framework.

3.4 Discussion

The contribution here is a generic EHR data analytics framework. We develop a conceptual framework that is consistent with the healthcare data analytics paradigm. We show this consistency by describing framework components and their corresponding capabilities. While the structure of data analysis systems should be driven by the data and the desired output, the five components described will be the foundation of this structure. A comprehensive framework with capabilities that not only allows analysis for deriving answers to clinical questions but also integrates the complexity of the clinical question itself [126] provides a platform for generating evidence for clinical guidelines.

4 MIMIC-PURDUE: A DATA EXTRACTION AND PREPROCESSING API

4.1 Introduction

Working with EHR data can be challenging because of its disorganized, redundant, and error-prone nature [82]. In addition to these challenges, EHRs can originate from various sources such as different vendors with different database schemas, introducing issues such as sparsity, complexity, incompatibility, and heterogeneity [127]. All of these challenges can make data extraction and preprocessing difficult. Extracting and preprocessing large volumes of EHR data accurately when some or all of these challenges are present in the data can be time-consuming and require extensive knowledge of the underlying data elements [72]. To extract data efficiently, a tool that addresses such challenges needs to be designed and implemented in a way that does not require extensive human interaction. To address these challenges, we introduce *MIMIC-Purdue* for storage, extraction and preprocessing MIMIC-III data. This novel and unified extraction API was developed in Python and provides a structured way to formulate complex queries and interfaces directly with a PostgreSQL schema of the MIMIC-III DB.

4.2 Related Work

Common Data Models Clinical research networks target unlocking the value of clinical data by structuring data, providing analytic tools, providing visualization capabilities, or a combination of the above. EHR data sharing networks such as *i2b2* [55] and *OHDSI* [22] provide platforms for clinical research with informatics solutions and are well-suited for performing simple inquiries regarding clinical characterization as well as patient and population level analysis. For example, the i2b2 system, consists

of two components: 1. the *Hive*, a back-end infrastructure for security and success management, and 2. the Workbench, a querying and mining tool [55]. The system is optimized to identify cohorts of patients. The i2b2 web-based Query and Analysis Tool allows users to create Boolean query combinations and returns a summary count of patients matching the query [128]. However, the i2b2 system, similar to others, are less effective when more complex queries are presented, such as those that are convoluted and granular in detail. Though networks such as i2b2 and OHDSI may use different approaches to structure observational and clinical data, all of the aforementioned platforms share the concept of a common data format or model to achieve some standard structure for their datasets. Each platform requires an ETL process to convert data into a desired syntactic and semantic standard, usually defined as a platform-specific CDM [72].

Post et al. extended their software, Eureka!, a metadata-driven ETL tool that can be customized for different data marts (or networks). They showed the value of their tool with an application to multiple EHR DBs [87]. While their approach could potentially reduce the burden of ETL processes, there are concerns of data loss, as some of the concepts may not be mapped correctly as a result of lack of support for inclusion and exclusion criteria. While, there are other organization-specific tools to help with the ETL process such as WhiteRabbit [129], a tool for the OMOP CDM, and Integrated Data Repository Toolkit [130], an i2b2 setup and administration tool, the lack of a standard model can lead to further segregation of healthcare data sources. For example, for MIMIC-III to be used within the i2b2 community, it must first be transformed to the i2b2 star schema data format. This requires data cleaning and mapping of concepts and vocabulary, among other procedures. A separate process, however, must be executed to use the OMOP CDM. This requirement to execute a new process to first prepare the data is present for each new network that is chosen. Such a taxing conversion processes can discourage data holders by forcing them to choose between networks, increasing disparity across the available data sources that are from these networks.

One of the main advantages for constructing a common data model is to provide structure to heterogeneous data and data sources. This is the most widely applied solution used by existing networks to represent data uniformly [131]. While these platforms are useful for mitigating EHR data challenges such as inconsistency, heterogeneity, and interoperability limit the amount of information that is represented. This is because the process of transforming datasets to fit the data models causes information loss and can create inaccurate information as a result of the encoding and standardized relationships among data elements [132]. The configuration and implementation designs of the data source systems can result in complexity issues when attempting to store EHRs using such data models [133]. These source systems vary in the way they are created, which impacts the extracted data when it is imported to systems such as i2b2. For instance, Deshmukh *et al.* [133] exposed unique challenges when attempting to reproduce data captured via structured clinical documentation because of the way different clinical systems were created.

One solution to address the challenge of standardization is to extend current platforms/models by applying data profiling, assessment data quality in source systems, during the ETL process [134]. Current systems use data models to correlate disparate data sources, such as claims and clinical data, to hospital data. Creating an additional feature that defines the underlying data source, providing the context to how it was derived can preserve the uniformity of the models. For example, in i2b2, both claims and clinical EHR data are transformed into the common star schema [55]. During this transformation, procedures, facts, observations, etc. are not differentiated between these two EHR data sources. When a researcher attempts to perform a more in-depth analysis on EHR data, the extended functionality will allow the intended and correct data to be presented.

Data Harmonization Data harmonization, the process of combining data from various sources, of common data models and APIs for both individual and multiple data sources is another idea that have been developed for EHR data extrac-

tion [135–137]. Fast Healthcare Interoperability Resources (FHIR) [138] has received attention by adopting a standard for exchanging and accessing healthcare information electronically. Their EHR standard structure is defined by modular components, called resource references, to combine resources together through extensions of the FHIR RESTful API. FHIR can be used alone or together with existing standards and data models such as i2b2 and OMOP [139].

The Patient-Centered Informatics Common: Standard Unification of Research Elements (PIC-SURE) at Harvard Medical School, [140, 141] has been developing an open source infrastructure to incorporate multiple heterogeneous patient level data including clinical, omics (i.e., genomics, proteomics, metabolomics), and environmental data. The core idea of PIC-SURE is to aggregate distributed data resources of various types and protocols within a single communication interface to perform queries and computations across different resources such as i2b2 and OHDSI. For this purpose, the PIC-SURE team developed the Big Data to Knowledge (BD2K) PIC-SURE RESTful API. The PIC-SURE API is resource-driven and allows new resources to be integrated. Any action executed by users will be passed to the resource interfaces, where each resource interface translates the action to the supported protocols. A query constitutes several clauses and in PIC-SURE clauses are limited to only be of types *Select*, *Where* and *Join*. However, to perform more in-depth and insightful analysis, supporting complex query generation and execution as well as post-processing is necessary. The PIC-SURE API has been leveraged to access data through integration with programming packages for accessing and analyzing various types of clinical data [142]. Gutiérrez-Sacristán *et al.*, [142] developed Rcupcake, an R package, for analyzing different databases through the BD2K PIC-SURE RESTful API.

While the standardized structure and extraction methods provided by OMOP, i2b2, FHIR, and PIC-SURE are beneficial for harvesting data from various sources and achieving interoperability, they limit functionality regarding longitudinal analysis of healthcare data. The "one size fits all" epidemic that data warehouses have

adopted is not conducive for the multitude of types and sources of data. The way the data models structure data from different sources (e.g., claims, EHR) may result in loss of information. For instance, a patient with COPD may have their blood oxygen level checked at a medical lab in July, an ICU visit in October, and a prescription re-filled at a pharmacy in November. Claims data will contain operational information about the patient's interaction with the healthcare system (e.g., procedures, diagnoses, providers), while EHR data will contain more clinical information about the patient such as lab results and vital sign measurements.

Data may contain erroneous, incomplete, or duplicate information as a result of the manner in which it was entered or stored [91]. Failure to remove or correct such data may result in inaccurate analysis results. For example, Just *et al.* [143] discusses the effects of duplication and erroneous data in EHRs on data analysis, such as various entries of lab tests, frequency of invalid fields, and missing data within the care setting. Thus, data extraction and processing (e.g., querying, extraction, and transformation) are critical components for data analysis [112]. Regardless of the data model or structure chosen for EHRs, preprocessing must be performed upon extraction before data can be analyzed.

4.3 Methodology

4.3.1 Data Storage and Access

The MIMIC-Purdue storage platform is a PostgreSQL relational database management system (RDBMS), built with MIMIC-III data. The DB is stored on a Dell PowerEdge R430, Linux-based machine with 16 cores, 8 terabytes (TB) disk, and 64 gigabytes (GBs) of random access memory (RAM), located at the Regenstrief Center for Healthcare Engineering (RCHE) at Purdue University. Access to the system is granted to researchers at RCHE upon completing the data use agreement [99] for MIMIC-III. By storing the MIMIC-III data in a central location, users (i.e., RCHE researchers) are granted unlimited remote access to the system. Users can interact

with the data via command-line interface, DB (e.g., DataGrip) and analysis software (e.g., Excel, Tableau), or through other web-based tools (e.g., Adminer). The system (Figure 4.1) has been integrated with SciDB, a column-oriented database management system, that allows additional functionality for exploration, visualization, clinical and waveform database integration, as well as complex analysis with distributed computing architecture [144]. For example, each user is assigned their own directory on the MIMIC-Purdue server. The directories are used to develop interactive web applications in R or Python, allowing direct interaction with applications written by researchers at RCHE.

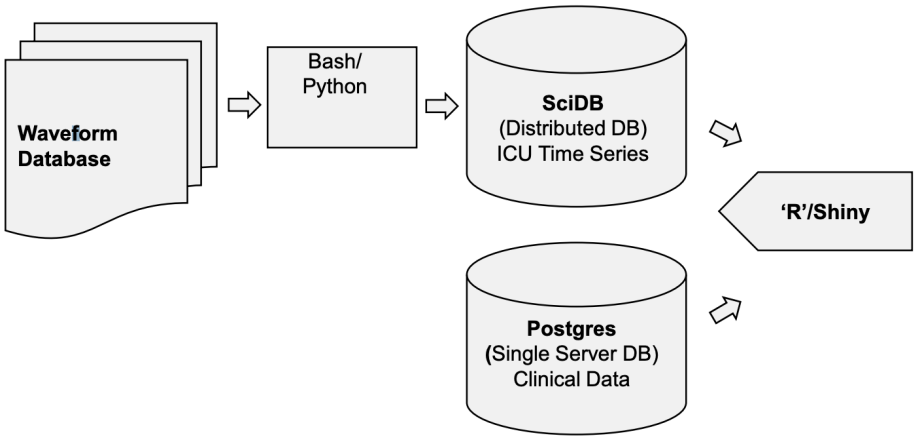


Figure 4.1.: MIMIC-Purdue: Software architecture and data flow of the MIMIC-Purdue DB system [144].

The MIMIC-Purdue storage system is equipped with data security mechanisms, such as firewalls, to adhere to data privacy and storage requirements of deidentified patient data. The user access is constrained with data use agreements, password authentication, as well as public-private key mechanisms.

The MIMIC-III DB is 80GB, and attempting to query the DB from a web-based tool can slow performance. This is mainly because data needs to be downloaded and then imported into the web-based application that is being used for analysis or generating visualizations. Thus, we access the MIMIC-III database through a remote

connection to the PostgreSQL DB via Python code with embedded SQL queries. This method provides a user-friendly interface to the DB for extracting data rather than using a command prompt. Furthermore, by querying the DB through Python, we remove the need to download query results and import them for analysis.

4.3.2 Data Extraction

To extract data from the MIMIC-Purdue DB a connection must be established using required credentials (e.g., user-name, password), and a query must be executed to retrieve results. The MIMIC-Purdue extraction API is an extension of methods for extracting clinical notes for input to SemEHR, an information and extraction system for clinical notes [145]. The main contribution of the MIMIC-Purdue extraction API is added flexibility in data extraction by providing enhanced functionality to execute dynamic and complex queries. Table 4.1 provides definitions for the two query types, static and dynamic, that are accessible to the user through a collection of wrapper functions.

Static queries are pre-constructed and require user input for an identifier such as subject_id, hadm_id, and icd9_code. Dynamic queries are queries that are not pre-constructed and are executed by passing a custom SQL statement as a string to the applicable API function. These queries are customizable, allowing for specific extraction and manipulation of the data. Allowing dynamic queries can potentially be harmful if permissions and privileges for the DB are not in place. The MIMIC-Purdue DB configuration restricts modification of the DB itself (e.g., inserting and deleting tables). Both dynamic and static queries can be complex or simple. Complex queries are queries that include combining data from multiple tables, and simple queries gather data from a single table. The MIMIC-Purdue API also includes wrapper functions to connect and disconnect from the DB. The DB connection is made by creating a secure shell (SSH) tunnel and taking parameters for user credentials to establish a connection to the MIMIC-Purdue.

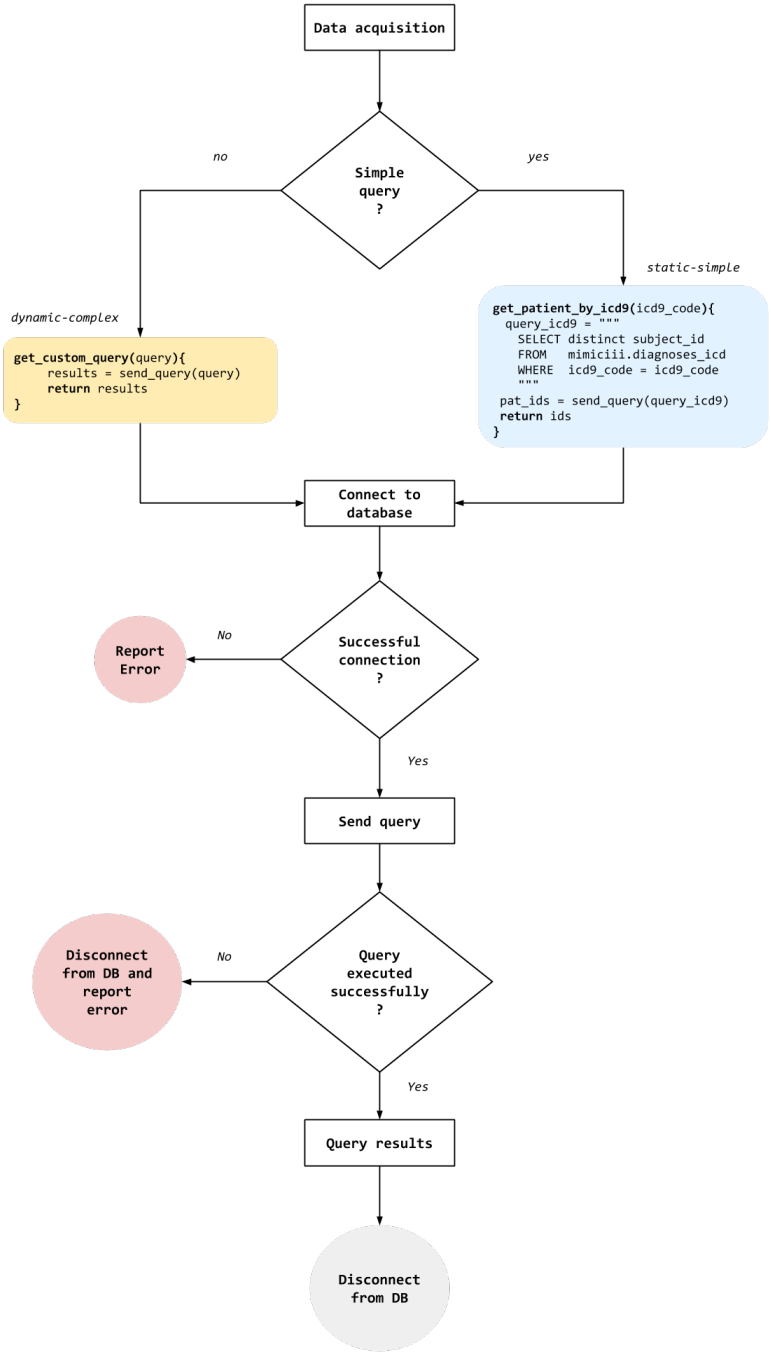


Figure 4.2.: Flow for data extraction API.

Table 4.1.: Definitions of query types for MIMIC-Purdue data extraction API.

Query Type	**Name**	**Definition**
Static	*static-simple*	pre-constructed; extracts data from a single table; optional input value for field name in WHERE condition
	static-complex	pre-constructed; extracts data from multiple tables using JOIN; requires an input value for field name in WHERE condition
Dynamic	*dynamic-simple*	partially pre-constructed and customizable; requires an input value for SELECT, FROM, and WHERE
	dynamic-complex	constructed by the user; customizable query with user input

Figure 4.2 illustrates the process of extracting data with the MIMIC-Purdue extraction API. When there is data acquisition, either a static or dynamic query function is called and the required parameters are passed. Next, a function call to connect to the MIMIC-Purdue DB is made. If the connection is not successful because of reasons such as incorrect DB credentials, failed network connection, or system failure, the process is terminated and an error message describing the connection error is displayed. Otherwise, the query is sent to the DB. If the query contains misspelled or undefined table names or variable/field names, the query will fail, terminating the process by disconnecting from the DB and displaying an error. Upon the execution of a successful query, results are stored as a Python DataFrame, a 2-dimensional labeled data structure that is later used for preprocessing. The database connection is closed and the extraction process is complete. The API also provides functions that enables query results to be "pickled". Pickle, a Python package, allows query results to be

serialized to files on disk and deserialized back into the program when needed. This is beneficial when the same data is needed repeatedly or when extracting data for cohorts. Also, this method reduces a portion of the consumption of network services and database resources when connecting to the DB multiple times for the same data acquisition.

4.3.3 Data Preprocessing

Once data has been extracted it enters a preprocessing stage. During this stage, the query results are checked for null values and for temporal data acquisitions all erroneous and implausible data records are removed. This is done through various mechanisms that the MIMIC-III DB includes as well as ones we define. The `CHARTEVENTS` table contains all charted (e.g., lab results, vital signs, interventions) data collected for patients during their hospital stay. The table includes a field, `error`, that indicates if an error occurred during the time the measurement was recorded. Thus, we use this field to identify and remove data that have been labeled erroneous within the DB.

Temporal data acquisitions are those that return query results based on time-stamps. For example, a simple extraction of a list of patient `subject_id` values is not validated for errors, while a more complex, temporal data extraction that includes time-stamps is checked for valid dates and times. Any temporal extractions are checked for erroneous and implausible data values. For example, if query results return data from a patient with a time-stamp for a measurement that was recorded after their discharge time-stamp, the patient data is excluded from the extracted data. The API also includes functionality to relabel the data fields using a dictionary of identifiers (or `item_id`) and corresponding names to replace duplicate names for the same measurement type. For example, there are four identifiers for respiratory rate in the MIMIC-III DB, each having a different name (e.g., RR, Resp Rate, Respiratory Rate). Once extracted, any respiratory rate identifier is renamed to Respiratory Rate.

In addition to renaming identifiers, measurements for the same variable are all converted to the same unit. For example, the measurement for weight is listed in the DB in kilograms, pounds, and ounces for different patient records. For consistent analysis, we convert all measurements for weight to pounds. Our preprocessing stage also applies clinical knowledge to identify and remove clinical measurement values that are potentially inaccurate by implementing range and consistency checks. For example, if a measurement for a patient's oxygen saturation is greater than 100%, which is physically impossible, the measurement is considered erroneous and discarded.

4.4 Discussion

Our approach to data extraction coupled with the preprocessing methods provided in the API allows for more granular data extraction and aggregation, as well as structured and semi-direct interaction with the raw data. The flexibility of the MIMIC-Purdue API has the ability to extract data from the DB by providing functions that have predefined queries and accept user input. Even more, it allows users to issue completely customizable, complex queries for dynamic data acquisition, which is a major difference from the capabilities of existing systems. Additionally, the methods presented for storing and extracting data reduces the risk of loss of information beyond the point of its original structure. That is, we do not define nor use a specific data model as done in previous work [22, 55]. Rather, we store the MIMIC-III data in the form (i.e., relational DB) in which it is provided by the provider (Laboratory for Computational Physiology at MIT) [99].

4.4.1 Limitations

Generally, common limitations of data-driven research are quality and accuracy of the data received [146]. While beyond our control, it is possible that faulty data entered at the source (e.g., hospital), as a result of human error, is included in the analysis. Another limitation of our approach is that each time a function is called

to extract data from MIMIC-Purdue, a new DB connection is established and closed when the function returns. If a large number of function calls are made, performance issues can arise. The API includes functionality to serialize files so that they can be saved on disk. This helps with potentially reducing the number of DB connections when gathering data for the same patients repeatedly to maintain persistence in the data that is being gathered. The extraction API does not have input validation for dynamic queries issued by the user. While the MIMIC-Purdue DB prohibits queries for data modification in the DB, an improperly constructed (e.g., misspelled table names), dynamic query which will fail will be sent to the DB as long as the DB connection has been established. This also occurs when interacting with DBs directly. Constructing dynamic queries could potentially be a usability limitation for users who are unfamiliar with SQL. Though the extraction API is built for the MIMIC-III DB schema, SQL is the underlying language used to extract the data. The same extraction approach can thus be for other types of SQL DBs, but must be modified to fit the corresponding DB schema.

4.4.2 Future work

To mitigate the limitation of connecting the DB each time an extraction function is called, a wrapper class for API functions can be created that keeps information such as connection credentials, manages DB sessions, and organizes logic necessary to make API calls as objects. Input validation solutions for dynamic queries can be integrated by developing a Python dictionary that translates the underlying MIMIC-III DB schema into the API. That is, parameters passed to the API functions will be checked against the MIMIC-III tables and field names so that there is input validation for dynamic-simple queries. Functionality for automatic preprocessing of the data can be incorporated by creating a rule-based engine that checks for known, potential errors in the data such as those discussed in the preprocessing section.

5 PACE: PATIENT AGGREGATED CARE EVENTS

5.1 Introduction

Information extraction for large clinical databases can be time-consuming when applied to an individual patient, and even more so when applied to large cohorts of patients [147]. A patient's EHR tells a story of their medical history. This medical history is composed of temporal clinical data, data collected over time, that can be used to understand the care process and the health state of a patient. Large amounts of raw data points, containing clinical information, such as timestamp observations, are stored in EHR DBs [21]. Because of the high volume of data, critical information can be buried within the EHR, causing an information overload. This makes it difficult to represent a patient's medical history coherently, as this information is typically scattered across different DBs such as pharmacy, ICU, and Emergency medicine, or multiple tables within a single DB [21]. Moreover, performing analysis can be challenging, given the representation and structure of the data. Attempts to analyze several variables with this information overload often leads to overlooking and misinterpreting critical information [20] that could potentially impact the care process. For example, vital signs (i.e., heart rate, blood pressure, and respiratory rate) are measurements of the body's basic functionality. They are a subset of clinical variables that are measured repeatedly and stored in EHRs. Such measurements are a source of critical information that is used to detect and monitor a patient's overall health state. This critical information can be overlooked when attempting to observe, gather, and analyze patient EHRs without proper representation [20]. Thus, to understand a patient's health journey and make more informed decisions, clinicians need a comprehensive view of the available and necessary information in a patient's EHR.

One method to provide comprehensive views is to use longitudinal time-series data to capture the temporal dynamics of the care process. The representation of patients' EHR data can be enhanced by aggregating and organizing time-series information. The data can then be used to create visualizations that provide comprehensive views of raw, clinical data. The ability to capture temporal granularity can provide more in-depth and descriptive analysis.

While several tools [85, 118] have been created to generate visualizations of longitudinal clinical data, they lack the ability to capture granular temporal dynamics of data, limiting the type of data they can analyze and the visual representations they can generate [21]. This chapter presents *Patient Aggregated Care Events (PACE)*, a novel tool for constructing and visualizing entire medical histories of both individual patients and patient cohorts. Our approach transforms each patient's EHR into a *caretrail*, a chronological collection of events, occurring during a patient's hospital encounter, integrated with clinical domain knowledge. We demonstrate the novelty of our approach by identifying patient cohorts using the MIMIC-Purdue extraction and preprocessing API, defining health states using recommendations from clinical guidelines and knowledge from clinical experts, and generating caretrails integrated clinical domain knowledge. This form of integration provides insight into the clinical condition of patients as they move between various health states during their hospitalization. We further demonstrate the usefulness of caretrails by providing visual representations of treatments, outcomes, and clinical measurements for patients diagnosed with AECOPD.

5.2 Related Work

Several clinical research networks have provided tools and afforded others opportunities to develop tools that leverage CDMs to perform large-scale observational clinical studies using retrospective data [148]. One tool is OHDSI's ATLAS, a web-based tool that shows a patient's health care records in a timeline view and allows users to create

cohorts based on specific conditions, drug exposures, etc [148]. ACHILLES, another OHDSI tool, is a browser-based, exploratory, interactive framework, designed to provide visualizations of pre-extracted summary statistics from OMOP CDM formatted datasets [85]. ACHILLES has two main components: (1) an R package to generate summary statistics and (2) an HTML/JavaScript website for exploring and visualizing the results. i2b2 developed a web-based Query and Analysis Tool that allows users to create simple Boolean query combinations to identify patient cohorts [128]. For example, one could construct a simple query that calculates the number of COPD patients who have been administered antibiotics. These discovery or exploratory queries provide insight into cohort statistics, but more granular analytics, specifically dealing with time, patient health states, and visual representations are not possible using this system.

Using temporal data to create timelines and other visualizations is a common approach for longitudinal analytics [149]. A timeline should capture granular temporal information such as the exact time a patient was administered antibiotics or the exact time blood samples were extracted. The i2b2 system provides a timeline view that gives limited information regarding timing of clinical events. In the timeline view, observations are plotted as vertical bars and grouped by categories according to times when the observations occurred [23]. However, observations (e.g., lab tests or diagnoses) regarding a patient are recorded using their start and end dates, but are not representative of the patient's entire medical record. Furthermore, i2b2 does not offer analytical tools to perform tasks such as statistical analysis or pattern discovery on query results. For example, to perform data analysis on a selected cohort, i2b2 can be used to extract the necessary data, but to perform additional analysis, other tools must be used. These tools could be made available to medical professionals and analysts as part of a group of i2b2 plug-ins or developed independently by researchers, but they are not native to the tool [150]. An example of a plugin for i2b2 is TimeAlign, a visual analysis tool for visualizing multiple patient records in a linear timeline [151], that was derived from another i2b2 tool, Lifelines2 [152]. TimeAlign

generates visualizations of categorical events (e.g., diagnosis, intervention, discharge), allowing temporal relationships between events to be aligned across patient timelines.

While these tools are useful for browsing and exploratory analytics, they do not allow either granular or comprehensive representations of clinical data. Also, their analytic capabilities are limited because they are based on CDMs, which can cause data loss during the ETL process that converts data into a standard format [153].

Other efforts have been made for analyzing longitudinal EHR data [159–161]. Table 5.1 compares characteristics of several tools used for summarizing EHR data and visual analytics. Specifically, these techniques aggregate patient data and create temporal event sequences [122, 159, 162]. For example, HARVEST, allows interactive temporal visualization for longitudinal EHRs by extracting clinical notes and aggregating information from multiple care settings [156, 163]. The idea of aggregating temporal sequences is beneficial for creating comprehensive views of EHRs; however, tools developed based on these techniques, such as [157] and [164], are better for use with discrete timing data rather than continuous timing data. Further, they do not allow for visual analysis of clinical conditions, defined using the underlying data points.

5.3 Methodology

5.3.1 Rule-based State Coding Engine

There are several terms used to describe a patient's condition in the hospital. While terms such as mild, moderate, or severe, are all used in care settings when treating and managing patients, they are broadly defined, lack detail, and are not context specific [165]. Clinical domain knowledge is needed to interpret the underlying meaning of these terms, as they are not intuitive and are typically associated with specific medical events or diseases. This domain knowledge can be obtained from clinical guidelines, medical literature, and medical experts. For example, GOLD guidelines classify exacerbations based on the additional therapy needed, as discussed in Chapter

Table 5.1.: Comparison of longitudinal visual analytics tools. Check-marks indicate the tool has a capability. Amber colored background indicates the PACE tool.

	ViTA-Lab [154]	**KNAVE-II** [155]	**HARVEST** [156]	**EventAction** [157]	**Care Pathway** Explorer [158]	**PACE**
EHR data aggregation	✓	✓	✓	✓	✓	✓
Full patient summaries	✓	✓	✓			✓
Domain knowledge integration						✓
Granular temporal extraction	✓	✓				✓
Timeline visualizations	✓	✓	✓	✓		✓

2.1.2. Specifically, severe exacerbations are those that require hospitalization and are identified by symptoms that are used to describe a patient's clinical condition in more detail. These symptoms are observed by assessing the patient's physical appearance, performing medical exams, and/or clinical measurements. GOLD guidelines suggest that severe exacerbations may be associated with acute respiratory failure (ARF), which is a buildup of fluid in the air sac of the lungs that inhibits the release of oxygen into the blood [2, 4, 47]. There are two types of ARF that can be identified via arterial blood gas (ABG) analysis for partial pressure of oxygen (PaO_2), partial pressure of carbon dioxide ($PaCO_2$), and pH, defined as [166]:

- Type 1 (hypoxemic): $PaO_2 \leq 60$ and $PaCO_2 \leq 50$.
- Type 2 (hypercapnic): $PaCO_2 > 50$ and $pH \leq 7.25$.

This information describes the condition (i.e., severe AECOPD), provides context (i.e., possible ARF), and includes details (i.e., ABG measurements) for determining the patient's state of health. This granularity can be hidden within EHR data and current tools do not expose this type of information.

Health States

We use clinical domain knowledge to define ***health states*** that represent the clinical conditions in which a patient can be categorized. For our approach, this knowledge derives from recommendations in the GOLD clinical guidelines as well as input from a COPD expert [1] on our team of researchers to identify and define health states on the basis of established clinical criteria in terms of the variables listed in Table 5.2 [2]. Health states (e.g., Type 1 ARF) are identified and defined by setting constraints on time-varying clinical variables (e.g., $PaCO_2$, PaO_2, pH). They are represented by a numerical state identifier (column 1 in the table), which can be used to obtain the text name from the dictionary of health states (i.e., definitions of

[1]Marvi Bikak, MD, Indiana University School of Medicine, Indianapolis, Indiana, USA

Table 5.2.: Description of health and outcome states.

State Identifier	**State Name**	**State Type**	**Definition**
0	NARF	Health	no ARF
1	ARF_1	Health	$PaO_2 \leq 60$ $PaCO_2 \leq 50$
2	ARF_2	Health	$PaCO_2 > 50$ $pH \leq 7.25$
			PaO2≤60 $PaCO_2 > 50$ $pH \leq 7.25$
3	Discharge	Outcome	discharge from hospital
4	Death	Outcome	in-hospital death

clinical conditions) provided by the API. For example, Type 1 ARF, a defined health state, is given the name, ARF_1, and can be obtained using the numerical identifier *1* from within the code.

Outcome States

Outcome states, similar to health states, are also represented by a numerical identifier. However, outcome states, such as death and discharge, are defined without the need for clinical knowledge. They are used to describe the final state of a patient when they are released from the care setting, indicating the completion of the patient's hospital stay.

5.3.2 Data Aggregation: Patient Histories

Each individual patient in MIMIC has unique characteristics and properties associated with their hospital stay and other medical encounters. This information is scattered across multiple tables and can be difficult to aggregate. For example, a single patient can have records for multiple hospital stays and each hospital stay can have several ICU stays, where data is collected for each specific stay. Thus, extracting data for a single patient can be a convoluted process. Figure 5.1 is an illustration of this process regarding data extraction for a single patient with multiple hospital stays in MIMIC, as well as the tables that must be queried to obtain the desired information.

From these various tables, we curate data from the MIMIC-Purdue DB using the MIMIC-Purdue data extraction and preprocessing API (Chapter 4). The flexibility of MIMIC-Purdue extraction and preprocessing API allows for simpler extraction that would otherwise be a convoluted or impossible task with existing systems [55, 145]. Our approach aggregates this scattered data for individual patients and organizes ***patient histories***. Patient histories are comprised of various data points such as

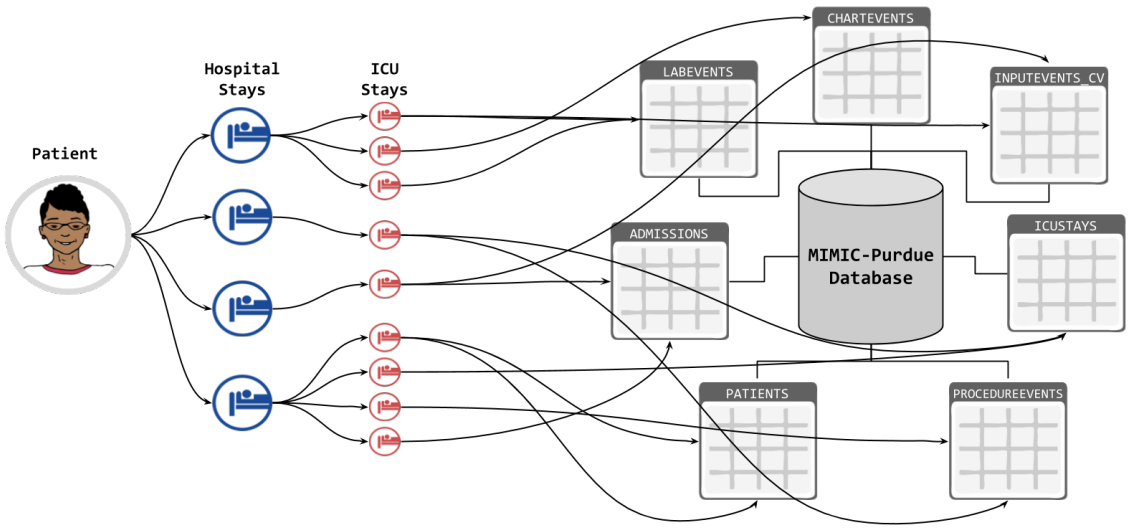


Figure 5.1.: Conceptual process of curating data for a patient.

demographics (e.g., ethnicity, gender), clinical measurements (e.g., heart rate, blood pressure), and other information documented during a patient's hospital stay.

Data Structure

We develop an object in Python (using a Python class), *Patient*, to construct and structure patient histories. This approach allows us to create multiple objects of the Patient class, each representing a unique patient and storing different values for their individual characteristics and properties. An instance of the class is created by passing a subject_id to initialize the instance. Once initialized, class methods and other functions from the MIMIC-Purdue extraction and preprocessing API are called to extract and store both static and time-series information for the corresponding patient. Figure 5.2 illustrates the process of extracting data from MIMIC-Purdue and aggregating the two data types for a patient.

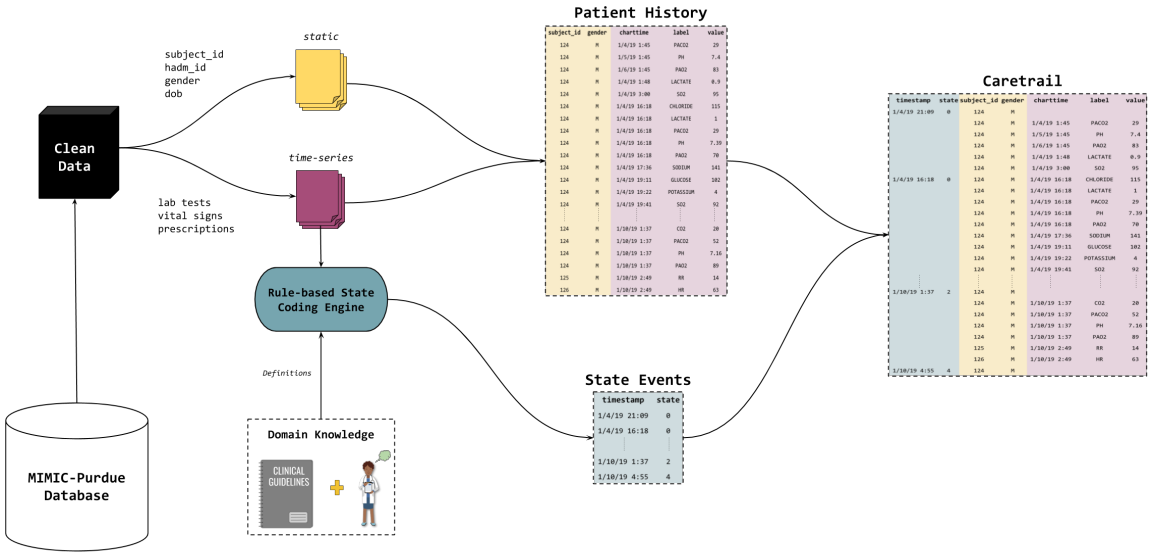


Figure 5.2.: Flow of data aggregated with the *Patient* class and identification of health and outcome states for caretrail creation.

Data Types

Patient histories contain two types of data: (1) *static* and (2) *time-series* (i.e., discrete and continuous). Static data includes data points that do not change, such as ethnicity and date of birth. Time-series data consist of any data point with an associated timestamp, such as measurements for vital signs and administration of medication. Each hospital admission for a patient in MIMIC has a timestamp field (i.e., `dischtime` and `deathtime`) that are used to determine if the patient was discharged from the hospital or died within the given hospitalization session. Though these types of data points do not change, they are unique and only recorded once for each hospital admission. Thus, such data points are considered to be both static and time-series, as they do not change and have a corresponding time-stamp.

Time-series data can be discrete or continuous. For example, in MIMIC blood pressure in measured at least once every hour for patients whose records contain such measurements [99]. There is a timestamp and value associated with each blood pressure measurement, as both the timestamp and measurement values change over time. Thus, blood pressure measurement is considered continuous time-series data. Antibiotics such as *vancomycin*, may be administered to patients periodically during their hospital stay, where each dose administered is associated with a timestamp. A patient may be given various dosage amounts of the same drug, which can be viewed as continuous data. However, for simplicity and to focus on clinical measurements and events, interventions that are considered to be infrequent (e.g., medications, meals, hygienic care) are considered to be discrete data.

5.3.3 Caretrail Generation

Caretrails are generated by encoding patient histories with clinical domain knowledge that is represented by health as well as information contained in the EHR (e.g., hospital discharge time) represented by outcome states. Once health states have been defined, given a combination of clinical variable constraints, the corresponding clinical

variables are first located in the MIMIC-Purdue DB and included in patient histories. As discussed in Section 2.2.5, some clinical variables are measured repeatedly and scattered across several tables. For example, some clinical measurements are duplicated between the `LABEVENTS` and `CHARTEVENTS` tables, and some are only found in one of these tables. Even more, each of these tables has a corresponding dictionary table that contains definitions for all measurements and items in the `LABEVENTS` and `CHARTEVENTS` tables. We process this information by aggregating both tables and removing duplicate measurements as well as relabeling multiple labels of the same measurement to have the same name (e.g., respiratory failure, RR). We extract the time-series data (i.e, data points with corresponding timestamps) from the patient histories to identify health and outcome states within the patient histories. Patient histories are then encoded with states, forming temporal paths, represented by pairs of ordered timestamps and state sequences. Figure 5.2 illustrates the process of generating a caretrail by aggregating data with the *Patient* class, and processing the time-series data to identify health and outcome states.

5.3.4 Patient Cohorts

The *Patient* class allows us to create multiple objects for each patient. Aggregating an individual patient's EHR data with other patient's EHR data creates patient cohorts. Members of an individual cohort share a common characteristic that matches a DB query. This is done by using functions from the MIMIC-Purdue extraction API to identify and extract patients based on similar characteristics and their unique identifiers. The identifiers are then used as inputs for the *PACE* tool to create *Patient* objects and generate corresponding caretrails for each individual patient to form a cohort of patients for conducting studies. For example, identifying and extracting information for distinct patient populations using ICD-9 codes has been shown to have good recall, precision, and specificity [147]. The MIMIC-Purdue API has a function that returns a list of subject_ids for patients who have a given ICD-9 diagnosis asso-

ciated with their hospital stay. The results returned by the function can be stored in a list where each element is an instance of the *Patient* class. Each *Patient* object corresponds to a `subject_id` that is returned from the ICD-9 query and the list of *Patient* objects is a patient cohort, with each member sharing the same ICD-9 code. Additionally, both query results and patient objects can be saved for later access and analysis. Caretrails generated for individual patients as well as patient cohorts are "pickled" to save them to disk.

5.3.5 Visualizations

All visual representations of the data that are extracted and aggregated are developed in Python. These visualizations of patient data are created to generate a comprehensive view of EHRs to help medical professionals understand the data. Visualizations include both longitudinal and descriptive plots as well as graphs that display statistics regarding both individual patients and patient cohorts. Interactive, graphical timelines are also created to provide a holistic view of patient histories and caretrails. Visualizations of patient histories are represented by timelines that include clinical measurements (e.g., vital signs, labs) and interventions (e.g., drugs administration). Clinical measurements and interventions are plotted using line graphs and scatter data points, respectively. Visualizations of caretrails also include color-coded health and outcome states as vertical lines or bars. These states are placed in the background, while data contained in the patient history is included as an overlay (see Figure 5.6 below for a caretrail example).

5.4 Caretrail Analysis

To analyze the potential impact of caretrails, we aggregate clinical data for in-hospitalized AECOPD patients. We create visualizations of granular clinical data and observe changes between health states and assess the care process and progression of exacerbations in regards to treatments and outcomes.

5.4.1 Patient Selection

We construct complex queries using the MIMIC-Purdue extraction and preprocessing API to identify AECOPD patients. Each patient visit is associated with multiple diagnoses, which are labeled according to ICD-9 diagnosis codes and ordered by priority. While there are a number of COPD-related ICD-9 codes, there are only three that are used to identify AECOPD-related visits [167,168]. We define a primary diagnosis of AECOPD as a hospital admission having at least one of the following ICD-9 codes as either primary, secondary, or tertiary diagnosis : 491.21 (obstructive chronic bronchitis with acute bronchitis), 491.22 (obstructive chronic bronchitis with acute exacerbation), and 494.1 (bronchiectasis with acute exacerbation) [9]. Inclusion criteria was limited to these three codes to analyze patients with ARF secondary to AECOPD primarily, without other confounding etiologies that may be present. A total of 697 patients satisfied this criteria (Figure 5.3).

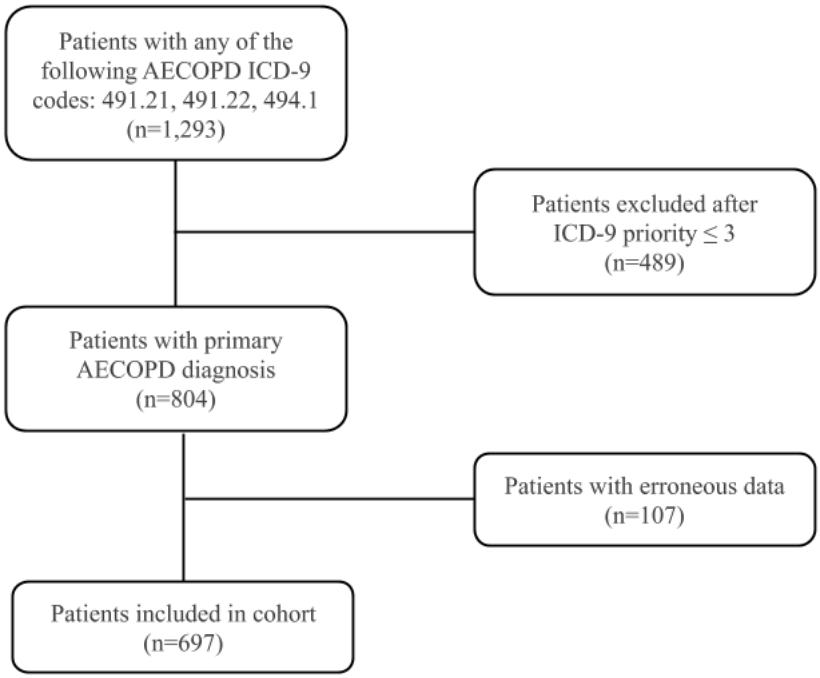


Figure 5.3.: Cohort selection criteria for AECOPD patients.

5.4.2 Descriptive Visualizations

Descriptive visualizations provide summary representations of the AECOPD patient cohort being analyzed. The overall death rate for the entire cohort of 697 patients was 8.5%. 181 of the patients were administered antibiotics and had a death rate of 4.4%, and 516 were not administered and had a death rate of 9.9%. Figure 5.4 is a histogram depicting the initial time in hours from admission for AECOPD patients that were administered antibiotics. This illustration reveals that on average, patients were administered antibiotics within the first 27 hours of their hospital admission.

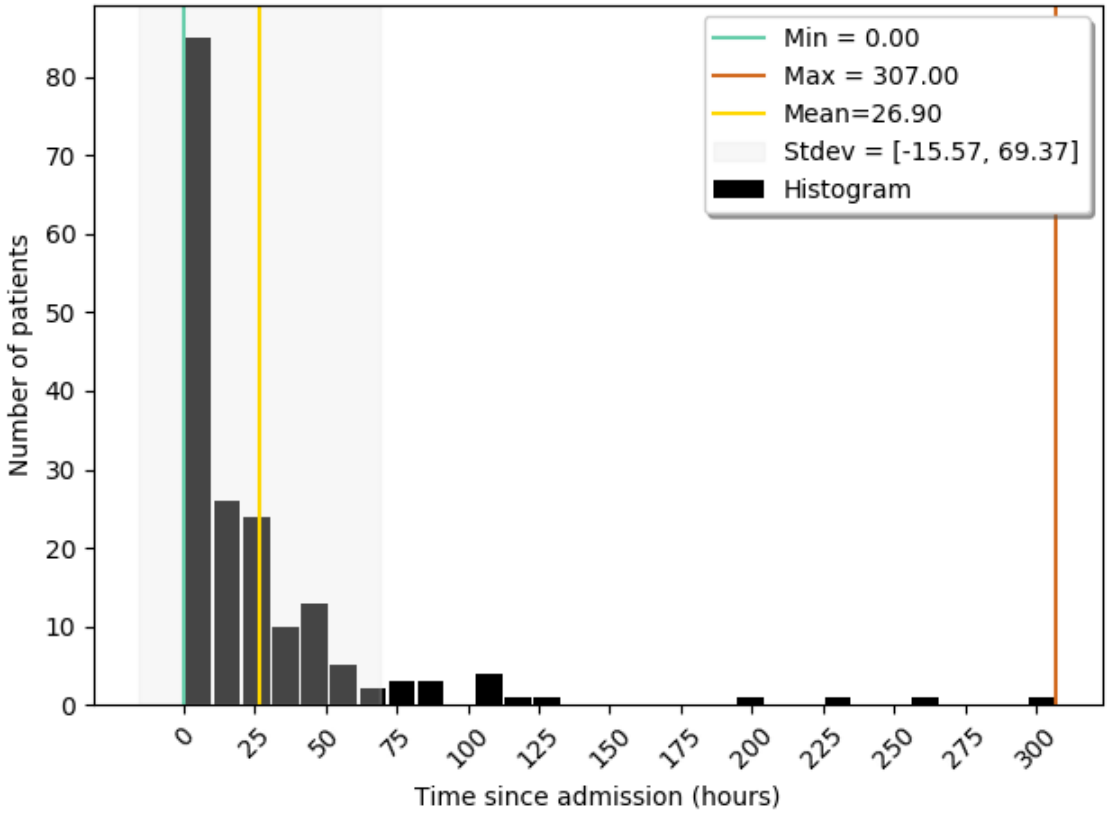


Figure 5.4.: Histogram of initial timing of antibiotics as hours since hospital admission.

5.4.3 Longitudinal Visualizations

Figure 5.5 is a visual representation of a patient history for an AECOPD patient. This illustration provides a view of clinical measurements (i.e., PaO_2, $PaCO_2$, and pH) that are used to classify an exacerbation, additional variables that are used to assess a patient's condition, and the times antibiotics were administered. The additional variables include: heart rate (HR), systolic blood pressure (SBP), respiratory rate (RR), and carbon dioxide (CO2). The figure reveals a significant decrease in PaO_2, which is indicative of ARF.

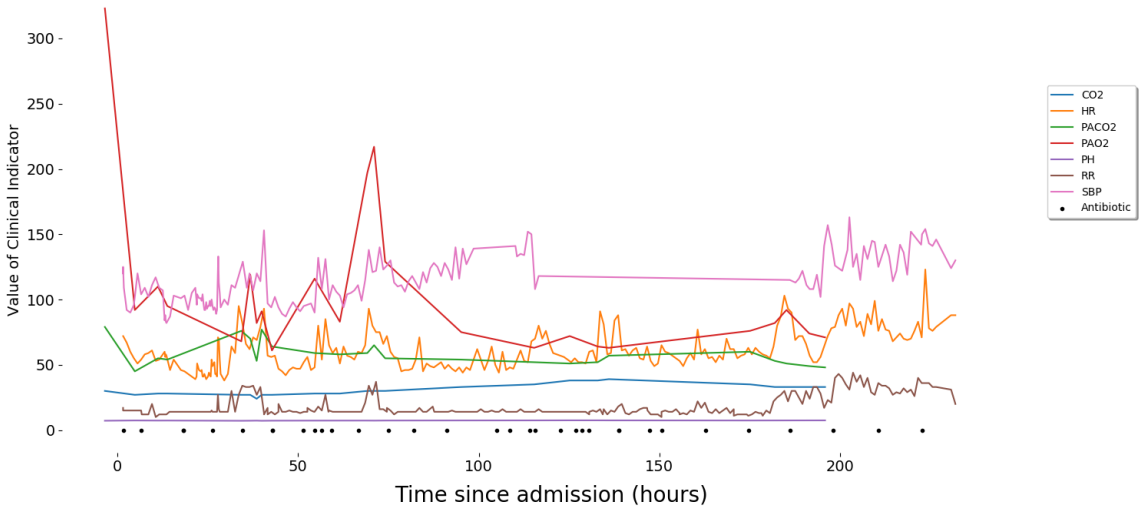


Figure 5.5.: An AECOPD patient history, including clinical measurements, represented by horizontal lines, and administration times of antibiotics, represented by small data points.

Similar to Figure 5.6, this same information is illustrated as a caretrail for the same patient. As depicted in Figure 5.6, the caretrail provides more granular details that include color-coded visuals of the health state of the patient at the exact time the antibiotic was administered. The caretrail reveals that the patient was in the ARF_2 state when their PaO_2 significantly decreased. The patient began in the NARF state,

changing between NARF and ARF_2 throughout their hospital encounter, with a final outcome of hospital discharge.

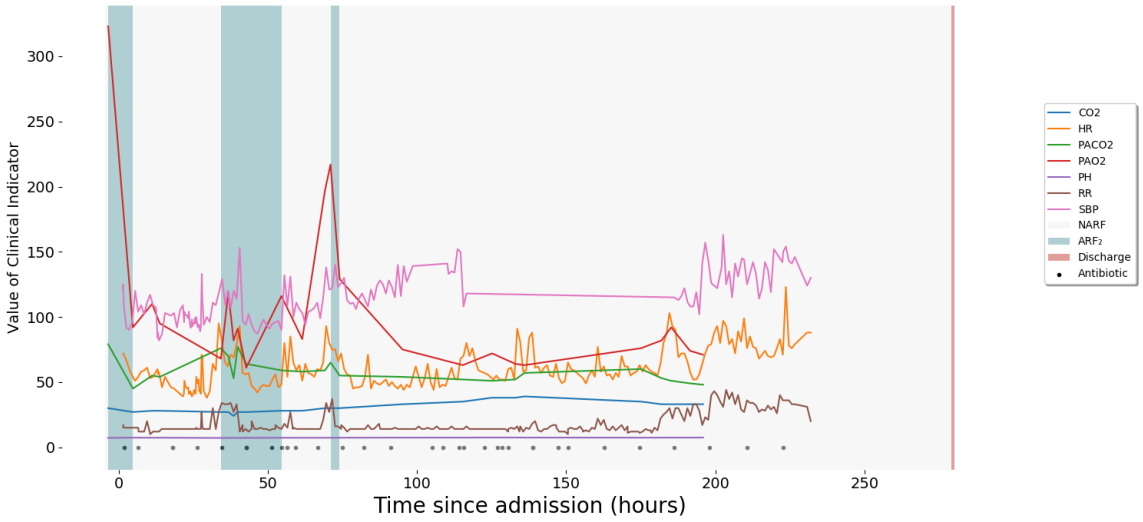


Figure 5.6.: Caretrail for an AECOPD patient. Health states are represented by vertical lines. Vital signs are represented by horizontal lines plots, and administration times of antibiotics are represented by small data points.

5.5 Discussion

Aggregating retrospective data can be valuable for understanding what happened in the past, providing insight for more advanced analytics to predict what will happen in the future. We present a novel tool for constructing caretrails by aggregating retrospective EHR data, integrating clinical domain knowledge to define health states, and providing visualizations of granular timing information through enhanced representations of EHR data. Our tool not only aggregates data for individual patients, but is also equipped with features for aggregating data for patient cohorts that can be stored for reuse. While previous systems [23, 148] have features for identifying patient cohorts, our tool extends beyond the capabilities of these systems by allowing cohorts to be identified from user-defined data acquisitions (i.e., complex queries)

and storing the extracted data for reuse. We demonstrate this capability by identifying AECOPD patients and specifying the granularity of the clinical diagnosis (i.e., primary). Additionally, the tool provides a color-coded scheme that correspond to the states produced by the rule-based state coding engine. This feature enhances the timeline visualizations by highlighting the health state of a patient during various observations and interventions.

Caretrails can be used by medical professionals and analysts to observe trends in EHR data. Figure 5.6 shows a caretrail for a patient who was in the ARF_2 health state when antibiotics were initially administered. For example, the analysis of the caretrails revealed that AECOPD patients who were administered antibiotics during their hospital stay had a lower mortality rate than those who were not. Using this information gathered from the data, the caretrails can be used as input for computational models for predicting patient outcomes (e.g., mortality) based on antibiotic administration, or even more, the timing of antibiotic administration. Previously, these trends would require some medical knowledge to read and understand the output. Moreover, capturing the granularity of such trends potentially required deep understanding of the data and technical skills for processing the data.

5.5.1 Limitations

The methodologies presented can also be applied to both common and more complex clinical conditions. That is, we defined health states that encompassed several clinical variables. However, although we defined states based on clinical experience and clinical guidelines, given the vast space of clinical conditions and corresponding clinical criteria that defines the conditions, some health states may not be well represented. For example, in our state space NARF is defined as no acute respiratory failure, which means that the only criteria that determines if a patient is in NARF is that the constraints used to define ARF_1 and ARF_2 are not met. In this case, NARF does not reveal other conditions of the patient. To address this, information

from various data sources such as past research studies [23], clinical trials [169], and medical devices can be combined to define additional clinical states. Time-variant measurements (e.g., vital signs and labs) used to identify health states are susceptible to missing values and varying timestamps. ABGs are measured by obtaining invasive arterial blood samples [170]. Thus, the clinical variables (i.e., ABG measurements) we used to define health states are typically measured simultaneously. However, when defining other clinical conditions, it is not always the case that the measurements used for defining the condition are measured simultaneously. Thus, creating additional health states may require mechanisms to handle clinical variables that are not measured together. Also, if there is a value present for at least one of the clinical variables, we assume the other variables were not measured. This is a limitation of missing data, which is common in using retrospective clinical measurements for analysis. Applying approximations such as grouping multiple measurements for variables not measured simultaneously through window-based segmentation methods (e.g., fixed-sized time windows), as well as imputing missing data are techniques that have been used to mitigate such limitations [28, 171, 172].

5.5.2 Future work

The work presented in this chapter can be extended to include functionality for enhancing the rule-based state coding engine for automatically defining and discovering additional health states. The rule-based state coding engine can be equipped with definitions for known clinical conditions, and advanced machine learning techniques such as deep learning can be used to potentially discover new conditions. Patient histories can be processed through the engine to automatically identify health states within a patient history. Another opportunity for extending this work is to extend the interactive visualizations to allow direct manipulation of caretrails through interactive dashboards. Such dashboards can provide visualizations for clinical variables that are specified by selecting from a list of available clinical measurements.

6 MARKSIM: TIME-BASED MODELING AND SIMULATION

6.1 Introduction

Existing software systems [22, 23] do not allow for longitudinal processes such as chronic disease progression to be observed directly, nor do they support analysis of how patients transition from one health state to another as a direct cause of an intervention (e.g., drug administration, oxygen therapy, surgery). Several previous studies have developed approaches that identify medical trends for a patient population with chronic conditions [25, 173, 174]. Markov chain models are one approach that have been used to estimate healthcare costs, utilization, and disease progression over time [25–27]. They provide support for decision-making under uncertainty by approximating patient transitions through a set of "health states," each of which corresponds to a clinical event [24]. For example, Bartolomeo *et al.*, [175] leveraged Markov model techniques to model patient care pathways, defining states based on age, gender, clinical status, and information from the GOLD guidelines. Sood *et al.*, [176] also used the GOLD guidelines to asses the diagnosis of COPD using the spirometry test.

While using Markov chains has been helpful for critical illnesses and chronic conditions such as COPD, directly using EHR data in Markov models presents certain challenges. In particular, Markov chains assume a one-step time-invariant transition. The assumption that values are collected with some fixed sampling frequency is typical with such time-series models [28]. However, EHRs contain highly-dimensional, time-variant, clinical data observed at irregular time intervals. Hence, the nature of EHR data limits the ability to represent the granular timing of transitions, which can lead to process misrepresentation in the model.

We present a novel approach that systematically captures the effect of interventions during medical encounters, and hence, may support evidence generation for clinical guidelines in a systematic and principled way. A Markov Chain Monte Carlo (MCMC) modeling and simulation package, *MarkSIM*, in Python that encodes clinical conditions as computable definitions of health states using raw EHR data is described. The MCMC package calculates transition probabilities (or model parameters) as a function of time using exact timing information extracted from patient caretrails to construct Markov Chains. Monte Carlo sampling methods are then applied to the model parameters to perform simulations. We use this approach to estimate the efficacy of antibiotics treatment for in-hospitalized AECOPD patients. We demonstrate the functionality of *MarkSIM* by modeling the pathway of AECOPD using retrospective EHR data collected during critical care encounters. Recommendations from the national GOLD guidelines [2] and physicians are incorporated to define model components. We estimate outcomes for two questions related to antibiotic treatment for AECOPD patients, namely: i) the impact of antibiotic administration on in-hospital death and ii) the impact on in-hospital death based on the initial timing of antibiotic administration. We perform Monte Carlo simulations, which are validated by comparing estimated model outcomes to actual outcomes from the EHR data, and conduct sensitivity analyses to evaluate the robustness of model projections.

6.2 Related Work

Several traditional algorithms and modeling techniques have been developed for time-series data, data observed at unequal time intervals, such as EHRs [122, 172]. Such data points are irregularly sampled within individual patient records and across other individual patients. Past techniques typically abstract patterns using state representations (e.g., mild, moderate, severe) [122]. Sherman *et al.*, [177] extracted longitudinal clinical data as irregularly sampled, timestamp features to predict in-hospital mortality and hypokalemia. Shah *et al.*, [172] proposed a finite-state machine-based

approach for predicting COPD exacerbations, focusing on telemonitoring of patients. They used a combination of three vital signs (i.e., peripheral capillary oxygen saturation (SpO_2), respiratory rate, and pulse rate) to define states that reflect seven day time periods. Their work found that using combinations of variables as such can be used to define model symptoms of exacerbations and improve the understanding of how symptoms worsen, leading to exacerbation events. Their techniques extract the temporal dynamics of the data by grouping the vital sign measurements into sets and applying least squares criterion to each set in seven day intervals. This allowed them to predict the length (i.e., number of days) of an exacerbation. Their ideas can be leveraged to capture more granularity by decreasing the time period in which they observe vital sign measurement sets to capture the exact time a patient enters a particular state in the model.

Dynamic Bayesian networks have been used to model EHRs for forecasting patient outcomes [178], reconstruct medical states of patients using lab tests [179], and analyze interactions between predictors and outcomes [122]. Inoue *et al.*, [180] proposed an approach for modeling disease progression based on the age at which the disease was detected and disease status at diagnosis. Huang *et al.*, [181] combined Bayesian hidden Markov model methods and hierarchical clustering to group patient paths by related treatment behaviors. Although they used timestamp information to define treatment events in inpatient EHRs as a timestamp and event type pair, they made the assumption that each event was regularly recorded. Such assumptions dismiss the idea that clinical time-series data is recorded irregularly. Bayesian networks are atemporal in that they only consider single points in time and lack the ability to capture granular time information [122].

Markov chain models have also been used to model temporal data. Wang *et al.*, [182] presents a Markov jump process to model transition behaviors between disease stages in COPD patients. Bueno *et al.*, [183] identified clusters of states by using hidden Markov models, associating similar observations and defining transition patterns. Rodina-Theocharaki *et al.*, [184] developed Markov chains to predict the number of

end-stage renal disease patients. They defined state-based treatment interventions and estimated model parameters using statistical data extracted from annual public health reports that spanned nearly a decade. The authors [26] further extended their work by combining the Markov chain models with Monte Carlo techniques to include cost calculations for treatments, based on predictions from the previous work. They developed MCMC methods to predict the number of patients with end-stage renal disease and perform a cost-effectiveness analysis for renal replacement therapy treatment. In both studies, the modeling approaches were applied to temporal data that represented equal time intervals. While their modeling techniques are suited for such data, temporal data such as time-series observations are among the most challenging models to develop [185].

Another technique to handle irregularly sampled data is by converting data into observation sequences and using a window-based segmentation approach, which segments time-series data to fixed-sized windows [28]. This approach was introduced by Liu *et al.*, [28] who developed models by combining a linear dynamical system and a Gaussian process model by defining time windows and extending model methods to treat observations as a function of time. Similar to other work discussed, the limitation of their work is that their approach only works with univariate time-series data.

Previous modeling strategies that address the task of analyzing and mining temporal, time-series data are limited by their ability to work with multivariate time-series data observed at irregular intervals [186]. Such strategies mainly support irregular univariate time-series data and target providing insight to questions where time is not relevant. Thus, more approaches that can capture and model the granular temporality of multiple observations collected over uneven time periods are needed.

6.3 Methodology

6.3.1 Markov Chain Model Formulation

ChainX, the Markov chain modeling component of *MarkSIM*, is a Python class that contains methods for structuring a Markov chain model. The Markov chain model approach is formulated by defining a set of states and computing transition probabilities to represent the likelihood of moving between states. There are two types of model states:

1. *Health states* are mutually exclusive (transient) states that are defined using clinical domain knowledge. These states are defined using clinical domain knowledge to represent the clinical conditions in which a patient can be categorized.

2. *Outcome states* are absorbing states that are defined without the need for clinical knowledge. They are used to describe the final state of a patient when they are released from the care setting, indicating the completion of the patient's hospital stay. Thus, once a patient enters an outcome state, no other transitions can occur.

An instance of the *ChainX* class is created by passing a model name (optional), the total number of states, a list of identifiers for health states, and a list of identifiers for outcome states. Table 5.2 describes health and outcome states, as well as the corresponding identifiers.

Input Data

Once a *ChainX* object is created, caretrails (Section 5.3.3) are generated from longitudinal EHR data extracted from the MIMIC-Purdue DB to form granular, temporal sequences to use as inputs for the model. The temporal sequences extracted from caretrails are represented by a chronological set of events, $h_k = \{e_0, e_1, \ldots, e_n\}$,

that occurred during a patient's hospital stay. Each event is represented by a pair, $e = (s_i, t_i)$, where t_i is the time (i.e., timestamp) the state, s_i, was entered by patient, k.

Transitions

Generally, Markov models are developed with time-invariant parameters such as age – a standard time-parameter across patients and EHRs. Our time-variant data includes clinical observations that are measured at different times during a patient's hospital encounter and across patient populations. Timestamps from the granular temporal sequences, $H = \{h_k, h_{k+1}, \ldots, h_n\}$, extracted from the caretrails are used to compute transitions as a function of time using class methods, giving us the flexibility needed to model temporal data with irregularly sampled observations. The sequences are aggregated to construct the state transition matrix as:

$$T = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \ldots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \ldots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{d1} & x_{d2} & x_{d3} & \ldots & x_{dn} \end{bmatrix} \tag{6.1}$$

where $N = \{0, 1, 2, \ldots\}$ denotes the total number of states and x_{ij} denotes the total number of hours spent in state i before moving to state j (i.e., a transition). That is, for every sequential pair of transition events for a patient, we calculate the actual time between them in hours as $(t_{i+1} - t_i)$.

Transition Types

There are two possible transitions illustrated in Figure 6.1 that can occur. The occurrence of each of the transition types triggers a different action for updating T. For simplicity, we use *ChainX* attribute names to explain the action that is triggered when a transition occurs. These attributes names and descriptions are:

- `current_state` - current state patient is in (also denoted by s_i)
- `current_time` - timestamp corresponding to current state (also denoted by t_i)
- `next_state` - state to which a patient transitions (also denoted by s_{i+1})
- `next_time` - timestamp corresponding to next state (also denoted by t_{i+1})
- `continued_time` - temporary attribute to store timestamp when a transition occurs that results in the patient remaining in the same state (i.e., type 1 transition)

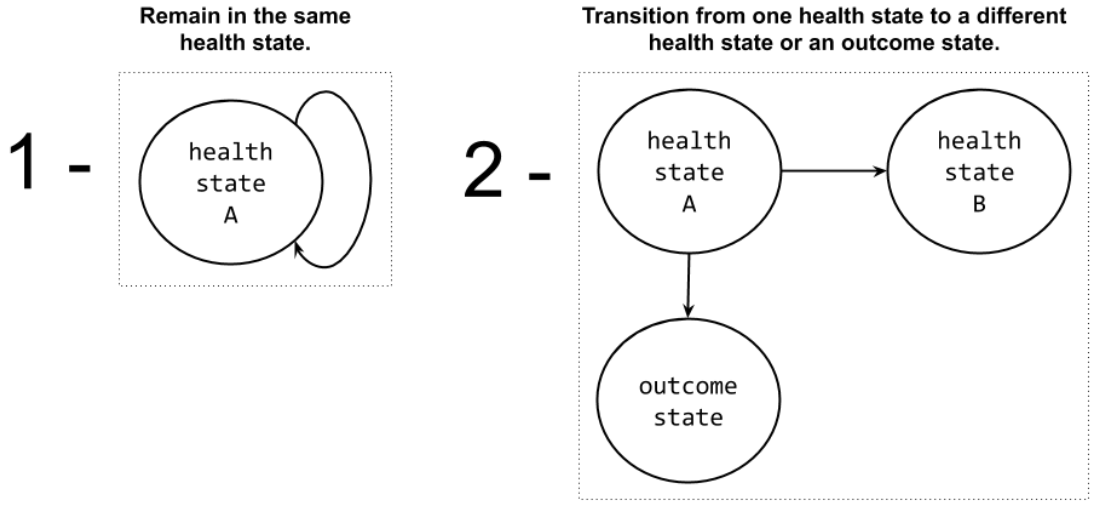


Figure 6.1.: Possible transitions for Markov chain model.

For transition type 2, the difference in hours of the `current_time` and `next_time` is calculated to update T, and the `current_state` and `current_time` attributes are updated to reflect the values of `next_time` and `next_state`. Type 1 transitions are more complex. Several attributes are maintained to obtain the full duration that a patient spends in a state. When the patient enters a state consecutively, the `current_time` and `current_state` attributes are not updated until a type 2 transition occurs. The `continued_time` attribute is updated to the timestamp of

each subsequent transition. For example, consider the following temporal sequence for a patient:

$$h_0 = \{(2,\ 1/9/19\ 3{:}00),(0,\ 1/9/19\ 5{:}00),\ (0,\ 1/9/19\ 6{:}30),(3,\ 1/9/19\ 7{:}00)\}$$

The patient first starts in state 2, and the `current_time` and `current_state` attributes are set to 2 and 1/9/19 3:00, respectively. T is not updated, as the patient has yet to make a transition. Following the sequence, the first transition is of Type 2, where the patient enters state 0. T is updated as $x_{20} = x_{20} + (\texttt{next\_time} - \texttt{current\_time})$, and the `current_state` and `current_time` attributes are set to 0 and 1/9/19 5:00, respectively. The next transition is a type 1 transition, and the patient remains in state 0. Here, T is updated as $x_{00} = x_{00} + (\texttt{next\_time} - \texttt{current\_time})$, and continued_time is updated with the value of `next_time` (i.e., 1/9/19 6:30). The next and final transition in the sequence is of type 2, and the patient enters state 3. T is updated as $x_{03} = x_{03} + (\texttt{next\_time} - \texttt{current\_time})$. Note that the value of `current_time` is 1/9/19 5:00, which is the initial time that the patient entered state 0. Because state 3 is an outcome state, another caretrail for a new patient is aggregated, and the steps for computing T are repeated until the last caretrail is aggregated. Listing 6.1 is a Python code snippet, demonstrating the programming logic for handling the occurrence of both transition types.

```
1 def transitionX(next_state, next_time):
2     if current_state == next_state: # Type 2 transtion
3         if continued_time is not None:
4             T[current_state][next_state] = next_time -
                  continued_time
5             continued_time = next_time
6         else:
7             T[current_state][next_state] = next_time -
                  current_time
8             continued_time = next_time
```

```
9     else: # Type 2 transtion
10        T[current_state][next_state]= next_time - current_time
11        current_state = next_state
12        current_time = next_time
13        continued_time = None
```

Listing 6.1: Example Python code for handling transitions.

Transition Probabilities

A transition probability matrix (TPM) for the Markov model is estimated using the ratio of the number of hours spent in a specific state before transitioning over the total number of hours spent in that state, which is extracted from T (i.e., the transition matrix). Each transition probability (or model parameter) represents the estimated likelihood of a patient changing from one state to another during their hospital stay. Specifically, we calculated transition probabilities by:

$$p_{ij} = \frac{x_{ij}}{\sum_{i=0}^{n} x_i} \tag{6.2}$$

where p_{ij} is the probability of transitioning from state i to state j at any given time. We set p_{ij} equal to the percentage of hours spent by individuals in state i and ending in state j relative to the total amount of time spent in state i. Note that $p_{ii} = 1$ for an outcome state. Once the TPM is created it is stored in a *ChainX* attribute for later use in the simulation. While the TPM are stored as matrices, *MarkSIM* includes functionality to generate visualizations of state transitions from the TPM. Figure 6.2 is a diagram for a 5-state Markov chain model, generated using *MarkSIM*.

Initial State Probabilities

The initial state counts are stored in a *ChainX* list, where each index represents a health state. Once all caretrails gave been aggregated for building the model, the list

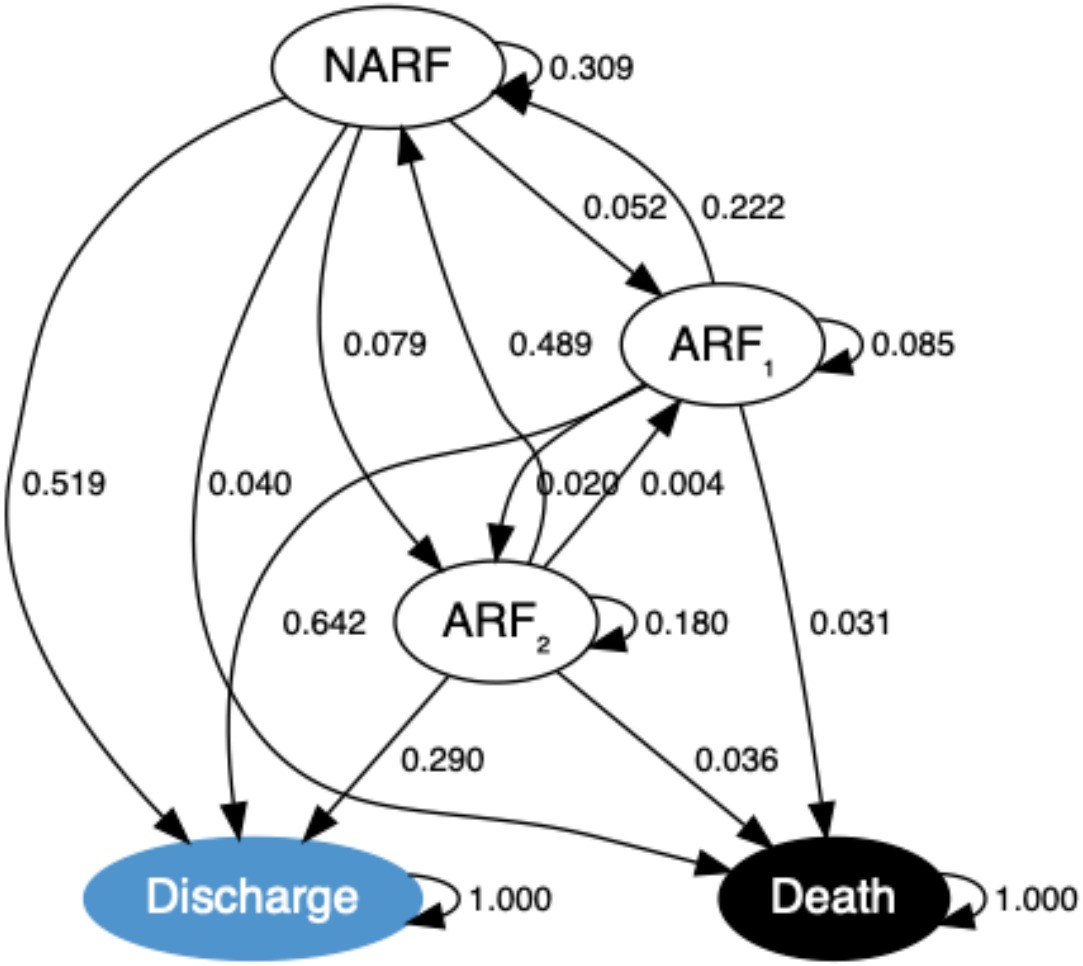


Figure 6.2.: Visualizations generated by *MarkSIM* of a 5-state Markov chain model for AECOPD patients.

is used to calculate initial state probabilities. The model has a corresponding initial probability vector that represents the quantities of patients in each health state at the beginning of their caretrail.

6.3.2 Monte Carlo Simulation

Monte Carlo simulations are performed to estimate the outcomes for patients based on clinical questions. *SiMzy*, the Monte Carlo simulations component of *MarkSIM*, is a Python class that contains methods for structuring a simulation. The class instantiated by passing a *ChainX* object of a previously built Markov chain model. The Monte Carlo approach simulates the transitions of individual patients over time. Simulations for patients are terminated by specifying an outcome state or the number of transitions for each patient. For each model generated, the simulation is run for 100 replications (or cohorts), each of size 1000 patients. Each patient is assigned to an initial health state based on the initial state probabilities. The occurrence of transitioning between two states is determined by generating a uniform random number on the interval [0,1] and selecting the next state based on the possible range of values from the transition probabilities. Because the probability of a patient leaving an outcome state is zero, once a patient enters an outcome state, the process ends, and a new patient is simulated. The mean, minimum, maximum, and variance of the patient outcomes for the 100 replications is calculated and used for simulation statistics and model validation. To validate the model, we obtain a baseline from the raw EHR data and compare the observed outcomes (mean and variance) to results estimated from the simulation.

6.3.3 Sensitivity Analysis

Probabilistic sensitivity analysis is conducted to quantify the effects of changes in model parameters (i.e., transition probabilities) and quantify confidence levels. We use variance reduction to assess the variability of model parameters by changing their

values. A small change randomly chosen between plus or minus 10% of each transition probability value is applied to the model parameters using,

$$p'_{ij} = (p_{ij} - (.1 * p_{ij})) + (r * (.1 * p_{ij})) \tag{6.3}$$

and normalized using the row sum for each state. A new TPM is obtained by normalizing the generated values so that the row sums equalled 1. The outcomes are then re-computed from the simulation using the new TPM.

6.4 MarkSIM Evaluation

We evaluate the value of *MarkSIM* by developing models and performing simulations to answer both atemporal and temporal clinical questions. Formulating clinical questions and identifying relevant results are steps used to conduct clinical studies. Results from such studies are evaluated by clinical experts and used as recommendations for clinical guidelines. These questions are well-formulated to provide a basis for collecting and analyzing data to derive answers. We use retrospective EHR data, and thus, must structure the data to fit the question as opposed to structuring the question to fit the data. We pose the following two questions of interest:

i. What is the impact of administration of antibiotics on death for hospitalized AECOPD patients?

ii. What is the initial timing of antibiotics for hospitalized AECOPD patients and the impact on death?

AECOPD has been widely studied as a major outcome in clinical studies and research [172]. While these studies have been mainly focused on identifying and predicting COPD exacerbations, we are interested in analyzing the progression of an exacerbation and the impact of treatment interventions on outcomes.

6.4.1 Cohort Selection

We construct complex queries using the MIMIC-Purdue extraction API to identify AECOPD patients. Each patient visit is associated with multiple diagnoses, which were labeled according to ICD-9 diagnosis codes and ordered by priority. While there are a number of COPD-related ICD-9 codes, there are only a few that are used to identify AECOPD-related visits [167,168]. We defined a primary diagnosis of AECOPD as a hospital admission having at least one of the following ICD-9 codes [9] as either primary, secondary, or tertiary diagnosis:

ICD-9 Code	**Diagnosis**
491.21	Obstructive chronic bronchitis with (acute) exacerbation
491.22	Obstructive chronic bronchitis with acute bronchitis
494.1	Bronchiectasis with acute exacerbation

Inclusion criteria is limited to these three codes to analyze patients with ARF secondary to AECOPD primarily, without other confounding etiologies that may be present. Based on this selection criteria (Figure 6.3), we identified 697 unique AECOPD patients with at least one ICU admission.

Antibiotics Treatment

We compile a comprehensive list of antibiotics names and codes used to treat COPD by extracting information from the NDC directory [167] and the National Library of Medicine's RXNorm database [187]. We perform an exhaustive search within the group of AECOPD patients to compile another list of antibiotics that were used to treat the patients. These two lists are compared to identify patients who were administered antibiotics that are specifically used to treat COPD. Once compared, a COPD clinical expert verified the list. The antibiotics used to treat a subset of patients in our cohort were: *doxycycline, azithromycin, levofloxacin, oseltamivir, vancomycin, trimethoprim, fluoroquinolones, vibramycin, ofloxacin, clarithromycin, telithromycin, amoxicillin, cefuroxime, pipericillin-tazobactam, cefepime.*

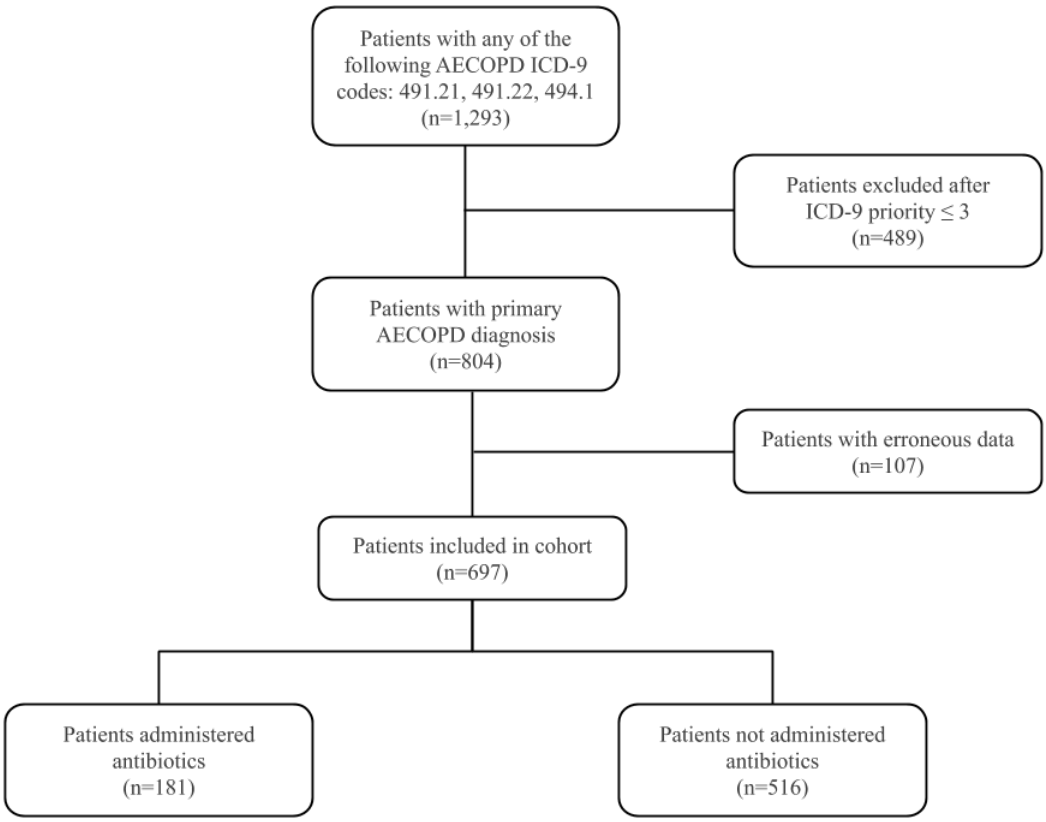


Figure 6.3.: Cohort selection criteria for AECOPD patients administered and not administered antibiotics.

6.4.2 Model Formulation

The Markov chain model consists of three health states: NARF, ARF_1, and ARF_2, and two outcome states: discharge and death. Clinical criteria used to define the health states are provided in Table 5.2. For the purposes of this research and after consulting with a clinician, the definition for ARF_2 encompasses clinical variable constraints for type 2 (hypercapnic) and/or type 1 (hypoxemic) and type 2 respiratory failure. Figure 6.2 is an illustration of the 5-state first-order Markov model for AECOPD based on clinical criteria. Nodes represent health and outcome states. Health states are colored white and appear above the outcome states, which are blue and black. Edges are labeled with values that represent the probability of returning to the same state or transitioning to a different state. Once an outcome had been entered, transitions can no longer occur. Thus, edges returning to outcome states are labeled with probability, 1.000. Section 5.3.1 discuses the details of defining and identifying health and outcome states.

6.4.3 Clinical Question 1: Antibiotics Treatment

Overview

The first clinical question we on which we focus is: "What is the impact of antibiotics on death for hospitalized AECOPD patients?"

> **Rationale:** Antibiotics treatment has been shown to reduce treatment failures in hospitalized AECOPD patients with severe exacerbations [188]. We verify this as well as assess the validity and robustness of models generated using *MarkSIM*.
>
> **Clinical Significance:** This is the first study to evaluate the effect of antibiotics treatment of severe AECOPD requiring hospitalization using probabilistic modeling and simulation retrospective EHR data. This study facilities simula-

tions of various scenarios to observe the changes regarding the administration of antibiotics and the impact of the changes.

Methods: Patients from the AECOPD cohort are classified into four groups, ALL (the full population), ANTIBIOTICS (those for whom antibiotics were administered), and NO ANTIBIOTICS (those who did not receive antibiotics). For simplification and to observe individual diagnoses of AECOPD, a caretrail is generated for each distinct hospital stay with a primary diagnosis of AECOPD. Markov chain models are generated for each group. Monte Carlo simulations are performed to estimate the outcomes from antibiotic administration on AECOPD patients in the ICU. The simulation is run to compare the in-hospital death for two populations, those who received antibiotics and those who did not.

Results

After preprocessing the data, a total of 697 AECOPD patients were categorized in the ALL group. This group has an 8.5% death rate. Breaking down the ALL group, there are a total of 181 ANTIBIOTICS AECOPD patients with a death rate of 4.1%, and 516 NO ANTIBIOTICS AECOPD patients with a 9.9% death rate. Table 6.1 contains the characteristics of the patient groups, including the percentage of patients who received antibiotics. Notably, the NO ANTIBIOTICS group had the highest death rate (9.9%). Appendix B.1 shows the transition probabilities for each group, which were derived from the Markov chain models created from the raw data. The TPM for HALF is computed using the TPMs from the ANTIBIOTICS and NO ANTIBIOTICS group. The HALF TPM is used to conduct an analysis for administering antibiotics to 50% of the population. We used the initial state probabilities from the ALL group to simulate each of the four models.

Figure 6.4 shows the estimated percentage deaths for AECOPD patients in the ICU as a function of the percentage that received antibiotics. The x-axis is ordered by

Table 6.1.: Descriptive characteristics of patients included. ALL represents the entire AECOPD cohort, and the other groups are subsets of ALL.

	ALL AECOPD (n = 697)	ANTIBIOTICS (n = 181)	NO ANTIBIOTICS (n = 516)
Age in years [a]	70.8 (11.2)	70.6 (10.8)	70.9 (11.3)
Women [b]	357 (51.2)	91 (50.3)	266 (51.6)
Death [b]	59 (8.5)	8 (4.4)	51 (9.9)
LOS, days [c]	8.6	6.9	9.3
Initial antibiotic [b]			-
vancomycin	58 (8.3)	58 (32.0)	-
azithromycin	47 (6.6)	47 (26.0)	-
levofloxacin	38 (5.5)	38 (21.0)	-
Initial timing of antibiotics, hours (%)	27.4	27.4	-

[a] mean (standard deviation)
[b] count (%)
[c] mean

the percentage patients administered antibiotics and labeled by group. The top and bottom horizontal lines are the maximum and minimum percentage of patient deaths, respectively. The black data point represents the average percentage of patient deaths, and the data point labeled, *observed*, represents the actual percentage of AECOPD patients who died when antibiotics were administered. The actual number of patients

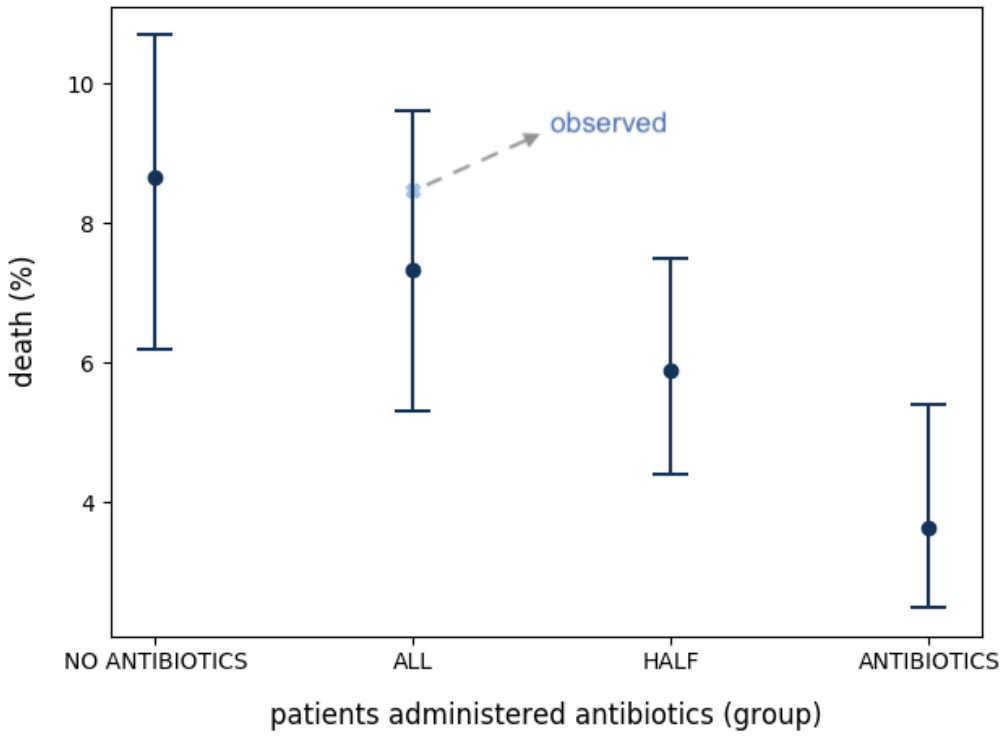


Figure 6.4.: Simulation results for the estimated percentage of AECOPD deaths based on antibiotics administration. ALL represents the entire AECOPD cohort, and the other groups are subsets of ALL.

that received antibiotics in the EHR was 26.0% with a death rate of 8.5% for the full cohort. This falls within the death rate interval [5.3% (ANTIBIOTICS group), 9.6% (NO ANTIBIOTICS group)] estimated by the simulation. As the number of patients administered antibiotics increases, the percentage of deaths decreases. Specifically, when all patients are given antibiotics, the death rate decreased exactly 50% compared

to not administering antibiotics to anyone. Figure 6.5 is a whisker plot illustrating the death percentage estimated using sensitivity analysis. Outliers are represented by data points. The median death percentage is represented by the horizontal line inside of the box. The top and bottom horizontal lines represent values maximum and minimum values, respectively.

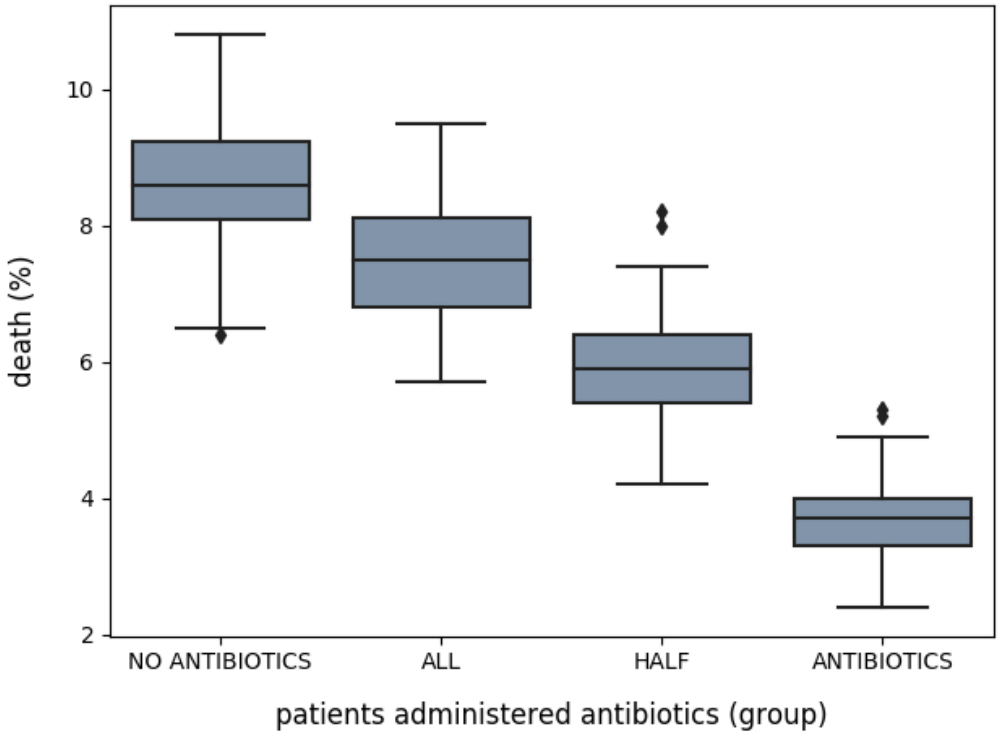


Figure 6.5.: Sensitivity analysis for estimated percentage of AECOPD deaths based on antibiotics administration. ALL represents the entire AECOPD cohort, and the other groups are subsets of ALL.

6.4.4 Clinical Question 2: Timing of Antibiotics Treatment

Overview

The second clinical question on which we focus: "What is the initial timing of antibiotics for hospitalized AECOPD patients and the impact on death?"

Rationale: Initiating antibiotics at the onset of AECOPD hospitalizations may improve short-term outcomes, shorten recovery time, reduce the risk of early relapse, treatment failure, and hospitalization duration [2,10]. The clinical question targets providing evidence for the validity of this claim. The question is posed by a clinical COPD expert on our team who is interested in determining the optimal time at which antibiotics should be administered.

Clinical Significance: This is the first study to evaluate the effect of initial timing of antibiotics treatment of severe AECOPD requiring hospitalization using retrospective EHR data. This study facilities simulations of various scenarios to observe the changes regarding the initial timing of antibiotics administration and the impact of the changes.

Methods: Patients from the AECOPD cohort who were administered antibiotics within the following specified time-frames are classified into four groups: i) within 6 hours of hospital admission, ii) between 6 and 24 hours of admission, iii) between 24 and 48 hours of admission, and iv) after 48 hours of admission. For simplification and to observe individual diagnoses of AECOPD, a caretrail is generated for each distinct hospital stay with a primary diagnosis of AECOPD. Markov chain models are generated for each group. Initial state probabilities from the ALL (the full population) group to simulate each of the four models. Monte Carlo simulations are performed to estimate the outcomes from the initial timing of antibiotic administration on AECOPD patients in the ICU. The simulation is run to compare the in-hospital death rate for two populations, those that received antibiotics and those that did not.

6.4.5 Results

The average time to antibiotic administration was 27 hours, and 32% of ANTIBIOTICS patients were administered *vancomycin* as the initial antibiotic Table 6.1. Appendix B.2 shows the transition probabilities for each group, which were derived from the Markov chain models created from the raw data.

Figure 6.6 shows the estimated percentage ICU patients with AECOPD who died based on the initial timing of antibiotics administration. The x-axis represents the time from admission when patients were administered antibiotics, labeled by group. The y-axis represents the percentage who died. The top and bottom horizontal lines are the maximum and minimum percentage of patient deaths, respectively, and the black data point represents the average percentage of patient deaths.

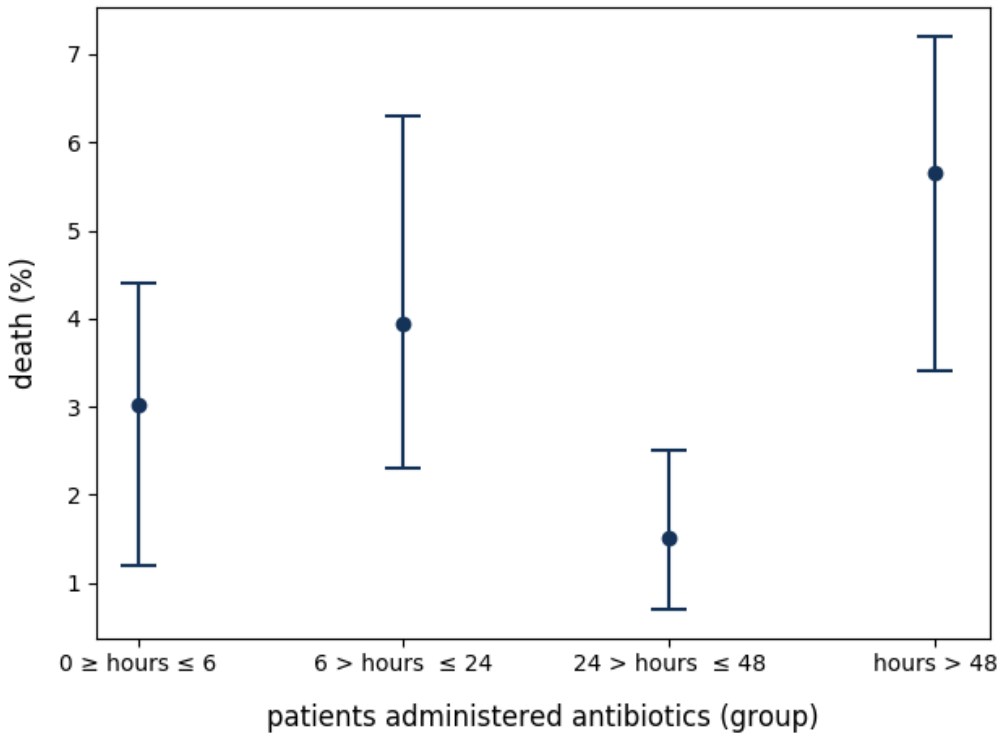


Figure 6.6.: Simulation results for the estimated percentage of AECOPD deaths based on the initial timing of antibiotics administration.

As shown, there is a 5.5% mortality in the group of cohorts that received antibiotics after 48 hours versus 1.8% in the group that received antibiotics between 24 and 48 hours. The sensitivity analyses were performed by the perturbation of each probability in the transition probability matrix and then observing how that perturbation changed the outcome (death). Figure 6.7 is a whisker plot illustrating the death percentage estimated using sensitivity analysis. Outliers are represented by data points. The median death percentage is represented by the horizontal line inside of the box. The top and bottom horizontal lines represent values maximum and minimum values, respectively.

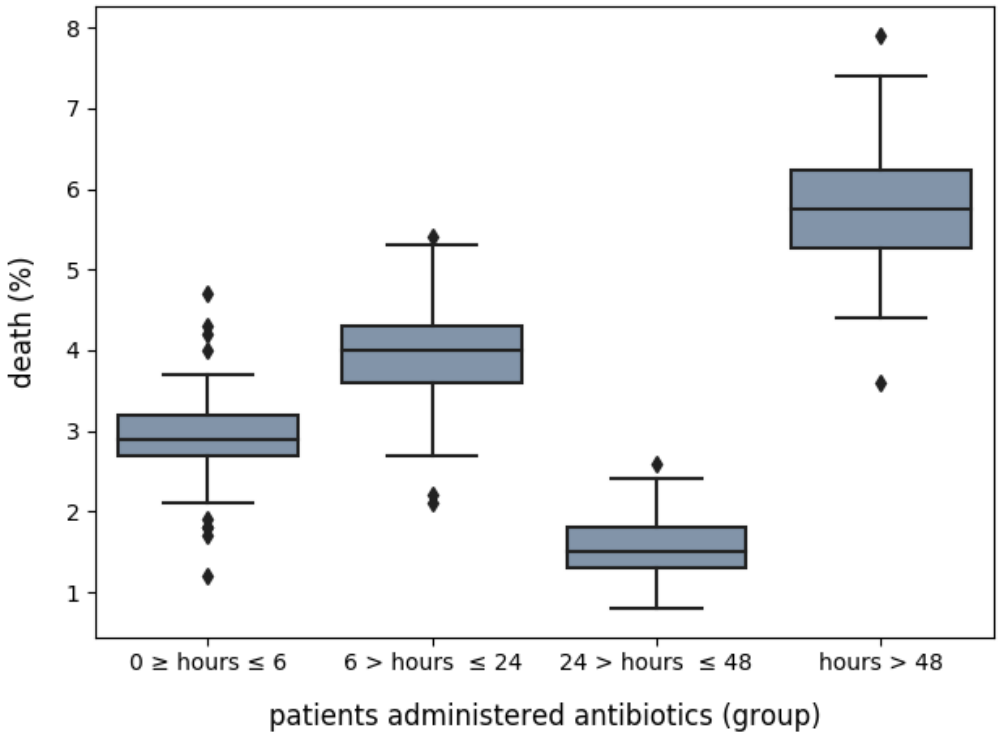


Figure 6.7.: Sensitivity analysis for estimated percentage of AECOPD deaths based on the initial timing of antibiotics administration.

6.5 Discussion

Physicians frequently make critical decisions under inexact science [189], weighing the risks and benefits of patient treatments and outcomes. Uncertainty is inevitable in medical reasoning and decision-making, especially in critical care settings where the complete patient history may not be available. This work introduced an approach to systematically capture, analyze, and estimate the effect of treatment interventions on outcomes through modeling and estimating uncertainties using EHR data. This approach captures the duration of time that a patient spent in a particular state and allows extraction of temporal inferences from mathematical models using EHR data. The novelty of our approach for generating models is evident by two main contributions:

(1) ***Definition of model states.***

Model states are defined based on clinical conditions and outcomes. Integrating clinical domain knowledge from guidelines and clinical experts into the modeling process, allows us to conduct analyses that are representative of the best available evidence and captures clinical practice experience.

(2) ***Estimation of model parameters.***

The function we developed for estimating model parameters (i.e., transition probabilities) uses timestamps to calculate the exact time a patient enters and leaves a state, as well as the total amount time a patient spends in a state. This approach introduces a new level of granularity to stochastic, temporal models.

We demonstrate the value of this approach by estimating the impact of the administration and timing of antibiotics for AECOPD on in-hospital mortality. While we applied *MarkSIM* to EHRs by generating population models (i.e., models built with data from multiple patients), the tool can also be used to generate individual patient models. Such models will only be valuable when large amounts of clinical data are accessible for an individual patient.

6.5.1 Limitations

One limitation of our study is that all of our data comes from the ICU. Additional patient history outside of the ICU might have helped to refine the findings. Another limitation is the process of selecting patient cohorts. The retrospective EHR data used required modeling patients who were assigned an AECOPD ICD-9 code. Though ICD-9 codes have been used for identifying patients when conducting retrospective clinical studies, accurately identifying COPD patients can be difficult in the absence of patient medical history [168]. COPD is diagnosed by lung function tests and staged according to the progression of the disease. Simply using ICD-9 codes may not be sufficient to differentiate an individual with COPD from an individual with COPD-like symptoms (e.g., wheezing, or noisy breathing, chest pain) who was given a COPD ICD-9 diagnosis [190]. Such symptoms may be documented as free-text, unstructured data in EHR clinical notes.

We acknowledge the limitations of modeling EHR data as Markov chains in that state transitions in first-order Markov chains only depend on the current state. Thus, such modeling does not account for confounding factors that previously impacted a patient prior to entering their current state. While Markov models are valuable when history is not important or available, they potentially lack accuracy when history is important. [191] While our model shows administering antibiotics to everyone increases survival likelihood, in practice there are many factors that may contribute to such decisions by physicians, including the probability of development of antibiotic resistance. Considering other factors such as severity of disease and comorbidities, among others, can also be used to reduce model assumptions.

6.5.2 Future Work

The methods presented can be translated to real-world scenarios by extending the methods for modeling and simulation. This includes additional functionality to perform one-way sensitivity analyses. This will allow the identification health states

that lead to undesired outcomes or states that are not conducive to improving the patient's condition. For example, if the model shows that patients who enter ARF_2 have a higher probability of death than those who do not, a simulation can be run using model parameters that have been recomputed to decrease likelihood of entering ARF_2. This can lead to the development of other clinical studies and techniques (e.g., causal inference) to understand factors that influence the health state of patients and the underlying cause for them entering particular states, Even more, understanding the impact that health states have on patients can drive improved treatment selection and quality of care for patients.

The development of clinical data networks such as PCORnet [54], OHDSI [22], and i2b2 [23], allow for the integration of various clinical and research data sources into a single repository, increasing the amount and availability of clinical data. These platforms offer tools for cohort discovery and analytics that are currently being used for clinical research. Specifically, in addition to using ICD-9 codes, i2b2's NLP feature for cohort discovery can be used for identifying AECOPD patient by extracting medications, smoking status, and diagnosis from clinical notes. While these tools allow simple data analysis, they do not offer support for granular temporal data analysis.

Our methods can be integrated into such existing systems to allow more complete analysis, and moreover, supplementing real-time analysis of data streams [169] from medical systems in care settings. Our methodology can also be applied to both common and more complex clinical conditions. That is, we defined health states for our model that encompassed several clinical variables. However, given the vast space of clinical conditions and corresponding clinical criteria that defines the conditions, some health states may not be well-represented. For example, in our state space NARF is defined as no acute respiratory failure, which means that the only criteria that determines if a patient is in NARF is that the constraints used to define ARF_1 and ARF_2 are not met. In this case, NARF does not reveal other conditions of the patient. To address this, information from various data sources such as past

research studies [23], clinical trials [169], and medical devices can be combined to define additional clinical states.

Clinical data warehouses [22, 23, 54] have proven to be valuable for overcoming the barriers of data access and availability as they provide access and structure of medical data from multiple sources as well as tools for clinical discoveries through data analysis. Our work can be integrated as an API for platforms such as i2b2 [23] for complex data analytics and generalized to fit clinical data standards such as FHIR [192] for integration with healthcare systems.

”

7 CONCLUSIONS

The highest quality of evidence for clinical guidelines comes from systematic reviews and meta-analyses, which are developed by combining published results from clinical studies such as observational and experimental studies [38]. The data used in these studies is collected to analyze clinical problems or experiments; therefore, applying data collected for specific research goals or documentation to other clinical research can be challenging. Thus, generating additional evidence beyond the scope of the original study may not be possible. While EHR data is originally collected for billing purposes, it has been leveraged for secondary use, proving to be valuable for clinical research [173]. Thus, in this research, we show that it is feasible to enhance representation of electronic health records to aggregate patients' entire medical histories with temporal trends and support complex clinical questions to drive clinical guidelines for chronic obstructive pulmonary disease. To show the feasibility of enhancing EHR representation, we present and discuss several topics.

First, we introduce a comprehensive framework for generating clinical evidence using EHR data by enhancing data representation, aggregating clinical data for patients, and developing models for predicting outcomes. Our work extends the current state of practice by embedding medical knowledge into EHR visualizations, allowing more in-depth analysis, which can lead to improved decision-making and understanding the progression of clinical diseases and their severity. We provide a tool to aggregate patients' entire EHRs for better understanding of their medical journey. Once the health state of a patient is identified, visualizations corresponding to other clinical indicators provide more insight into the patient's overall health state. Identifying cause or other confounding variables can be done using these visual representations for modeling health states.

In addition to enhancing visualizations, we demonstrate the process of modeling aggregate observations of EHR data and how this can be enhanced through integrating temporal aspects of time-variant data. We transform raw, clinical data to develop statistical models that illustrate how patients move between health states. We define health states using combinations of clinical variables measured during hospital visits.

We explore and approximate the results of our model by performing Monte Carlo simulations to estimate probabilistic quantities such as the amount of time spent in a health state. We show that our model produces valid results by performing formal goodness-of-fit testing and comparing outcome results to the real data. Modeling longitudinal clinical data that has repeated measurements of symptoms [183] that are used to define clinical conditions allows for a more complete assessment of the evolution of conditions such as AECOPD.

Finally, to show the impact of these enhancements, we demonstrate how these models can be used to answer clinical questions that cannot be interpreted through simple descriptive statistics. Specifically, our study examines the efficacy of antibiotics treatment for in-hospitalized patients with chronic obstructive pulmonary disease. Chronically ill patients are susceptible to frequent changes in health status, [193] especially when under care in an ICU. For AECOPD patients in the ICU, clinical indicators can change suddenly, causing a patient to move between health states repeatedly. Our model is used to answer questions regarding the impact of antibiotics on hospital discharge and death for patients with an acute exacerbation of COPD. Using clinical evidence, we answer these questions and provide more in-depth analysis than prior work by capturing granular time information within the analysis.

In our study we observe an overall reduced mortality rate (Figure 6.4) in patients with AECOPD admitted to the ICU who received antibiotics in comparison to those who did not. These findings are consistent with previous evidence that supports the use of antibiotics in critically ill patients given they frequently have community acquired pathogens [194]. Antibiotics were shown to reduce all-cause mortality in critically ill patients and also reduce treatment failure after 4 weeks of discharge [5].

Studies [2, 10, 195] have found that antibiotic therapy reduces mortality for COPD exacerbations requiring intensive care, and reduces treatment failure in the inpatient setting. Our findings suggest that there may be a reduction in mortality rate when initiation of antibiotics occurs earlier in settings of AECOPD patients with severe respiratory failure warranting an ICU admission. Thus, developing clinical questions and identifying relevant results closely aligns with the steps taken by the GOLD committee.

Techniques for coding can be improved to make analysis of EHRs both simpler and more useful by developing automated data cleaning tools that integrate clinical domain knowledge for identifying erroneous and implausible data points within the EHR. For example, clinical domain knowledge can be used to define the normal and abnormal ranges for a clinical measurements such as heart rate. The tool will use the definitions to implement rules that trigger the automatic identification or removal of measurements that are implausible (i.e., outside of both the normal and abnormal ranges) within a patient's EHR. Techniques for testing and evaluation could also be improved to make analysis of EHRs both simpler and more useful. Such techniques could express analysis results as mathematical expressions that represent potential biases, such as confounding, selection, and population characteristics, within the underlying data. Such methods for testing and evaluation can eventually lead to a common framework for understanding the results across different datasets, data collection methods, and populations. For example, testing and evaluation can be performed using silico models with various populations/datasets (e.g. UK biobank [196]), cross-validation, and prospective randomized trials.

Generating evidence by modeling and visualizing real-world data may advance the adoption of research results into day to day clinical practice [74]. This research provides enhancements to each stage of the EHR analysis system by incorporating temporal data systematically. Future work in both clinical studies and big data analysis will continue to enhance EHR representation, helping to improve patient treatment and analytic capabilities for both data analysts and medical professionals.

REFERENCES

[1] NIH Research Portfolio Online Reporting Tools (RePORT). FACT SHEET - Chronic Obstructive Pulmonary Disease (COPD). `https://report.nih.gov/nihfactsheets/viewfactsheet.aspx?csid=77` (accessed July 18, 2019).

[2] Global Initiative for Chronic Obstructive Lung Disease (GOLD). *Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease 2018 Report*, 2018. `https://goldcopd.org/wp-content/uploads/2017/11/GOLD-2018-v6.0-FINAL-revised-20-Nov_WMS.pdf` (accessed March 27, 2019).

[3] Earl S. Ford. Trends in mortality from COPD among adults in the United States. *Chest*, 148(4):962 – 970, 2015.

[4] Stephan Budweiser, Rudolf A Jörres, and Michael Pfeifer. Treatment of respiratory failure in COPD. *International journal of chronic obstructive pulmonary disease*, 3(4):605, 2008.

[5] Daniela J Vollenweider, Harish Jarrett, Claudia A Steurer-Stey, Judith Garcia-Aymerich, and Milo A Puhan. Antibiotics for exacerbations of chronic obstructive pulmonary disease. *Cochrane Database of Systematic Reviews*, 12, 2012.

[6] Robert Wilson, Sanjay Sethi, Antonio Anzueto, and Marc Miravitlles. Antibiotics for treatment and prevention of exacerbations of chronic obstructive pulmonary disease. *Journal of Infection*, 67(6):497 – 515, 2013.

[7] Anthony J Guarascio, Shauntá M Ray, Christopher K Finch, and Timothy H Self. The clinical and economic burden of chronic obstructive pulmonary disease in the USA. *ClinicoEconomics and outcomes research: CEOR*, 5:235, 2013.

[8] Maryam A Hakim, Frances L Garden, Matthew D Jennings, and Claudia C Dobler. Performance of the lace index to predict 30-day hospital readmissions in patients with chronic obstructive pulmonary disease. *Clinical Epidemiology*, 10:51–59, 2018.

[9] Prasadini N Perera, Edward P Armstrong, Duane L Sherrill, and Grant H Skrepnek. Acute exacerbations of COPD in the United States: inpatient burden and predictors of costs and mortality. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, 9(2):131–141, 2012.

[10] Kristina Vermeersch, Maria Gabrovska, Joseph Aumann, Ingel K Demedts, Jean-Louis Corhay, Eric Marchand, Hans Slabbynck, Christel Haenebalcke, Michiel Haerens, Shane Hanon, et al. Azithromycin during acute COPD exacerbations requiring hospitalization (bace): a multicentre, randomized, double-blind, placebo-controlled trial. *American journal of respiratory and critical care medicine*, 2019.

[11] National Heart Lung and Blood Institute. How serious is COPD, 2017. http://www.lung.org/lung-health-and-diseases/lung-disease-lookup/copd/learn-about-copd/how-serious-is-copd.html (accessed April 11, 2019).

[12] Lise Poissant, Jennifer Pereira, Robyn Tamblyn, and Yuko Kawasumi. The impact of electronic health records on time efficiency of physicians and nurses: a systematic review. *Journal of the American Medical Informatics Association*, 12(5):505–516, 2005.

[13] José M Quintana, Cristóbal Esteban, Anette Unzurrunzaga, Susana Garcia-Gutierrez, Nerea Gonzalez, Irantzu Barrio, Inmaculada Arostegui, Iratxe Lafuente, Marisa Bare, Nerea Fernandez-de Larrea, et al. Predictive score for mortality in patients with COPD exacerbations attending hospital emergency departments. *BMC medicine*, 12(1):66, 2014.

[14] S. Suissa and P. J. Barnes. Inhaled corticosteroids in COPD: the case against. *European Respiratory Journal*, 34(1):13–16, 2009.

[15] Samy Suissa. Randomized trials built on sand: examples from COPD, hormone therapy, and cancer. *Rambam Maimonides medical journal*, 3(3), 2012.

[16] Leonardo M. Fabbri, Piera Boschetto, and Cristina E. Mapp. COPD guidelines. *American Journal of Respiratory and Critical Care Medicine*, 176:527–528, 2007. PMID: 17823358.

[17] M Hassan Murad. Clinical practice guidelines: a primer on development and dissemination. In *Mayo Clinic Proceedings*, volume 92, pages 423–433. Elsevier, 2017.

[18] Danny Epstein, Yuval Barak-Corren, Yoni Isenberg, and Gidon Berger. Clinical decision support system: A pragmatic tool to improve acute exacerbation of COPD discharge recommendations. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, pages 1–7, 2019.

[19] Charles Safran, Meryl Bloomrosen, W Edward Hammond, Steven Labkoff, Suzanne Markel-Fox, Paul C Tang, and Don E Detmer. Toward a national framework for the secondary use of health data: an american medical informatics association white paper. *Journal of the American Medical Informatics Association*, 14(1):1–9, 2007.

[20] Jesus J Caban and David Gotz. Visual analytics in healthcare opportunities and research challenges. *Journal of the American Medical Informatics Association*, 22(2):260–262, 2015.

[21] Rimma Pivovarov and Noémie Elhadad. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association*, 22(5):938–947, 2015.

[22] Observational Health Data Sciences and Informatics. Ohdsi: Observational health data science and informatics, 2018. https://www.ohdsi.org (accessed July 2, 2019).

[23] Shawn N Murphy, Griffin Weber, Michael Mendis, Vivian Gainer, Henry C Chueh, Susanne Churchill, and Isaac Kohane. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association*, 17(2):124–130, 2010.

[24] Daniel Abler, Vassiliki Kanellopoulos, Jim Davies, Manjit Dosanjh, Raj Jena, Norman Kirkby, and Ken Peach. Data-driven markov models and their application in the evaluation of adverse events in radiotherapy. *Journal of radiation research*, 54(suppl_1):i49–i55, 2013.

[25] Christopher H Jackson, Linda D Sharples, Simon G Thompson, Stephen W Duffy, and Elisabeth Couto. Multistate markov models for disease progression with classification error. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(2):193–209, 2003.

[26] A Rodina-Theocharaki, K Bliznakova, and N Pallikarakis. Markov chain monte carlo simulation for projection of end stage renal disease patients in greece. *Computer Methods and Programs in Biomedicine*, 107(1):90–96, 2012.

[27] Emma Tan, Ruud Boessen, David Fishwick, Rinke Klein Entink, Tim Meijster, Anjoeka Pronk, Birgit van Duuren-Stuurman, and Nick Warren. A microsimulation model for the development and progression of chronic obstructive pulmonary disease. *Respiratory medicine*, 109(12):1521–1531, 2015.

[28] Zitao Liu and Milos Hauskrecht. Clinical time series prediction: Toward a hierarchical dynamical system framework. *Artificial intelligence in medicine*, 65(1):5–18, 2015.

[29] Gerard J Criner, Jean Bourbeau, Rebecca L Diekemper, Daniel R Ouellette, Donna Goodridge, Paul Hernandez, Kristen Curren, Meyer S Balter, Mohit Bhutani, Pat G Camp, et al. Prevention of acute exacerbations of COPD: American college of chest physicians and canadian thoracic society guideline. *Chest*, 147(4):894–942, 2015.

[30] Centers for Disease Control and Prevention. Chronic Diseases in America . `https://www.cdc.gov/chronicdisease/index.htm` (accessed July 18, 2019).

[31] World Health Organization. Burden of COPD, 2017. `http://www.who.int/respiratory/copd/burden/en/` (March 31, 2019).

[32] Jean Joel Bigna, Angeladine Malaha Kenne, Serra Lem Asangbeh, and Aurelie T Sibetcheu. Prevalence of chronic obstructive pulmonary disease in the global population with hiv: a systematic review and meta-analysis. *The Lancet Global Health*, 6(2):e193 – e202, 2018.

[33] Marilyn J Field and Kathleen H Lohr. Clinical practice guidelines: Directions for a new program, institute of medicine, washington dc, 1990.

[34] Steven Woolf, Holger J Schünemann, Martin P Eccles, Jeremy M Grimshaw, and Paul Shekelle. Developing clinical practice guidelines: types of evidence and outcomes; values and economics, synthesis, grading, and presentation and deriving recommendations. *Implementation Science*, 7(1):61, 2012.

[35] Patricia B Burns, Rod J Rohrich, and Kevin C Chung. The levels of evidence and their role in evidence-based medicine. *Plastic and reconstructive surgery*, 128(1):305, 2011.

[36] *Research Hub: Evidence Based Practice Toolkit: Levels of Evidence*, 2019 (accessed June 3, 2019). `https://libguides.winona.edu/c.php?g=11614&p=61584`.

[37] Joan Vlayen, Bert Aertgeerts, Karin Hannes, Walter Sermeus, and Dirk Ramaekers. A systematic review of appraisal tools for clinical practice guidelines: multiple similarities and one common deficit. *International Journal for Quality in Health Care*, 17(3):235–242, 2005.

[38] George Weisz, Alberto Cambrosio, Peter Keating, Loes Knaapen, Thomas Schlich, and Virginie J Tournay. The emergence of clinical practice guidelines. *The Milbank Quarterly*, 85(4):691–727, 2007.

[39] Amir Qaseem, Timothy J Wilt, Steven E Weinberger, Nicola A Hanania, Gerard Criner, Thys van der Molen, Darcy D Marciniuk, Tom Denberg, Holger Schunemann, Wisia Wedzicha, et al. Diagnosis and management of stable chronic obstructive pulmonary disease: a clinical practice guideline update from the american college of physicians, american college of chest physicians, american thoracic society, and european respiratory society. *Annals of internal medicine*, 155(3):179–191, 2011.

[40] Klaus F RABE. Guidelines for chronic obstructive pulmonary disease treatment and issues of implementation. *Proceedings of the American Thoracic Society*, 3(7):641–644, 2006.

[41] Physiopedia. Formulate an answerable question — physiopedia, 2017. `https://www.physio-pedia.com/index.php?title=Formulate_an_answerable_question&oldid=174716` (accessed April 17, 2019).

[42] Connie Schardt, Martha B Adams, Thomas Owens, Sheri Keitz, and Paul Fontelo. Utilization of the pico framework to improve searching pubmed for clinical questions. *BMC Medical Informatics and Decision Making*, 7:16–16, 2007.

[43] Jørgen Vestbo, Suzanne S Hurd, Alvar G Agustí, Paul W Jones, Claus Vogelmeier, Antonio Anzueto, Peter J Barnes, Leonardo M Fabbri, Fernando J Martinez, Masaharu Nishimura, et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: Gold executive summary. *American journal of respiratory and critical care medicine*, 187(4):347–365, 2013.

[44] National Guideline Clearinghouse (NGC). Care of the hospitalized patient with acute exacerbation of COPD, 2016. `https://www.guideline.gov/summaries/summary/50686/care-of-the-hospitalized-patient-with-acute-exacerbation-of-copd?q=copd` (accessed July 2, 2019).

[45] The Ohio State University Wexner Medical Center, 2017. `https://evidencebasedpractice.osumc.edu/Documents/Guidelines/{COPD}.pdf` (accessed December 4, 2018).

[46] Julia AE Walters, Daniel J Tan, Clinton J White, and Richard Wood-Baker. Different durations of corticosteroid therapy for exacerbations of chronic obstructive pulmonary disease. *The Cochrane Library*, 2014.

[47] Neil MacIntyre and Yuh Chin Huang. Acute exacerbations and respiratory failure in chronic obstructive pulmonary disease. *Proceedings of the American Thoracic Society*, 5(4):530–535, 2008.

[48] Karen Y He, Dongliang Ge, and Max M He. Big data analytics for genomic medicine. *International journal of molecular sciences*, 18(2):412, 2017.

[49] Jorge Blasco, Thomas M. Chen, Juan Tapiador, and Pedro Peris-Lopez. A survey of wearable biometric recognition systems. *ACM Comput. Surv.*, 49(3):43:1–43:35, sep 2016.

[50] Harsh Kupwade Patil and Ravi Seshadri. Big data security and privacy issues in healthcare. In *Big Data (BigData Congress), 2014 IEEE International Congress on*, pages 762–765. IEEE, 2014.

[51] Nishita Mehta and Anil Pandit. Concurrence of big data analytics and healthcare: A systematic review. *International Journal of Medical Informatics*, 114:57–65, 2018.

[52] Hoerbst A. Electronic health records. a systematic review on quality requirements., January 2010.

[53] Taxiarchis Botsis, Gunnar Hartvigsen, Fei Chen, and Chunhua Weng. Secondary use of ehr: data quality issues and informatics opportunities. *Summit on Translational Bioinformatics*, 2010:1, 2010.

[54] Rachael L Fleurence, Lesley H Curtis, Robert M Califf, Richard Platt, Joe V Selby, and Jeffrey S Brown. Launching pcornet, a national patient-centered clinical research network. *Journal of the American Medical Informatics Association*, 21(4):578–582, 2014.

[55] Partners Healthcare. i2b2: Informatics for integrating biology & the bedside, 2018. `https://www.i2b2.org/about/index.html` (accessed July 2, 2019).

[56] Acquiring and using electronic health record data, 2019. `https://rethinkingclinicaltrials.org/resources/acquiring-and-using-electronic-health-record-data/` (accessed May 31, 2019).

[57] Sharukh Lokhandwala and Barret Rush. Objectives of the secondary analysis of electronic health record data. In *Secondary Analysis of Electronic Health Records*, pages 3–7. Springer, 2016.

[58] John PA Ioannidis. Why most clinical research is not useful. *PLoS medicine*, 13(6):e1002049, 2016.

[59] Jessica L. Stump. Henrietta lacks and the hela cell: Rights of patients and responsibilities of medical researchers. *The History Teacher*, 48(1):127–180, November 2014.

[60] RB Ness. Influence of the hipaa privacy rule on health research. *JAMA*, 298(18):2164–2170, 2007.

[61] Kathy L Hudson. Genomics, health care, and society. *New England Journal of Medicine*, 365(11):1033–1041, 2011.

[62] Chunhua Weng, Paul Appelbaum, George Hripcsak, Ian Kronish, Linda Busacca, Karina W Davidson, and J Thomas Bigger. Using ehrs to integrate research with patient care: promises and challenges. *Journal of the American Medical Informatics Association*, 19(5):684–687, 2012.

[63] Vince Stanford. Pervasive health care applications face tough security challenges. *IEEE pervasive computing*, 1(2):8–12, 2002.

[64] Mohammad Adibuzzaman, Poching DeLaurentis, Jennifer Hill, and Brian D Benneyworth. Big data in healthcare–the promises, challenges and opportunities from a research perspective: A case study with a model database. In *AMIA Annual Symposium Proceedings*, volume 2017, page 384. American Medical Informatics Association, 2017.

[65] George Hripcsak, Meryl Bloomrosen, Patti FatelyBrennan, Christopher G Chute, Jim Cimino, Don E Detmer, Margo Edmunds, Peter J Embi, Melissa M Goldstein, William Ed Hammond, Gail M Keenan, Steve Labkoff, Shawn Murphy, Charlie Safran, Stuart Speedie, Howard Strasberg, Freda Temple, and Adam B Wilcox. Health data use, stewardship, and governance: ongoing gaps and challenges: a report from amia's 2012 health policy meeting. *Journal of the American Medical Informatics Association*, 21(2):204–211, 2014.

[66] Asaph Azaria, Ariel Ekblaw, Thiago Vieira, and Andrew Lippman. Medrec: Using blockchain for medical data access and permission management. In *Open and Big Data (OBD), International Conference on*, pages 25–30. IEEE, 2016.

[67] Miriam Reisman. Ehrs: the challenge of making electronic data usable and interoperable. *Pharmacy and Therapeutics*, 42(9):572, 2017.

[68] Robert Steinbrook. Health care and the american recovery and reinvestment act. *New England Journal of Medicine*, 360(11):1057–1060, 2009.

[69] *The Standard Health Record Collaborative*, 2019. `http://standardhealthrecord.org/` (accessed June 20, 2019).

[70] Tsung-Ting Kuo, Hyeon-Eui Kim, and Lucila Ohno-Machado. Blockchain distributed ledger technologies for biomedical and health care applications. *Journal of the American Medical Informatics Association*, 24(6):1211–1220, 2017.

[71] Dipak Kalra and David Ingram. *Electronic Health Records*, pages 135–181. Springer London, London, 2006.

[72] Ted D Wade. Traits and types of health data repositories. *Health information science and systems*, 2(1):4, 2014.

[73] Wei-Qi Wei and Joshua Denny. Extracting research-quality phenotypes from electronic health records to support precision medicine. *Genome Medicine*, 7, 04 2015.

[74] Tomas Skripcak, Claus Belka, Walter Bosch, Carsten Brink, Thomas Brunner, Volker Budach, Daniel Büttner, Jürgen Debus, Andre Dekker, Cai Grau, et al. Creating a data exchange strategy for radiotherapy research: towards federated databases and anonymised public datasets. *Radiotherapy and Oncology*, 113(3):303–309, 2014.

[75] Jeffrey G Klann, Aaron Abend, Vijay A Raghavan, Kenneth D Mandl, and Shawn N Murphy. Data interchange using i2b2. *Journal of the American Medical Informatics Association*, 23(5):909–915, 2016.

[76] Carlo Batini, Stefano Ceri, Shamkant B Navathe, et al. *Conceptual database design: an Entity-relationship approach*, volume 116. Benjamin/Cummings Redwood City, CA, 1992.

[77] Xixuan Feng, Arun Kumar, Benjamin Recht, and Christopher Ré. Towards a unified architecture for in-rdbms analytics. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 325–336. ACM, 2012.

[78] Veronika Abramova, Jorge Bernardino, and Pedro Furtado. Experimental evaluation of nosql databases. *International Journal of Database Management Systems*, 6(3):1, 2014.

[79] John Klein, Ian Gorton, Neil Ernst, Patrick Donohoe, Kim Pham, and Chrisjan Matser. Performance evaluation of nosql databases: A case study. In *Proceedings of the 1st Workshop on Performance Analysis of Big Data Systems*, pages 5–10. ACM, 2015.

[80] Rick Cattell. Scalable sql and nosql data stores. *Acm Sigmod Record*, 39 (4):12–27, 2011.

[81] Michael Stonebraker and Lawrence A Rowe. *The design of Postgres*, volume 15 (2). ACM, 1986.

[82] Alex Milinovich and Michael W Kattan. Extracting and utilizing electronic health data from epic for research. *Annals of translational medicine*, 6(3), 2018.

[83] Sandra Karcher, Egon L Willighagen, John Rumble, Friederike Ehrhart, Chris T Evelo, Martin Fritts, Sharon Gaheen, Stacey L Harper, Mark D Hoover, Nina Jeliazkova, et al. Integration among databases and data sets to support productive nanotechnology: Challenges and recommendations. *NanoImpact*, 9:85–101, 2018.

[84] William Goossen, Anneke Goossen-Baremans, and Michael Van Der Zel. Detailed clinical models: a review. *Healthcare informatics research*, 16(4):201–214, 2010.

[85] George Hripcsak, Jon D Duke, Nigam H Shah, Christian G Reich, Vojtech Huser, Martijn J Schuemie, Marc A Suchard, Rae Woong Park, Ian Chi Kei Wong, Peter R Rijnbeek, et al. Observational health data sciences and informatics (ohdsi): opportunities for observational researchers. *Studies in health technology and informatics*, 216:574, 2015.

[86] Lemuel R Waitman, Judith J Warren, E LaVerne Manos, and Daniel W Connolly. Expressing observations from electronic medical record flowsheets in an i2b2 based clinical data repository to support research and quality improvement. In *AMIA Annual Symposium Proceedings*, volume 2011, page 1454. American Medical Informatics Association, 2011.

[87] Andrew R Post, Akshatha K Pai, Richard Willard, Bradley J May, Andrew C West, Sanjay Agravat, Stephen J Granite, Raimond L Winslow, and David S Stephens. Metadata-driven clinical data loading into i2b2 for clinical and translational science institutes. *AMIA Summits on Translational Science Proceedings*, 2016:184, 2016.

[88] Rupa Makadia and Patrick B Ryan. Transforming the premier perspective® hospital database into the observational medical outcomes partnership (omop) common data model. *eGEMs*, 2(1), 2014.

[89] Tom Pollard, Franck Dernoncourt, Samuel Finlayson, and Adrian Velasquez. *Data Preparation*, pages 101–114. Springer International Publishing, Cham, 2016.

[90] Erhard Rahm and Hong Hai Do. Data cleaning: Problems and current approaches. *Bulletin of the Technical Committee on*, page 3, 2000.

[91] Danieleand Wu Joy Tzung-yu Malley, Brianand Ramazzotti. *Data Preprocessing*, pages 115–141. Springer International Publishing, Cham, 2016.

[92] Allan F Simpao, Luis M Ahumada, Jorge A Gálvez, and Mohamed A Rehman. A review of analytics and clinical informatics in health care. *Journal of medical systems*, 38(4):45, 2014.

[93] Wullianallur Raghupathi and Viju Raghupathi. An overview of health analytics. *J Health Med Informat*, 4(132):2, 2013.

[94] Anima Singh, Girish Nadkarni, Omri Gottesman, Stephen B Ellis, Erwin P Bottinger, and John V Guttag. Incorporating temporal ehr data in predictive models for risk stratification of renal function deterioration. *Journal of biomedical informatics*, 53:220–228, 2015.

[95] Jonathan A Cook and Gary S Collins. The rise of big clinical databases. *British Journal of Surgery*, 102(2):e93–e101, 2015.

[96] Pascal R Fuchshuber, William Greif, Chantal R Tidwell, Michael S Klemm, Cheryl Frydel, Abdul Wali, Efren Rosas, and Molly P Clopp. The power of the national surgical quality improvement programachieving a zero pneumonia rate in general surgery patients. *The Permanente Journal*, 16(1):39, 2012.

[97] Leo Anthony Celi, Roger G Mark, David J Stone, and Robert A Montgomery. big data in the intensive care unit. closing the data loop. *American journal of respiratory and critical care medicine*, 187(11):1157, 2013.

[98] Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific Data*, 5:180178, September 2018.

[99] Alistair Ew Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016.

[100] Angelina Kurniati. Data quality issues with using the mimic-iii data for process mining in healthcare, 2017. `https://www.researchgate.net/publication/316545308_Data_Quality_Issues_with_Using_the_MIMIC-III_Data_for_Process_Mining_in_Healthcare` (accessed July 2, 2019).

[101] Alistair EW Johnson, David J Stone, Leo A Celi, and Tom J Pollard. The mimic code repository: enabling reproducibility in critical care research. *Journal of the American Medical Informatics Association*, 25(1):32–39, 2018.

[102] Bonnie L Westra, Beverly Christie, Steven G Johnson, Lisiane Pruinelli, Anne LaFlamme, Jung In Park, Suzan G Sherman, Matthew D Byrne, Piper Ranallo, and Stuart Speedie. Expanding interprofessional ehr data in i2b2. *AMIA Summits on Translational Science Proceedings*, 2016:260, 2016.

[103] Cansu Sen, Thomas Hartvigsen, Elke Rundensteiner, and Kajal Claypool. Crest-risk prediction for clostridium difficile infection using multimodal data mining. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 52–63. Springer, 2017.

[104] Gerhard Nierhaus. *Markov Models*, pages 67–82. Springer Vienna, Vienna, 2009.

[105] Matthieu Komorowski and Jesse Raffa. Markov models and cost effectiveness analysis: applications in medical research. In *Secondary Analysis of Electronic Health Records*, pages 351–367. Springer, 2016.

[106] Hakan Demirtas, Rawan Allozi, Yiran Hu, Gul Inan, and Levent Ozbek. Joint generation of binary, ordinal, count, and normal data with specified marginal and association structures in monte-carlo simulations. In *Monte-Carlo Simulation-Based Statistical Modeling*, pages 3–15. Springer, 2017.

[107] Joseph A Maxwell. *Qualitative research design: An interactive approach*, volume 41. Sage publications, 2012.

[108] Bikash Kanti Sarkar. Big data for secure healthcare system: a conceptual design. *Complex & Intelligent Systems*, 3(2):133–151, 2017.

[109] Chun-Wei Tsai, Chin-Feng Lai, Han-Chieh Chao, and Athanasios V Vasilakos. Big data analytics: a survey. *Journal of Big data*, 2(1):21, 2015.

[110] Mandeep Kaur Saggi and Sushma Jain. A survey towards an integration of big data analytics to big insights for value-creation. *Information Processing & Management*, 54(5):758–790, 2018.

[111] Thomas H Davenport, Jeanne G Harris, David W De Long, and Alvin L Jacobson. Data to knowledge to results: building an analytic capability. *California management review*, 43(2):117–138, 2001.

[112] Yichuan Wang and Nick Hajli. Exploring the path to big data analytics success in healthcare. *Journal of Business Research*, 70:287 – 299, 2017.

[113] Sijia Liu, Yanshan Wang, Andrew Wen, Liwei Wang, Na Hong, Feichen Shen, Steven Bedrick, William Hersh, and Hongfang Liu. Create: Cohort retrieval enhanced by analysis of text from electronic health records using omop common data model, 2019.

[114] Honghan Wu, Giulia Toti, Katherine I Morley, Zina M Ibrahim, Amos Folarin, Richard Jackson, Ismail Kartoglu, Asha Agrawal, Clive Stringer, Darren Gale, et al. Semehr: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. *Journal of the American Medical Informatics Association*, 25(5):530–537, 2018.

[115] Genevieve Gorrell, Xingyi Song, and Angus Roberts. Bio-yodie: A named entity linking system for biomedical text. *arXiv preprint arXiv:1811.04860*, 2018.

[116] Nitesh V Chawla and Darcy A Davis. Bringing big data to personalized healthcare: a patient-centered framework. *Journal of general internal medicine*, 28(3):660–665, 2013.

[117] Hamzeh Khazaei, Carolyn McGregor, J Mikael Eklund, and Khalil El-Khatib. Real-time and retrospective health-analytics-as-a-service: a novel framework. *JMIR medical informatics*, 3(4):e36, 2015.

[118] OHDSI: Observational Health Data Science and Informatics. ATLAS, 2018. `http://www.ohdsi.org/web/atlas` (accessed July 2, 2019).

[119] Wullianallur Raghupathi and Viju Raghupathi. Big data analytics in healthcare: promise and potential. *Health information science and systems*, 2(1):3, 2014.

[120] Ashwin Belle, Raghuram Thiagarajan, SM Soroushmehr, Fatemeh Navidi, Daniel A Beard, and Kayvan Najarian. Big data analytics in healthcare. *BioMed research international*, 2015, 2015.

[121] Cheryl Bagley Thompson. Descriptive data analysis. *Air medical journal*, 28(2):56–59, 2009.

[122] Pranjul Yadav, Michael Steinbach, Vipin Kumar, and Gyorgy Simon. Mining electronic health records (ehrs): A survey. *ACM Computing Surveys (CSUR)*, 50(6):85, 2018.

[123] Ari Moskowitz and Kenneth Chen. *Defining the Patient Cohort*, pages 93–100. Springer International Publishing, Cham, 2016.

[124] Andrew T Jebb and Louis Tay. Introduction to time series analysis for organizational research: Methods for longitudinal analyses. *Organizational Research Methods*, 20(1):61–94, 2017.

[125] Berk Ekmekci, Charles E McAnany, and Cameron Mura. An introduction to programming for bioscientists: a python-based primer. *PLoS computational biology*, 12(6):e1004867, 2016.

[126] E Dantony, MH Elsensohn, A Dany, E Villar, C Couchoud, and René Ecochard. Estimating the parameters of multi-state models with time-dependent covariates through likelihood decomposition. *Computers in biology and medicine*, 69:37–43, 2016.

[127] Clemens Scott Kruse, Rishi Goswamy, Yesha Raval, and Sarah Marawi. Challenges and opportunities of big data in health care: a systematic review. *JMIR medical informatics*, 4(4), 2016.

[128] Evan T Sholle, Marcos A Davila, Joseph Kabariti, Julian Z Schwartz, Vinay I Varughese, Curtis L Cole, and Thomas R Campion Jr. A scalable method for supporting multiple patient cohort discovery projects using i2b2. *Journal of biomedical informatics*, 84:179–183, 2018.

[129] OHDSI: Observational Health Data Science and Informatics. WhiteRabbit for ETL design, 2018. `https://www.ohdsi.org/analytic-tools/whiterabbit-for-etl-design/` (accessed July 2, 2019).

[130] Christian R Bauer, Carolin Knecht, Christoph Fretter, Benjamin Baum, Sandra Jendrossek, Malte R
"uhlemann, Femke-Anouska Heinsen, Nadine Umbach, Bodo Grimbacher, Andre Franke, et al. Interdisciplinary approach towards a systems medicine toolbox using the example of inflammatory diseases. *Briefings in bioinformatics*, 18(3):479–487, 2017.

[131] ST Rosenbloom, RJ Carroll, JL Warner, ME Matheny, and JC Denny. Representing knowledge consistently across health systems. *Yearbook of medical informatics*, 26(01):139–147, 2017.

[132] Mark D. Danese, Marc Halperin, Jennifer Duryea, and Ryan Duryea. The generalized data model for clinical research. *bioRxiv*, 2017.

[133] Vikrant G Deshmukh, Stéphane M Meystre, and Joyce A Mitchell. Evaluating the informatics for integrating biology and the bedside system for clinical research. *BMC medical research methodology*, 9(1):70, 2009.

[134] Ranjit Singh and Kawaljeet Singh. A descriptive classification of causes of data quality problems in data warehousing. *International Journal of Computer Science Issues (IJCSI)*, 7(3):41, 2010.

[135] Vijay Gadepally, Peinan Chen, Jennie Duggan, Aaron Elmore, Brandon Haynes, Jeremy Kepner, Samuel Madden, Tim Mattson, and Michael Stonebraker. The bigdawg polystore system and architecture. In *High Performance Extreme Computing Conference (HPEC), 2016 IEEE*, pages 1–6. IEEE, 2016.

[136] Jeffrey G Klann, Matthew AH Joss, Kevin Embree, and Shawn N Murphy. Data model harmonization for the all of us research program: Transforming i2b2 data into the omop common data model. *PloS one*, 14(2):e0212463, 2019.

[137] Jeffrey G Klann, Lori C Phillips, Christopher Herrick, Matthew A H Joss, Kavishwar B Wagholikar, and Shawn N Murphy. Web services for data warehouses: OMOP and PCORnet on i2b2. *Journal of the American Medical Informatics Association*, 25(10):1331–1338, 07 2018.

[138] Mark L. Braunstein. *FHIR*, pages 179–203. Springer International Publishing, Cham, 2018.

[139] Marten Smits, Ewout Kramer, Martijn Harthoorn, and Ronald Cornet. A comparison of two detailed clinical model representations: Fhir and cda. *European Journal for Biomedical Informatics*, 11(2), 2015.

[140] Harvard University. Bd2k pic-sure restful api, 2018. `http://bd2k-picsure.hms.harvard.edu` (accessed July 2, 2019).

[141] Alex A.T. Bui and John Darrell Van Horn. Envisioning the future of big data biomedicine. *Journal of Biomedical Informatics*, 69:115 – 117, 2017.

[142] Alba Gutiérrez-Sacristán, Romain Guedj, Gabor Korodi, Jason Stedman, Laura I Furlong, Chirag J Patel, Isaac S Kohane, and Paul Avillach. Rcupcake: an r package for querying and analyzing biomedical data through the bd2k pic-sure restful api. *Bioinformatics*, 34(8):1431–1432, 2017.

[143] Beth Haenke Just and Katherine Lusk. Keep it clean. optimizing ehrs starts with ensuring data quality. *Journal of AHIMA*, 77(6):42, 2006.

[144] Mohammad Adibuzzaman, Ken Musselman, Alistair Johnson, Paul Brown, Zachary Pitluk, and Ananth Grama. Closing the data loop: An integrated open access analysis platform for the mimic database. *Computing in cardiology*, 43:137, 2016.

[145] Honghan Wu. Cogstack-semehr. `https://github.com/CogStack/CogStack-SemEHR/blob/master/mimicdao.py` (Accessed December 10, 2018).

[146] Avita Katal, Mohammad Wazid, and RH Goudar. Big data: issues, challenges, tools and good practices. In *2013 Sixth international conference on contemporary computing (IC3)*, pages 404–409. IEEE, 2013.

[147] Raymond Francis Sarmiento and Franck Dernoncourt. Improving patient cohort identification using natural language processing. In *Secondary analysis of electronic health records*, pages 405–417. Springer, 2016.

[148] Benjamin S Glicksberg, Boris Oskotsky, Nicholas Giangreco, Phyllis M Thangaraj, Vivek Rudrapatna, Debajyoti Datta, Remi Frazier, Nelson Lee, Rick Larsen, Nicholas P Tatonetti, et al. Romop: a light-weight r package for interfacing with omop-formatted electronic health record data. *JAMIA Open*, 2(1):10–14, 2019.

[149] Zhou Yuan, Sean Finan, Jeremy Warner, Guergana Savova, and Harry Hochheiser. Longitudinal visual analytics for unpacking the cancer journey. *bioRxiv*, page 444356, 2018.

[150] Matteo Gabetta, Ivan Limongelli, Ettore Rizzo, Alberto Riva, Daniele Segagni, and Riccardo Bellazzi. Bigq: a nosql based framework to handle genomic variants in i2b2. *BMC Bioinformatics*, 16(1):415, Dec 2015.

[151] Timealign: Exploratory visual analysis plugin for temporal events. `https://community.i2b2.org/wiki/display/timealign/TimeAlign` (accessed June 17, 2019).

[152] Discovering temporal categorical patterns across multiple records. `http://www.cs.umd.edu/hcil/lifelines2/` (accessed June 17, 2019.

[153] Toan C Ong, Michael G Kahn, Bethany M Kwan, Traci Yamashita, Elias Brandt, Patrick Hosokawa, Chris Uhrich, and Lisa M Schilling. Dynamic-etl: a hybrid approach for health data extraction, transformation and loading. *BMC medical informatics and decision making*, 17(1):134, 2017.

[154] Denis Klimov, Alexander Shknevsky, and Yuval Shahar. Exploration of patterns predicting renal damage in patients with diabetes type ii using a visual temporal analysis laboratory. *Journal of the American Medical Informatics Association*, 22(2):275–289, 2014.

[155] Yuval Shahar, Dina Goren-Bar, Maya Galperin, David Boaz, and Gil Tahan. Knave-ii: A distributed architecture for interactive visualization and intelligent exploration of time-oriented clinical data. *Proceedings of Intelligent Data Analysis in Medicine and Pharmacology, Protaras, Cyprus*, 2003.

[156] Jamie S Hirsch, Jessica S Tanenbaum, Sharon Lipsky Gorman, Connie Liu, Eric Schmitz, Dritan Hashorva, Artem Ervits, David Vawdrey, Marc Sturm, and Noémie Elhadad. Harvest, a longitudinal patient record summarizer. *Journal of the American Medical Informatics Association*, 22(2):263–274, 2014.

[157] Fan Du, Catherine Plaisant, Neil Spring, and Ben Shneiderman. Eventaction: Visual analytics for temporal event sequence recommendation. In *2016 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 61–70. IEEE, 2016.

[158] Adam Perer, Fei Wang, and Jianying Hu. Mining and exploring care pathways from electronic medical records with visual analytics. *Journal of Biomedical Informatics*, 56(Supplement C):369 – 378, 2015.

[159] Taowei David Wang, Catherine Plaisant, Alexander J Quinn, Roman Stanchak, Shawn Murphy, and Ben Shneiderman. Aligning temporal data by sentinel events: discovering patterns in electronic health records. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 457–466. ACM, 2008.

[160] Vivian L West, David Borland, and W Ed Hammond. Innovative information visualization of electronic health record data: a systematic review. *Journal of the American Medical Informatics Association*, 22(2):330–339, 2014.

[161] Mona Hosseinkhani Loorak, Charles Perin, Noreen Kamal, Michael Hill, and Sheelagh Carpendale. Timespan: Using visualization to explore temporal multi-dimensional data of stroke patients. *IEEE transactions on visualization and computer graphics*, 22(1):409–418, 2015.

[162] Fan Du, Ben Shneiderman, Catherine Plaisant, Sana Malik, and Adam Perer. Coping with volume and variety in temporal event sequences: Strategies for sharpening analytic focus. *IEEE transactions on visualization and computer graphics*, 23(6):1636–1649, 2016.

[163] Jesus J Caban and David Gotz. Visual analytics in healthcare–opportunities and research challenges, 2015.

[164] Krist Wongsuphasawat, John Alexis Guerra Gómez, Catherine Plaisant, Taowei David Wang, Meirav Taieb-Maimon, and Ben Shneiderman. Lifeflow: visualizing an overview of event sequences. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1747–1756. ACM, 2011.

[165] A Jay Block. What is" guarded condition" anyway? *Chest*, 105(3):649–650, 1994.

[166] Samuel J. Stratton. Acute respiratory failure, 2019. `https://bestpractice.bmj.com/topics/en-us/853` (accessed June 10, 2019).

[167] Centers for Disease Control and Prevention. ICD - Classification of Diseases, Functioning, and Disability. `https://www.cdc.gov/nchs/icd/index.htm` (accessed May 28, 2019).

[168] Brian D Stein, Adriana Bautista, Glen T Schumock, Todd A Lee, Jeffery T Charbeneau, Diane S Lauderdale, Edward T Naureckas, David O Meltzer, and Jerry A Krishnan. The validity of international classification of diseases, ninth revision, clinical modification diagnosis codes for identifying patients hospitalized for COPD exacerbations. *Chest*, 141(1):87–93, 2012.

[169] Shawn D Aaron, Katherine L Vandemheen, François Maltais, Stephen K Field, Don D Sin, Jean Bourbeau, Darcy D Marciniuk, J Mark FitzGerald, Parameswaran Nair, and Ranjeeta Mallick. Tnf antagonists for acute exacerbations of COPD: a randomised double-blind controlled trial. *Thorax*, 68(2):142–148, 2013.

[170] Ivan Tomasic, Nikica Tomasic, Roman Trobec, Miroslav Krpan, and Tomislav Kelava. Continuous remote monitoring of COPD patientsjustification and explanation of the requirements and a survey of the available technologies. *Medical & biological engineering & computing*, 56(4):547–569, 2018.

[171] Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Clinical intervention prediction and understanding using deep networks. *arXiv preprint arXiv:1705.08498*, 2017.

[172] Syed Ahmar Shah, Carmelo Velardo, Andrew Farmer, and Lionel Tarassenko. Exacerbations in chronic obstructive pulmonary disease: identification and prediction using a digital health system. *Journal of medical Internet research*, 19(3):e69, 2017.

[173] Sunil Nair, Douglas Hsu, and Leo Anthony Celi. Challenges and opportunities in secondary analyses of electronic health record data. In *Secondary Analysis of Electronic Health Records*, pages 17–26. Springer International Publishing, 2016.

[174] Eric J Topol and Dick Hill. *The creative destruction of medicine: How the digital revolution will create better health care*. Basic Books New York, 2012.

[175] Nicola Bartolomeo, Paolo Trerotoli, and Gabriella Serio. A multistate model to evaluate COPD progression integrating drugs consumption data and hospital databases. *Epidemiology, Biostatistics and Public Health*, 12(2), 2015.

[176] Akshay Sood, Hans Petersen, Clifford Qualls, Paula M Meek, Rodrigo Vazquez-Guillamet, Bartolome R Celli, and Yohannes Tesfaigzi. Spirometric variability in smokers: transitions in COPD diagnosis in a five-year longitudinal study. *Respiratory research*, 17(1):147, 2016.

[177] Eli Sherman, Hitinder Gurm, Ulysses Balis, Scott Owens, and Jenna Wiens. Leveraging clinical time-series data for prediction: a cautionary tale. In *AMIA Annual Symposium Proceedings*, volume 2017, page 1571. American Medical Informatics Association, 2017.

[178] Xiongcai Cai, Oscar Perez-Concha, Enrico Coiera, Fernando Martin-Sanchez, Richard Day, David Roffe, and Blanca Gallego. Real-time prediction of mortality, readmission, and length of stay using electronic health record data. *Journal of the American Medical Informatics Association*, 23(3):553–561, 2015.

[179] Chaitanya Shivade, Preethi Raghavan, Eric Fosler-Lussier, Peter J Embi, Noemie Elhadad, Stephen B Johnson, and Albert M Lai. A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association*, 21(2):221–230, 2013.

[180] Lurdes Y T Inoue, Ruth Etzioni, Christopher Morrell, and Peter Müller. Modeling disease progression with longitudinal markers. *Journal of the American Statistical Association*, 103(481):259–270, 2008.

[181] Zhengxing Huang, Wei Dong, Fei Wang, and Huilong Duan. Medical inpatient journey modeling and clustering: a bayesian hidden markov model based approach. In *AMIA Annual Symposium Proceedings*, volume 2015, page 649. American Medical Informatics Association, 2015.

[182] Xiang Wang, David Sontag, and Fei Wang. Unsupervised learning of disease progression models. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 85–94. ACM, 2014.

[183] Marcos LP Bueno, Arjen Hommersom, Peter JF Lucas, Mariana Lobo, and Pedro P Rodrigues. Modeling the dynamics of multiple disease occurrence by latent states. In *International Conference on Scalable Uncertainty Management*, pages 93–107. Springer, 2018.

[184] A Rodina, K Bliznakova, and K Stavrianou. Prevalence prognosis of the end stage renal disease patients in greece. In *World Congress on Medical Physics and Biomedical Engineering, September 7-12, 2009, Munich, Germany*, pages 237–240. Springer, 2009.

[185] Zitao Liu and Milos Hauskrecht. A personalized predictive framework for multivariate clinical time series via adaptive model selection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1169–1177. ACM, 2017.

[186] Nancy Yesudhas Jane, Khanna Harichandran Nehemiah, and Kannan Arputharaj. A temporal mining framework for classifying un-evenly spaced clinical data. *Applied clinical informatics*, 7(01):1–21, 2016.

[187] National library of medicine. rxnorm. `https://www.nlm.nih.gov/research/umls/rxnorm/` (accessed Jan 12, 2019).

[188] Ernesto Crisafulli, Enric Barbeta, Antonella Ielpo, and Antoni Torres. Management of severe acute exacerbations of COPD: an updated narrative review. *Multidisciplinary respiratory medicine*, 13(1):36, 2018.

[189] Andrew M Fine, Ben Y Reis, Lise E Nigrovic, Donald A Goldmann, Tracy N LaPorte, Karen L Olson, and Kenneth D Mandl. Use of population health data to refine diagnostic decision-making for pertussis. *Journal of the American Medical Informatics Association*, 17(1):85–90, 01 2010.

[190] Colin R Cooke, Min J Joo, Stephen M Anderson, Todd A Lee, Edmunds M Udris, Eric Johnson, and David H Au. The validity of using icd-9 codes and pharmacy records to identify patients with chronic obstructive pulmonary disease. *BMC health services research*, 11(1):37, 2011.

[191] Tao Wu and David F. Gleich. Retrospective higher-order markov processes for user trails. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 1185–1194. ACM, 2017.

[192] Barbara Franz, Andreas Schuler, and Oliver Krauss. Applying fhir in an integrated health monitoring system. *EJBI*, 11(2):51–56, 2015.

[193] Judith E Nelson, Christopher E Cox, Aluko A Hope, and Shannon S Carson. Chronic critical illness. *American journal of respiratory and critical care medicine*, 182(4):446–454, 2010.

[194] Marc Miravitlles and Antonio Anzueto. Role of infection in exacerbations of chronic obstructive pulmonary disease. *Current opinion in pulmonary medicine*, 21(3):278–283, 2015.

[195] Shireen Mirza, Ryan D Clay, Matthew A Koslow, and Paul D Scanlon. COPD guidelines: a review of the 2018 gold report. In *Mayo Clinic Proceedings*, volume 93, pages 1488–1502. Elsevier, 2018.

[196] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

A RECOMMENDATIONS FROM GOLD GUIDELINES

A.1 ABCD Assessment Tool

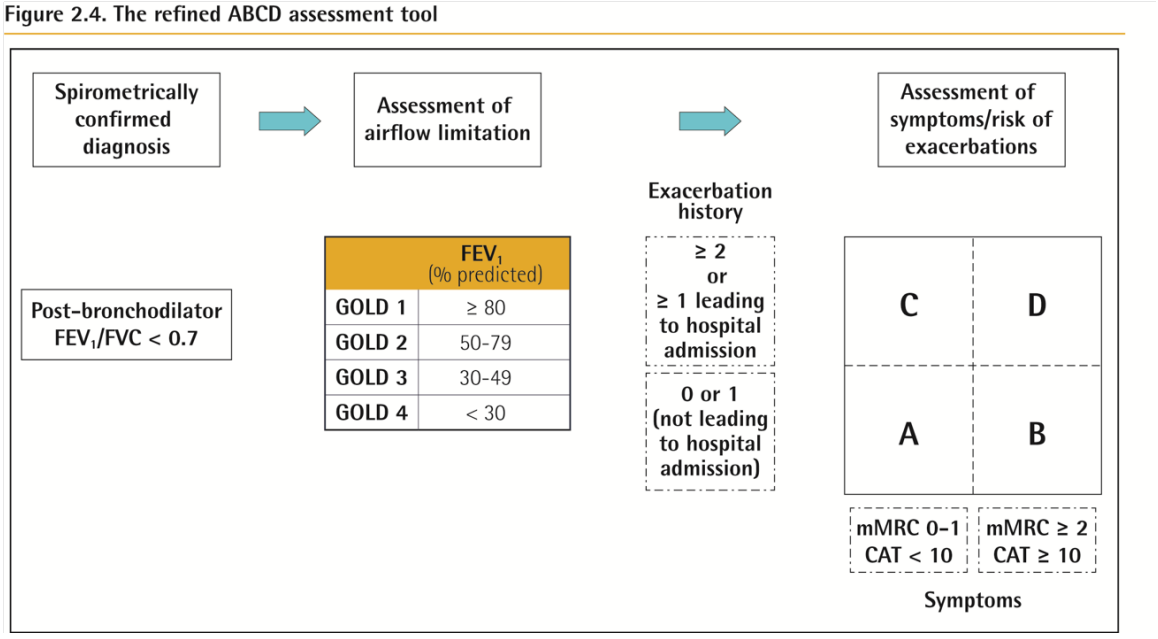


Figure A.1.: GOLD ABCD assessment tool. Adapted from [2].

A.2 Pharmacological Treatment Algorithm

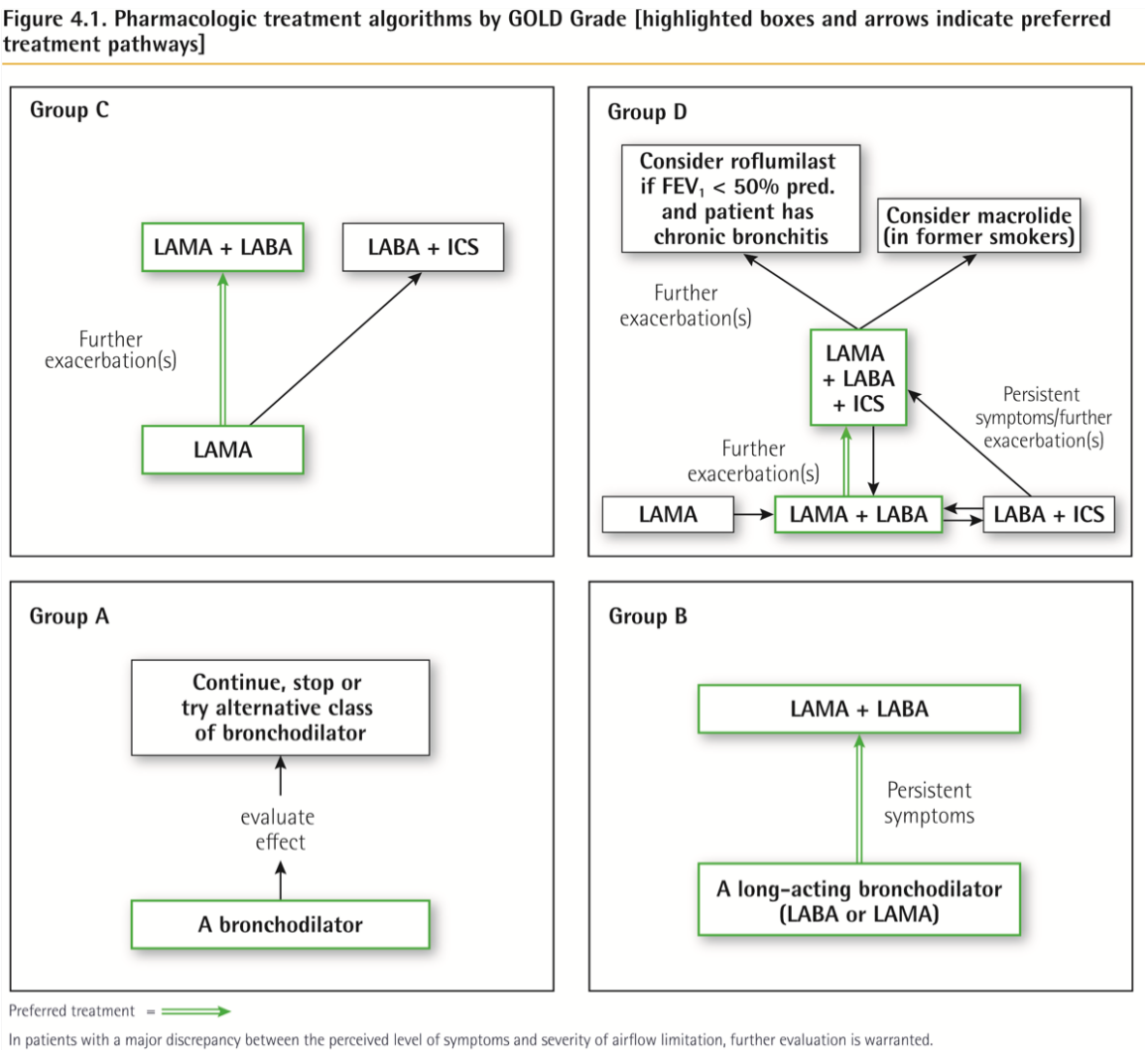


Figure A.2.: GOLD pharmacological treatment algorithm. Adapted from [2].

A.3 Classifications of Exacerbations

Table A.1.: Severity of exacerbations for AECOPD based on clinical signs. Adapted from [2].

Classification	**RR (bpm)**	**Use of accessory muscles**	**Change in mental status**	**Hypoxemia improved with supplemental O_2**	**$PaCO_2$ increased compared with baseline or elevated**
No respiratory failure	20-30	No	No	FiO_2 28-35%	None
Acute respiratory failure (non-life-threatening)	>30	Yes	No	FiO_2 25-30%	50-60 mmHg
Acute respiratory failure (life-threatening)	>30	Yes	Yes	not improved or requiring FiO_2 >40%;	60 mmHg or presence of acidosis (pH <7.25)

bpm - beats per minute

mmHg - millimeter of mercury

B TRANSITION PROBABILITY MATRICES

B.1 Model Parameters for Clinical Question 1

$$P_{\text{ALL}} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0.309 & 0.052 & 0.079 & 0.519 & 0.040 \\ 0.222 & 0.085 & 0.020 & 0.642 & 0.031 \\ 0.489 & 0.004 & 0.180 & 0.290 & 0.036 \\ 0.000 & 0.000 & 0.000 & 1.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 1.000 \end{pmatrix} \end{array}$$

$$P_{\text{ANTIBIOTICS}} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0.290 & 0.038 & 0.055 & 0.599 & 0.018 \\ 0.081 & 0.021 & 0.000 & 0.898 & 0.000 \\ 0.513 & 0.000 & 0.169 & 0.278 & 0.040 \\ 0.000 & 0.000 & 0.000 & 1.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 1.000 \end{pmatrix} \end{array}$$

$$P_{\text{HALF}} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0.302 & 0.047 & 0.070 & 0.549 & 0.032 \\ 0.174 & 0.064 & 0.013 & 0.729 & 0.021 \\ 0.499 & 0.003 & 0.176 & 0.285 & 0.038 \\ 0.000 & 0.000 & 0.000 & 1.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 1.000 \end{pmatrix} \end{array}$$

$$P_{\text{NO ANTIBIOTICS}} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0.314 & 0.056 & 0.085 & 0.498 & 0.046 \\ 0.267 & 0.106 & 0.027 & 0.560 & 0.041 \\ 0.484 & 0.005 & 0.183 & 0.293 & 0.035 \\ 0.000 & 0.000 & 0.000 & 1.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 1.000 \end{pmatrix} \end{array}$$

Figure B.1.: Model parameter estimations for AECOPD based on antibiotics administration presented by group. ALL represents the entire AECOPD cohort, and the other groups are subsets of ALL. Row and column labels correspond to health and outcome states in Table 5.2.

B.2 Model Parameters for Clinical Question 2

$$P_{0\geq\text{hours}\leq 6} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0.287 & 0.010 & 0.022 & 0.658 & 0.023 \\ 0.012 & 0.003 & 0.000 & 0.985 & 0.000 \\ 0.375 & 0.000 & 0.151 & 0.460 & 0.014 \\ 0.000 & 0.000 & 0.000 & 1.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 1.000 \end{pmatrix} \end{array} \qquad P_{6>\text{hours}\leq 24} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0.278 & 0.040 & 0.069 & 0.606 & 0.008 \\ 0.062 & 0.066 & 0.000 & 0.872 & 0.000 \\ 0.553 & 0.000 & 0.131 & 0.229 & 0.087 \\ 0.000 & 0.000 & 0.000 & 1.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 1.000 \end{pmatrix} \end{array}$$

$$P_{24>\text{hours}\leq 48} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0.239 & 0.007 & 0.061 & 0.681 & 0.011 \\ 0.115 & 0.004 & 0.000 & 0.881 & 0.000 \\ 0.672 & 0.000 & 0.328 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 1.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 1.000 \end{pmatrix} \end{array} \qquad P_{\text{hours}>48} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0.350 & 0.095 & 0.069 & 0.453 & 0.034 \\ 0.201 & 0.011 & 0.000 & 0.788 & 0.000 \\ 0.844 & 0.000 & 0.156 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 1.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 1.000 \end{pmatrix} \end{array}$$

Figure B.2.: Model parameter estimations for AECOPD based on initial timing (hours) of antibiotics administration presented by group. Row and column labels correspond to health and outcome states in Table 5.2.

VITA

Amber M. Johnson is a Jackson, MS native, and a proud graduate of William H. Lanier High School. She is the daughter of Martha Murrell and Alvin Johnson, both Jackson State University (JSU) graduates. While at Lanier, Amber was a standout student, both academically and athletically, following the path paved by her mother and older cousins, Gwendolyn Jones-Thibodaux and Genina Johnson, who both paved the way as former standout Lanier scholars and athletes. Johnson attributes her sharp mathematical and computational skills to her mother, a Mathematics major at JSU, and former basketball coach and mentor, Erica Stringfellow-Smith.

Upon graduating from Lanier, Johnson received a full basketball and cross-country scholarship to Tougaloo College, where she majored in Computer Science and became a member of Zeta Phi Best Sorority, Inc. After her second year at Tougaloo, she took her talents to Mississippi's neighboring state, Memphis, TN, to join The LeMoyne-Owen College (LOC) family, where she received her Bachelors degree in Computer Science in 2011. While at LOC, Amber served as President of the Student-Athlete Advisory Committee (SAAC) and a member of several other organizations.

Amber went on to receive her M.S. in Computer Science from JSU in 2013 as a Louis Stokes Mississippi Alliance for Minority Participation (LSMAMP) Bridge to the Doctorate scholar. Upon graduating from JSU, Amber received multiple fellowships, including the National Physical Science Consortium (NPSC), to study at the worlds first Computer Science program (Est. 1962) at Purdue University. While at Purdue, Amber served as a member of several organizations on campus, and most notably, as President of the Computer Science Graduate Student Board, Amber piloted programs intended to create community within the Computer Science department, such as a peer mentoring program as well as Social Power Hour, research competition where

graduate students are able to share their research and helped to enhance their public speaking and presentation skills.

Amber also piloted her own dissertation research, “Generating Evidence COPD Clinical Guideline Using EHRs”, where she worked as a Graduate Student Data Scientist at the Regenstrief Center for Healthcare Engineering (RCHE) at Purdues Discovery Park. Her work was inspired by the her aunt, Ethel Mae Cooper, a pillar of the Jackson-Georgetown community, who suffered from COPD for nearly 30 years before passing in 2014, just 1 month before the start of Johnson’s second year at Purdue. Johnson served in many capacities while at Purdue, including a mentor and instructor for both Girls Who Code and Black Girls Rock Tech. This year, Purdue celebrates its 150th year anniversary, with a theme 150 Years of Giant Leaps , and with great honor and determination, on August 3rd, 2019, Amber M. Johnson will most certainly take a giant leap, becoming the first African American woman to graduate with a PhD in Computer Science from Purdue University.