Validation for Drug Repurposing Candidates

by

Malvika Pillai

Dissertation Proposal for the Degree of Doctor of Philosophy
The University of North Carolina at Chapel Hill
Carolina Health Informatics Program

August 31, 2020

Contents

1 Introduction **2**

1.1 Problem definition and Motivation 4

1.2 Research Aims 6

2 Background and Related Work **6**

2.1 EMR data use in validation 6

2.2 EMR data use in drug candidate prediction 7

2.3 Limitations 8

3 Methodology **9**

3.1 Aim 1. Produce a computational phenotyping algorithm using electronic medical records. 9

3.2 Aim 2. Build a pipeline for retrospective clinical record analysis to validate drug repurposing candidates. 17

4 Timeline **24**

References **26**

Introduction

The traditional process for drug development can take approximately 12 to 16 years and cost approximately $1 to $2 billion [1]. The process consists of the following stages: drug discovery and development, pre-clinical development, Phase I-III clinical trials, and regulatory approval. Due to the high cost and time burden of the traditional process, alternative options for drug development must be explored. Drug repurposing or repositioning is the process of applying known drugs/compounds that are already on the market to new disease indications and has been successfully used to expedite this process. Repositioned drugs are exempt from the stages prior to Phases II and III of the clinical trials and FDA approval process reducing time and cost (Figure 1). For example, a liberal estimate for cost and number of years required to reposition a drug is approximately $300 million and approximately 6 years [1]. Putting potential drugs on the market faster can have positive downstream effects on population health outcomes, and the decreased cost makes drug repositioning attractive to researchers and patients. Due to the delays and barriers of going from a molecule to an approved drug, there has been a national push toward drug repositioning.

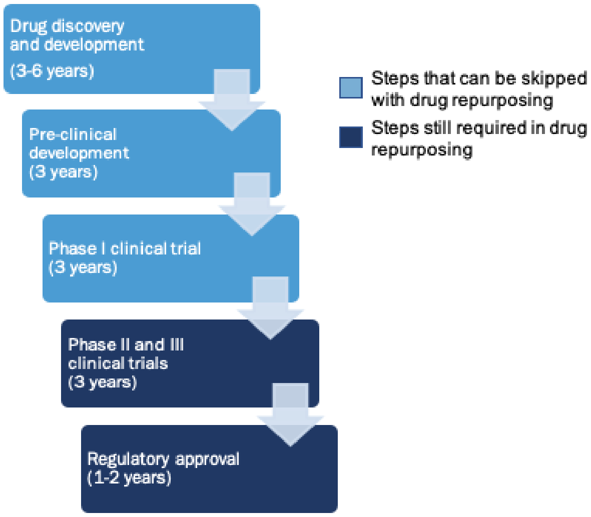

Drug discovery and development (3-6 years)
Pre-clinical development (3 years)
Phase I clinical trial (3 years)
Phase II and III clinical trials (3 years)
Regulatory approval (1-2 years)
Steps that can be skipped with drug repurposing
Steps still required in drug repurposing

Figure 1: Traditional Drug Development Process

Over the past 60 years, there has been significant increase in spending for drug development, with few drugs approved. A computing term, Moore's Law, is the idea that as the number of transistors on a microchip (i.e., computing power) doubles every two years, the monetary cost of computers is halved. The term "Eroom's Law" (i.e., the inverse of Moore's Law) is used to describe the inverse correlation of increased monetary input into drug development and the number of drugs approved remaining flat or decreasing [1]. However, recent evidence has shown the dismantling of "Eroom's Law" due to the following factors: an increase in genetics-based drug development, better use of information (i.e., decision-making), and less stringent thresholds for FDA approval [2]. Drug repurposing falls under all of these overarching factors that have indicated an increase in drugs coming out to market. Genetics-based prediction is one of the most common methods used to identify drug repurposing candidates. By using existing information and not wasting time or effort doing research that others have already done, drug repurposing can lead to better decision-making. Lastly, although recent evidence points to less stringent thresholds for FDA approval by way of the Orphan Drug Act [2], there are also fast-track approval pathways for drug repurposing candidates [3]. Therefore, drug repurposing can help minimize the disparity between increased spending for drug development and number of drug approvals.

From the perspective of monetary returns on drug research, according to BCC Research, the global market for drug repurposing reached $24.4 billion in 2015 and was projected to reach over $30 billion in 2020 [4, 5]. Many successful attempts of drug repurposing have been accidentally discovered side effects or extensive, time intensive research on particular drug properties [6]. Sildenafil was originally developed to treat angina and was repurposed, by chance, to treat erectile dysfunction. Global sales for sildenafil for erectile dysfunction totaled $2.05 billion in 2012 [7]. Minoxidil was developed to treat hypertension and was repurposed for hair loss through identification of hair growth as an adverse side effect. Global sales for minoxidil for hair loss were $860 million in 2016 [4]. Both sildenafil and minoxidil were repurposed through retrospective clinical analysis [4, 6].

Often, successful examples of drug repurposing have been by chance, but recent approaches that are more direct are being explored in the field. Computational drug repurposing consists of using computational approaches for systematic data analysis that can lead to forming drug repurposing hypotheses [4]. Omics-based repurposing, for example, has been shown to increase success in clinical development of a drug candidate [2]. -Omic information can provide a comprehensive view of a set of molecules and insight into the functions of a cell, tissue, or organism. The most mature -omic field, genomics, focuses on identifying genetic variants associated with disease, response to treatment, and more [8]. However, often times the translation from research to clinical development is hindered by a lack of information bridging the two. In computational drug repurposing, researchers often output a series of drug-disease associations or drug-target interactions; of which, some results are true positives and many are false positives. Narrowing the candidate list is important to identify the strongest candidates that have the highest chance of successfully treating a condition, and this can be done through drug candidate validation (i.e., providing independent supporting evidence). The various types of supporting evidence that researchers have considered as validation were described in detail in the previous literature review [9], and of all computational validation methods, retrospective clinical analysis was found to be the strongest.

Electronic medical records (EMR) contain an overview of a patient's health that can be used to bridge the gap between drug repurposing research and clinical implementation. Retrospective clinical analysis, and more specifically, EMR validation is a powerful method to bridge the gap between research and clinical development. The combination of structured components of the EMR and unstructured clinical notes contain information that can provide a comprehensive, longitudinal view of patient health. In related work, EMR data has been used to predict the probability of treatment success using statistical approaches [10, 11]. To do so, researchers identify patient populations, separate patients as cases and controls, and predict disease improvement caused by treatment with a drug repurposing candidate.

For patient population identification and case-control separation, there are various approaches to computationally phenotype conditions [12], but in drug repurposing studies the main identification approach is searching EMR databases with ICD-CM billing codes [10]. Although billing codes have been widely used in the past, a comprehensive search strategy would include other sources of information to ensure that all patients who may have a disease diagnosis are accounted for in a sample. For example, in the case of a patient who has breast cancer, the EMR would include billing codes, images, biopsy results, and more variables which could be used to define a disease diagnosis. In research on EMR validation for drug repurposing candidates, computational phenotyping approaches must be considered to construct comprehensive search strategies for patient population identification.

To predict probability of treatment success for validation, studies have predominantly used data from the structured components of the EMR, and some have supplemented missing structured data with information from clinical notes. However, many challenges in analyzing the unstructured components (e.g., variability of natural language expressions) have made analysis of clinical free-text difficult and computationally intensive. For drug repurposing, the notes contain medical reasoning behind prescriptions as well as documentation of any adverse side effects. Advances in clinical natural language processing (NLP) like in named entity recognition (NER) can facilitate large-scale analysis of unstructured clinical notes as well, broadening the scope of EMR data that can be accessed and analyzed. The predictive task has previously been solved with finely focused condition-specific models, indicating a need for a generalizable method for EMR validation of drug repurposing candidates. Machine learning models have been successfully used in EMR validation in related work, and deep learning models have produced promising results in other predictive contexts [13, 14]. In comparison to using statistical models, using machine or deep learning approaches may make EMR validation algorithms more generalizable.

Three condition cases will be considered in the study: breast cancer, oral cancer, and primary ciliary dyskinesia (PCD). Breast cancer is a highly prevalent and widely studied condition in drug repurposing and will serve as a proof of concept for the EMR validation algorithms that will be created. 1 in 8 women (13% of women) receive a breast cancer diagnosis in the United States [15]. Oral cancer is a less commonly researched cancer, which needs early detection and treatment, and affects an estimated 10.5 adults in 100,000 (0.0105% of all adults)[16]. PCD is a rare, genetic disease, that also needs early detection and treatment, and affects an estimated 1 in 16,000 people (0.00625% of people)[17]. Having various condition cases will test the EMR validation algorithms as they differ greatly in terms of medical need and prevalence.

In the proposed study, I will develop algorithms for the cohort extraction and disease improvement prediction stages of EMR validation for drug repurposing candidates. The aim will be to produce algorithms for computational phenotyping and improvement prediction, presenting a way for researchers conducting drug repurposing studies to validate their results with EMR. Section 1 (p. 2) includes the problem definition, motivation for this work, and research aims. Section 2 (p. 6) describes studies using EMR for validation and drug candidate prediction as well as the limitations in current work. Section 3 (p. 9) describes the plan for the proposed work. Section 4 (p. 24) is the timeline. References are included at the end of the document (p. 26).

Problem Definition And Motivation

Between 2007 and 2009, drug repurposing led to the launch of 30-40% of new drugs, which addresses the time and cost burden of drug development but also presents opportunities to address unmet medical need [18]. For example, rituximab was developed as a treatment for various cancers but was repurposed to treat rheumatoid arthritis. From the cost perspective, global sales for rituximab in 2012 were greater than $7 billion [19], where approximately 17% of sales were targeted for rheumatoid arthritis [20]. From the medical need perspective, rheumatoid arthritis is a complex disease for which its pathogenesis is only partially understood. For conditions with poorly characterized pathophysiology, drug repurposing is often the only route for drug development. Lopez-Olivo *et al* (2015)[21] showed that the usage of rituximab for rheumatoid arthritis has had positive impact on patient quality of life. 70 of 100 people who took rituximab in combination with methotrexate, the standard treatment, perceived their general health to be better in comparison with 36 of 100 people who took the standard treatment, methotrexate, alone [21]. Drug repurposing is not only aimed at reducing time and cost burden for drug developers, it is also a critical method to meet medical need.

Past retrospective clinical analysis successes have been random events, motivating systematic approaches. With the increased proliferation of EMR systems, the volume of EMR data is predicted to grow astronomically [22]. The power of health data creates an opportunity to explore clinical records and validate drugs by identifying cases in which clinicians have prescribed drugs for purposes other than their intended or cases in which patients are taking multiple drugs that have unprecedented interactive effects. Previous successful applications of drug repurposing from retrospective clinical analysis were not conducted with systematic computational analysis [23, 24, 25]; however, many successes from this method motivate the creation of automated, computational approaches.

Although EMR data is powerful, quickly growing, and has been used successfully in the past, there are many factors contributing to its complexity. The physician workflow consists of four overarching components: information review, patient assessment, EMR documentation, and care delivery. For a single patient visit, EMR documentation should include information verbally provided by the patient, previous written documentation (e.g., family history), and documentation of care (e.g., diagnostic strategy, treatment plan) [26]. To provide context, if there is a female patient who is 24 years of age, she may have at least 1 to 3 visits yearly of different types (e.g., annual exam, emergency), which would constitute 24 to 72 visits over her current lifetime, with each visit having its own documentation. If the patient only visited one healthcare system in her lifetime, all visits would be documented in one EHR system, assuming the system had been instituted before her first visit or that the system contains legacy records. However, even in the simplistic example provided, there are many intersecting components of EMR data that are being generated over time (e.g., laboratory results, medical imaging), demonstrating the vast, dense, and longitudinal nature of EMR data.

Along with the complexity of EMR data, using EMR for clinical research has been hindered by the lack of support for data manipulation provided by electronic health record (EHR) systems. The original purpose of EHR was to support clinical care and billing. Workflows for clinical research were integrated as a secondary purpose; however, significant progress has been made since the inception of EHR, including the Meaningful Use incentives put forth in 2009. Consequently, various common data models have been instituted to allow researchers easier access to EMR and help with data integration, but these efforts are still in progress. Lack of data interoperability and data integration are a few of many issues persisting with EMR use for research [27].

EMR complexity and the lack of support for data manipulation in EHR lend to the use of machine learning methods for data extraction and analysis. Traditionally, statistical methods have been used to perform retrospective clinical analysis. However, in dealing with high-dimensional data, machine learning methods can outperform traditional statistical approaches. Machine learning uses data-driven and statistical rules in order to transform feature representations of input data into desired outputs. It can be described as an extension of traditional statistical approaches [28]. Ideal machine learning tasks are aimed at developing systems that are too expensive in terms of processing time or power or too difficult to program explicitly as standard computational algorithms. There are drawbacks to machine learning, however, that can be addressed with deep learning approaches. Feature engineering (i.e., transforming raw data into a form understandable by the machine) is needed for machine learning approaches. However, deep learning consists of representation learning methods, where the machine can be fed raw data, detect representations of the data, and complete the prediction task. The feature representations generated are done using general procedures, so domain expertise is not required in the process, allowing for a more generalizable approach [13]. For computational phenotyping, a high level of transparency is required, so only machine learning approaches will be used for cohort extraction and case/control prediction. For treatment success prediction, both machine and deep learning approaches will be explored. While deep learning methods are not as transparent as machine learning methods, they have can achieve higher performance in some cases, as demonstrated in research areas [29, 30, 31]. To leverage the full potential of EMR, machine and deep learning methods can be used to take patient-level data variables and predict viability of drug repurposing candidates.

Research Aims

Aim 1: Produce a computational phenotyping algorithm using electronic medical records.
Aim 2: Build a pipeline for retrospective clinical record analysis to validate drug repurposing candidates.

Background And Related Work

Over the past decades, there has been increasing implementation of electronic health record (EHR) systems, allowing for a large amount of data to be produced on the patient and population levels. In terms of drug repurposing, EHRs can provide longitudinal information that can be used to predict drug outcomes and validate drug candidates [10]. Given a drug candidate and its target indication, various methods have been used to connect the two.

A review was conducted following the *PRISMA Statement* for systematic reviews [32] to identify key literature associated with drug repurposing and validation [9]. Subsequently, studies using electronic health records for either drug repurposing candidate prediction or validation were selected. After 2386 articles were screened, 732 were reviewed in full, and 10 studies using clinical records as a data source were selected. Of the 10 studies, 5 used clinical records in validation and 5 used clinical records in drug candidate prediction. The studies using clinical records in validation and prediction are described in detail in terms of prediction task, dataset, and assumptions. Sample size estimates from literature are shown in Table 1.

EMR Data Use In Validation

In studies using clinical records for validation, the validation methods used included Cox proportional hazard analysis [10, 11, 24], other statistical analysis [25], and off-label use extraction [23]. Of all the studies, Xu *et al* (2015)[10] is the only study that did not include any candidate prediction and only sought to validate a drug repurposing hypothesis. The study used a stratified Cox proportional hazards model to validate the association between metformin use, which is originally meant for type 2 diabetes mellitus treatment, and cancer mortality. In the study, diabetic individuals with breast, colorectal, lung, or prostate cancer were identified and divided into four groups based on disease and medication statuses. Consequently, clinical covariates were retreived from structured components of the EMR using data extraction algorithms and retrieved from clinical narratives using NLP algorithms. Then, the statistical model was used to examine the effect of metformin use on cancer survival for each diabetes group [10].

Other studies using Cox proportional hazards models aimed to associate predicted drug use with treatment success [11, 24, 25]. Khatri *et al* (2013)[11] identified therapeutics to combat acute rejection in organ transplantation and used models to associate statin use with graft survival. The study adjusted for donor and recipient ages, repeat transplantation, and year [11]. Gayvert *et al* (2016)[24] focused on drug repurposing for cancer and used retrospective cohort analysis with EMR to validate the association between dexamethasone treatment and prostate cancer. The study used Kaplan-Meier survival analysis and used the Cox proportional hazards test to test for significance. A logistic regression model was then developed to assess the relationship between treatment (e.g., dexamethasone and control) and prostate cancer diagnosis, independent of prostate cancer confounders. Using the logistic regression model, the study found that dexamethasone had a protective effect against prostate cancer. Xu *et al* (2018)[25] and Gottlieb *et al* (2014)[23] did not provide detailed methodologies for their validation processes. Xu *et al* (2018)[25] provided background for patient record extraction, cohort selection, and stated t-test p-values along with derived conclusions. Gottlieb *et al* (2014)[23] extracted off-label uses from EMR but did not provide a methodology for the process.

EMR Data Use In Drug Candidate Prediction

In studies using clinical records for drug candidate prediction, both statistical analysis methods [33, 34, 35] and machine learning methods [36, 37] were used. The statistical methods used were fixed effect models and machine learning methods like logistic regression, random forest, and neural networks for classification.

Koren *et al* (2018)[36] used machine learning methods to predict computational drug repurposing candidates for hypertension from electronic health records. The dataset used contained 30,705 patients. The study used logistic regression as a form of propensity score matching in order to predict treatment success for potential anti-hypertensive agents. For cohort identification, Koren *et al* (2018)[36] only included patients that had at least two initial systolic and diastolic blood pressure values in a given timeframe. Low *et al* (2017)[37] used both gene expression and EMR data to predict drug candidates for breast cancer patients. The study constructed a logistic regression model with pairwise interactions and used lasso regularization. In the EHR analysis, the study differentiated between individual and combination effects of drug exposure. Demographic, tumor, and treatment variables from patient records were processed into a matrix to account for concomitant drug exposures and possible pairwise combinations that met inclusion criteria were outputted. All variables were included in the logistic regression model. The task was structured as prediction of binary 5-year mortality, and results on a 10% holdout validation set were presented (90% area under the curve (AUC), 40% sensitivity, 99% specificity) [37]. The study included 1,212 cases (i.e., dead) and 8,733 controls (i.e., alive), with a 10%/90% data split in response variables. Low *et al* (2017)[37] further differentiated between variables associated with survival in the EHR. Variables associated with lower mortality included lower tumor stage and living in a neighborhood of the top 20% in socioeconomic status in California. Variables associated with higher mortality included: advanced tumor stage, having triple negative breast cancer (TNBC), and older age at diagnosis [37]. The study did not differentiate groups by breast cancer subtype in the primary classification but consequently conducted a subgroup analysis. Two synergistically beneficial pairs were found for breast cancer treatment: anti-inflammatory agents with lipid modifiers as well as anti-inflammatory agents with anticancer hormone antagonists.

Three studies used variations of fixed effect models for prediction. Paik *et al* (2015)[35] combined EMR laboratory test results and genomic signatures from public databases to construct a bipartite network for drug repurposing. The study calculated drug-drug and disease-disease similarities using clinical and genomic signatures to create two similarity matrices each for drug-disease association prediction. Similarities between pairs were represented as edge widths. Kuang *et al* (2016)[33] proposed a continuous self-controlled case series (CSCCS) model for computational drug repurposing. The use case presented in this study is to look for drugs that can control fasting blood glucose levels, which are important for diabetes regulation. To identify off-label usage, Kuang *et al* (2016)[33] examined fasting blood glucose levels before and after any drug was prescribed to a patient. The CSCCS model was derived from the linear fixed effect model to take drug prescription history into consideration by differentiating between drugs prescribed for longer or shorter durations. To account for different effects of drugs associated with impacting fasting blood glucose levels, the study separated the drugs into three categories: decrease levels, increase levels, and irrelevant/possible discoveries [33]. The study did not provide details on how the EHR data was extracted or how the cohort was identified. In another study conducted by the same group, Kuang *et al* (2016)[34] used baseline regularization and a variant to extend the one-way fixed effect model. The baseline regularization model assumes that there is a baseline state for fasting blood glucose level and that based on various drug exposures, there is an exposure state for fasting blood glucose level. Based on these assumptions, the study constructed a fixed effect model with regularization on baseline parameters. Like Kuang *et al* (2016)[33], the study did not include any details on cohort identification and EHR data extraction [34].

Table 1: EMR sample sizes in literature

Study	Sample size (in patients)
EMR Use in Validation	
Khatri *et al* (2013)	2,515
Xu *et al* (2015)	42,165
Gayvert *et al* (2016)	–
Gottlieb *et al* (2014)	–
Xu *et al* (2018)	–
EMR Use in Prediction	
Paik *et al* (2015)	530,000
Koren *et al* (2018)	30,000
Kuang *et al* (2016)[33] & (2016)[34]	64,515
Low *et al* (2017)	9,945

Limitations

Many studies rely on the linear fixed effect model in order to predict drug candidates. Statistical approaches to causal inference are very powerful; however, machine learning algorithms are able to outperform classical statistical techniques in cases with high-dimensional data. In addition, many studies focus solely on using structured data (e.g., ICD-CM billing codes) from the EMR as they are more accessible than data from clinical notes. The use of ICD-CM billing codes does not provide enough granularity in defining a disease diagnosis to draw conclusions on whether or not particular groups of patients would benefit from taking a specific repurposed drug. The exception is the work conducted by Xu *et al* (2015) as they used NLP algorithms to extract data from clinical notes. The clinical notes contain background information like patient occupation, duration of symptoms, and medical reasoning for prescriptions given. The study focused on filling missing data from structured fields with information extracted from the clinical notes; however, the study did not use any extra information from the clinical notes to influence prediction. Other weaknesses of all studies discussed are the lack of mechanistic basis for treatment success and analysis on which groups of patients a drug should be targeted toward.

Methodology

– Study setting

The study will be conducted at the UNC Health Care System, where UNC Hospitals is a public, academic medical center that serves patients across North Carolina. All clinical, research and administrative data from UNC Health Care is housed in a central data repository called the Carolina Data Warehouse for Health (CDW-H). Data in the CDW-H consists of over 5 million unique patients with over 1 million active patients from 2004 onward and can be accessed by investigators with approval from the Institutional Review Board (IRB). As of 2014, UNC Health Care transitioned into the current EMR system and converted into the ICD-10 coding system in 2015. The data in CDW-H consists of legacy data and data from the current EMR system. While some structured data from the EMR can be de-identified, the unstructured clinical notes are considered identifiable due to HIPAA indicators found in the notes and require IRB approval for access. After IRB approval, a CDW-H Project Request form will be submitted to the North Carolina Translational and Clinical Sciences Institute (NC TraCS), which is an honest broker between researchers and the CDW-H, and considered based on feasibility, scope, and time and cost estimates. An NC TraCS data analyst will then extract and process the data for use [38].

The Carolina Mammography Registry (CMR) will be used as an additional data source for breast cancer if needed. The CMR is a source for community-based mammography screenings in North Carolina. Previous research has connected the CDW-H to the CMR, and if needed for the breast cancer condition case, the CMR and CDW-H will both be connected.

– Subjects

The conditions that will be considered are breast cancer, oral cancer, and primary ciliary dyskinesia (PCD). The conditions considered vary in prevalence, targeted population, and degree of medical need, providing a spectrum of test cases for the EMR phenotyping and validation algorithms proposed. For PCD, patients with a diagnosis in the date range, July 1, 2004 to May 10, 2020, will be included. IRB approval has been obtained to access data for patients with likelihood of PCD. For breast cancer and oral cancer, IRB approval has not been obtained; therefore, all patients diagnosed with breast cancer or oral cancer in the date range, July 1, 2004 to the date of IRB application submission, will be included in the study.

Aim 1. Produce A Computational Phenotyping Algorithm Using Electronic Medical Records.

3.1.1 Significance

Past studies [10] for validation of drug repurposing candidates have solely used ICD-CM diagnosis codes to identify patient cohorts, but these codes are meant for billing. Many rare conditions do not have specific ICD-CM billing codes and fall under an “Other” category. For example, PCD does not have a specific ICD-CM code. Instead, it falls under an umbrella

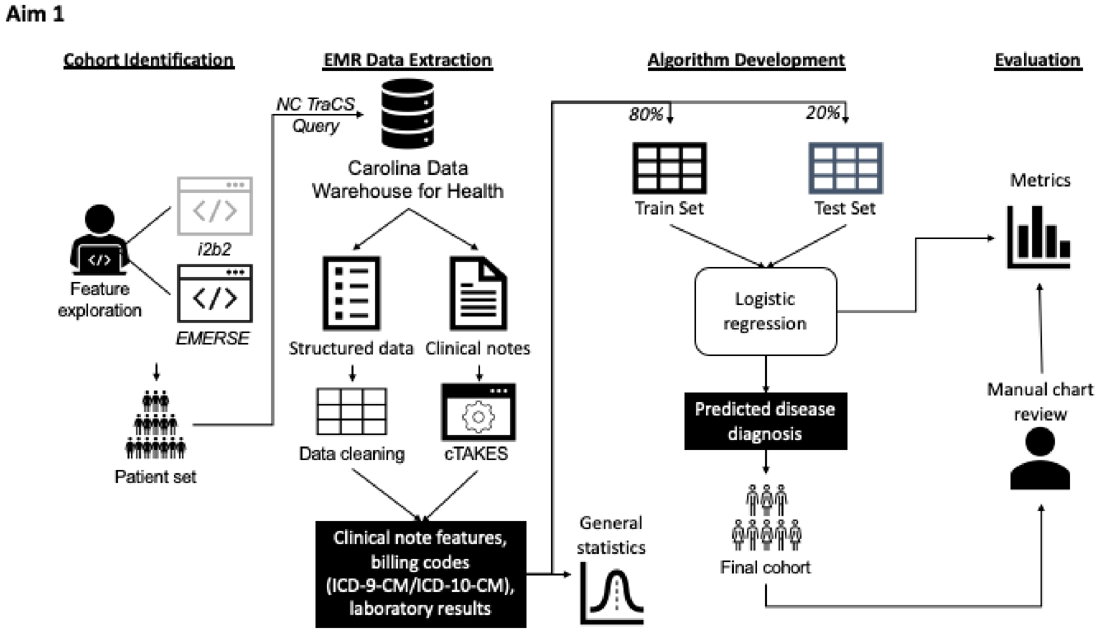


Figure 2: Aim 1 Flowchart

code called "Q34.8: Congenital pulmonary airway malformation". Breast cancer and oral cancer both have specific ICD-CM codes. However, breast cancer ICD-CM codes do not provide enough granularity about molecular subtype. Using other sources of information to supplement the diagnosis codes will make the extracted disease diagnosis more reliable. This aim is to incorporate different elements of the EMR to identify patient populations in a way that will enable drug repurposing validation analysis for researchers without much additional cohort manipulation. A previous study, Pfaff *et al* (2020)[39], has demonstrated success identifying PCD patients with high confidence, but the study only used clinical notes as a data source. This aim will take both structured and unstructured elements of the EMR to computationally phenotype patients. EMR provide a comprehensive view of a patient's health, and properly identifying a cohort is an important step for any retrospective clinical analysis study, especially within drug repurposing research.

3.1.2 Study Design

A computational phenotyping algorithm will be created to thoroughly identify patient populations for drug repurposing validation studies using EMR. Figure 2 shows the study design from cohort identification to evaluation. The goal is to computationally phenotype patients in three use cases (i.e., condition cases). EMR phenotyping prior to case/control prediction will be divided into two tasks: (1) extracting structured information and (2) extracting unstructured information. Structured information is comprised of patient demographics, billing codes, laboratory tests, medications, treatment, and vitals. Unstructured information consists of clinical notes and images. Since the goal is to identify patient populations for drug repurposing validation, medications will not be used as a source of information influencing identification. Only billing codes, laboratory tests, and clinical notes will be used for computational phenotyping. Due to differences in sample size from i2b2 cohort estimation, the process of extracting a disease diagnosis will differ between the conditions chosen. For example, for rare diseases like PCD, diagnosis itself is an ongoing research area, so a PCD-like diagnosis may also be considered. The aim is to computationally phenotype the conditions for cohort extraction generally before elaborating or fine tuning based on each condition. The process for extracting patient information will initially be the same for all condition cases, but fine tuning will then be done for each condition case. For example, oral-specific information will only be extracted for the oral cancer condition case. The final output of the aim will be a cohort of individuals divided into cases (i.e., patient has disease) and controls (i.e., patient does not have disease) or divided by case type (e.g., triple negative breast cancer).

3.1.3 Methods

3.1.3 (a) General approach

Before EMR extraction, cohort sizes will be approximated using Informatics for Integrating Biology and the Bedside (i2b2)[40]. i2b2 is a web application that is a view of UNC Health Care data, and it allows for the investigation of de-identified, aggregate data. The Electronic Medical Record Search Engine (EMERSE)[41] allows users to search through unstructured, identified clinical notes from the EHR and will be used to narrow down the starting patient set (Figure 2). i2b2 provides information on structured, discrete data in CDW-H, while EMERSE provides key background information that allows users to identify patient cohorts based on characteristics like patient symptoms (e.g., wet cough) and social history (e.g., tobacco use). After cohort exploration, an analyst from NC TraCS will retrieve the EMR from CDW-H.

To identify disease diagnoses, computational phenotyping methods can be used to leverage various types of information found in the EMR. A combination of clinical notes, billing codes (ICD-9-CM and ICD-10-CM), and laboratory results will be used to extract patient records of a specific disease diagnosis. The clinical free-text is used to document clinical events and contains key information that may not be included in structured data. The clinical Text Analysis and Knowledge Extraction System (cTAKES) is an NLP system for information extraction from clinical free-text that uses rule-based and machine learning techniques in various modules that allow for named entity recognition of different clinical entities [42]. cTAKES will be used to process and annotate the clinical free-text. Billing codes and laboratory results will be extracted from structured fields in the EMR.

3.1.3 (b) Condition case: Breast cancer

Female patients with a breast cancer diagnosis will be extracted. Breast cancer in males will not be considered in this research. Patients will be considered as having a breast cancer diagnosis if an abnormal mammogram is found in the CDW-H. If more information is needed, records will also be checked in the CMR.

Patients with different types of breast cancer must be grouped to account for variations in treatment strategy. The two factors associated with subtype classification are hormone receptors (HR) and human epidermal growth factor 2 (HER2). Patients will either be positive (+) or negative (-) for having either HR or HER2 or both affect tumor growth. There are four female breast cancer subtypes, shown in Table 2 [43].

Table 2: Breast cancer subtypes

Subtype	**Details**	**Frequency of invasive breast cancer**
Luminal A	HR+/HER2-	30-40%
Triple Negative/Basal-like	HR-/HER2-	15-20%
Luminal B	HR+/HER2+ or HR+/HER2-	20-30%
HER2-enriched	HR-/HER2+	12-20%

Molecular subtypes of breast cancer play a significant role in treatment identification. For example, for patients with TNBC, since the growth of the cancer is not associated with HER2 protein, progesterone or estrogen receptors, drugs targeting these receptors would not be suitable for treating TNBC.

3.1.3 (c) Condition case: Oral cancer

Patients with an oral cancer diagnosis will be extracted. A diagnosis will consist of having cancers of the tongue, lips, floor of the mouth, palate, gum, salivary gland or other unspecified parts of the mouth. Risk factors that will be considered as patient clinical covariates include tobacco use, alcohol use, testing positive for human papillomavirus (HPV), age and more. Patient records will need to be examined to understand the original structured record and clinical note formatting for oral cancer. During regular dental examinations, a dentist will assess characteristics of the mouth. If change in texture, discoloration, ulcers, growths, lymph node enlargement, or fixed lymph nodes are detected, there may be suspicion of oral cancer. If there are any abnormal characteristics, the following details are recorded: location, size of lesion, consistency of tissue, and induration. If oral cancer is suspected in any areas of the mouth previously described, a biopsy will be ordered. If cancer is confirmed, possible treatments include excision, chemotherapy, radiation therapy, and topical medications. Accurate identification of the area of the mouth with cancer plays a significant role in deciding which treatment to use [44].

Based on the dental care path, characteristics of the mouth associated with oral cancer can be used as features to discriminate between cases and controls in a cohort. The characteristics listed are found in the clinical notes of a patient record. EMERSE will be used to explore the clinical notes for oral cancer characteristics. If a biopsy has been done, the biopsy results would be listed in both the clinical notes and in the laboratory results in structured data. With a positive biopsy, it can also be assumed that a patient would receive a corresponding ICD-CM code associated with having cancer in part of the mouth, but the positive biopsy laboratory result will be the main determinant of case or control status for a record.

3.1.3 (d) Condition case: PCD

Based on ICD-9-CM and ICD-10-CM codes, patients with the following diagnoses will be extracted: PCD, cystic fibrosis (CF), and non-CF bronchiectasis (PCD/CF/BR). Since PCD has no ICD-CM code, the general code will be used. Patients with CF and non-CF bronchiectasis will be included because current management of PCD is based on studies for these conditions. In addition, rare diseases generally have small cohort sizes, and to facilitate analysis of high-dimensional data like EMR with various methods including machine learning techniques and traditional statistical approaches, sufficient cohort sizes are needed to produce results without significant bias or variance. By incorporating patients with PCD/CF/BR that have similar pulmonary phenotypes, the sample size will be more suited for the usage of various techniques.

The Clinical Annotation Research Kit (CLARK) has already been used at UNC to identify undiagnosed individuals from the CDW-H with high likelihood of PCD with high sensitivity (0.88) and specificity (1.0) [39]. Since there is an existing computational phenotyping effort for PCD at UNC, the feature list from that effort and gold standard, annotated notes from subject matter experts will be requested for this work. A few identified discriminating features from Pfaff *et al* (2020)[39] include: “situs inversus”, “denies shortness of breath”, and “ear tubes”.

EMERSE will be used to find patients with abnormal pulmonary phenotypes based on discriminating features. i2b2 will be used to identify all patients with a PCD/CF/BR diagnosis code. An NC TraCS expert will extract patient records from the CDW-H using the patient set resulting from the i2b2 query and the patient set exported from EMERSE. The goal behind using both the structured data and clinical notes for cohort extraction is that the starting cohort will be more comprehensive than it would if only one data source were used.

3.1.3 (e) Predictive modeling

Logistic regression will be used for all condition cases. However, based on the condition case, the outcome variables will vary. Logistic regression is a classifier that makes predictions based on the linear distribution of features. In other words, it creates a dividing hyperplane with a linear classifier to predict the probability of an instance (i.e., record) belonging to its given output class. Multivariate logistic regression is used to assess the association between independent exposure variables and an outcome variable, while accounting for confounding factors.

3.1.4 Analytic Plan

3.1.4 (a) General approach

Cohort size will be estimated based on ICD-9-CM and ICD-10-CM codes using i2b2. Estimated cohort size will be assessed based on diagnoses made in the respective timeframes for each condition case. EMERSE will be used to narrow down the patient set. After estimation using i2b2 and EMERSE as well as IRB approval, a request will be made to the CDW-H to extract the EMR. Clinical notes, laboratory results, and ICD-CM codes will be extracted from the EMR for each condition case.

The data variables mentioned in Section 3.1.3 (p. 11) will be queried from the CDW-H. Separate CSV’s will be used for clinical notes, laboratory results, and billing codes. The data will be placed in a CSV format and manipulated in the following steps:

i. Ensure CSV headers and row labels are uniform.

ii. Identify rows with missing data variables, and output percentage of rows with missing data as well as number of missing data values within each record. If <10% of rows have missing data, remove all patients with missing variables. If there are any data variables where >50% of patients have missing data, the data variable list must be refined or a different source for the variable must be found.

iii. Correct for data variable type issues (e.g., converting lab value from string to integer).

iv. Remove any unreadable characters.

For clinical notes, extra pre-processing steps will be taken in order to make the data easier to analyze. The notes will be converted into text representations for each patient. The process for preparing the notes may change based on how NC TraCS provides the data. In addition, some of the notes will need to be read to understand which sections will be necessary for identifying a diagnosis. For each patient record, the following steps will be taken:

i. Clinical notes will be separated by datestamp.

ii. Each note will be assigned section tags, if they are not already.

iii. Within a section, the text will be processed with cTAKES. The following functions within cTAKES will be used: separating text by sentence, detecting negation (e.g., not metastatic), and named entity recognition.

3.1.4 (b) Condition case: Breast cancer

All female patients in the CDW-H with abnormal mammograms will be used as a starting cohort. The task for breast cancer is for an algorithm to be able to detect the molecular subtype based on the three data sources provided. The clinical notes will include pathology notes, which will be the most indicative of breast cancer subtype. All relevant laboratory results, billing codes, and clinical text features will be used as features for the algorithm.

3.1.4 (c) Condition case: Oral cancer

If a patient record includes a positive biopsy result, the record will be included in the study. In addition, all patient dental records will be queried for the abnormal characteristics listed in Section 3.1.3 in the CDW-H using EMERSE. If any abnormal characteristics are found within the clinical notes for a patient, the patient record will be included in the study. Therefore, the starting cohort will consist of all patients with abnormal mouth characteristics and positive biopsy results listed in their patient records. The task for oral cancer is for an algorithm to be able to detect whether the patient has oral cancer or not. The predictive model will be trained on cases with positive biopsy results, and tested on cases with suspicion of an oral cancer diagnosis. The location of the cancer is also important for making treatment decisions; therefore, feature extraction is important for this condition case.

3.1.4 (d) Condition case: PCD

All patients with a PCD/CF/BR diagnosis or history of abnormal respiratory phenotypes will be used as a starting cohort. An i2b2 query will be constructed using ICD-CM codes and date filters. An EMERSE search will consist of finding abnormal respiratory events (e.g., chronic sinusitis). The task for PCD is for an algorithm to be able to detect PCD and PCD-like diagnoses. For the PCD true cases, the records will be examined for PCD laboratory results in the structured data and in the pathology notes. UNC also has PCD mutation testing with a CPT code, so provided the information is in the CDW-H, the testing results can also discriminate PCD true cases. This will be a binary classification, with the positive class being PCD (1) and the negative class being PCD-like (0).

3.1.4 (e) Data analysis and predictive modeling

General statistics will be used to gain an understanding of distributions in each condition case cohort. For each patient characteristic, the following will be outputted: the number (N) or mean number of patients and the percentage or standard deviation (SD)

of patients within each patient characteristic category. For example, age at diagnosis could be divided into five groups with an N and SD for each of the five groups.

The dataset will be split into stratified train and test sets, where 80% of the data will be used for training and 20% will be used for testing. Cross-validation will be used on the training set to train the model. Cross-validation is an evaluation method that can avoid overfitting of the model. 10-fold cross-validation consists of using 90% of the training set for training and 10% of the training set for testing. The portion of the training set used for testing will rotate across the folds, until every 10% slice of the training set has been used for testing once. In the algorithm tuning process, feature engineering will be used to identify meaningful features from the data, using feature importance as a measure. After training, the model will be tested on the 20% of the data held out as a test set. The Scikit Learn package will be used to conduct all analysis in Python [45].

3.1.5 Measures for evaluation

Odds ratio and adjusted odds ratios will be used for confounding factors. The percentage of missing data will be used to assess data completeness. A stratified random sample of clinical records from the control and treatment groups will be taken and two experts will independently review the medical records to assess algorithm performance and confirm case/control or case type classification. Any discrepancies will be resolved through discussion between the reviewers. Cohen's kappa coefficient will be used to measure inter-rater reliability:

$$k = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$

where $Pr(a)$ indicates the agreement that is present and $Pr(e)$ indicates the agreement by chance [46].

The prediction algorithm performances will be assessed with the following evaluation metrics: accuracy, precision (i.e., specificity), recall (i.e., sensitivity), and F-score.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F - score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

where TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives.

3.1.6 Potential Challenges and Limitations

For each condition case, there may be different potential challenges. With breast cancer, screening and diagnosis is already well-established, so it is not necessary to find whether a patient has breast cancer or not. The study can begin with patients who have already confirmed having breast cancer. However, discriminating between different cancer subtypes is an important research area. A potential challenge for identifying a multi-class problem is that many algorithms function better with binary classification, and there is more support from various Python packages for binary classification. Although this is a potential challenge, this can be mitigated by adapting functions that are currently written for binary classification to multi-class classification. With oral cancer, there is a similar challenge because it is necessary to know which part of the mouth is affected when deciding on treatment. However, if documentation is clear, this could also be solved with a rule-based approach. An additional limitation is the disconnection between dental records and records in the CDW-H. All cohort estimates will be made using i2b2, meaning that all estimates will reflect what data is in CDW-H. Therefore, dental records will not be included; however, a future direction is to connect the CDW-H and dental records to ensure the cohort contains all possible patients. For PCD, the major challenge is that diagnosis itself is an ongoing research area, so identifying and evaluating PCD-like phenotypes will determine the success of the approach. For evaluating the case/control or case type classification, the aim is to have two subject matter experts review a stratified random sample of patient records. To do so, the subject matter experts need to be paid, so I am applying to grants such as the NC TraCS $2,000 grant. The back-up option for evaluation is to find existing gold standard datasets for testing.

Aim 2. Build A Pipeline For Retrospective Clinical Record Analysis To Validate Drug Repurposing Candidates.

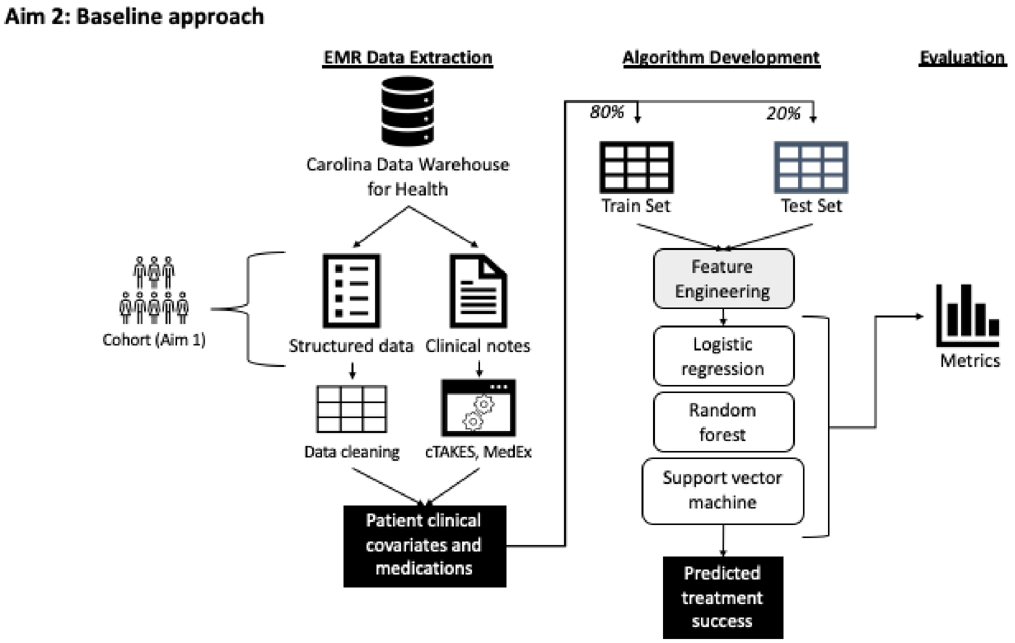


Figure 3: Aim 2 Flowchart- Baseline Approach

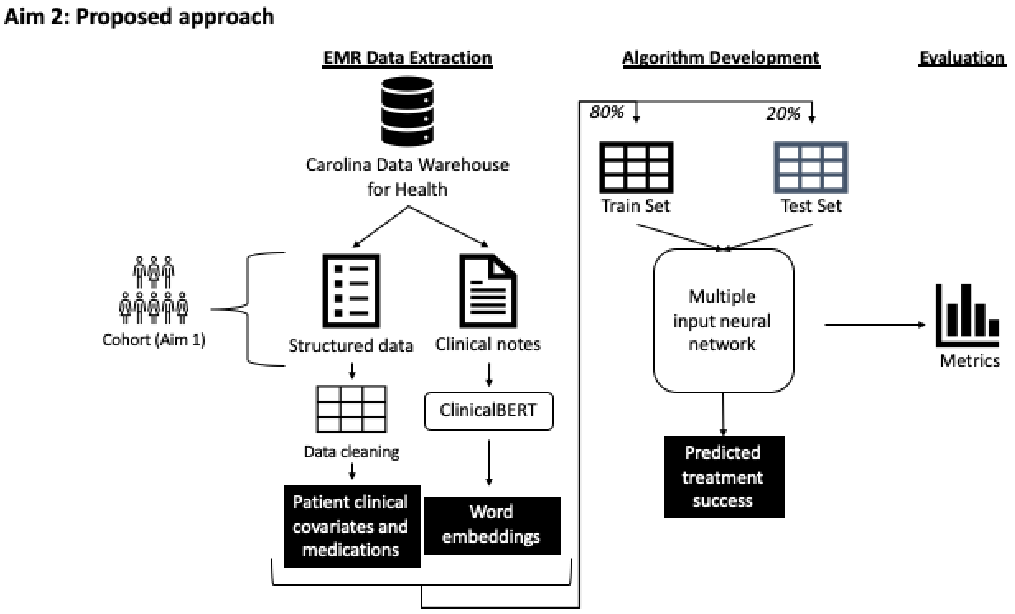


Figure 4: Aim 2 Flowchart- Proposed Approach

3.2.1 Significance

EMR validation of drug repurposing candidates has historically been done using statistical approaches and logistic regression as a machine learning approach. Traditional machine learning approaches, including logistic regression, will be explored as a baseline for the proposed approach. Machine learning models work well with small sample sizes, are transparent, and are easily interpretable. However, they require feature engineering and annotated datasets. The proposed approach builds off of the existing studies to explore deep learning methods for EMR validation. As shown in other domains, deep learning approaches can achieve higher performance than classical machine learning algorithms, do not require feature engineering, and can be generalized to various datasets. However, they are not as interpretable as machine learning algorithms, are computationally expensive, and are greedy in terms of sample size required to achieve high accuracy. The aim for the proposed study is to create an EMR validation algorithm that can be generalized to different condition cases, and the proposed methods will allow for that generalization.

3.2.2 Study Design

The study will be conducted to validate drug candidate predictions in three use cases (i.e., conditions). Steps for the study will consist of: data extraction, algorithm development, and evaluation, as shown in Figures 3 and 4. All relevant information will be extracted from EMR records for patients of each condition case. The case/control distinction in the cohort of patients can either be from the Aim 1 output (Figure 2), or the controls can be "normal" patients without any symptoms of each condition case. The cohort will be defined in detail after data exploration. The data will be processed into a format suitable for statistical and machine learning analysis. General features will be included for all patient records, regardless of condition cases. Condition-specific features will be identified and added for each condition case in the baseline approach (Figure 3). The proposed method will take a more generalizable, representation learning approach (Figure 4).

The task for this study is to predict the probability of treatment success. The classification task will be binary (i.e., disease improvement, no disease improvement), where the definition of true positives would be indicators of positive outcomes. For all condition cases, vital signs can be used. For breast cancer and oral cancer, tumor size shrinkage will be the indicator for disease improvement. For PCD, increased airway health, as marked by sputum cultures and spirometry testing, will be the indicator for disease improvement. Machine learning algorithms will be used as a baseline for the task, and deep learning methods will be explored in the proposed approach.

3.2.3 Methods

3.2.3 (a) Cohort size estimation

The input for the study is a list of drug-disease repurposing predictions. Before EMR extraction, cohort size estimation will be done in the same way as Aim 1. Cohort sizes will be assessed using i2b2 [40]. i2b2 is a web application that is a view of UNC Health Care data, and it allows for the investigation of de-identified, aggregate data.

3.2.3 (b) Baseline approach

Data extraction and manipulation

EMR data extraction will be divided into two tasks: (1) extracting medications used (2) extracting patient clinical covariates. The process of extracting medications will be the same across condition cases. The process for extracting patient clinical covariates will initially be the same for all condition cases, but fine tuning will then be done for clinical covariates specific to each condition cases. For example, age at diagnosis and gender will be extracted for all condition cases. However, cancer-specific covariates, such as tumor stage, will only be extracted for cancer condition cases.

For clinical covariate fields with missing data, NLP algorithms will be used to extract the information from unstructured clinical notes. cTAKES will be used to extract any covariates missing in structured fields from clinical narratives. Similarly, although the structured data contains a medication list, medication data is often recorded in clinical free-text. From medication orders and clinical narratives, an NLP algorithm, MedEx, will be used to identify medications. MedEx is a rule-based NLP system that extracts medication information such as drug names, dose, route, and frequency from clinical free-text [47]. The results from MedEx will be organized and filtered by which drugs are meant for the given indication, which are evidence of off-label usage, and which are due to a patient having multiple conditions.

Predictive modeling

The task will consist of predicting the probability of treatment success for the predicted drug based on relationships between features from the EMR. The task can be described as prediction of the outcome variable where 1 = disease improvement and 0 = no disease improvement. The experiment will use three machine learning algorithms for prediction: logistic regression, random forest [48], and support vector machine [49]. Feature engineering with feature importance or feature ablation will be used to tune the algorithms. Feature importance techniques assign each feature a score based on how useful it is in predicting the output variable. Feature ablation is the process of removing features individually to identify a feature set that will provide optimal performance.

Logistic regression, where the outcome variable is disease improvement and input variables are patient clinical covariates and medications, will be used as a baseline for this analysis as it has previously been used for drug repurposing candidate validation [36]. Logistic regression is a classifier that makes predictions based on the linear distribution of features. In other words, it creates a dividing hyperplane with a linear classifier to predict the probability of an instance (i.e., record) belonging to its given output class. SVM is similar to a one-layer neural network and functions by identifying the optimal hyperplane for classification. With a linear decision function, if the points are linearly separable in 2D, SVM works like linear regression. In this way, SVM is similar to logistic regression. However, if the points are not linearly separable in 2D, the SVM functions by mapping to higher dimension spaces and finding linear separation. The same logic applies to using other decision functions for SVM classification. Random forest is an ensemble machine learning method that uses bootstrap aggregation (bagging) and feature randomness techniques to create uncorrelated decision trees. The classifier then uses the series of trees to predict individual outputs and selects the output with the highest number of votes as the final prediction. Based on data received, other algorithms may be tested against the three highlighted in this proposal as well.

3.2.3 (c) Proposed approach

For data extraction and treatment success probability prediction, the proposed approach will use deep neural networks. A neural network is a series of processing nodes that are densely connected into layers. Most neural networks are feed-forward, meaning that data moves in one direction through the network from the input layer to the output layer. The advantage of representation learning methods, like deep learning methods, is in their data processing capabilities. Traditional machine learning algorithms require extensive data pre-processing and feature engineering, while representation learning methods allow a machine to take raw data and learn representations from them. In particular, deep learning methods are useful for learning the intricacies of high-dimensional data, like EMR data [13].

Data extraction and manipulation

EMR data extraction will consist of extracting structured data and extracting data from the clinical notes. Structured data will be extracted from the CDW-H by an NC TraCS analyst and be used as part of the input for the predictive model without extensive pre-processing. Like in the baseline approach, patient clinical covariates and medications used will be extracted from structured data. From the unstructured portion of the EMR, a deep neural network will be used to extract and manipulate information from the clinical notes. The input will be the clinical notes, and the output of the model will be a series of word embeddings, which a predictive model would be able to process.

A Bidirectional Encoder Representations from Transformers (BERT) model[50] will be used. BERT is a deep neural network that uses transformer architecture to learn text embeddings. BERT functions by taking a sequence of words and learning the contextual relationships in the sequence. There are two components in BERT: a transformer encoder layer and a classification layer. For each sequence of words (i.e., tokens), the following information is needed:

- Masked tokens
- Tokens at the beginning <cls> and ending <sep> of the sequence, where cls is used for classification and sep is used for separation.
- Sentence embedding
- Positional embedding for each token

In the case of NER, each masked token is a named entity and the output of the BERT model would be the NER label. Pre-trained embedding models are becoming more useful for various tasks, but for biological and clinical tasks, domain-specific knowledge for pre-training can improve performance in comparison to using general knowledge.

ClinicalBERT is a BERT model that has been pre-trained on clinical notes [51, 14]. Alsentzer *et al* (2019)[51] describes how clinical notes have different linguistic characteristics in comparison to general and biomedical articles, creating a need for models trained on clinical narratives. The study used MIMIC III narratives and discharge summaries to train BERT models with clinical-specific contextual embeddings. They have made the pre-trained models publicly available on Github [52]. Huang *et al* (2019)[14] further developed a ClinicalBERT model and compared the performance of ClinicalBERT to other commonly used word embedding models such as word2vec to show the improvement in performance. In comparing pearson correlation between cosine similarity of embeddings from clinical text models and physician ratings of medical concepts, ClinicalBERT and word2vec achieved 0.670 and 0.553 pearson correlations, respectively. In addition, the study found that word2vec did not perform as well with “out of vocabulary” words in comparison to ClinicalBERT, providing more motivation for using a ClinicalBERT model over other word embedding models. The traditional BERT model trained on clinical narratives will be used for EMR data manipulation from the clinical notes and compared to the baseline approach.

Predictive modeling

The task will consist of predicting the probability of treatment success for a predicted drug where the outcome variables are 1 = disease improvement and 0 = no disease improvement. The probability of treatment success will be considered the prediction probability, which is the algorithm confidence for a given prediction. The experiment will use a multiple input neural network for prediction. Generally, neural networks use a single data type for prediction. For example, the ClinicalBERT model described previously will only take clinical notes, which are text data, as its input. However, for predicting disease improvement, discrete data variables (e.g., age, gender, race) are necessary. A multiple input neural network will be able to use the word embeddings generated by the ClinicalBERT model alongside the discrete data variables for prediction.

3.2.4 Analytic Plan

3.2.4 (a) Cohort size estimation

Cohort size will be estimated based on ICD-9-CM and ICD-10-CM codes using i2b2. Estimated cohort size will be assessed based on a condition case diagnosis made within a given timeframe, where condition case diagnosis refers to a diagnosis for any of the condition cases (i.e., breast cancer, oral cancer, PCD/CF/BR). The patient set selected will then be queried for a disease diagnosis for the original indication of a drug candidate. For example, if the drug candidate is metformin, diabetes mellitus, the original indication, will be queried in the patient set. For a patient set with the condition case diagnosis, the set will be queried for any use of the drug candidate after the diagnosis date. If the drug candidate has not been administered for individuals with a condition case diagnosis, the next drug repurposing candidate will be assessed. For patient sets that include both drug candidates and a condition case diagnosis, any patients with contraindications prior to diagnosis for the drug repurposing candidate will be removed from the cohort. For example, if the drug candidate is metformin, it cannot be taken by patients with chronic kidney disease. Therefore, patients with a condition case diagnosis and chronic kidney disease will be removed from the cohort [10]. Contraindications will be identified with ICD-9-CM and ICD-10-CM codes in i2b2. For the predictions associated with sufficient cohort sizes, EMR will be extracted.

After estimating cohort size using i2b2 and IRB approval, a request will be made to CDW-H to extract the EMR. The components extracted from the EMR would include structured data (e.g., the problem list, patient demographics), unstructured clinical notes, and laboratory results. Disease diagnosis, and features like patient clinical covariates and medications will be extracted from these components. Patient clinical covariates that will be extracted include: age at disease diagnosis, height, weight, gender, race, smoking status, zipcodes, as well as diabetes and high blood pressure diagnoses.

3.2.4 (b) Baseline approach

Data extraction and manipulation

The clinical notes for each patient will be annotated with cTAKES [42]. The following functions will be used: sentence boundary detector, tokenizer, normalizer, part-of-speech (POS) tagger, shallow parser, and NER with negation and status annotators. All of these functions will be used within the cTAKES system, as opposed to Scikit Learn[45], in order to account for properties specific to clinical notes. In cTAKES, the sentence boundary detector predicts punctuation type at the end of a sentence. The tokenizer separates words by spaces and also merges tokens to account for various data types like date and range. The normalizer is used to map mentions of the same word that have different string representations. The POS tagger and shallow parser are used to add sentence structure annotations. The NER component draws from SNOMED CT[53], Unified Medical Language System (UMLS)[54], and RxNORM[55] to identify and annotate terms in the clinical notes. The negation and status annotation portions of NER search for words that indicate negation (e.g., not bleeding) and status (e.g., family history of) respectively. The medications administered will then be extracted from structured fields. MedEx will be used to identify medication data not in structured fields and provide context for medication use from the clinical free-text.

Predictive modeling

The training and testing process described in Section 3.1.4 will be used. The dataset will be split into stratified train and test sets, where 80% of the data will be used for training and 20% will be used for testing. Cross-validation will be used on the training set to train the model. Cross-validation is an evaluation method that can avoid overfitting of the model. 10-fold cross-validation consists of using 90% of the training set for training and 10% of the training set for testing. The portion of the training set used for testing will rotate across the folds, until every 10% slice of the training set has been used for testing once. In the algorithm tuning process, feature engineering will be used to identify meaningful features from the data. Feature importance will be assessed for logistic regression and random forest. For SVM, feature importance can only be assessed if the SVM has a linear kernel. Linear and non-linear kernels will be tested during training, and feature importance will be reported if a linear kernel is selected. If a linear kernel is not selected, feature ablation will be used to assess the importance of data features. After training, the three models will be tested on the 20% of the data held out as a test set. The Scikit Learn package will be used to conduct all analysis in Python [45].

3.2.4 (c) Proposed approach

The development process for the data extraction and manipulation and predictive modeling sections will be the same. The dataset will be split into three sets: train (70%), development (10%), and test (20%). The development set will be used to fine tune hyperparameters, and the test set will be used to assess overall algorithm performance. All analysis will be done using the Keras framework[56] which is built on top of TensorFlow[57] in Python.

Data extraction and manipulation

The pre-trained ClinicalBERT model by Alsentzer *et al* (2019)[51] will be obtained from the project Github page [52]. The clinical notes for patients of each condition case will be divided into sections as described in Section 3.1.4. Dimensions of pre-trained word embedding models, number of epochs, batch size, learning rate, and max predictions per sequence will be considered for fine tuning the ClinicalBERT model. These parameters were also considered in Alsentzer *et al* (2019)[51].

Predictive modeling

The word embeddings from the ClinicalBERT model and structured, discrete data from the EMR will be included as inputs for the model. Instead of stacking ClinicalBERT and a separate neural network with discrete data, a multiple input neural network will be used to combine the two feature spaces for prediction. Similar to adding metadata for text input, a layer of numerical features will be concatenated to word embeddings. This approach will be compared to one-hot encoding all discrete features and using the vectors alongside word embeddings in one overarching feature space. Hyperparameters such as number of epochs, batch size, learning rate, and dropout will be fine tuned. Number of layers will also be explored.

3.2.5 Measures of evaluation

The prediction algorithm performances will be assessed with the following evaluation metrics: precision (i.e., specificity), recall (i.e., sensitivity), area under the receiver operating curve (AUROC), and area under the precision recall curve (AUPR). A ROC curve shows the trade-off between true positive and false postive rates. A precision recall curve shows the trade-off between precision and recall at various thresholds. To define true positives, a cut-off will be taken after making a distribution plot of values from disease improvement indicators (as described in Section 3.2.2).

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

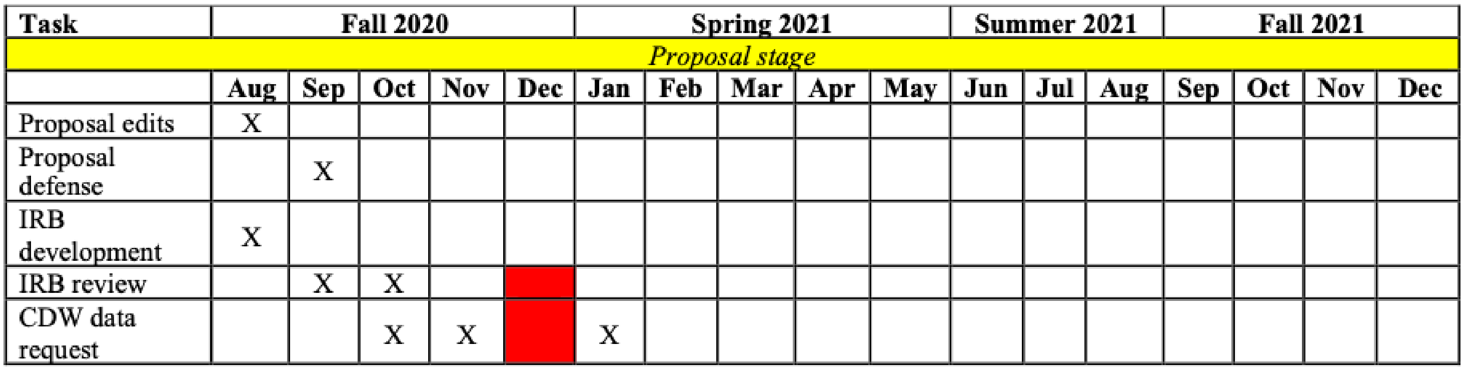
where TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives.

3.2.6 Potential Challenges and Limitations

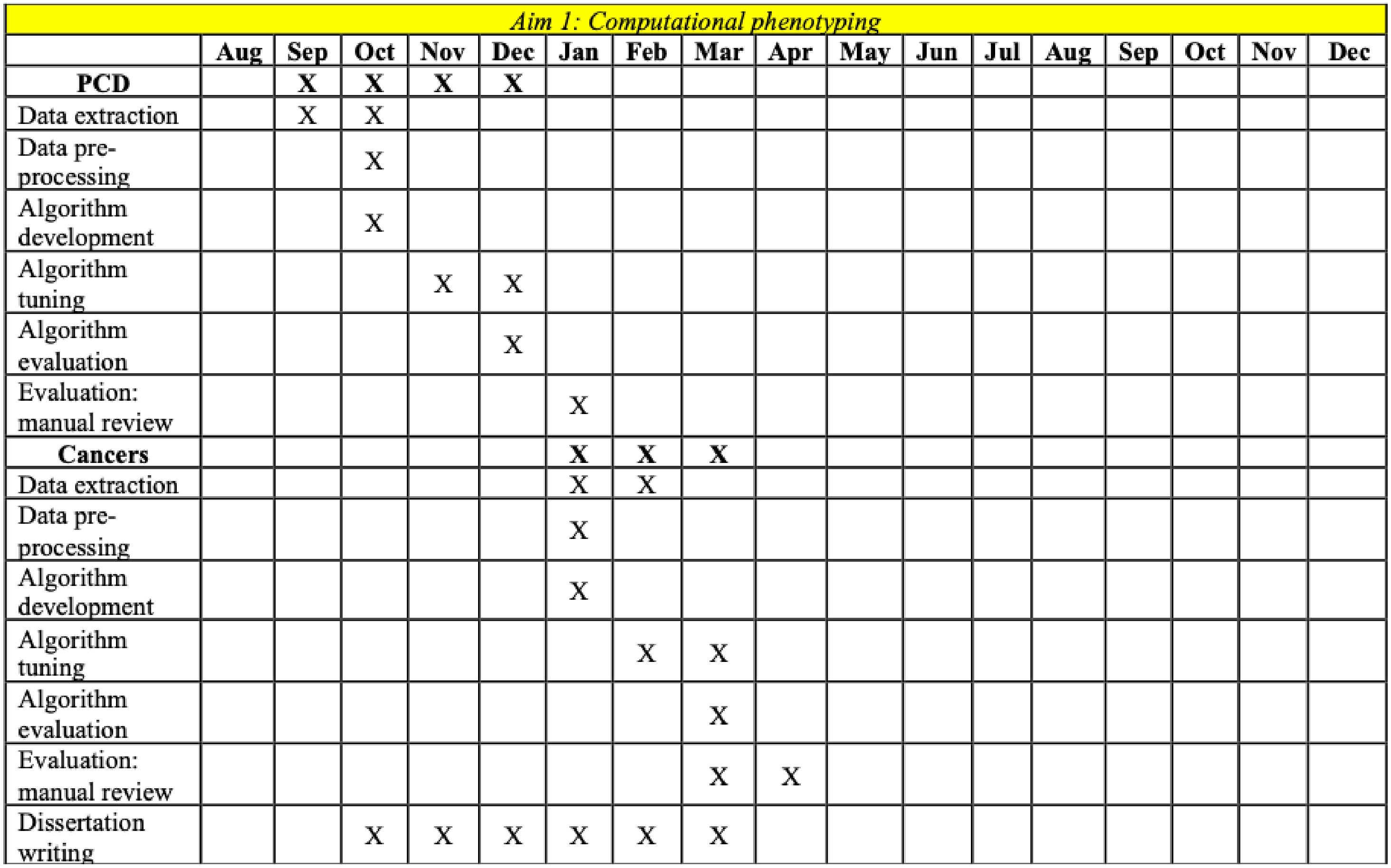
From an overall study perspective, a limitation of the work is that pharmacy fill records will not be used as a data source. Using fill records would provide more solid evidence for when each medication exposure took place, but medication exposures extracted from EMR have been shown to indicate timeline with high performance in the related work.

From a methodological perspective, a known weakness of neural networks is that they are not as interpretable as machine learning models. For the proposed approach, attention weights will be examined to interpret the model. In addition, a comparison will be made to the baseline approach to improve interpretability. Processing power is also a concern for deep learning methods. For training the ClinicalBERT model, Alsentzer *et al* (2019)[51] required 17-18 days of runtime on a GeForce GTX TITAN X 12 GB GPU. The advantage to using a pre-trained model is that processing cost is greatly reduced. The proposed study will be conducted using a computer with a GeForce GTX 1650 4 GB GPU. However, if more computing power is necessary, other options such as different training techniques and cluster computing will be explored.

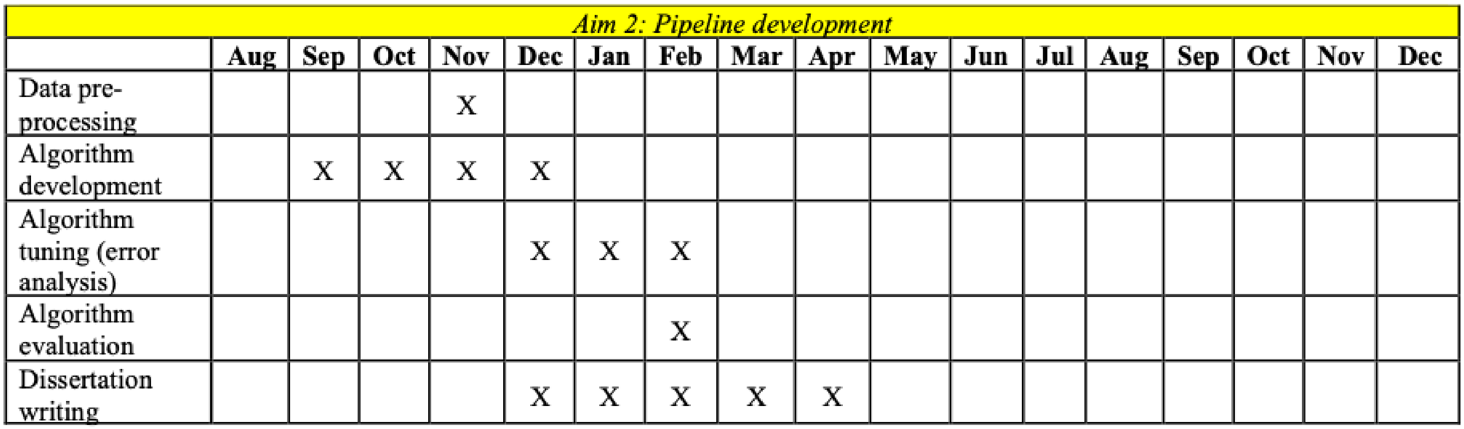
Timeline



Task	**Fall 2020**					**Spring 2021**					**Summer 2021**			**Fall 2021**			
Proposal stage																	
	Aug	**Sep**	**Oct**	**Nov**	**Dec**	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
Proposal edits	X																
Proposal defense		X															
IRB development	X																
IRB review		X	X		🟥												
CDW data request			X	X	🟥	X											

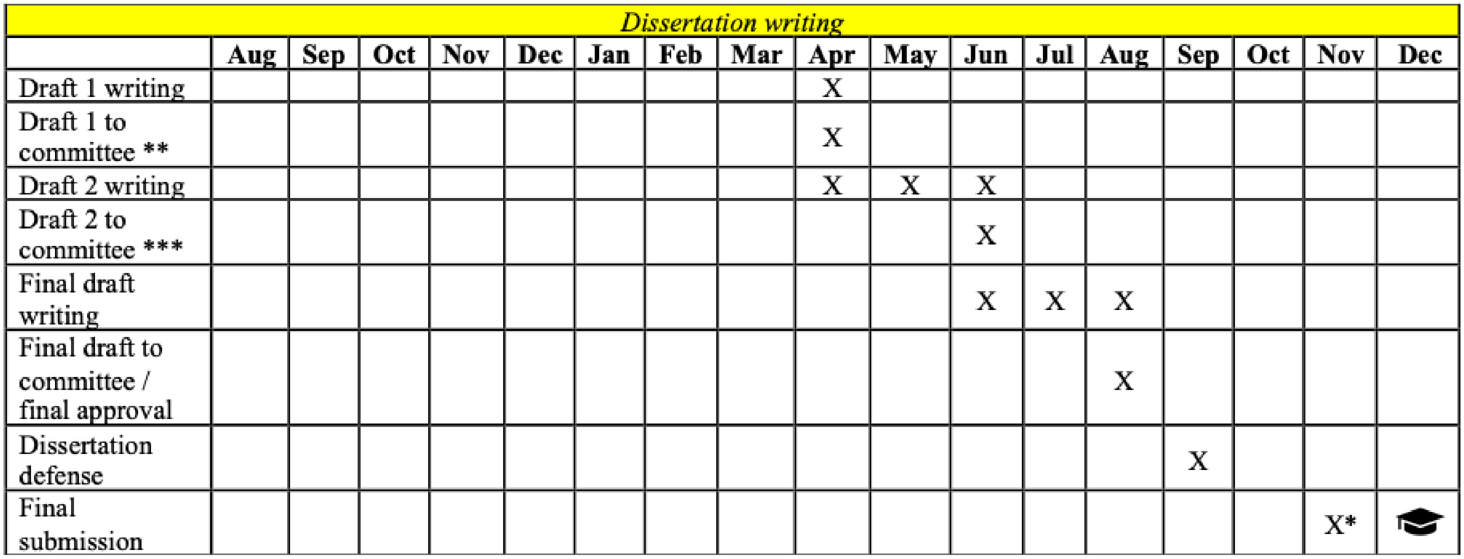


Aim 1: Computational phenotyping																	
	Aug	**Sep**	**Oct**	**Nov**	**Dec**	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
PCD		**X**	**X**	**X**	**X**												
Data extraction		X	X														
Data pre-processing			X														
Algorithm development			X														
Algorithm tuning				X	X												
Algorithm evaluation					X												
Evaluation: manual review						X											
Cancers						**X**	**X**	**X**									
Data extraction						X	X										
Data pre-processing						X											
Algorithm development						X											
Algorithm tuning							X	X									
Algorithm evaluation								X									
Evaluation: manual review								X	X								
Dissertation writing			X	X	X	X	X	X									



Aim 2: Pipeline development

	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Data pre-processing				X													
Algorithm development		X	X	X	X												
Algorithm tuning (error analysis)					X	X	X										
Algorithm evaluation							X										
Dissertation writing					X	X	X	X	X								



Dissertation writing

	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Draft 1 writing									X								
Draft 1 to committee **									X								
Draft 2 writing									X	X	X						
Draft 2 to committee ***											X						
Final draft writing											X	X	X				
Final draft to committee / final approval													X				
Dissertation defense														X			
Final submission																X*	🎓

* For Fall 2020, the deadline for electronic submissions to the Graduate School is: **November 18, 2020 before 4 pm** for December graduation.

** Draft 1 will consist of- Ch 1: Introduction, Ch 2: Literature Review, Ch 3: Methods (Aim 1)

*** Draft 2 will consist of- Ch 1: Introduction, Ch 2: Literature Review, Ch 3: Methods, Ch 4: Results, Ch 5: Discussion

References

[1] N. Nosengo, "Can you teach old drugs new tricks?" *Nature*, vol. 534, no. 7607, pp. 314–316, jun 2016. [Online]. Available: http://dx.doi.org/10.1038/534314a

[2] M. Ringel, J. Scannell, M. Baedeker, and U. Schulze, "Breaking eroom's law." *Nature reviews. Drug Discovery*, 2020.

[3] F. F. Track, "Breakthrough therapy, accelerated approval and priority review," 2013.

[4] S. Pushpakom, F. Iorio, P. A. Eyers, K. J. Escott, S. Hopper, A. Wells, A. Doig, T. Guilliams, J. Latimer, C. McNamee, A. Norris, P. Sanseau, D. Cavalla, and M. Pirmohamed, "Drug repurposing: progress, challenges and recommendations." *Nature Reviews. Drug Discovery*, vol. 18, no. 1, pp. 41–58, 2018. [Online]. Available: http://www.nature.com.libproxy.lib.unc.edu/doifinder/10.1038/nrd.2018.168

[5] "Global markets for drug repurposing: PHM175A |BCC research." [Online]. Available: https://www-bccresearch-com.libproxy.lib.unc.edu/market-research/pharmaceuticals/drug-repurposing-markets-report.html

[6] P. Sun, J. Guo, R. Winnenburg, and J. Baumbach, "Drug repurposing by integrated literature mining and drug-gene-disease triangulation." *Drug Discovery Today*, vol. 22, no. 4, pp. 615–619, 2017. [Online]. Available: http://dx.doi.org.libproxy.lib.unc.edu/10.1016/j.drudis.2016.10.008

[7] "Pfizer's expiring viagra patent adversely affects other drugmakers too." [Online]. Available: https://www.forbes.com/sites/investor/2013/12/20/pfizers-expiring-viagra-patent-adversely-affects-other-drugmakers-too/#2a45335668d4

[8] J. M. Pulley, J. P. Rhoads, R. N. Jerome, A. P. Challa, K. B. Erreger, M. M. Joly, R. R. Lavieri, K. E. Perry, N. M. Zaleski, J. K. Shirey-Rice, and D. M. Aronoff, "Using what we already have: uncovering new drug repurposing strategies in existing omics data." *Annual Review of Pharmacology and Toxicology*, vol. 60, pp. 333–352, jan 2020. [Online]. Available: http://dx.doi.org/10.1146/annurev-pharmtox-010919-023537

[9] M. Pillai and D. Wu, "Validation strategies for computational drug repurposing: a review," Feb 2020.

[10] H. Xu, M. C. Aldrich, Q. Chen, H. Liu, N. B. Peterson, Q. Dai, M. Levy, A. Shah, X. Han, X. Ruan, M. Jiang, Y. Li, J. S. Julien, J. Warner, C. Friedman, D. M. Roden, and J. C. Denny, "Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality." *Journal of the American Medical Informatics Association*, vol. 22, no. 1, pp. 179–191, jan 2015. [Online]. Available: http://dx.doi.org/10.1136/amiajnl-2014-002649

[11] P. Khatri, S. Roedder, N. Kimura, K. De Vusser, A. A. Morgan, Y. Gong, M. P. Fischbein, R. C. Robbins, M. Naesens, A. J. Butte, and M. M. Sarwal, "A common rejection module (CRM) for acute rejection across multiple organs identifies novel therapeutics for organ transplantation." *The Journal of Experimental Medicine*, vol. 210, no. 11, pp. 2205–2221, oct 2013. [Online]. Available: http://dx.doi.org/10.1084/jem.20122709

[12] N. Hiob and S. Lessmann, "Health analytics: a systematic review of approaches to detect phenotype cohorts using electronic health records," 2017.

[13] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[14] K. Huang, J. Altosaar, and R. Ranganath, "Clinicalbert: Modeling clinical notes and predicting hospital readmission," *arXiv preprint arXiv:1904.05342*, 2019.

[15] C. E. DeSantis, J. Ma, M. M. Gaudet, L. A. Newman, K. D. Miller, A. Goding Sauer, A. Jemal, and R. L. Siegel, "Breast cancer statistics, 2019," *CA: a cancer journal for clinicians*, vol. 69, no. 6, pp. 438–451, 2019.

[16] "Oral cancer incidence by age, race, and gender," Jul 2018. [Online]. Available: https://www.nidcr.nih.gov/research/data-statistics/oral-cancer/incidence

[17] M. W. Leigh, A. Horani, B. Kinghorn, M. G. O'Connor, M. A. Zariwala, and M. R. Knowles, "Primary ciliary dyskinesia (PCD): A genetic disorder of motile cilia." *Translational science of rare diseases*, vol. 4, no. 1-2, pp. 51–75, jul 2019. [Online]. Available: http://dx.doi.org/10.3233/{TRD}-190036

[18] M. Rastegar-Mojarad and R. Prasad, "Toward a complete database of drug repurposing candidates extracted from social media, biomedical literature, and genetic data," in *2015 International Conference on Healthcare Informatics*. IEEE, oct 2015, pp. 494–494. [Online]. Available: http://ieeexplore.ieee.org/document/7349746/

[19] U. Storz, "Rituximab," *mAbs*, vol. 6, no. 4, pp. 820–837, 2014. [Online]. Available: https://dx.doi.org/10.4161/mabs.29105

[20] C. Helfand, "Rituxan/mabthera," May 2014. [Online]. Available: https://www.fiercepharma.com/special-report/rituxan-mabthera

[21] M. A. Lopez-Olivo, M. Amezaga Urruela, L. McGahan, E. N. Pollono, and M. E. Suarez-Almazor, "Rituximab for rheumatoid arthritis." *Cochrane Database of Systematic Reviews*, vol. 1, p. CD007356, jan 2015. [Online]. Available: http://dx.doi.org/10.1002/14651858.{CD007356}.pub2

[22] S. Health, "Harnessing the power of data in health." 2017. [Online]. Available: https://med.stanford.edu/content/dam/sm/sm-news/documents/StanfordMedicineHealthTrendsWhitePaper2017.pdf

[23] A. Gottlieb and R. B. Altman, "Integrating systems biology sources illuminates drug action." *Clinical Pharmacology and Therapeutics*, vol. 95, no. 6, pp. 663–669, jun 2014. [Online]. Available: http://dx.doi.org/10.1038/clpt.2014.51

[24] K. M. Gayvert, E. Dardenne, C. Cheung, M. R. Boland, T. Lorberbaum, J. Wanjala, Y. Chen, M. A. Rubin, N. P. Tatonetti, D. S. Rickman, and O. Elemento, "A computational drug repositioning approach for targeting oncogenic transcription factors." *Cell reports*, vol. 15, no. 11, pp. 2348–2356, jun 2016. [Online]. Available: http://dx.doi.org/10.1016/j.celrep.2016.05.037

[25] C. Xu, D. Ai, D. Shi, S. Suo, X. Chen, Y. Yan, Y. Cao, R. Zhang, N. Sun, W. Chen, J. McDermott, S. Zhang, Y. Zeng, and J.-D. J. Han, "Accurate drug repositioning through non-tissue-specific core signatures from cancer transcriptomes." *Cell reports*, vol. 25, no. 2, pp. 523–535.e5, oct 2018. [Online]. Available: http://dx.doi.org/10.1016/j.celrep.2018.09.031

[26] C. M. Pugh, "Electronic health records, physician workflows and system change: defining a pathway to better healthcare," *Annals of Translational Medicine*, vol. 7, no. S1, pp. S27–S27, 2019. [Online]. Available: https://dx.doi.org/10.21037/atm.2019.01.83

[27] A. H. Nordo, H. P. Levaux, L. B. Becnel, J. Galvez, P. Rao, K. Stem, E. Prakash, and R. D. Kush, "Use of ehrs data for clinical research: Historical progress and current applications," *Learning Health Systems*, vol. 3, no. 1, p. e10076, 2019. [Online]. Available: https://dx.doi.org/10.1002/lrh2.10076

[28] A. L. Beam and I. S. Kohane, "Big data and machine learning in health care." *The Journal of the American Medical Association*, vol. 319, no. 13, pp. 1317–1318, apr 2018. [Online]. Available: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2017.18391

[29] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2018.

[30] N. Razavian, J. Marcus, and D. Sontag, "Multi-task prediction of disease onsets from longitudinal laboratory tests," in *Machine Learning for Healthcare Conference*, 2016, pp. 73–100.

[31] A. Sathyanarayana, S. Joty, L. Fernandez-Luque, F. Ofli, J. Srivastava, A. Elmagarmid, T. Arora, and T. Shahrad, "Correction of: sleep quality prediction from wearable data using deep learning," *JMIR mHealth and uHealth*, vol. 4, no. 4, p. e130, 2016.

[32] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, and P. Group, "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement." *PLoS Medicine*, vol. 6, no. 7, p. e1000097, jul 2009. [Online]. Available: http://dx.doi.org/10.1371/journal.pmed.1000097

[33] Z. Kuang, J. Thomson, M. Caldwell, P. Peissig, R. Stewart, and D. Page, "Computational drug repositioning using continuous self-controlled case series." *KDD : proceedings / International Conference on Knowledge Discovery & Data Mining. International Conference on Knowledge Discovery & Data Mining*, vol. 2016, pp. 491–500, aug 2016. [Online]. Available: http://dx.doi.org/10.1145/2939672.2939715

[34] ——, "Baseline regularization for computational drug repositioning with longitudinal observational data." *IJCAI : proceedings of the conference / sponsored by the International Joint Conferences on Artificial Intelligence*, vol. 2016, pp. 2521–2528, jul 2016. [Online]. Available: https://www-ncbi-nlm-nih-gov.libproxy.lib.unc.edu/pubmed/28392671

[35] H. Paik, A.-Y. Chung, H.-C. Park, R. W. Park, K. Suk, J. Kim, H. Kim, K. Lee, and A. J. Butte, "Repurpose terbutaline sulfate for amyotrophic lateral sclerosis using electronic medical records." *Scientific Reports*, vol. 5, p. 8580, mar 2015. [Online]. Available: http://dx.doi.org/10.1038/srep08580

[36] G. Koren, G. Nordon, K. Radinsky, and V. Shalev, "Machine learning of big data in gaining insight into successful treatment of hypertension." *Pharmacology research & perspectives*, vol. 6, no. 3, p. e00396, apr 2018. [Online]. Available: http://dx.doi.org/10.1002/prp2.396

[37] Y. S. Low, A. C. Daugherty, E. A. Schroeder, W. Chen, T. Seto, S. Weber, M. Lim, T. Hastie, M. Mathur, M. Desai, C. Farrington, A. A. Radin, M. Sirota, P. Kenkare, C. A. Thompson, P. P. Yu, S. L. Gomez, G. W. Sledge, A. W. Kurian, and N. H. Shah, "Synergistic drug combinations from electronic health records and gene expression." *Journal of the American Medical Informatics Association*, vol. 24, no. 3, pp. 565–576, may 2017. [Online]. Available: http://dx.doi.org/10.1093/jamia/ocw161

[38] T. N. C. T. and C. S. N. T. Institute, "CDW-h frequently asked questions." [Online]. Available: https://tracs.unc.edu/index.php/services/informatics-and-data-science/cdw-h/cdw-h-faq

[39] E. R. Pfaff, M. Crosskey, K. Morton, and A. Krishnamurthy, "Clinical annotation research kit (CLARK): computable phenotyping using machine learning." *JMIR medical informatics*, vol. 8, no. 1, p. e16042, jan 2020. [Online]. Available: http://preprints.jmir.org/preprint/16042/accepted

[40] "i2b2." [Online]. Available: https://tracs.unc.edu/index.php/services/informatics-and-data-science/i2b2

[41] "Emerse." [Online]. Available: https://tracs.unc.edu/index.php/services/informatics-and-data-science/emerse

[42] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, "Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications." *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 507–513, oct 2010. [Online]. Available: http://dx.doi.org/10.1136/jamia.2009.001560

[43] S. M. Fragomeni, A. Sciallis, and J. S. Jeruss, "Molecular subtypes and local-regional control of breast cancer," *Surg Oncol Clin N Am*, vol. 27, no. 1, pp. 95–120, 2018.

[44] R. Pillai and M. Pillai, "Dental care path for oral cancer," Jun 2020.

[45] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Édouard Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011. [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html

[46] M. E. Gianinazzi, C. S. Rueegg, K. Zimmerman, C. E. Kuehni, G. Michel, and S. P. O. Group, "Intra-rater and inter-rater reliability of a medical record abstraction study on transition of care after childhood cancer." *Plos One*, vol. 10, no. 5, p. e0124290, may 2015. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0124290

[47] H. Xu, S. P. Stenner, S. Doan, K. B. Johnson, L. R. Waitman, and J. C. Denny, "MedEx: a medication information extraction system for clinical narratives." *Journal of the American Medical Informatics Association*, vol. 17, no. 1, pp. 19–24, feb 2010. [Online]. Available: http://dx.doi.org/10.1197/jamia.M3378

[48] A. Liaw, M. Wiener *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[49] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.

[50] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[51] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical bert embeddings," *arXiv preprint arXiv:1904.03323*, 2019.

[52] E. Alsentzer, "Bio-clinicalbert," 2020. [Online]. Available: https://github.com/huggingface/transformers/tree/master/model_cards/emilyalsentzer/Bio_ClinicalBERT

[53] "SNOMED CT." [Online]. Available: https://www.nlm.nih.gov/healthit/snomedct/index.html

[54] "Unified medical language system (UMLS)." [Online]. Available: https://www.nlm.nih.gov/research/umls/

[55] "Rxnorm." [Online]. Available: https://www.nlm.nih.gov/research/umls/rxnorm/

[56] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[57] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/