COMPUTATIONAL PHENOTYPING AND DRUG REPURPOSING FROM ELECTRONIC MEDICAL RECORDS

Malvika Pillai

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Carolina Health Informatics Program in the Graduate School.

Chapel Hill
2022

Approved by:

Di Wu

Javed Mostafa

Jaime Arguello

Alexander Tropsha

Amanda Seyerle

ABSTRACT

Malvika Pillai: Computational Phenotyping and Drug Repurposing from Electronic Medical Records
(Under the direction of Di Wu)

Using electronic medical records (EMR) for research involves selecting cohorts and manipulating data for tasks like predictive analysis. Computational phenotyping for cohort characterization and stratification is becoming increasingly important for researchers to produce clinically relevant findings. There are significant amounts of time and effort devoted to manual chart abstraction by subject matter experts and researchers, which creates a large bottleneck for progress in clinical research. I focus on developing computational phenotyping pipelines, and I also focus on using EMR for drug repurposing in breast cancer. Drug repurposing is defined as the process of applying known drugs that are already on the market to new disease indications. Using EMR data for drug repurposing has the unique advantage of being able to observe a patient cohort over time and see drug effects on outcomes. In this dissertation, I present work on computational phenotyping and EMR-based drug repurposing. First, I use embedding models and foundational natural language processing methods to predict oral cancer risk with pathology notes. Second, I use natural language processing methods and transfer learning for breast cancer cohort selection and information extraction. Third, I present a pipeline for producing drug repurposing candidates from EMR and provide supporting evidence for predictions with biomedical literature and existing clinical trials.

To my parents, Ranjini and Rajendran Pillai. Thank you for being the wind beneath my wings.

ACKNOWLEDGEMENTS

I would like to sincerely thank my advisor, Dr. Di Wu for all her mentorship and support throughout my doctoral journey. Thank you for the countless hours you spent to help shape me into a better researcher. This dissertation would not have been possible without your continuous faith in me.

I am also very grateful for my wonderful committee members, Drs. Javed Mostafa, Alexander Tropsha, Jaime Arguello, and Amanda Seyerle. Thank you for your guidance and feedback that helped me shape my dissertation. I would also like to especially thank Dr. Javed Mostafa for his unwavering support and guidance. I am incredibly grateful for your mentorship and advice on both research and life. Your encouragement and optimistic thinking made this PhD journey an enjoyable one. I would also like to offer my special thanks to mentors that have helped me grow: Drs. Shiva Das for being a wonderful mentor and for guidance in both research and life, Saif Khairat for great conversations and helpful suggestions, and Kimberly Robasky for insightful advice.

I would like to recognize my funding source: The National Library of Medicine Institutional Training Grant for Research Training in Biomedical Informatics and Data Science (T15-LM012500), which provided funding for my training and professional growth. I am also thankful for the professional services from North Carolina Translational and Clinical Sciences Institute. Also, special thanks to the Carolina Health Informatics Program (CHIP) and the past program coordinators, Lindsey Womack and Shikha Yadhav.

My PhD journey was also shaped by the many colleagues and friends that supported me along the way. I would like to thank those in my cohort, especially Karthik Adapa for being a pillar of support and for countless working sessions to keep each other motivated, Terika McCall for her moral support and consistent willingness to help, Ashley Griffin for her continuous optimism and great conversations, and Rachel Stemerman for realistic conversations and brainstorming sessions. Special thanks to one of my fellow interns at the National Library of Medicine, Protiva Rahman, for advice on life, career, and professional growth. I would also like to thank my past roommates Brittany Allen, Ashtyn Neuwirth, Amy Martin, and Emma Kikerkov for helping me maintain a work-life balance. My friends from college, Minita Patel, Akila Khan, and Nadia Khan, thank you for listening to me brain dump and for working out problems with me. To all my friends- we laughed, we cried, and now we celebrate. Thank you again for being a part of this journey with me.

Lastly, I must express my most sincere appreciation and gratitude for my family. Amma and Acha, this achievement would not have been possible without you, and this is for you. I have many things to thank you both for, but with respect to my education, thank you for believing in me and always encouraging me to reach greater heights. Michu, thank you for being the best sister and always having my back. I cannot thank you all enough for your unwavering love and support

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ASCO	American Society of Clinical Oncology
ASOD	Adams School of Dentistry
CDW-H	Carolina Data Warehouse for Health
CPT	Current Procedural Terminology
dEMR	Dental electronic medical records
EHR	Electronic health record
EMERSE	Electronic Medical Record Search Engine
ER	Estrogen receptor
HER2	Human epidermal growth factor receptor 2
hEMR	Hospital electronic medical records
HR	Hormone receptor
i2b2	Informatics for Integrating Biology & the Bedside
ICD	International Classification of Diseases
ICD-CM	International Classification of Diseases-Clinical Modification
IRB	Institutional Review Board
NCCR	North Carolina Cancer Registry
NC TraCS	North Carolina Translational and Clinical Sciences Institute
OMOP	Observational Medical Outcomes Partnership
OMPL	Oral and Maxillofacial Pathology Lab
PCORnet	National Patient-Centered Clinical Research Network
PPV	Positive predictive value

PR	Progesterone receptor
TF-IDF	Term frequency-Inverse document frequency
UMLS	Unified Medical Language System
UNC	University of North Carolina at Chapel Hill

1. INTRODUCTION

1.1. **Problem Definition**

EMR data is powerful, quickly growing, and has been used successfully for clinical research in the past, but 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) (1). To provide context, if there is a female patient who is 26 years of age, she may have at least 1 to 3 visits yearly of different types (e.g., annual exam, emergency), which would constitute 26 to 78 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.

Using EMR for clinical research has been hindered by the lack of support for data manipulation in electronic health record (EHR) systems as well as missing data. Before the HITECH Act was instituted in 2009, very few (1.5%) hospitals in the US had EHR systems that were considered relatively comprehensive (2).

Even today, hospital systems are working toward building comprehensive EHR systems. Since 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 Meaningful Use incentives were put forth in the HITECH Act. Consequently, various common data models like those from the National Patient-Centered Clinical Research Network (PCORnet) and the Observational Medical Outcomes Partnership (OMOP) have been instituted to allow researchers easier access to EMR and help with data integration, but these efforts are still in progress. In addition, many hospital systems transitioned from legacy systems or no systems at all to industry-based systems like Epic and Cerner, leading to issues with data integration and missing data across patient lifespans. Lack of data interoperability and data integration are a few of many issues persisting with EMR use for research (3).

EHR contain written data in structured and unstructured formats. Structured data can range from including terminology-based codes (e.g., International Classification of Diseases-Clinical Modification (ICD-CM) diagnosis codes, Current Procedural Terminology (CPT) procedure codes), local codes, or no codes. Unstructured data can range from machine readable free-text (e.g., progress notes) to scanned reports (e.g., PDF). For instance, in the case of a biopsy, there will be a CPT code associated with the procedure, an ICD-CM code if a diagnosis was made, and a pathology report (free-text) dictating the findings of the biopsy. If a genetic test was done, results would predominantly be found in PDF reports. With large amounts of data in various types, computational phenotyping approaches can be used for cohort stratification. Computational phenotyping can be defined as the transformation of EHR data into meaningful variables for cohort selection and stratification (i.e., selecting a set of

patients and dividing them into groups of patients like them) (4). Computational phenotyping can be done for a variety of purposes, and in this work two purposes are highlighted: predicting disease risk for an individual and extracting covariates for repurposed drug treatment prediction. Figure 1 depicts a high-level picture of computational phenotyping for risk prediction.

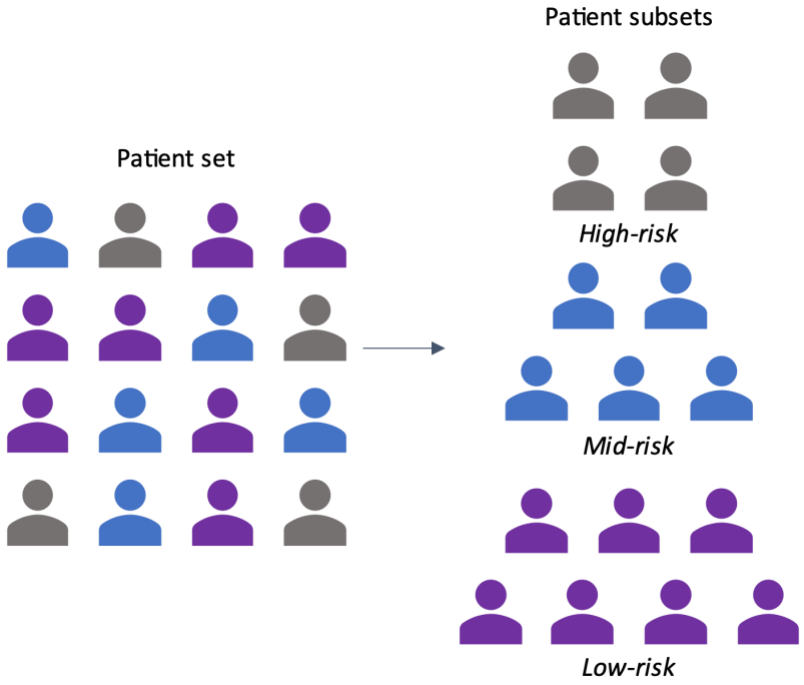


Figure 1. Computational phenotyping for risk prediction diagram

EMR data can also 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 longitudinal view of patient health including information on care for a disease of interest as well as for any co-morbidities. In addition, clinicians can prescribe off-label drugs to patients, which can enable researchers to simulate clinical trials to validate

whether the drugs are working to treat a disease of interest or not. In related work, EMR data has been used to predict the probability of treatment success using statistical approaches (5, 6). 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.

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 to transform feature representations of input data into desired outputs. It can be described as an extension of traditional statistical approaches (7). 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 (8). For computational phenotyping, both machine and deep learning approaches will be explored. For treatment success prediction, statistical and machine learning approaches will be compared. While deep learning methods are not as transparent as machine learning methods, they can achieve higher

performance in some cases, as demonstrated in research areas (9-11). To leverage the full potential of EMR, machine and deep learning methods can be used for cohort stratification and to take patient-level data variables and predict viability of drug repurposing candidates.

1.2. **Research Aims**

The purpose of this dissertation is divided into two components: using electronic medical records to 1) produce a computational phenotyping algorithm and 2) build a pipeline to validate drug repurposing candidates. The computational phenotyping tasks in this research were defined as: 1) using embedding models and foundational natural language processing methods to predict oral cancer risk with pathology notes, and 2) using natural language processing methods and transfer learning for breast cancer cohort selection and information extraction. A pipeline for validating drug repurposing candidates using electronic medical records was prototyped with breast cancer patients. The aims of this dissertation are:

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.

This dissertation is organized in six chapters. In Chapter 2, I described literature in the drug repurposing validation space. In Chapter 3, I presented a computational phenotyping approach for oral cancer risk prediction. In Chapter 4, I presented computational phenotyping approaches for breast cancer information extraction and outcome prediction. In Chapter 5, I proposed an approach for validating drug repurposing candidates for breast cancer. In Chapter 6, I presented my conclusions and directions for future research.

REFERENCES

1. Pugh CM. Electronic health records, physician workflows and system change: defining a pathway to better healthcare. Annals of Translational Medicine. 2019:27.

2. Kim E, Rubinstein SM, Nead KT, Wojcieszynski AP, Gabriel PE, Warner JL. The Evolving Use of Electronic Health Records (EHR) for Research. Seminars in Radiation Oncology. 2019;29(4):354-61.

3. Nordo AH, Levaux HP, Becnel LB, Galvez J, Rao P, Stem K, et al. Use of EHRs data for clinical research: Historical progress and current applications. Learning Health Systems. 2019;3(1):e10076.

4. Ma J, Zhang Q, Lou J, Ho JC, Xiong L, Jiang X, editors. Privacy-preserving tensor factorization for collaborative health data analysis. Proceedings of the 28th ACM International Conference on Information and Knowledge Management; 2019.

5. Xu H, Aldrich MC, Chen Q, Liu H, Peterson NB, Dai Q, et al. Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality. J Am Med Inform Assoc. 2015;22(1):179-91.

6. Khatri P, Roedder S, Kimura N, De Vusser K, Morgan AA, Gong Y, et al. A common rejection module (CRM) for acute rejection across multiple organs identifies novel therapeutics for organ transplantation. J Exp Med. 2013;210(11):2205-21.

7. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. JAMA. 2018;319(13):1317-8.

8. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-44.

9. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinformatics. 2017;19(6):1236-46.

10. Razavian N, Marcus J, Sontag D. Multi-task Prediction of Disease Onsets from Longitudinal Laboratory Tests. In: Finale D-V, Jim F, David K, Byron W, Jenna W, editors. Proceedings of the 1st Machine Learning for Healthcare Conference; Proceedings of Machine Learning Research: PMLR; 2016. p. 73--100.

11. Sathyanarayana A, Joty S, Fernandez-Luque L, Ofli F, Srivastava J, Elmagarmid A, et al. Sleep quality prediction from wearable data using deep learning. JMIR mHealth and uHealth. 2016;4(4):e6562.

2. LITERATURE REVIEW

2.1. **Introduction**

Drug discovery and development is a cost and time burdensome process that has stagnated the entry of new drugs into the market. The traditional process for drug development can take approximately 12 to 16 years and cost approximately $1 to $2 billion (12) (Figure 2). Pre-clinical research consists of laboratory and animal testing of a drug compound. Consequently, Phases I through III clinical trials determine drug safety, efficacy, and therapeutic effect, respectively. Then, in America, the drug candidate is pushed for Food and Drug Administration (FDA) review. Due to the high cost and time burden of the traditional process of drug development, finding whether an existing drug can be repurposed for treatment of a different disease that this drug hasn't been indicated to in the drug label is an alternative, more cost-effective option to address many of the barriers to getting a drug to the market.

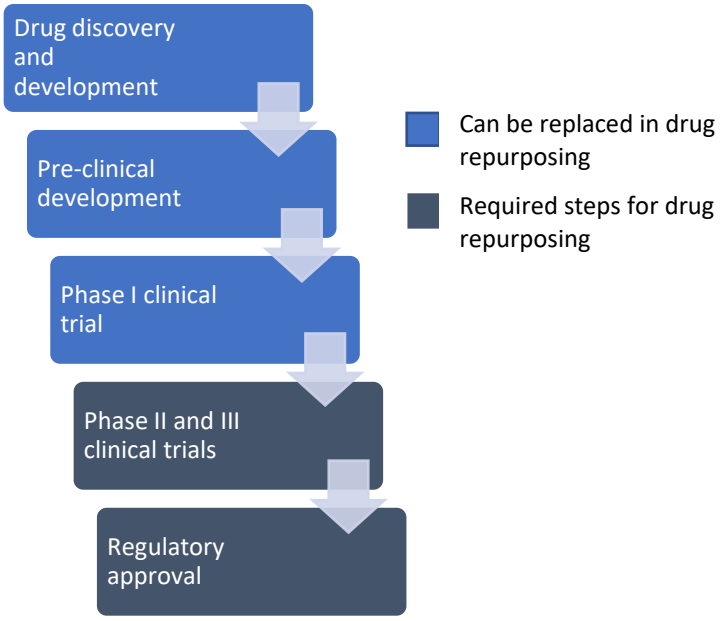


Figure 2. Traditional drug development process

Drug repurposing is defined as the process of applying known drugs/compounds that are already on the market to new disease indications. Repurposed drugs can be exempt from the prior phases leading to Phases II and III clinical trials and FDA approval process, reducing time and cost. For example, a liberal estimate for cost and number of years required to repurpose a drug is approximately $300 million for approximately 6 years (12). The risk of failure is lower for repurposed drugs because candidates for late-stage repurposing have already been proven safe through preclinical models and in humans (13). Based on prior preclinical testing, drug repositioning shortens the processing time and reduces the cost to find a drug for the different disease, that will positively affect downstream effects on population health outcomes, at the patient level.

Many previous successful attempts to repurpose drugs have been accidentally discovered side effects or extensive, time intensive research on particular drug properties(14). For example, sildenafil was originally developed to treat angina and was repurposed, by chance,

to treat erectile dysfunction. Minoxidil was developed to treat hypertension and was repurposed for hair loss through identification of hair growth as an adverse side effect. Both sildenafil and minoxidil were repurposed through retrospective clinical analysis (13, 14). Due to the serendipitous nature of previous discoveries, there has been a push toward data-driven repurposed drug development, a method that allows for more consistent hypothesis generation, that also responds to the recent availability of large-scale biomedical datasets (e.g., risk single nucleic polymorphisms (SNPs) identified in genome wide association studies (GWAS) and protein interaction databases) and clinical datasets (e.g., electronic medical records (EMRs)). Computational drug repurposing consists of using computational approaches for systematic data analysis that can lead to forming drug repurposing hypotheses. Essentially, the rigorous drug repurposing pipeline mainly involves making connections between two components, the existing drugs and the diseases that need drug treatments. The connection is built based on the features collected via biological experiments or clinically that can represent or describe these two components through computational tools, particularly when the feature datasets are large and high dimensional. It also involves later steps of validation (Figure 3).

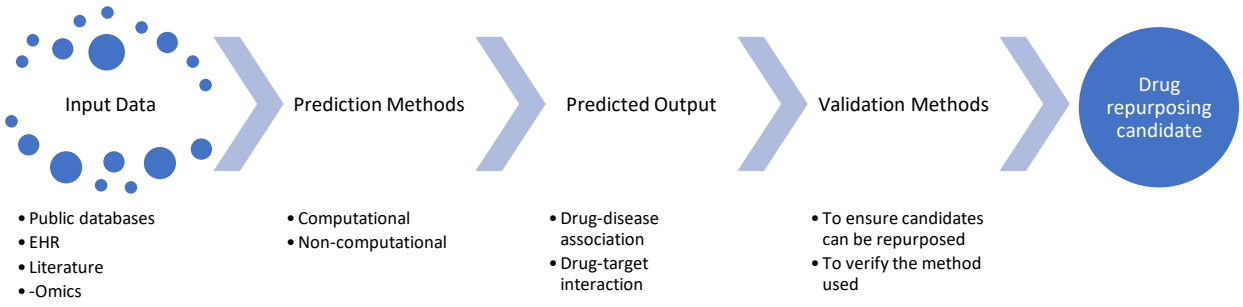


Figure 3. Drug repurposing workflow

The push toward data-driven drug repositioning has led to an increase in computational drug repositioning efforts. Conservative drug development consists of 'one drug, one target' research that does not evaluate off-target effects or multiple drug indications (14). Computational approaches are essentially to build direct or indirect connections between known drugs and diseases at a high-throughput scale in an automated way. We define the following main steps for a more complete drug repurposing pipeline for a disease. First, in the prediction step, people use the drug-disease connection to predict repurposed drug candidates computationally, producing the predicted repurposed drug candidate. Second, in the validation step, to remove some false positives, people use independent information that has not been used in the prediction step such as previous experimental/clinical studies, or independent resources/aspects of data (e.g., protein interaction data and gene expression data) about the drug-disease connection. If further supporting evidence is provided in this step, it builds better confidence of repurposed drugs, producing a validated repurposed drug candidate. In addition, false positive candidates may be removed from the repurposing list.

This review aims to answer the research question: how do researchers provide validation for drug repurposing candidate predictions from computational methods? In this review, we examine types of validation for drug candidates in computational drug repurposing studies. We provide a survey of the types of validation, which are divided into computational and non-computational approaches. We compare validation approaches within each category and describe the trade-offs of using each approach. We propose that the strongest forms of validation are those which provide the most evidence to push a drug candidate along the drug

development pipeline and provide recommendations for researchers deciding on how to validate drug repurposing candidates.

2.2. **Methods**

2.2.1. Search strategy

The methodology presented in the *PRISMA Statement* for systematic reviews (15) was used to create the search strategy. A comprehensive search was conducted across three databases: PubMed, Web of Science, and ACM Digital Library for all relevant articles pertaining to computational methods for drug repurposing. Both peer-reviewed journal articles and conference proceedings were included in the review. The search was conducted on September 12, 2019, with the query: (drug repurpos* OR drug reposition*) AND (computational OR computation OR computations OR algorithm OR algorithms OR network OR networks OR machine learning OR deep learning OR prediction OR predictions). Table 1 provides a breakdown of the results of the query per database, and the PRISMA flow diagram is presented in Figure 4.

2.2.2. Inclusion and exclusion criteria

The inclusion criteria for this review were: (1) the paper focused on drug repurposing candidate prediction and (2) the paper used a computational method for prediction. A study was excluded from the review if it: (1) did not include validation of predictions, (2) did not relate to drug repurposing, (3) was a non-computational paper, (4) was not an independent study (i.e., a review or perspective), (5) was not a full paper (i.e., an abstract for a poster), (6) was a duplicate paper, and (7) was not research for humans.

2.2.3. Study evaluation and data extraction

Covidence software was used for article screening (16). Extracted data included: number of citations, whether the paper was condition-specific, computational method used, and validation method used. Quality assessment was conducted with a citation analysis.

Table 1. Breakdown of Query Results

Query	PubMed	Web of Science	ACM Digital Library	Total Number of Studies Found
(drug repurpos* OR drug reposition*) AND (computational OR computation OR computations OR algorithm OR algorithms OR network OR networks OR machine learning OR deep learning OR prediction OR predictions)	996	1144	946	3086

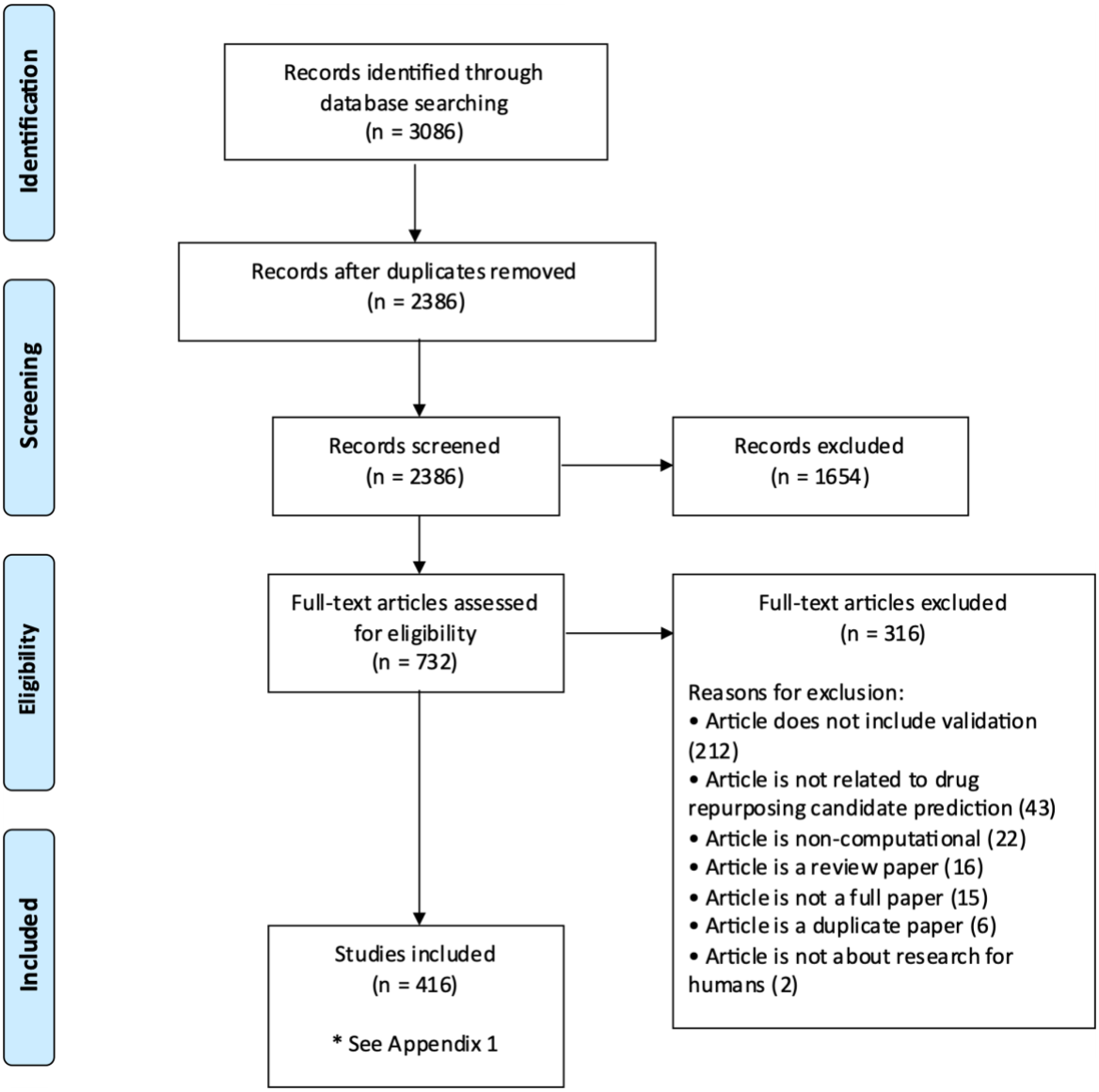


Figure 4. PRISMA flow diagram

2.3. **Results**

2.3.1. Overview of literature search results from all three databases

The search across PubMed, Web of Science, and ACM Digital Library identified 3086 articles. After filtering out duplicates, 2386 articles were included in the screening process. In abstract screening, 1654 studies were excluded for either not being related to drug repurposing candidate prediction, not using a computational method, not being research for humans, or not

being an independent study (i.e., a review or perspective). 732 studies were assessed for full-text eligibility. In full-text screening, 212 papers did not contain a validation method, 43 were not about drug repurposing candidate prediction, 22 were non-computational, 16 were review papers, 15 were not full papers, 6 were duplicate papers, and 2 were not research for humans.

2.3.2. Types of computational drug repurposing validation

For studies to push drug repositioning candidates forward in the drug discovery process, a drug candidate requires validation (i.e., supporting evidence). Two kinds of validation will be discussed: computational validation and non-computational validation. Computational validation methods found consist of retrospective clinical analysis, literature support, public database search, testing with external datasets, and online resource search. Non-computational validation methods found consist of in vitro, in vivo, or ex vivo experiments, drug repurposing clinical trials, personalized patient treatment, and expert review of predictions. Many studies use multiple forms of validation. Studies using both computational and non-computational validation are described in detail (See 2.3.5).

2.3.3. Computational validation

266 studies only contained computational validation.

2.3.3.1. Retrospective clinical analysis

Validation with retrospective clinical analysis can be divided into two categories: studies using EHR or insurance claims to validate drug repurposing candidates and studies searching for existing clinical trials. Both forms of validation are used on their own and in combination with other forms of validation. Brown et al (17) presented a clinical quantitative phenotyping

approach to identify interactions between diagnostic phenotypes and prescription drug use from a combination of four National Health and Nutrition Examination Survey (NHANES) datasets and validated three cases with longitudinal insurance claims data. The study used association testing to find sufficient evidence within the claims data to support drug predictions. Studies that search for existing clinical trials to validate drug candidate predictions generally use the clinical trials database (clinicaltrials.gov) to find trials that are testing the potential of predictions made within the studies. Evaluation datasets can also be compiled from the database to test performance of a drug repositioning system on a larger scale. Having existing clinical trials as support is vital information about a drug candidate because that indicates that the drug has already passed through hurdles in the drug discovery process (18).

There is no clear weakness in this approach and it's the strongest computational evidence toward regulatory approval for a given indication. This does not mean that preclinical evidence is not valuable, and it will be discussed in later sections. Knowing the phases (I-III) of clinical trials is important to evaluate how much validation is provided. This is because, if a study has shown that a drug has passed through Phase I clinical trials to treat a disease, this drug still needs to pass the remaining clinical trials to proceed to the FDA approval process. While some studies differentiated by clinical trial phase (19), others extracted drug-disease connection from clinical trials into datasets without specifying the clinical trial phases (20). EHR or insurance claims data, as a part of retrospective clinical analysis, have traditionally been used to examine off-label usage of drugs and finding off-label usage is another strong form of validation because it provides evidence that a drug has efficacy in humans for a given indication (17, 21-23). However, there are privacy and data accessibility issues when considering using

clinical records for validation, unlike many of the publicly available validation methods described in the review.

2.3.3.2. Literature support

166 studies solely used literature support and over half of the studies in the review mention using literature to support drug candidate predictions in conjunction with other validation methods. There has been tremendous growth in the amount of biomedical literature published with PubMed alone comprising of over 30 million citations (24), which allows for different kinds of methods to extract information. The types of literature support are grouped into three categories: literature search, survey, and mining.

Literature search validation uses a tool like PubMed to manually find relevant articles containing connections between old drugs and new uses. If there are no methods described for extracting literature and only a citation available, it is assumed that the authors used a literature search. Methods of prediction in studies that use literature for validation range across gene expression analysis, network or matrix manipulation, machine and deep learning, structure-based modeling or screening, and text or data mining models. The extent of literature support provided varies across studies, irrespective of the method used for prediction.

Literature search is the most prevalent method of validation found in the review, and literature search validation is conducted in various ways throughout studies in the review (25-30). The strength and weakness of this approach come from the studies selected as previous evidence. For example, Grenier et al (31) found existing clinical trials to support four of six predictions, while literature was found validating the predictions with human cell lines and animal models. The literature evidence was described in detail, making this strong literature support because

the drug-disease association mentioned in literature had been directly tested. However, if previous literature did not contain experimental evidence, compared to an in-depth description with case studies, providing citations for prediction without explanation or providing a low-quality citation can be considered weaker validation. For example, Zador et al (32) did not find research directly connecting drug repurposing candidates to atypical meningiomas but provided rationale with case studies for why the application should be studied further using previous literature as support. In contrast, Peng et al (33) primarily used public databases for validation but supported one drug candidate with a reference that indicated the drug-disease association had been tested in a clinical trial. The reference used as support was another drug repositioning study (34), rather than the original study or clinical trial.

Literature survey validation is defined as a person verifying a set of literature search results as true connections, which is more in-depth than a literature search. Two studies only used a literature survey to validate predictions. A literature survey can be time consuming, but it is the most thorough literature support described in this review (35, 36). Surveying literature consists of experts reading studies and deciding if the literature can be considered validation for predictions. The difference between a literature search and a literature survey is the expert opinion included, which ensures the quality of the supporting evidence provided. For example, Tan et al (35) had three experts read through the literature and include a study as validation if the majority of experts agreed. Using expert opinion increases the confidence in the drug repurposing candidate and its validation.

Literature mining validation uses computational algorithms to analyze literature and verify connections. 9 studies mentioned using literature mining to validate predictions.

Literature mining is the quickest approach to investigate previous evidence; however, all literature mining validation methods in the review used co-occurrence to illustrate the extent of evidence for a drug-disease co-mention (37-41). Using term-occurrence for literature mining only provides basic information on whether a drug and disease have been mentioned together. While co-occurrence can be effective, it does not take the kind of interaction into account. For example, a drug could have been reported to increase the risk of a disease like in Cheng et al (23) where retrospective clinical analysis showed that although there was evidence of two drugs in connection with coronary artery disease, one drug increased risk, and the other drug decreased risk and was used for further examination. Examining drug-disease co-occurrence does not provide this information, but it can demonstrate that the pair has been studied previously.

2.3.3.3. Public database search

8 studies only used public databases to validate predictions. Public databases can be data sources for predictive models, but after model training and testing, public databases are useful sources of supporting evidence for drug repurposing predictions. They are useful for both drug-disease and drug-target interaction (DTI) prediction, and many of the databases used for validation comprise the external datasets discussed in 2.3.3.4. In this review, searching the clinical trials database is considered different from searching for drug indications in other databases (See 2.3.3.1).

Public database search is a form of validation that is frequently used in combination with other methods of validation (42-46). The strength of using public databases comes from the type of information provided in the database and the frequency at which the database is

updated. The three most commonly used databases, DrugBank (47), KEGG (48), and CTD (49), are manually curated from various sources, which differentiates public database search from literature search and builds trust in the quality of associations described in the databases. The three databases are also updated regularly. Since the databases are reputed and well referenced in the scientific community, the weakness in this validation approach comes from how supporting evidence is extracted and examined within studies. For example, Luo et al (50) used KEGG and CTD to validate predictions, but when no support was found in either database for a given indication, the study did not reference any other evidence to validate a prediction. In addition, after showing the drug-disease associations that had support from either or both databases, the study did not provide further explanation to describe the supporting evidence. This is in contrast to how Peng et al (33) used case studies to describe evidence found in various databases and clearly explained how the predicted drugs and targets could interact.

2.3.3.4. External dataset support

15 studies only used external datasets to validate predictions. External dataset validation is defined as when a study uses an independent dataset separate from the data used in training the predictive model to evaluate drug repurposing predictions. The independent datasets are generally comprised of data from public databases such as those previously described. When testing with external datasets, the evaluation metrics used are especially important. Across the studies using external datasets for validation, area under the receiver operating characteristic (ROC) curve (AUC or AUROC) is the most common metric used, where higher AUC indicates better performance. The ROC curve is created by plotting true positive rate against false positive rate. Other commonly used metrics include precision and area under

the precision-recall curve (AUPR). Also, the types of external datasets vary depending on the prediction task. For example, many DTI prediction and drug-disease association prediction studies use external datasets to evaluate performance, but the type of dataset and extent of validation used for both tasks differ.

External dataset support is primarily used in drug repurposing studies using network analysis or machine learning methods for prediction (51-54). Validating with benchmark datasets like the Gottlieb et al (55) or Yamanishi et al (56) datasets is useful for comparing performance across prediction methods. However, using these benchmark datasets, especially for DTI prediction, does not provide enough supporting evidence to repurpose a drug for a given indication. For example, Xia et al (57) only used external datasets to show how the proposed prediction model could outperform others, but no further examination of drug candidates was included that could connect a drug to a new indication. Another limitation of benchmark datasets is that they require updates to account for additional knowledge since publication, and some studies have overcome this limitation by using them for training rather than as validation. For example, Keum et al (58) trained a model on the Yamanishi et al (56) datasets and tested on a dataset with updated DTI's from the DrugBank (59), KEGG BRITE (60), and DsigDB (61) databases. Studies also included validation using external datasets comprised of drug-disease pairs from clinical trials. As mentioned previously, studies that used clinical trials datasets provided metrics rather than detailed information about the clinical trials, which is exemplified in Wang et al (20) where performance is reported using precision-recall curves and mean average precision (MAP).

2.3.3.5. Online resource support

Online resource search is defined as using websites with drug or condition information that are generally directed at consumers for validation. It is the weakest form of validation in this review and is only used in combination with other validation methods (62-65). Commonly used websites include drugs.com and webMD.com. The information compiled on websites such as drugs.com may have reputed sources; however, without the source being described in a study, it is not possible to understand whether the validation presented is substantiated or not. Therefore, only stating that a prediction was mentioned on a website is not thorough validation, and further discussion is necessary. For example, Zhang et al (63) mainly used online resources for validation but also evaluated the prediction model with benchmark datasets and referenced DrugBank and biomedical literature to support drug-disease associations. Xu et al (65) also used online resources for secondary validation after evaluating model performance on a dataset compiled from literature and clinical trials.

2.3.4. Non-computational validation

123 studies only contained non-computational validation. Non-computational validation consists of expert review of predictions (66), experimental support (22, 53, 67-71), drug repurposing clinical trials (72), and personalized patient treatment (73, 74). A drug repurposing clinical trial is defined as a clinical trial that resulted from a drug repurposing effort. Personalized patient treatment is defined as using patient biological information to inform clinical decisions for the same patients. This is also referred to as precision medicine, where treatment is tailored to the individual patient.

2.3.4.1. <u>Experimental validation</u>

119 studies only used an in vitro, in vivo, or ex vivo experiment to validate predictions. Experimental validation for drug candidates is crucial in the preclinical development stage of drug development. Therefore, computational drug repurposing studies that validate candidates experimentally satisfy criteria for early-stage repurposing, making this approach strong. With satisfactory experimental validation, there is enough supporting evidence to pursue a drug repurposing clinical trial. The weakness in this approach is based on the effort required to complete the experiments. In comparison to searching a public database or finding another study with evidence in literature, conducting in vitro assays or examining drug performance in an in vivo model is much more time and cost intensive. Methods used for prediction in the studies using experimental validation in this review were network analysis, gene expression analysis, structural modeling, machine learning, and similarity-based approaches.

2.3.4.2. Other non-computational validation

Four studies performed alternate forms of experimental validation. Grammer et al (72) performed a clinical trial to validate a drug candidate prediction, Laganà et al (74) and Velez et al (73) conducted studies with personalized patient treatment, and Bakal et al (66) conducted an expert review of predictions. Of all the approaches used, expert review of predictions can be considered the weakest because it is based on experts describing what could be used in clinical practice. As it is based on human expertise, experiments need to be conducted to verify the hypotheses made. A drug repurposing clinical trial is a strong from of validation and stronger than experimental support. It satisfies the Phase I clinical trial requirement to determine the safety of a drug candidate in humans. The precision medicine approach is also strong validation along with other drug repurposing clinical trials that may or not have biological basis. Laganà et

al (74) used a genomic and transcriptomic approach, while Velez et al (73) used a proteomic approach to predict drug candidates for their patient populations. The drugs predicted for use were already approved for treatment in humans but for other conditions.

2.3.5. Both computational and non-computational validation

27 studies include both computational and in vitro, in vivo, or ex vivo experimental validation. The goal of using combinations of validation is to provide multi-faceted support for drug candidates to push drug repurposing candidates through the drug development process and inform clinical trials. Including experimental support satisfies the preclinical development stage in the drug development process.

Cheng et al (23) used a network analysis approach by constructing a protein-protein interactome for drug-disease association prediction. The study selected high confidence drug-cardiovascular outcome associations, quantified by distance-based z-score, to illustrate the effectiveness of the approach. Propensity score matching and sensitivity analyses were conducted to validate four associations using the Truven MarketScan and Optum Clinformatics databases, and half of the drugs, hydroxychloroquine and carbamazepine, were found to decrease and increase the risk of coronary artery disease respectively. An in vitro assay using human aortic endothelial cells was used to validate the connection between hydroxychloroquine and decreased CAD risk. Similarly, Gayvert et al (22) used a network analysis approach and validated candidates with existing clinical trial support, in vitro experiments, and retrospective clinical analysis. The approach was used to predict drug-transcription factor interactions, specifically to find drugs that could inhibit oncogenic transcription factors. Gene expression analysis was then used to prioritize drug candidates and

identified dexamethasone as a potential drug candidate. In vitro experiments validated the connection between dexamethasone and ERG transcription factor activity inhibition. An existing clinical trial associated dexamethasone and prostate cancer treatment, and this drug-disease association was consequently validated with retrospective clinical analysis using survival analysis, Cox proportional hazards test, and logistic regression with EHR from Columbia University Medical Center. Chen et al (69) used virtual screening of two databases, CMap and LINCS, to predict drug candidates for hepatocellular carcinoma (HCC), and used existing clinical trials, in vitro assays, and an in vivo mouse model to validate predictions. Unlike other studies that searched for clinical trials with predicted drug-disease associations, a list of drugs from the clinical trials database was used to rank predictions and provide confidence in the approach. However, the candidate with the highest score had not been tested in preclinical models yet, so the study validated the connection between niclosamide ethanolamine and HCC using in vitro assays and an in vivo mouse model.

2.3.6. Prediction methods used with validation types

In each drug repurposing candidate prediction study in this review, there are two methods used: a prediction method and a validation method. All the studies were divided into categories based on validation type. Within each category, the most used prediction methods were examined. Given a validation type like retrospective clinical analysis, the aim of the comparison is to understand which prediction methods can be validated using this validation method. Some examples of input data for prediction in Figure 3 are also used by different studies for their validation as described in this review. Table 2 shows the distribution of validation methods used and the most common prediction methods used with each validation

subtype. The most used prediction methods across all validation subtypes were network analysis (e.g., non-neural networks like protein-protein interaction network), gene expression analysis (e.g., connectivity mapping (75-80)), traditional machine or deep learning, and structural modeling. Another less popular prediction method, structural modeling is an overarching category that includes virtual screening, pharmacophore modeling, and molecular docking.

Network analysis was the most frequently used overall with 55% (229) of studies using the method to predict drug repurposing candidates. Gene expression analysis was used for prediction mostly in studies using retrospective clinical analysis, literature support, or experimental support for validation, while in other validation categories, it was not among the top three methods used. 12% (8) of studies (65) using retrospective analysis, 11% (26) of studies (233) using literature support, and 25% (36) of studies (146) using experimental support applied gene expression analysis in prediction. Machine or deep learning methods were used for prediction in studies using the following validation types: 13% (10) of studies (65) using retrospective clinical analysis, 9% (22) of studies (233) using literature support, 11% (3) of studies (28) using external dataset support, 20% (8) of studies (40) using public database search, and 5% (8) of studies (146) using experimental support. While prediction methods were similar across validation types, a few were different from the majority. 32% (47) of studies (146) using experimental support for validation used structural modeling for drug repurposing candidate prediction, which is in comparison to 14% (4) of studies (28) using external datasets, 11% (26) of studies (233) using literature support, and 2% (1) of studies (65) using retrospective clinical analysis as validation. The third most common drug repurposing candidate prediction

method (6%, 3 studies) associated with studies using public database search for validation was matrix factorization, which is in comparison to 7% (2) of studies (28) using external datasets, 2% (5) of studies (233) using literature support, and 2% (1) of studies (65) using retrospective clinical analysis as validation.

Table 2. Prediction methods used with validation subtypes in literature.

* For validation subtypes associated with more than 3 studies, the prediction methods are ranked in terms of frequency of use. For validation subtypes where only one prediction method was used across all associated studies, a ranking was not included.

Validation Type	Validation Subtype	Count of Studies Using Validation Subtype (% out of 416)	Top Prediction Methods Used with Validation Subtype
Computational	Retrospective clinical analysis	65 (15.6%)	1. * Network analysis 2. Gene expression analysis 3. Machine or deep learning
	Literature support	233 (56.0%)	1. Network analysis 2. Gene expression analysis 3. Structural modeling
	External dataset support	28 (6.7%)	1. Network analysis 2. Structural modeling 3. Machine learning
	Public database search	40 (9.6%)	1. Network analysis 2. Machine or deep learning 3. Matrix factorization
	Online resource search	5 (1.2%)	• Network analysis
Non-computational	Experimental support	146 (35.1%)	1. Structural modeling 2. Network analysis

			3. Gene expression analysis
	Drug repurposing clinical trial	1 (0.2%)	• Gene expression analysis
	Personalized patient treatment	2 (0.5%)	• Hierarchical clustering • Whole-exome sequencing, Targeted panel sequencing, RNA sequencing
	Expert review of predictions	1 (0.2%)	• Machine learning

2.4. **Discussion**

This review examined how researchers define and provide validation for computational drug repurposing candidates. Computational drug repurposing provides a systematic method for connecting approved drugs with new indications, reducing the amount of time and cost of drug development. 628 studies using computational approaches for drug repurposing were identified in this review, showing the vast amount of research that has been conducted in this area. However, predicting drug candidates without providing independent support does not provide enough evidence for a drug to be further pursued. 416 of 628 studies, which is roughly two thirds of the studies, contained validation for predictions. Validation is needed to demonstrate the significance of the prediction, and the review has shown that the number of studies including validation have increased over time (Figure 5). There are various levels of validation though, and not all types will allow for drug candidates to progress faster in the drug development process.

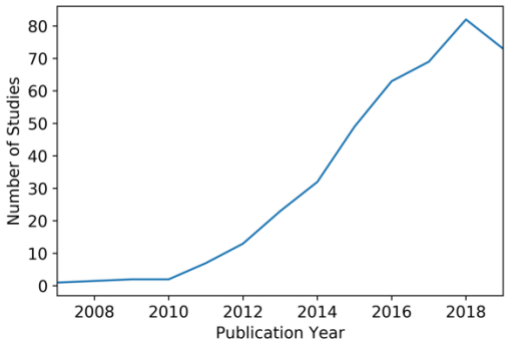


Figure 5. Number of studies including validation over time

Nine types of validation are described in the review, and each has its strengths and weaknesses. They also overlap across studies in that some are used as input data for prediction and used by different studies as validation. Computational validation methods consist of retrospective clinical analysis, literature support, public database search, testing with external datasets, and online resources. The computational methods are ranked in order based on their strength of validation from validation methods that provide enough support for a drug to continue to clinical trials to methods that do not provide enough support to move a drug through the drug development process (Figure 2). Non-computational validation methods are in vitro, in vivo, or ex vivo experiments, drug repurposing clinical trials, personalized patient treatment, and expert review of predictions. Many studies use multiple forms of validation. Studies using both computational and non-computational validation are described in detail (See 2.3.5). The strengths and weaknesses are based on how much we perceive that the form of validation will allow for a drug to be pushed to market faster. Computational and non-computational validation are difficult to compare because non-computational validation is a required component of the drug development process unless it has already been conducted in past studies.

This review focuses on providing a broad overview of methods researchers use to validate drug repurposing candidates by categorizing them and describing their uses. The validation methods used were analyzed within their categories and all nine categories were compared to each other. There were five types of computational validation found and four types of non-computational validation found. There are varying definitions of validation in drug repurposing, and this is the first review to search for computational drug repurposing studies and consequently explore how researchers provide independent supporting evidence as validation without prediction method-focused exclusion criteria. In addition, this review examines the most common validation types used in combination with different prediction methods, providing guidance for researchers who wish to select a validation method based on their chosen prediction method and the trade-off between strength and time/cost needed for each validation method.

2.4.1. Recommendations

Given a computational drug repurposing study, there are various types of validation to choose from. Strength of validation is determined in this review based on how close a drug candidate is to regulatory approval in the traditional drug development process after completing a given form of validation. In terms of strength, non-computational approaches like experimental validation, clinical trials, and personalized patient treatment are far stronger than any computational approaches for validation. The non-computational approaches can be considered “true” validation, but they are more time and cost intensive. Computational approaches can be ranked as follows: retrospective clinical analysis, literature survey, literature search, literature mining, public database support, external dataset support, and online

resource support. Retrospective clinical analysis is the strongest form of computational validation as it consists of off-label usage and clinical trial support. Off-label usage can be used to demonstrate a drug candidate's effect on humans; however, aside from identifying off-label usage, analysis must be conducted to identify whether the drug effect was positive or negative on a given condition. Clinical trial support for a drug can indicate drug safety and efficacy, depending on the trial stage. Searching for clinical trials is a more straightforward approach than identifying off-label usage as clinical trials are more systematic and have clearly defined results for a drug's effect on a given condition. As both off-label usage and clinical trials can be used to demonstrate a drug candidate's effect on humans, the evidence can be considered stronger than validation with animal models, leading to late-stage repurposing. Based on review findings, computational evidence should be explored prior to conducting non-computational validation.

2.4.2. Limitations of the review

This review explores how studies provide validation for computational drug repurposing, and a key limitation is the diversity in how researchers interpret validation. For example, within literature search support, studies can provide detailed case studies or provide citations as validation. Both are considered validation although one is more thorough than the other. In addition, some validation methods are also used for prediction. For example, retrospective clinical analysis is a commonly used method of prediction (81-84), but studies have also used it for hypothesis validation. Although the overlap between prediction and validation methods can be a limitation, this is mitigated if the prediction method and validation method are not the same in one study.

2.4.3. Conclusions

Validation for computational drug repurposing provides confidence in predicted drug repurposing candidates, the extent of which varies based on the form of validation used. Studies using computational and non-computational validation approaches are described in this review. All non-computational validation methods can be summarized as expert opinion, animal testing, and clinical testing. Animal and clinical testing are undertaken in the traditional drug development process and are still required for the repurposed drug development process if there is no evidence of them already being completed. All computational validation methods can be summarized as either finding overlaps between predicted associations and an accepted form of evidence or using analytical metrics to evaluate model performance. Finding overlaps with predicted associations and evidence like literature or public databases with case studies can provide confidence by satisfying parts of the drug development process for early or late-stage repurposing. Literature support and public databases are used regardless of prediction method. Using analytical metrics can provide confidence in the predictions through statistical significance and inform further non-computational validation for early-stage repurposing. Studies with DTI prediction, for example, mainly have external dataset validation, which primarily uses analytical metrics to substantiate results. As the main goal of drug repurposing is to reduce the amount of time and money that goes into pushing a drug to market, the main goal of validation is to shorten that process even further.

REFERENCES

12. Nosengo N. Can you teach old drugs new tricks? Nature News. 2016;534(7607):314.

13. Pushpakom S, Iorio F, Eyers PA, Escott KJ, Hopper S, Wells A, et al. Drug repurposing: progress, challenges and recommendations. Nat Rev Drug Discov. 2019;18(1):41-58.

14. Brown AS, Patel CJ. A review of validation strategies for computational drug repositioning. Brief Bioinformatics. 2018;19(1):174-7.

15. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS medicine. 2009;6(7):e1000097.

16. Innovation VH, Innovation VH. Covidence. Melbourne, Australia; 2017.

17. Brown AS, Rasooly D, Patel CJ. Leveraging Population-Based Clinical Quantitative Phenotyping for Drug Repositioning. CPT Pharmacometrics Syst Pharmacol. 2018;7(2):124-9.

18. Baker NC, Ekins S, Williams AJ, Tropsha A. A bibliometric review of drug repurposing. Drug Discov Today. 2018;23(3):661-72.

19. Yeung T-L, Sheng J, Leung CS, Li F, Kim J, Ho SY, et al. Systematic Identification of Druggable Epithelial-Stromal Crosstalk Signaling Networks in Ovarian Cancer. J Natl Cancer Inst. 2019;111(3):272-82.

20. Wang Q, Xu R. Drug repositioning for prostate cancer: using a data-driven approach to gain new insights. AMIA Annu Symp Proc. 2017;2017:1724-33.

21. Gottlieb A, Altman RB. Integrating systems biology sources illuminates drug action. Clin Pharmacol Ther. 2014;95(6):663-9.

22. Gayvert KM, Dardenne E, Cheung C, Boland MR, Lorberbaum T, Wanjala J, et al. A computational drug repositioning approach for targeting oncogenic transcription factors. Cell Rep. 2016;15(11):2348-56.

23. Cheng F, Desai RJ, Handy DE, Wang R, Schneeweiss S, Barabv°si A-Ls, et al. Network-based approach to prediction and population-based validation of in silico drug repurposing. Nat Commun. 2018;9(1):2691.

24. Information BNCfB. PubMed Help - PubMed Help - NCBI Bookshelf. 2019.

25. Luo H, Wang J, Li M, Luo J, Peng X, Wu F-X, et al. Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm. Bioinformatics. 2016;32(17):2664-71.

26. Wain LV, Shrine N, Artigas MaS, Erzurumluoglu AM, Noyvert B, Bossini-Castillo L, et al. Genome-wide association analyses for lung function and chronic obstructive pulmonary disease identify new loci and potential druggable targets. Nat Genet. 2017;49(3):416-25.

27. Luo Y, Zhao X, Zhou J, Yang J, Zhang Y, Kuang W, et al. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. Nat Commun. 2017;8(1):573.

28. Zhou H, Gao M, Skolnick J. Comprehensive prediction of drug-protein interactions and side effects for the human proteome. Sci Rep. 2015;5:11090.

29. Fang J, Cai C, Chai Y, Zhou J, Huang Y, Gao L, et al. Quantitative and systems pharmacology 4. Network-based analysis of drug pleiotropy on coronary artery disease. Eur J Med Chem. 2019;161:192-204.

30. Wang Q, Chen R, Cheng F, Wei Q, Ji Y, Yang H, et al. A Bayesian framework that integrates multi-omics data and gene networks predicts risk genes from schizophrenia GWAS data. Nat Neurosci. 2019;22(5):691-9.

31. Grenier L, Hu P. Computational drug repurposing for inflammatory bowel disease using genetic information. Comput Struct Biotechnol J. 2019;17:127-35.

32. Zador Z, King AT, Geifman N. New drug candidates for treatment of atypical meningiomas: An integrated approach using gene expression signatures for drug repurposing. PLoS ONE. 2018;13(3):e0194701.

33. Peng L, Zhu W, Liao B, Duan Y, Chen M, Chen Y, et al. Screening drug-target interactions with positive-unlabeled learning. Sci Rep. 2017;7(1):8087.

34. Mathur S, Dinakarpandian D. Drug repositioning using disease associated biological processes and network analysis of drug targets. AMIA Annu Symp Proc. 2011;2011:305-11.

35. Tan F, Yang R, Xu X, Chen X, Wang Y, Ma H, et al. Drug repositioning by applying 'expression profiles' generated by integrating chemical structure similarity and gene semantic similarity. Mol Biosyst. 2014;10(5):1126-38.

36. Cohen T, Widdows D, Schvaneveldt RW, Davies P, Rindflesch TC. Discovering discovery patterns with Predication-based Semantic Indexing. J Biomed Inform. 2012;45(6):1049-65.

37. Wei X, Zhang Y, Huang Y, Fang Y. Predicting drug,Äìdisease associations by network embedding and biomedical data integration. Data Technologies and Applications. 2019;53(2):217-29.

38. Zhao M, Yang CC, editors. Automated Off-label Drug Use Detection from User Generated Content. the 8th ACM International Conference; 2017 2017/08/20/. New York, New York, USA: ACM Press.

39. Lee T, Yoon Y. Drug repositioning using drug-disease vectors based on an integrated network. BMC Bioinformatics. 2018;19(1):446.

40. Zhao Q-Q, Li X, Luo L-P, Qian Y, Liu Y-L, Wu H-T. Repurposing of Approved Cardiovascular Drugs against Ischemic Cerebrovascular Disease by Disease-Disease Associated Network-Assisted Prediction. Chem Pharm Bull. 2019;67(1):32-40.

41. Oh M, Ahn J, Lee T, Jang G, Park C, Yoon Y. Drug voyager: a computational platform for exploring unintended drug action. BMC Bioinformatics. 2017;18(1):131.

42. Lotfi Shahreza M, Ghadiri N, Mousavi SR, Varshosaz J, Green JR. Heter-LP: A heterogeneous label propagation algorithm and its application in drug repositioning. J Biomed Inform. 2017;68:167-83.

43. Yan C, Wang J, Lan W, Wu F-X, Pan Y. SDTRLS: Predicting Drug-Target Interactions for Complex Diseases Based on Chemical Substructures. Complexity. 2017;2017:1-10.

44. Xuan P, Cao Y, Zhang T, Wang X, Pan S, Shen T. Drug repositioning through integration of prior knowledge and projections of drugs and diseases. Bioinformatics. 2019;35(20):4108-19.

45. Sharma A, Rani R. BE-DTI': Ensemble framework for drug target interaction prediction using dimensionality reduction and active learning. Comput Methods Programs Biomed. 2018;165:151-62.

46. Yan C-K, Wang W-X, Zhang G, Wang J-L, Patel A. Birwdda: A novel drug repositioning method based on multisimilarity fusion. J Comput Biol. 2019.

47. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018;46(D1):D1074-D82.

48. Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M. New approach for understanding genome variations in KEGG. Nucleic Acids Res. 2019;47(D1):D590-D5.

49. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, McMorran R, Wiegers J, et al. The Comparative Toxicogenomics Database: update 2019. Nucleic Acids Res. 2019;47(D1):D948-D54.

50. Luo H, Li M, Wang S, Liu Q, Li Y, Wang J. Computational drug repositioning using low-rank matrix approximation and randomized algorithms. Bioinformatics. 2018;34(11):1904-12.

51. Wu Z, Cheng F, Li J, Li W, Liu G, Tang Y. SDTNBI: an integrated network and chemoinformatics tool for systematic prediction of drug-target interactions and drug repositioning. Brief Bioinformatics. 2017;18(2):333-47.

52. Wang X, Pan C, Gong J, Liu X, Li H. Enhancing the Enrichment of Pharmacophore-Based Target Prediction for the Polypharmacological Profiles of Drugs. J Chem Inf Model. 2016;56(6):1175-83.

53. Napolitano F, Carrella D, Mandriani B, Pisonero-Vaquero S, Sirci F, Medina DL, et al. gene2drug: a computational tool for pathway-based rational drug repositioning. Bioinformatics. 2018;34(9):1498-505.

54. Xu J, Regan-Fendt K, Deng S, Carson WE, Payne PRO, Li F. Diffusion mapping of drug targets on disease signaling network elements reveals drug combination strategies. Pac Symp Biocomput. 2018;23:92-103.

55. Gottlieb A, Stein GY, Ruppin E, Sharan R. PREDICT: a method for inferring novel drug indications with application to personalized medicine. Mol Syst Biol. 2011;7:496.

56. Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. Bioinformatics. 2008;24(13):i232-40.

57. Xia L-Y, Yang Z-Y, Zhang H, Liang Y. Improved Prediction of Drug-Target Interactions Using Self-Paced Learning with Collaborative Matrix Factorization. J Chem Inf Model. 2019;59(7):3340-51.

58. Keum J, Nam H. SELF-BLM: Prediction of drug-target interactions via self-training SVM. PLoS ONE. 2017;12(2):e0171839.

59. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, et al. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res. 2014;42(Database issue):D1091-7.

60. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, et al. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 2006;34(Database issue):D354-7.

61. Yoo M, Shin J, Kim J, Ryall KA, Lee K, Lee S, et al. DSigDB: drug signatures database for gene set analysis. Bioinformatics. 2015;31(18):3069-71.

62. Udrescu Li, SbV¢rcea L, TopvÆrceanu A, Iovanovici A, Kurunczi L, Bogdan P, et al. Clustering drug-drug interaction networks with energy model layouts: community analysis and drug repurposing. Sci Rep. 2016;6:32745.

63. Zhang W, Yue X, Huang F, Liu R, Chen Y, Ruan C. Predicting drug-disease associations and their therapeutic function based on the drug-disease association bipartite network. Methods. 2018;145:51-9.

64. Zhang W, Yue X, Chen Y, Lin W, Li B, Liu F, et al., editors. Predicting drug-disease associations based on the known association bipartite network. 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2017 2017/11/13/: IEEE.

65. Xu R, Wang Q. A genomics-based systems approach towards drug repositioning for rheumatoid arthritis. BMC Genomics. 2016;17 Suppl 7:518.

66. Bakal G, Talari P, Kakani EV, Kavuluru R. Exploiting semantic patterns over biomedical knowledge graphs for predicting treatment and causative relations. J Biomed Inform. 2018;82:189-99.

67. Ganapathiraju MK, Thahir M, Handen A, Sarkar SN, Sweet RA, Nimgaonkar VL, et al. Schizophrenia interactome with 504 novel protein-protein interactions. NPJ Schizophr. 2016;2:16012.

68. Lee J-K, Liu Z, Sa JK, Shin S, Wang J, Bordyuh M, et al. Pharmacogenomic landscape of patient-derived tumor cells informs precision oncology therapy. Nat Genet. 2018;50(10):1399-411.

69. Chen B, Wei W, Ma L, Yang B, Gill RM, Chua M-S, et al. Computational discovery of niclosamide ethanolamine, a repurposed drug candidate that reduces growth of hepatocellular carcinoma cells in¬†vitro and in mice by inhibiting cell division cycle 37 signaling. Gastroenterology. 2017;152(8):2022-36.

70. Fuentealba Ma, DV∂nerta≈ü HM, Williams R, Labbadia J, Thornton JM, Partridge L. Using the drug-protein interactome to identify anti-ageing compounds for humans. PLoS Comput Biol. 2019;15(1):e1006639.

71. Turanli B, Karagoz K, Bidkhori G, Sinha R, Gatza ML, Uhlen M, et al. Multi-Omic Data Interpretation to Repurpose Subtype Specific Drug Candidates for Breast Cancer. Front Genet. 2019;10:420.

72. Grammer AC, Ryals MM, Heuer SE, Robl RD, Madamanchi S, Davis LS, et al. Drug repositioning in SLE: crowd-sourcing, literature-mining and Big Data analysis. Lupus. 2016;25(10):1150-70.

73. Velez G, Bassuk AG, Colgan D, Tsang SH, Mahajan VB. Therapeutic drug repositioning using personalized proteomics of liquid biopsies. JCI Insight. 2017;2(24).

74. LaganV† A, Beno I, Melnekoff D, Leshchenko V, Madduri D, Ramdas D, et al. Precision medicine for relapsed multiple myeloma on the basis of an integrative multiomics approach. JCO Precis Oncol. 2018;2018.

75. Zhang M, Luo H, Xi Z, Rogaeva E. Drug repositioning for diabetes based on 'omics' data mining. PLoS ONE. 2015;10(5):e0126082.

76. Chung F-H, Chiang Y-R, Tseng A-L, Sung Y-C, Lu J, Huang M-C, et al. Functional Module Connectivity Map (FMCM): a framework for searching repurposed drug compounds for systems treatment of cancer and an application to colorectal adenocarcinoma. PLoS ONE. 2014;9(1):e86299.

77. O'Reilly PG, Wen Q, Bankhead P, Dunne PD, McArt DG, McPherson S, et al. QUADrATiC: scalable gene expression connectivity mapping for repurposing FDA-approved therapeutics. BMC Bioinformatics. 2016;17(1):198.

78. Wen Q, O'Reilly P, Dunne PD, Lawler M, Van Schaeybroeck S, Salto-Tellez M, et al. Connectivity mapping using a combined gene signature from multiple colorectal cancer datasets identified candidate drugs including existing chemotherapies. BMC Syst Biol. 2015;9 Suppl 5:S4.

79. Powell TR, Murphy T, Lee SH, Price J, Thuret S, Breen G. Transcriptomic profiling of human hippocampal progenitor cells treated with antidepressants and its application in drug repositioning. J Psychopharmacol (Oxford). 2017;31(3):338-45.

80. Chen Y-T, Xie J-Y, Sun Q, Mo W-J. Novel drug candidates for treating esophageal carcinoma: A study on differentially expressed genes, using connectivity mapping and molecular docking. Int J Oncol. 2019;54(1):152-66.

81. Kuang Z, Thomson J, Caldwell M, Peissig P, Stewart R, Page D. Computational Drug Repositioning Using Continuous Self-Controlled Case Series. KDD. 2016;2016:491-500.

82. Low YS, Daugherty AC, Schroeder EA, Chen W, Seto T, Weber S, et al. Synergistic drug combinations from electronic health records and gene expression. J Am Med Inform Assoc. 2017;24(3):565-76.

83. Koren G, Nordon G, Radinsky K, Shalev V. Machine learning of big data in gaining insight into successful treatment of hypertension. Pharmacol Res Perspect. 2018;6(3):e00396.

84. Su EW, Sanger TM. Systematic drug repositioning through mining adverse event data in ClinicalTrials.gov. PeerJ. 2017;5:e3154.

3. COMPUTATIONAL PHENOTYPING FOR ORAL CANCER USING ELECTRONIC MEDICAL RECORDS

3.1. **Introduction**

In 2020, the World Health Organization estimated that there were over 400,000 new cases of oral cancer worldwide. Globally, the age-standardized disability-adjusted life years (DALYs) was estimated at 64.23 per 100,000 people (85). In the United States, oral cancer accounts for about 3% of all cancer incidences with a 66.2% 5-year relative survival. However, if diagnosed at early stages, oral cancer survival rates can be over 80% (86). Patients are typically screened for abnormalities associated with oral cancer by their dentist. If any abnormalities are detected, the patient is referred to a specialist for a biopsy, which is assessed by an oral pathologist who will examine the sample's histopathology and write a paragraph describing what they see under the microscope. In its early stages, it can be difficult to diagnose oral cancer based on the classical clinical features associated with malignancy. Later diagnosis can result from detection of cancer metastases (i.e., spreading) or significant abnormalities (e.g., lesions).

The typical clinical workflow for a patient who may have oral cancer begins with a dentist screening for oral cancer by assessing characteristics of the mouth. If the patient has a persistent lesion in the oral cavity, the dentist will refer them for a biopsy. An oral pathologist then conducts the biopsy assessment and writes a microscopic description that would explain how the sample looks under the microscope.

If the microscopic description does not contain clear indications that the lesion is malignant, the patient may not be referred to an oncologist. However, the lesion could be precancerous. Mortality and morbidity rates of oral cancer decrease with early diagnosis, and risk stratification is an open research area.

In this study, we aim to use the microscopic descriptions written by oral pathologists to estimate the probability of a sample being cancerous to aid clinicians in deciding whether to refer a patient to an oncologist. The significance of this work would be reducing the burden on a clinician with a machine learning system that will be able to provide a recommendation to triage cases.

3.2. **Prior Work**

Many research studies have focused on identify biomarkers as indicators of carcinogenesis (87-89). Machine learning approaches have increasingly been used in the oral cancer space, largely for oral cancer risk, survival, and prognosis prediction (90-93). A biopsy provides the gold standard diagnosis for oral cancer (86), and related work has also focused on developing machine and deep learning models on pathology images as virtual assistants to pathologists (94-96). However, to the best of our knowledge there is a gap in the literature in terms of machine learning approaches to tackle early diagnosis prediction in oral cancer using pathology notes. Researchers have been successful differentiating specific oral cancer types such as oral squamous cell carcinoma from benign and normal tissues using histopathological images, but after the biopsy report is written, a clinician will interpret the biopsy report to determine if a referral to an oncologist is warranted, making the description written by the pathologist useful for risk prediction. Pathology reports for all cancers are a tremendous source

of information on a patient's cancer with details like site, histology, and behavior. Various natural language processing (NLP) methods have been explored from an information extraction context in the cancer research. From rule-based (97-100) to machine learning (101-103) to deep learning (104-107) approaches, NLP has proven to be invaluable for pathology report classification. There is a lack of NLP studies focusing on oral cancer and very few on head and neck cancers, providing an opportunity to explore NLP methods for oral cancer.

3.3. **Methods**

3.3.1. Study Setting

The study was conducted at the UNC Adams School of Dentistry (ASOD) using dental electronic medical records (dEMR) at the Oral and Maxillofacial Pathology Lab (OMPL) at UNC ASOD. UNC contains the largest OMPL in NC, and the records used in this study were pathology notes written for biopsy samples collected from 2005 to 2020. Data from the OMPL required approval from the internal Institutional Review Board (IRB) in ASOD and was extracted by ASOD staff.

3.3.2. Subjects

All patients from the UNC ASOD OMPL in the date range, January 7, 2005 to January 27, 2020, were included in the starting cohort. Cohort stratification was done in the data processing phase.

3.3.3. Study Design

A computational phenotyping algorithm was created to thoroughly identify patient populations for drug repurposing validation studies using dEMR pathology notes. The goal was to computationally phenotype oral cancer patients using oral pathology notes. EMR

phenotyping was divided into two tasks: (1) extracting patients with cancer mentions in their records and (2) case/control prediction. Since the goal was to identify patient populations for drug repurposing validation, medications were not used as a source of information influencing identification. Only the pathology notes were used for computational phenotyping. The case/control prediction was conducted using a set of records that was labeled by a dentist. The expert labels were used to train the predictive models. Records that were labeled as possible cases were set aside for evaluation.

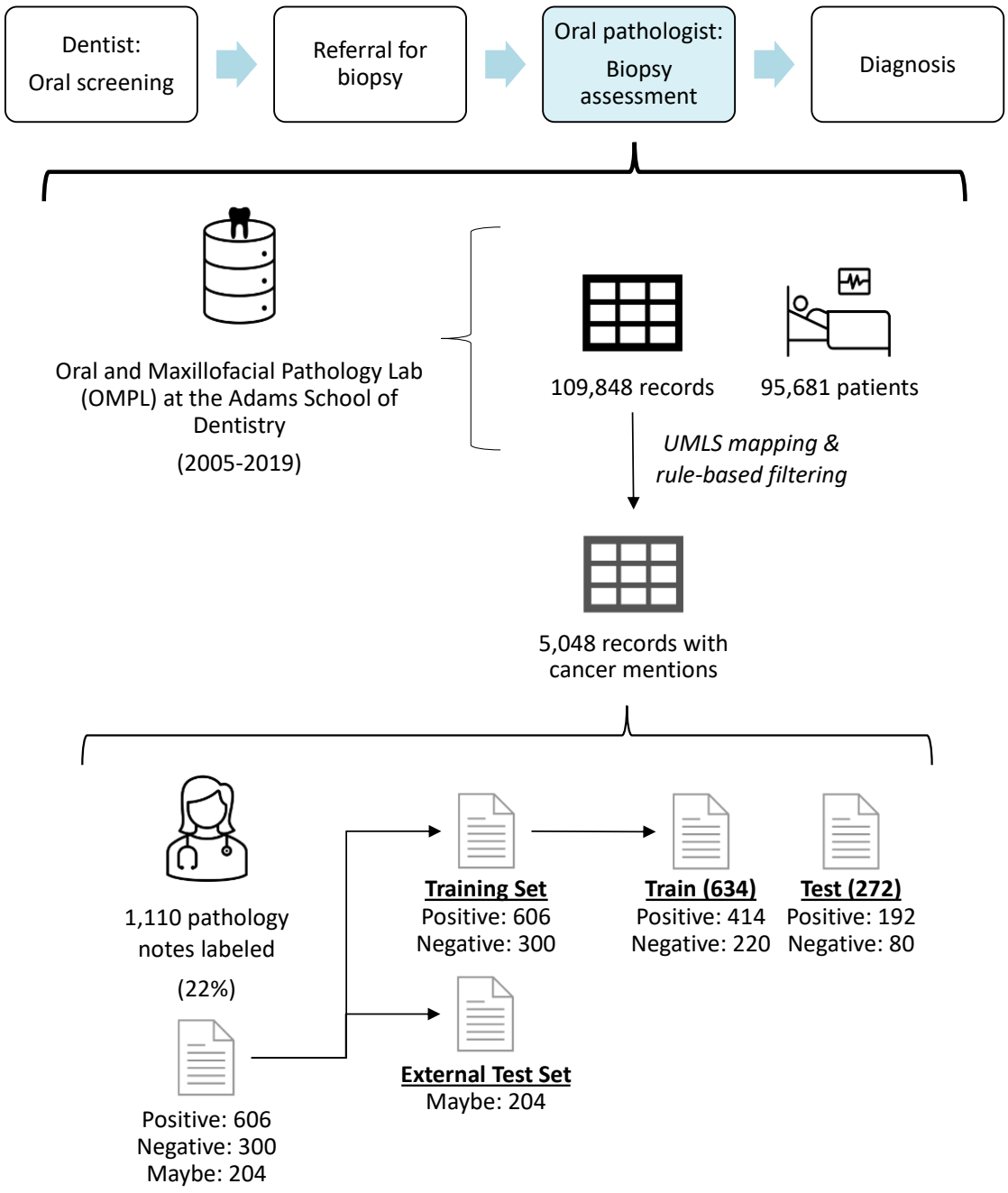


Figure 6. Oral cancer phenotyping data collection and preparation

3.3.4. Data preprocessing

All records from the OMPL were taken (109,848 records from 95,681 patients) as an initial cohort. Patients had at least one recorded visit and at most, five recorded visits. Each sample was treated as an independent encounter. The records consisted of the following

variables: patient ID, age, procedure date, clinical impression, microscopic description, and comments. Figure 6 shows the typical clinical workflow as well as how data was collected and prepared for analysis. A clinical impression is written by a clinician before a pathologist examines the sample and writes a microscopic description, where a clinician will indicate if oral cancer should be ruled out. The dataset contained biopsies for all types of oral conditions with the most common being fibromas, dentigerous cysts, papillomas, and mucoceles. Therefore, every record without a cancer mention in the clinical impression was filtered out. The overall set of records was narrowed down to records with cancer mentions (5,048 records) using char-3gram Unified Medical Language System (UMLS) Metathesaurus matching from the scispaCy package and rule-based filtering. For example, for a clinical impression stating, “rule out squamous cell carcinoma”, “squamous cell carcinoma” was identified with UMLS as cancer, and rules were used to include the case in the dataset. The pathology notes referenced in Figure 6 are the microscopic descriptions of the oral biopsy samples.

A dentist labeled 1,110 (22%) records as being cancerous (i.e., positive), non-cancerous (i.e., negative) or showing pre-cancerous symptoms (i.e., maybe). The data was separated into a dataset of positive/negative cases and an external test set of maybe cases. The motivation behind the dataset separation was that models should be trained on cases that show definite signs of having cancer or not having cancer. In this study, only the positive and negative cases were used to build and evaluate predictive models. From the dataset of positive and negative cases (906), the data were divided with a 70/30 split, and the training set was comprised of 414 positive cases and 220 negative cases, while the testing set was comprised of 192 positive cases

and 80 negative cases. In this binary predictive task, the microscopic descriptions were used to predict the clinician labels; no other variables were used in the classification.

3.3.5. Methods

3.3.5.1. Data transformation

Two transformation approaches were taken: a scispaCy (108) pipeline with term frequency-inverse document frequency (TF-IDF) and using BERT-based models like SciBERT (109), PubMedBERT (110), and BioBERT (111). SciBERT is a pretrained language model based on Bidirectional Encoder Representations from Transformers (BERT) (112) trained on biomedical and scientific full-text articles from Semantic Scholar. PubMedBERT and BioBERT are pretrained language model based on BERT that were trained on biomedical literature from PubMed. PubMedBERT was trained on article abstracts, and BioBERT was trained on abstracts and full-text articles. BERT is a deep neural network that uses transformer architecture to learn text embeddings. It functions by reading a sequence of words at once rather than from left to right or right to left and learning the contextual relationships in the sequence. However, the original BERT model was trained on general-purpose text, making SciBERT a better candidate for this work. SciBERT was implemented using the PyTorch AllenNLP “scibert-scivocab-uncased” model. PubMedBERT and BioBERT were both implemented from models posted on HuggingFace (113) with “microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract” and “dmis-lab/biobert-base-cased-v1.1” models. SpaCy (114) is an open-source library for natural language processing, where foundational techniques such as part-of-speech tagging can be included as pipes in pipelines. ScispaCy contains custom pipes and models for scientific document classification. Its named entity recognition and entity span detection models, syntactic parser, and part-of-

speech tagger have been trained on biomedical tasks, making it useful for predictive tasks in the biomedical domain, as shown in Figure 7.

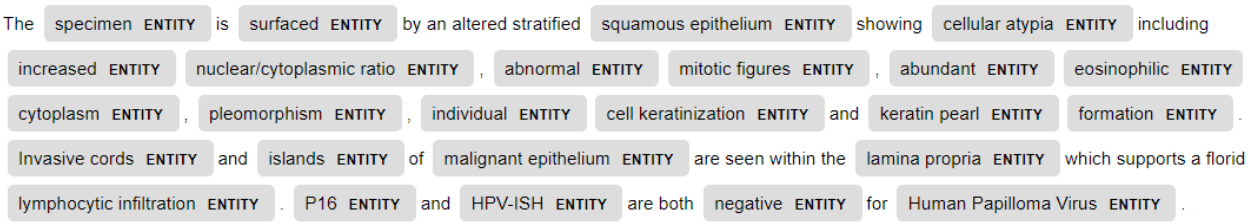


Figure 7. Named entities identified in a microscopic description of an oral pathology sample.

The scispaCy pipeline used consisted of using part-of-speech tags, lemmatization, dependency parsing, and named entity recognition. The scispaCy data preprocessing pipeline took a list of texts as an input and outputted the texts in a format suitable for analysis. The tagger placed part-of-speech tags (e.g., noun, adjective) on each word in the microscopic description. The lemmatizer took each word and reverted it to its base form (e.g., was/be, increased/increase). The parser used dependency parsing to analyze the grammatical structure of a sentence and see how each word depended on others in the sentence. The named entity recognition component took entities from the text and categorized them. For example, “human papilloma virus” would be identified as a named entity (Figure 7). The pipeline output was then passed to a term frequency-inverse document frequency (TF-IDF) vectorizer from scikit-learn (115). TF-IDF is a statistical score that weights each word in a document in a corpus. Term frequency measures the frequency of a word in a document. Document frequency measures the frequency of a word in a corpus of documents, while inverse document frequency measures how informative a term is.

```
('tok2vec', <spacy.pipeline.tok2vec.Tok2Vec at 0x22709cbb8b0>),
('tagger', <spacy.pipeline.tagger.Tagger at 0x22709ff2cc0>),
('attribute_ruler',
<spacy.pipeline.attributeruler.AttributeRuler at 0x2270a028400>),
('lemmatizer', <spacy.lang.en.lemmatizer.EnglishLemmatizer at
0x2270a025540>),
('parser', <spacy.pipeline.dep_parser.DependencyParser at 0x22709c43880>),
('ner', <spacy.pipeline.ner.EntityRecognizer at 0x22709c43be0>),
('abbreviation_detector',
<scispacy.abbreviation.AbbreviationDetector at 0x2276c467820>),
('negex', <negspacy.negation.Negex at 0x2276c467a30>)
```

Figure 8. spaCy natural language processing pipeline used for oral cancer classification.

3.3.5.2. Classification

Three supervised machine learning classifiers were used for prediction: random forest, support vector machine (SVM), and logistic regression. All models were implemented using scikit-learn (28). 10-fold cross-validation was used for training and hyperparameter tuning. The tuned models were then tested with the held-out test set. The rationale behind using both cross-validation and a test set was to present performance on unseen data after tuning. The random forest model had 100 estimators. SVM was implemented with a radial basis function kernel and scaled gamma value. Logistic regression was used with a stochastic average gradient solver, L2 penalty, and the maximum number of iterations was set to 50,000. The models were evaluated with the scispaCy tokenizer with TF-IDF and with SciBERT.

3.4. **Results**

The results between models and pipelines were compared using three metrics: precision (i.e., positive predictive value (PPV)), recall (i.e., sensitivity), and area under the receiver operating characteristic curve (ROC-AUC) using 10-fold cross validation (Tables 1, Figure 3) and on the test set (Table 2, Figure 4). From 10-fold cross validation, using the scispaCy tokenizer with TF-IDF, logistic regression had the best overall performance in terms of precision, recall,

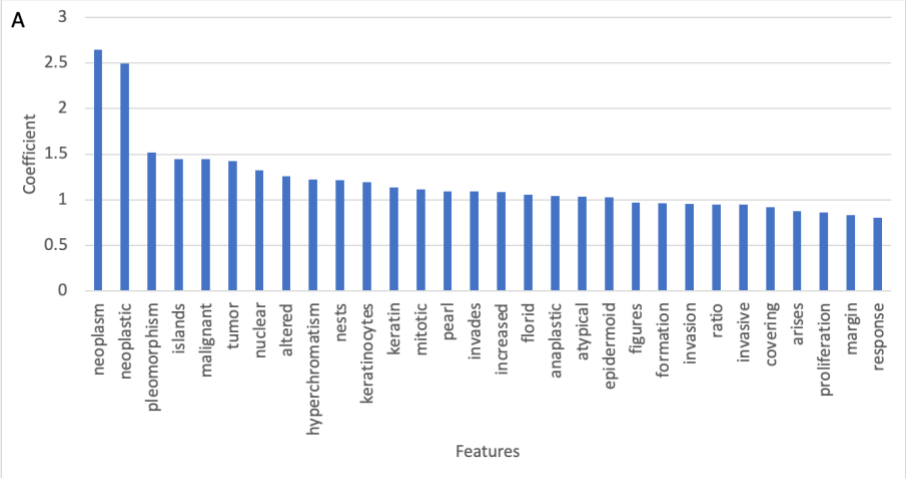
and ROC-AUC score (Table 1). However, in terms of precision, PubMedBERT with SVM performed as well as the scispaCy tokenizer. Out of the embedding models, PubMedBERT with SVM also achieved the highest ROC-AUC score (96.2%), and SciBERT with random forest achieved the highest recall (96.5%). The differences in performance between classifiers using scispaCy are more minute than the differences between classifiers using BERT-based models.

Table 3. Oral cancer classification results from 10-fold cross validation: precision, recall, and ROC-AUC score.

ScispaCy Tokenizer with TF-IDF			
Model	**Precision**	**Recall**	**ROC-AUC Score**
Random Forest	93.5%	97.4%	98.2%
SVM	93.7%	97.6%	98.4%
Logistic Regression	**93.9%**	**98.1%**	**98.9%**
SciBERT			
Model	**Precision**	**Recall**	**ROC-AUC Score**
Random Forest	87.2%	**96.5%**	95.2%
SVM	91.0%	92.4%	**95.5%**
Logistic Regression	**92.5%**	90.6%	95.4%
BioBERT			
Model	**Precision**	**Recall**	**ROC-AUC Score**
Random Forest	88.8%	**91.5%**	**94.5%**
SVM	89.3%	90.8%	93.2%
Logistic Regression	**90.6%**	91.0%	94.4%
PubMedBERT			
Model	**Precision**	**Recall**	**ROC-AUC Score**
Random Forest	90.4%	**93.6%**	95.8%

SVM	**93.7%**	88.5%	95.8%
Logistic Regression	92.1%	93.1%	**96.2%**

Logistic regression was the highest performing classifier in cross-validation, so feature coefficient values were examined to understand which features were most important. Figure 9 shows the top 30 weighted features, sorted by absolute value. From the negatively weighted features, interesting features include “acanthosis” (enlargement of the spinous layer of the skin) and “hyperparakeratotic” (outer keratin layer thickening), which are non-cancerous changes. From the positively weighted features, many terms are cancer-related like “neoplasm”, “proliferation”, “malignant”, and “tumor”. There are also less obvious terms like “pleomorphism”, “anaplastic”, and “invades”. The feature importances give insight into the relationships between the input features and the output variable, showing that the model is reasonable and performed as expected.



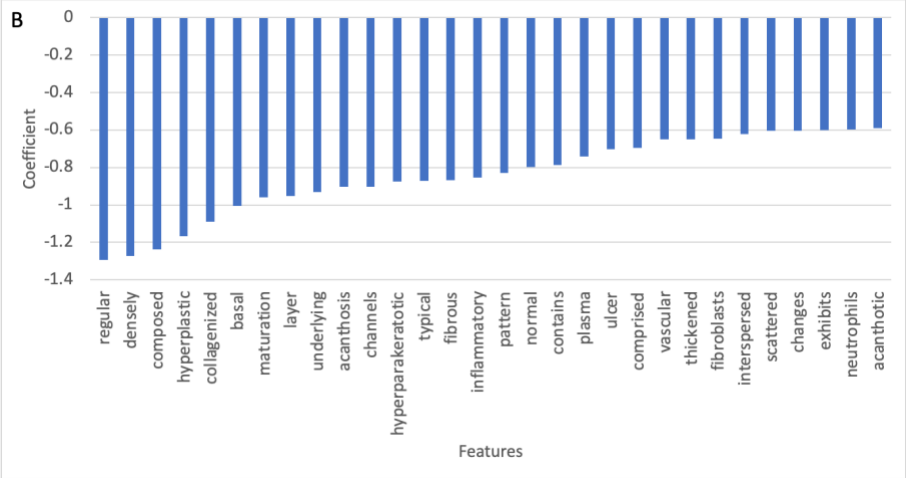


Figure 9. Oral cancer cross-validation results: most important features from logistic regression (A) top 30 positive features (B) top 30 negative features.

On the test set, using the scispaCy tokenizer with TF-IDF, SVM had the highest performance across precision, recall, and ROC-AUC score with minute differences in comparison to those of other classifiers (Table 2). SVM achieved the highest ROC-AUC score (98.3%) but had similar performance to logistic regression (98.1%). SVM also had similar precision (94.9%) to that of logistic regression (94.7%). Out of the BERT-based models, BioBERT with SVM achieved the highest ROC-AUC score (95.2%), and BioBERT with random forest achieved the highest recall (88.6%). BioBERT and PubMedBERT both had slightly better performance than SciBERT. Overall, the scispaCy tokenizer with TF-IDF outperformed the BERT-based models across all classifiers on the test set.

Table 4. Oral cancer classification results on the test set: precision, recall, ROC-AUC score.

ScispaCy Tokenizer with TF-IDF			
Model	**Precision**	**Recall**	**ROC-AUC Score**
Random Forest	94.4%	94.1%	97.4%
SVM	**94.9%**	**94.9%**	**98.3%**

Logistic Regression	94.7%	94.5%	98.1%
SciBERT			
Model	**Precision**	**Recall**	**ROC-AUC Score**
Random Forest	86.4%	86.0%	**93.7%**
SVM	87.7%	**87.9%**	93.5%
Logistic Regression	**87.8%**	**87.9%**	93.2%
BioBERT			
Model	**Precision**	**Recall**	**ROC-AUC Score**
Random Forest	**91.2%**	**88.6%**	95.8%
SVM	86.5%	85.9%	**95.2%**
Logistic Regression	87.1%	87.3%	95.6%
PubMedBERT			
Model	**Precision**	**Recall**	**ROC-AUC Score**
Random Forest	**89.2%**	**87.0%**	**93.9%**
SVM	86.1%	85.1%	92.1%
Logistic Regression	87.9%	86.4%	92.8%

3.5. **Discussion**

Early diagnosis of oral cancer is critical to improving morbidity and mortality rates, and machine learning models are increasingly being adopted in oral oncology for risk stratification. However, it is in nascent stages in comparison to efforts focusing on other cancers such as breast and prostate cancers (116). In this study, we developed a natural language processing classification pipeline to assess oral cancer risk from dental pathology notes to assist clinicians in deciding whether to refer patients to an oncologist. Clinicians are given vast amounts of information from which they must make decisions, indicating a need for clinical decision support tools. The end goal of this work is to provide clinicians with a probability of a sample being cancerous to help them in their decision-making.

We found that oral cancer risk can be predicted with reasonable success from microscopic descriptions in dental pathology notes. We were unable to find literature using dental pathology notes for cancer diagnosis models but can compare results to recent work on pathology report classification supporting national cancer surveillance. Qiu et al. (2017)(117) focused on classifying breast and lung ICD-O-3 codes from pathology reports and used word vector representations as input for a convolutional neural network, which outperformed using TF-IDF in classification. The underlying conclusion of the work done by Qiu et al. (117) was that using representation learning for classification may be more adaptable and have higher performance than using traditional approaches like TF-IDF. However, their study was focused on a 6-class and 12-class classification problem over various datasets from different institutions, while our work focused on a binary classification on a dataset from a single institution, lending to why using embedding models to represent the feature space could have enabled a higher performing model. Similarly, Alawad et al. (2020)(104) used word embeddings concatenated with CUIs from UMLS in a CNN for subsite and histology classification with a vast number of labels, where their feature combination outperformed standard models. Like with the work done by Qiu et al. (117), the classification task differences make it difficult to compare results to our study. There are also few studies using BERT or pre-trained BERT models to learn text representations for pathology tasks, so this work will be added to a growing body of literature in this space (118, 119).

In this study, our comparison between embedding models and traditional models, showed that traditional models can outperform embedding models for a specific task. Additionally, the prior related work was not geared toward building decision support tools

targeted at clinicians, and for clinical decision support tools, the final model decision may not be the most important model output in comparison to seeing the journey the model took to produce an output. Our use of traditional models enables explainability and transparency in model decisions, allowing clinicians to understand why a determination was made. While the end goal of this work is to produce a probability of a sample being cancerous or not to a clinician, future directions could also include creating a loop where significant terms or phrases are highlighted for a clinician to see along with a score and agree or disagree with the model prediction.

3.5.1. Limitations

All data in this work comes from a single institution and its oral pathologists who have received similar training and write pathology reports similarly, so we could not evaluate generalizability. Also, only one clinician annotated the data. In next steps, we plan to have one more clinician annotate the data and calculate inter-rater reliability. In addition, there may be selectivity bias in filtering for clinical impressions with cancer mentions, but we relied on clinician expertise from the clinical impression, assuming the clinician has reasonable belief that a lesion will warrant a biopsy as it is an invasive procedure that could be stressful for patients. There is scope to also include cases with other oral conditions that could develop into oral cancer (e.g., lichen planus) written in the clinical impression, which will be explored in future work.

REFERENCES

85. Ren Z-H, Hu C-Y, He H-R, Li Y-J, Lyu J. Global and regional burdens of oral cancer from 1990 to 2017: Results from the global burden of disease study. Cancer Commun (Lond). 2020;40(2-3):81-92.

86. Abati S, Bramati C, Bondi S, Lissoni A, Trimarchi M. Oral Cancer and Precancer: A Narrative Review on the Relevance of Early Diagnosis. International Journal of Environmental Research and Public Health. 2020;17(24):9160.

87. Kaur J, Jacobs R, Huang Y, Salvo N, Politis C. Salivary biomarkers for oral cancer and pre-cancer screening: a review. Clinical Oral Investigations. 2018;22(2):633-40.

88. Gherlone EF, Capparé P, Tecco S, Polizzi E, Pantaleo G, Gastaldi G, et al. A Prospective Longitudinal Study on Implant Prosthetic Rehabilitation in Controlled HIV-Positive Patients with 1-Year Follow-Up: The Role of CD4+ Level, Smoking Habits, and Oral Hygiene. Clinical Implant Dentistry and Related Research. 2016;18(5):955-64.

89. Li W-C, Huang C-H, Hsieh Y-T, Chen T-Y, Cheng L-H, Chen C-Y, et al. Regulatory Role of Hexokinase 2 in Modulating Head and Neck Tumorigenesis. Front Oncol. 2020;10.

90. Karadaghy OA, Shew M, New J, Bur AM. Development and assessment of a machine learning model to help predict survival among patients with oral squamous cell carcinoma. JAMA Otolaryngology–Head & Neck Surgery. 2019;145(12):1115-20.

91. Bur AM, Holcomb A, Goodwin S, Woodroof J, Karadaghy O, Shnayder Y, et al. Machine learning to predict occult nodal metastasis in early oral squamous cell carcinoma. Oral oncology. 2019;92:20-5.

92. Kim DW, Lee S, Kwon S, Nam W, Cha I-H, Kim HJ. Deep learning-based survival prediction of oral cancer patients. Sci Rep. 2019;9(1):6994.

93. Liu Y, Li Y, Fu Y, Liu T, Liu X, Zhang X, et al. Quantitative prediction of oral cancer risk in patients with oral leukoplakia. Oncotarget. 2017;8(28):46057-64.

94. Jeyaraj PR, Samuel Nadar ER. Computer-assisted medical image classification for early diagnosis of oral cancer employing deep learning algorithm. Journal of cancer research and clinical oncology. 2019;145(4):829-37.

95. Lu C, Lewis JS, Dupont WD, Plummer WD, Janowczyk A, Madabhushi A. An oral cavity squamous cell carcinoma quantitative histomorphometric-based image classifier of nuclear morphology can risk stratify patients for disease-specific survival. Modern Pathology. 2017;30(12):1655-65.

96. Shaban M, Khurram SA, Fraz MM, Alsubaie N, Masood I, Mushtaq S, et al. A novel digital score for abundance of tumour infiltrating lymphocytes predicts disease free survival in oral squamous cell carcinoma. Sci Rep. 2019;9(1):1-13.

97. Glaser AP, Jordan BJ, Cohen J, Desai A, Silberman P, Meeks JJ. Automated Extraction of Grade, Stage, and Quality Information From Transurethral Resection of Bladder Tumor Pathology Reports Using Natural Language Processing. JCO Clin Cancer Inform. 2018;2:1-8.

98. Soysal E, Warner JL, Wang J, Jiang M, Harvey K, Jain SK, et al. Developing Customizable Cancer Information Extraction Modules for Pathology Reports Using CLAMP. Stud Health Technol Inform. 2019;264:1041-5.

99. Nguyen AN, Moore J, O'Dwyer J, Philpot S. Automated Cancer Registry Notifications: Validation of a Medical Text Analytics System for Identifying Patients with Cancer from a State-Wide Pathology Repository. AMIA Annu Symp Proc. 2016;2016:964-73.

100. Lee J, Song H-J, Yoon E, Park S-B, Park S-H, Seo J-W, et al. Automated extraction of Biomarker information from pathology reports. BMC Med Inform Decis Mak. 2018;18(1):29.

101. Oleynik M, Patrão DF, Finger M. Automated classification of semi-structured pathology reports into ICD-O using SVM in Portuguese. Stud Health Technol Inform. 2017;235:256-60.

102. Yala A, Barzilay R, Salama L, Griffin M, Sollender G, Bardia A, et al. Using machine learning to parse breast pathology reports. Breast Cancer Res Treat. 2017;161(2):203-11.

103. Napolitano G, Marshall A, Hamilton P, Gavin AT. Machine learning classification of surgical pathology reports and chunk recognition for information extraction noise reduction. Artif Intell Med. 2016;70:77-83.

104. Alawad M, Gao S, Alamudun FT, Wu X-C, Durbin EB, Doherty J, et al. Multimodal Data Representation with Deep Learning for Extracting Cancer Characteristics from Clinical Text. Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States); 2020.

105. Yoon H, Gounley J, Young MT, Tourassi G, editors. Information Extraction from Cancer Pathology Reports with Graph Convolution Networks for Natural Language Texts. 2019 IEEE International Conference on Big Data (Big Data); 2019 9-12 Dec. 2019.

106. Saib W, Chiwewe T, Singh E. Hierarchical Deep Learning Classification of Unstructured Pathology Reports to Automate ICD-O Morphology Grading. arXiv preprint arXiv:200900542. 2020.

107. Gao S, Alawad M, Schaefferkoetter N, Penberthy L, Wu X-C, Durbin EB, et al. Using case-level context to classify cancer pathology reports. PLoS ONE. 2020;15(5):e0232840.

108. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: fast and robust models for biomedical natural language processing. arXiv preprint arXiv:190207669. 2019.

109. Beltagy I, Lo K, Cohan A. SciBERT: A pretrained language model for scientific text. arXiv preprint arXiv:190310676. 2019.

110. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH). 2021;3(1):1-23.

111. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2019;36(4):1234-40.

112. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805. 2018.

113. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:191003771. 2019.

114. Honnibal M, Johnson M, editors. An improved non-monotonic transition system for dependency parsing. Proceedings of the 2015 conference on empirical methods in natural language processing; 2015.

115. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011;12:2825-30.

116. Sultan AS, Elgharib MA, Tavares T, Jessri M, Basile JR. The use of artificial intelligence, machine learning and deep learning in oncologic histopathology. Journal of Oral Pathology & Medicine. 2020;49(9):849-56.

117. Qiu JX, Yoon H-J, Fearn PA, Tourassi GD. Deep learning for automated extraction of primary sites from cancer pathology reports. IEEE J Biomed Health Inform. 2017;22(1):244-51.

118. Zhou S, Wang N, Wang L, Liu H, Zhang R. CancerBERT: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. J Am Med Inform Assoc. 2022.

119. Santos T, Tariq A, Gichoya JW, Trivedi H, Banerjee I. Automatic Classification of Cancer Pathology Reports: A Systematic Review. Journal of Pathology Informatics. 2022;13:100003.

4. COMPUTATIONAL PHENOTYPING FOR BREAST CANCER USING ELECTRONIC MEDICAL RECORDS

4.1. **Introduction**

Among women, breast cancer accounts for 30% of all cancer cases. Take a female breast cancer patient in her mid 20's, Jane Doe, who has a family history of breast cancer and currently has recurrent breast cancer. Jane has a history of left breast ductal carcinoma in situ, which upstaged to a recurrent infiltrating ductal carcinoma. The carcinoma is grade 2, estrogen receptor (ER) positive, progesterone receptor (PR) positive, human epidermal growth factor receptor 2 (HER2) negative with an immunohistochemistry (IHC) score of 1+, and measures 1 cm in greatest dimension. She got a mastectomy done, and then she got a sentinel lymph node biopsy to understand if the cancer had spread into her lymphatic system. She chose to take adjuvant medication therapy with no adjuvant radiation therapy afterwards, but she later developed a chest wall recurrence. In her next stage of treatment, she will take adjuvant radiation therapy.

Jane and other women's cancer care can be broken down into the following key categories: risk assessment (e.g., family history), primary intervention (e.g., lifestyle counseling), detection (e.g., mammogram), diagnosis (e.g., biopsy), cancer or precursor treatment (e.g., surgery, radiation therapy, chemotherapy, adjuvant therapy), post-treatment survivorship (e.g., recurrence, surveillance), and end-of-life-care (e.g., palliative care) (120). In every stage of her care, there are copious amounts of data generated.

An overall data profile for her would include information on demographics, encounter details, diagnoses, procedures, vital signs, lab results, co-morbidities, images, prescriptions, medications administered, addresses, and if applicable, death and cause of death. Additionally, there are other types of data generated in her care process like clinical notes, MyChart messages, and referrals.

Risk factors for breast cancer can be broken into intrinsic (e.g., age, race, family history) and extrinsic groups (e.g., diet, lifestyle) (121). In EMR, intrinsic factors like demographics are captured very well; however, family history is in the clinical free-text with the structure of descriptions varying by the note. Age is a key risk factor because breast cancer is most commonly found in women around menopause (121). Extrinsic factors like diet and lifestyle are also only captured in the clinical free-text, depending on the clinician. Cohort stratification can be defined as selecting a set of patients and dividing patients into groups of patients like them.

Treatment for breast cancer is largely determined by the state of disease progression, cancer stage, molecular subtype, patient age, and patient preferences. Patients with newly diagnosed breast cancer can be divided by whether they have noninvasive (which has not spread to surrounding breast tissue) or invasive (in surrounding breast tissue or metastatic) breast cancer. Breast cancer staging is characterized by the Tumor, Node, Metastasis (TNM) system. For breast cancer staging, clinicians may use measures such as blood tests to measure protein markers and circulating tumor cells, or imaging such as mammograms, a magnetic resonance imaging (MRI), bone scans, computerized tomography (CT) scans, or positron emission tomography (PET) scans. These tests and scans are also used for monitoring response to therapy. Tumor markers are defined as biomarkers found in blood, urine, or body tissues

that can be elevated due to the presence of cancer cells or the body's response to cancer cells (122). Patients are treated after determining the molecular subtype of their carcinoma, which is based on the gene expression of cancer cells. The two factors associated with subtype classification are hormone receptors (HR) and HER2. Patients will either be positive (+) or negative (-) for having either HR or HER2 or both. There are four female breast cancer subtypes: luminal A (HR+/HER2-), luminal B (HR+/HER2+ or HR+/HER2-), HER2-enriched (HR-/HER2+), and triple negative/basal-like (HR-/HER2-). Estrogen and progesterone receptors are hormone receptors, and if a patient is either ER+ or PR+, the patient would be considered hormone receptor positive. HER2 status is determined by a HER2 expression score by immunohistochemistry in the following range: 0, 1+, 2+, 2+ with FISH amplified (2+FISH+), and 3+. According to the American Society of Clinical Oncology (ASCO) guidelines, patients with a score of 0, 1+, or 2+ are considered HER2 negative, and patients with a score of 2+FISH+ or 3+ are considered HER2 positive. Patients who have a score of 1+ or 2+ are also considered HER2 low in recent years due to advancements in treatments for HER2 low patients (123). Molecular subtype is assessed in breast cancer patients to select patients for targeted therapies (124).

The primary output of this work is a stratified cohort of individuals to later examine treatment effects in. Computational phenotyping is needed to accomplish this goal because the variables necessary to stratify the cohort like patient outcomes and cancer-related covariates are not clearly illustrated in the structured EMR. This work will be divided into two natural language processing tasks: predicting breast cancer recurrence and extracting tumor receptor status from the clinical notes. To measure disease improvement in cancer patients, survival-based outcomes like overall survival, progression-free survival (for metastatic cancer patients),

and recurrence-free survival (for nonmetastatic cancer patients). Recurrence-free survival of early-stage breast cancer patients was selected as the primary outcome in this work. Patient overall survival is included in the structured data, but recurrence-free survival and progression-free survival are not. The first task in this chapter is to detect whether a patient has breast cancer recurrence or not with a hybrid approach: (a) predict whether a patient has breast cancer recurrence from their clinical notes (b) detect breast cancer recurrence based on ICD-CM diagnosis codes. For cohort stratification, knowing the tumor receptor status of a patient's breast cancer is important to understand treatment pathways; however, tumor receptor status is not included as a structured variable in EMR data at UNC. Therefore, the second task for this work is to extract tumor receptor statuses for breast cancer patients from their clinical notes.

4.2. **Prior Work**

4.2.1. Breast cancer recurrence detection from EMR

The literature landscape on breast cancer recurrence prediction can be summarized in three categories: using administrative and structured EMR data to predict recurrence, using natural language processing to extract or predict recurrence from the EMR clinical notes, or using a combination of structured and unstructured data for recurrence prediction. Using structured data for recurrence prediction has many benefits. Models developed with structured data in common data models can more easily be transported to different settings in comparison to models developed with unstructured text. The strength of building models on clinical notes is that the clinical notes are incredibly rich and may contain useful information that is not in the structured data. For example, if a patient received a biopsy from a clinic other than her current hospital that she is receiving treatment from, the biopsy will not be logged in

her current hospital's system. However, the patient will have spoken to her clinician about it, and the clinician will likely mention that in her clinical notes. Select prior work study details are shown in

Table 5.

Several studies have used administrative health or structured EMR data to identify breast cancer recurrence (125-128), and two examples are discussed. Ritzwoller et al (2018) (125) developed a model to detect breast cancer recurrence and a model to estimate the recurrence timing. They used data from a common data model that included both health insurance claims data and EHR data. The gold standard data was compiled by tumor registrars as part of routine cancer surveillance, and they did not evaluate inter-rater reliability. They separated the data into a training set and two validation sets from three different sites, where data from the two sites were used for training and the first validation set, while the second validation set was held-out and only included data from the third site. The authors originally develop the RECUR algorithm in a previous study on detecting lung and colorectal cancer recurrence and adapted it to detect breast cancer recurrence (129). They used a multivariate logistic regression model with LASSO regularization. The cohort was limited to patients with stage I-III breast cancer who were 21 years of age or older with no previous cancer, who had completed local-regional therapy, had survived, and were followed for at least 30 days post-therapy. They were unable to account for patients have multiple recurrent events. The authors also paid careful attention to whether an event was a recurrence or a second primary. On the first validation set, they reported an AUC of 95.6%, sensitivity (i.e., recall) of 79.8%, and positive predictive value (PPV) (i.e., precision) of 68.0%. On the held-out validation set, they reported an AUC of 90.0%, sensitivity of 80.0%, and PPV of 65.6% (125).

Another study by Lambert et al (2021) (126) used administrative and structured data from a Canadian province with a universal health system to detect breast and colorectal cancer

recurrence. The authors separated their data based on time, where the training set was comprised of individuals diagnosed between 2004 and 2007, and the validation set included patients diagnosed between 2008 and 2012. They narrowed their cohort down to patients with stage I-III breast cancers that were ER-, PR- or HER2+. The authors decided to limit based on molecular subtype to manually review less records because the selected cancers reportedly have higher recurrence rates. Manual chart review was done by research assistants, and they evaluated inter-rater reliability. They developed four models, which were unweighted and weighted pre-specified variable and conditional inference tree algorithms. The pre-specified variable approach was defined as using variables with clinically meaningful cut-offs, and if a patient record was positive for a single variable according to the cut-off, the patient was marked as having recurrence. Conditional inference trees are a non-parametric class of decision trees which use a statistical measure to select variables rather than an information measure (e.g., information gain) (130). The unweighted pre-specified variable approach had better performance on the validation set than the weighted one with a sensitivity of 83.2% and PPV of 64.7%. The unweighted conditional inference tree algorithm had a sensitivity and PPV of 73.7%, and its weighted counterpart had a sensitivity of 68.5% and PPV of 75.4%.

Many studies have described using natural language processing techniques to detect breast cancer (131-135), and four are discussed. Carrell et al (2014)(131) developed an approach to identify breast cancer recurrence and estimated diagnosis dates within 30 days from a cohort of patients with early-stage incident breast cancer. They defined breast cancer recurrence as ipsilateral, regional, or metastatic diagnoses made during a follow-up period at least 120 days after the primary cancer diagnosis was made. Second primaries were defined as

recurrences in the contralateral breast and were not considered recurrences. The study created a custom dictionary for terms relevant to recurrent breast cancer diagnosis and used the clinical Text Analysis and Knowledge Extraction System (cTAKES), which uses rule-based and machine learning techniques, for recurrence detection. The system was tested on a manually reviewed held-out test set from a previous study and had a precision of 59%, recall of 92%, and F1-Score of 72%.

Zeng et al (2018)(132) aimed to detect local breast cancer recurrence from progress notes. They had an extensive manual review process starting with reviewing 50 breast cancer patient progress notes and extracting partial sentences indicating local recurrence. They processed the partial sentences with MetaMap to obtain UMLS concepts, which comprise what they defined as the positive concept set. On the rest of the progress notes, they used MetaMap to obtain UMLS concepts and only kept the concepts that fell in the positive concept set. Two authors (post-doc fellow and PhD student) divided and annotated 6,899 patient notes to identify local recurrences and found 569 (8.25%) local recurrences. Then, they selected a random sample of 201 notes indicating recurrence and 500 notes without recurrence, and two more authors (medical student and breast surgery fellow) annotated the random sample. The inter-rater reliability for the double annotation with Cohen's kappa score was 0.92. The 701 double-annotated notes were split into training and held-out test sets. Along with the note features, they also used the number of pathology notes a patient had as a feature in the model. Their model pipeline used chi-square feature selection, TF-IDF vectorization, and then grid search cross validation to train and tune hyperparameters for a support vector machine model. The authors reported performance with four feature variations: filtered MetaMap concepts

with pathology report counts, all MetaMap concepts, filtered MetaMap concepts only, and bag of words. Using filtered MetaMap concepts with pathology report counts had the highest performance with a precision of 0.74, recall of 0.84, F1-Score of 0.79, and AUC of 0.87.

Wang et al (2020)(133) focused on predicting distant breast cancer recurrence with a knowledge-guided convolutional neural network. The work came out of Yuan Luo's group at Northwestern University, which also published Zeng et al (2018)(132). The authors used the same annotations from their previous work but with different feature representations. They used bag of words, bag of UMLS concepts unique identifiers (CUIs), structured EMR data, word embeddings trained from MIMIC-III clinical notes, and CUI embeddings obtained using cui2vec. Performance was first evaluated on an array of machine learning classifiers with balanced class weighting to mitigate the class imbalance using the bag of words, structured EMR features, and CUIs. They developed a standard K-CNN configuration and implemented it with different feature configurations. The best model produced used word embeddings and structured EMR features, which resulted in a precision of 53.70%, recall of 46.80%, F1-Score of 50.00%, and ROC-AUC score of 88.80%.

Ling et al (2019)(134) define metastatic breast cancer recurrence as neoplasms that spread to other parts of the body after the initial breast cancer diagnosis, as opposed to Stage IV metastatic cancer that is diagnosed in the initial diagnosis. The primary goal of the study was to automate metastatic recurrence detection on a population-level, so the authors developed a distant supervision framework to avoid manually reviewing a large number of cases. For evaluation, two medical oncologists manually reviewed 146 female breast cancer patients' records, labeling the patients as having metastatic recurrence or not. They decided the

evaluation set size based on a previous study on validation for EMR-based phenotyping algorithms from the eMERGE network that compiled evaluation sets with 50-200 subjects (136). The distant framework had two steps: using the Clinical Event Recognizer (CLEVER)(137) to extract metastatic information from clinical notes to create distant labels and classifying patients as having a metastatic recurrence or not with logistic regression with L2 regularization. The study showed performance with three feature variations: structured cancer registry data only (e.g., age, race, tumor receptor status), NLP features only, a combination of structured cancer registry and NLP features. The model using a combination of features had the highest AUC of 92.5%, and it also had a sensitivity of 86.1%, PPV of 87.3%, and F1-Score of 86.7% on the evaluation set.

Table 5. Breast cancer recurrence detection from EMR prior work details

Study	Recurrence Detection Task	Type of data	Model	Precision	Recall	F1-Score	ROC-AUC	Number of Reviewers	Inter-rater Reliability (Kappa statistic)
Ritzwoller at al (2018)	Breast cancer recurrence	Administrative and structured EMR data	Logistic regression with LASSO regularization (test set)	68.00%	79.80%	-	95.60%	-	-
			Logistic regression with LASSO regularization (external site dataset)	65.60%	80.00%	-	90.00%	-	-
Lambert et al (2021)	Breast and colorectal cancer recurrence	Administrative and structured EMR data	Unweighted pre-defined variable algorithm	64.70%	83.20%	-	-	2	0.81
			Weighted pre-defined variable algorithm	61.40%	81.10%	-	-		
			Unweighted conditional inference tree	73.70%	73.70%	-	-		

			Weighted conditional inference tree	75.40%	68.50%	-	-		
Zeng et al (2018)	Breast cancer local recurrence	Progress notes	Support vector machine with filtered MetaMap concepts and pathology report counts	74.00%	84.00%	79.00%	87.00%	4	0.92 (calculated for final 2 reviewers)
			Support vector machine with all MetaMap concepts	66.00%	34.00%	45.00%	80.00%		
			Support vector machine with filtered MetaMap concepts	71.00%	78.00%	74.00%	84.00%		
			Support vector machine with bag of words	53.00%	43.00%	48.00%	74.00%		
Wang et al (2020)	Breast cancer metastatic recurrence	Progress notes, structured EMR data	Knowledge-guided convolutional neural network (best model)	53.70%	46.80%	50.00%	88.80%		

Ling et al (2019)	Breast cancer metastatic recurrence	ICD-9-CM diagnosis codes	Rule-based classifier	63.00%	93.00%	-	-	2	-
		Structured EMR data	Logistic regression (L2 regularization)	75.00%	54.20%	62.90%	78.90%		
		Clinical notes, radiology reports, pathology reports		87.30%	86.10%	86.70%	91.70%		
		Clinical notes, radiology reports, pathology reports, structured EMR data		87.30%	86.10%	86.70%	92.50%		

4.2.2. Breast cancer characteristic extraction from EMR notes

Several studies have focused on developing methods to extract phenotypes from cancer records, and more specifically to extract breast cancer characteristics (118, 138-140). Hochheiser et al (2016)(138) and Savova et al (2017)(141) developed DeepPhe, an information extraction system for extracting cancer phenotypes from EMR. The system has four levels: mentions, documents, episodes, and phenotypes. Tumor receptor status is collected at the mention level but interpreted at the phenotype level. Reason being, if a patient has recurrent cancer, for example, the first cancer may be estrogen receptor negative while the recurrence may be estrogen receptor positive. The DeepPhe phenotype level takes that into account and summarizes a "deep phenotype" for the patient. The system was compared against performance with human annotators in a series of interviews including contextual inquiries and information modeling interviews. For breast cancer, 137 data elements were discussed in the information modeling interviews. Research staff who collected the data elements regularly examined those extracted by the system and found that 112 (81.8%) of elements could only be manually abstracted at that point in time (138). The calculated inter-annotator agreement ranged from 46-100% between the subject matter experts, and the agreement between the system and the experts ranged from 20% to 96% (141).

Yala et al (2017)(139) collected 91, 5105 breast pathology reports and extracted 20 data elements from the reports. All annotations were from two databases from past studies. The first database contained annotations for various carcinomas and atypias, while the second database contained annotations for tumor characteristics like tumor receptor status. They developed a decision tree-based boosting classifier with n-gram features for each data element

(20 models), and they tested for accuracy, precision, recall, and F1-Score on a held-out set of 500 reports. F1-Scores for tumor receptor status models (ER status, PR status, HER2 status) are described here. The classification task had three classes: positive, negative, and unknown. The ER status model had an F1-Score of 98% on positive cases, 95% on negative cases, and 90% on unknown cases. The PR status model had an F1-Score of 97% on positive cases, 94% on negative cases, and 90% on unknown cases. The HER2 status model had an F1-Score of 87% on positive cases, 95% on negative cases, and 93% on unknown cases. Since the models did not achieve an F1-Score of 100%, the authors created an interface for researchers to use, and the workflow requires that researchers review and correct the system predictions. It is set up such that when researchers correct predictions, the data is fed back into the training set, so the model will continuously improve.

Zhou et al (2022)(118) developed CancerBERT, a fine-tuned BlueBERT model for extracting breast cancer phenotypes from EMR. BlueBERT was trained on MIMIC-III clinical notes and PubMed abstracts (142). They annotated 200 pathology reports and 50 clinical notes for 8 data elements, including tumor receptor type and tumor receptor status. The annotations were done by two graduate students with clinical and pharmacy backgrounds. Tumor receptor statuses were labeled as positive or negative. Based on issues with how the WordPiece tokenizer from the BlueBERT model tokenized entities, the authors decided to expand the BERT vocabulary with a knowledge-based method (terms selected by a researcher with a clinical background) and a frequency-based method (most frequent words selected) after extracting word lists from the training corpus. The authors also compared CancerBERT with various other BERT-based models. The best performance for tumor receptor status was an F1-Score of 90.1%

using the original vocabulary. Scores for each tumor receptor status were not presented in the study.

4.3. **Methods**

4.3.1. Study setting

The study was 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 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. In addition, the NC Cancer Registry (NCCR) was used to identify patients with confirmed cancer diagnoses.

Data in the CDW-H and NCCR all require IRB approval for access. After IRB approval, a CDW-H Project Request form was 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, after linking records between the CDW-H and NCCR, extracted and processed the data for use (143).

4.3.2. Subjects

All records from the CDWH for patients in the NCCR between April 4, 2014 and January 13, 2021 (approximately 7 years) were extracted, and patients designated as having breast cancer in the NCCR were included in the cohort. Patients who only got a diagnosis and not treatment at UNC were filtered out. Patients who received an initial diagnosis in the first two weeks of January 2021 were also filtered out.

4.3.3. Study design

This study takes a two-step approach to cohort stratification: (1) detect breast cancer recurrence (2) extract tumor receptor status. The first step uses a hybrid approach leveraging clinical notes and diagnosis codes to identify patients with breast cancer recurrence. The second step uses clinical notes to fine-tune a pre-trained named entity recognition pipeline to extract tumor receptor status for each patient in the cohort outputted from step 1. The final output of the aim is a cohort of individuals divided by recurrence-free survival status and by tumor receptor status. The workflow is show in Figure 10.

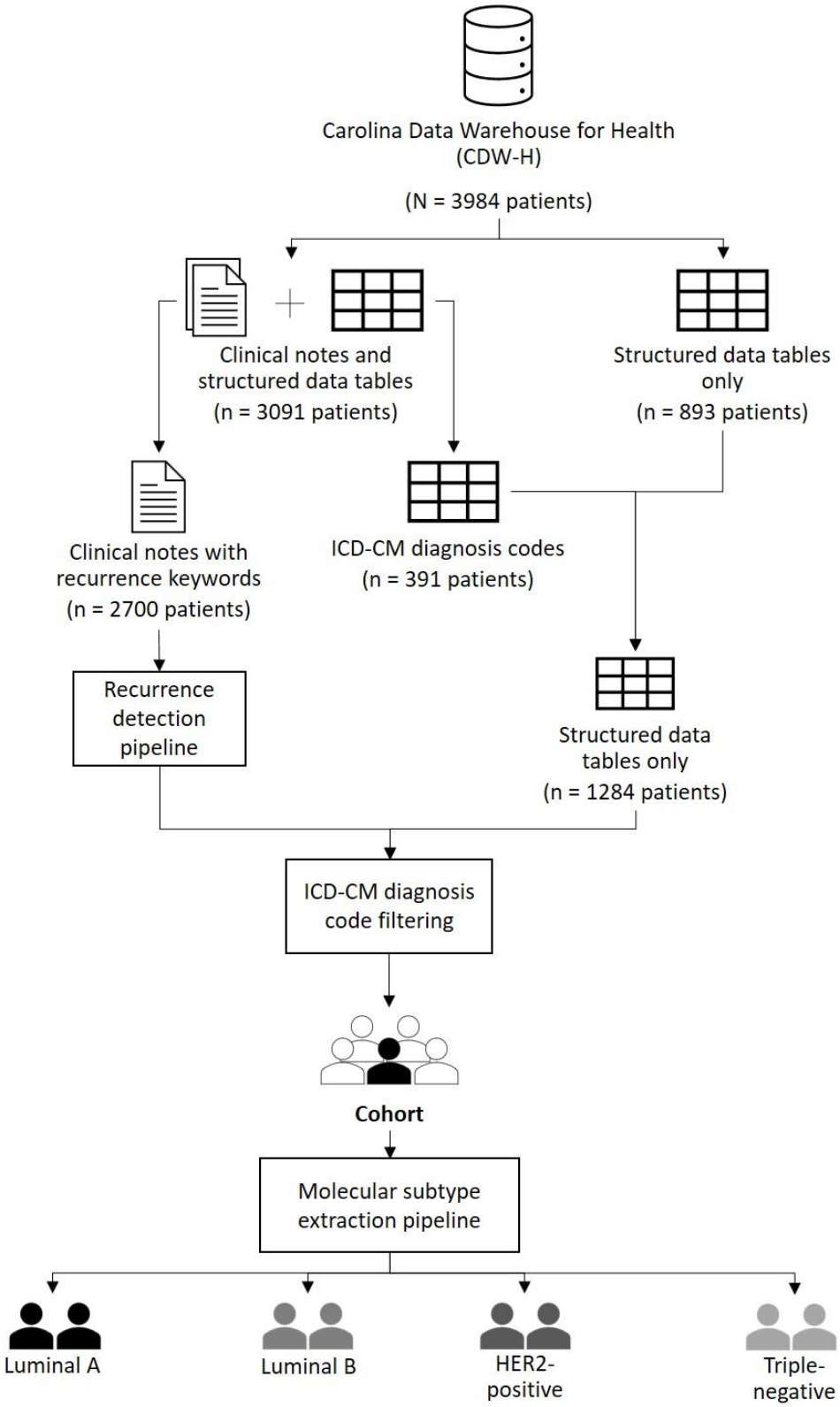


Figure 10. Breast cancer computational phenotyping workflow

4.3.4. Data preprocessing

Before IRB approval, the Informatics for Integrating Biology and the Bedside (i2b2) was used to estimate cohort size and basic patient demographics (e.g., patient counts per age group and biological sex). 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. After IRB approval, the unstructured clinical notes were analyzed using the Electronic Medical Record Search Engine (EMERSE)(144), a system which allows users to search through unstructured, identified clinical notes from hEMR. The notes for all patients with breast mammograms were examined to estimate cohort size.

All records of female patients with a breast cancer diagnosis recorded in the NCCR were extracted from the CDW-H between April 4, 2014 and January 13, 2021. The CDW-H records consisted of all PCORnet Common Data Model (CDM) data tables, a clarity data table, and unstructured clinical notes. The NCCR data included patient demographics, cancer staging and grading, HER2 status, and details on chemotherapy, radiotherapy, or surgeries conducted. Patient survival was extracted from the clarity data table from CDW-H and the North Carolina Department of Health and Health Services (NC DHHS) death database.

Structured information in the EMR is comprised of patient demographics, billing codes, laboratory tests, medications, treatments, encounters, diagnoses, procedures, and vitals. Unstructured information consists of clinical notes. The CDW-H data contained 863,193 clinical notes and 12,037,991 structured records. Clinical notes were divided into five .psv files and parsed using Python. Structured records were from 16 .csv files and were also parsed using Python. There were 15 files using the PCORnet CDM and one clarity file (with death dates from

NC DHHS. Structured CDW-H and NCCR data were first processed to obtain patient demographics. Age at diagnosis was first obtained from NCCR data and verified by calculating the difference between patient birth date and biopsy date from CDW-H data. Patient race was obtained from CDW-H data.

4.3.5. Recurrence detection

4.3.5.1. Clinical note processing

For patients with clinical notes, all notes were filtered down to those including recurrence keywords based on keywords mentioned in prior literature (134): “recurrent metastatic tnbc”, “distant relapse”, “distant recurrences”, “distant metastatic disease involving”, “regional recurrence”, “loco-regional failure”, “locally recur”, “in-breast recurrence”, “local recur”, “recur”, “rapid recurrence”, “multiple recurrences”, “recurrent disease”, “reoccurrence”, and “reoccurring”. The filtered notes were deidentified with the Protected Health Information filter (Philter) published by Norgeot et al (2020)(145). The deidentified notes were annotated using Prodigy, an annotation tool created by Explosion AI (146), by whether the patient had breast cancer recurrence or not. Annotation guidelines were developed to manage the process and ensure reproducibility. The guidelines along with example note snippets are described in Appendix 1. Notes from 1221 patients were labeled and used to develop a recurrence detection pipeline. Patients with Stage 0 carcinomas and Stage IV metastatic cancer diagnosed in the initial diagnosis were filtered out.

4.3.5.2. Natural language processing pipeline development

The annotated patient notes (1021 no recurrence, 200 recurrence) were separated into train (597, 49%), validation (257, 21%), and held-out test (367, 30%) sets. The recurrence

outcome variable was stratified across all sets, and the class distributions are shown in Table 6. The pipeline consists of training with 5-fold cross-validation to select a classification approach and tune the selected classifier and using the development set to evaluate performance. After the model was finalized, its performance was evaluated on the held-out test set. The training step of the pipeline consisted of comparing an array oversampling and undersampling methods (using imbalanced-learn in Python) against baseline performance with four machine learning classifiers (using scikit-learn (115) in Python).

Table 6. Training and evaluation set class distributions for recurrence detection with clinical notes

Train	
Recurrence	98 (16.42%)
No recurrence	499 (84.58%)
Validation	
Recurrence	42 (16.34%)
No recurrence	215 (83.66%)
Test	
Recurrence	60 (16.35%)
No recurrence	307 (83.65%)

The oversampling methods used include Synthetic Minority Oversampling Technique (SMOTE)(147) and Adaptive Synthetic Sampling Approach (ADASYN) (148). SMOTE creates synthetic minority samples from the minority class in a dataset based on the Euclidean distance of the majority class points and minority class neighbors using K-Nearest Neighbors (KNN)(147), and ADASYN creates synthetic minority samples from the minority class based on a weighted distribution of minority class samples (148).

The undersampling methods used are random undersampling, near miss version 3 undersampling (149), and Tomek Links (150). Random undersampling is a technique where

randomly selected samples from the majority class are removed from the dataset. The near miss undersampling version used calculates the Euclidean distance between majority and minority points and selects the majority class points with the least distance to the minority class points (149). Tomek Links undersampling removes the majority sample counterpart of a tomek link, which is a pair of points of opposite classes that are also nearest neighbors (150). The purpose of the undersampling method is to remove boundary instances that would confuse a model. One oversampling and undersampling combination approach, SMOTE Tomek Links, was also used. SMOTE Tomek Links first executes SMOTE to generate synthetic minority points. Then, it executes Tomek Links, where random points are selected from the majority class, and if their nearest neighbors are in the minority class, the points are removed.

Word-level TF-IDF vectorization with stopword removal was used to transform the clinical notes. Logistic regression, SVM with a linear kernel (linear SVM), SVM with a radial-basis function kernel (SVM), and random forest were compared during training.

4.3.5.3. Measures of evaluation

A random sample of notes from 50 patients was annotated by a graduate research assistant and medical student to confirm recurrence labels. Cohen's kappa coefficient was 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 (151).

Four metrics were examined as measures of evaluation: F1-Score, precision (i.e., positive predictive value), recall (i.e., sensitivity), and ROC-AUC score. ROC-AUC score is the

area under the receiver operating characteristic curve. A ROC curve plots the false positive rate against the true positive rate (i.e., recall), and the AUC represents the degree of separability between two classes.

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

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

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

$$False\ positive\ rate = \frac{FP}{TN + FP}$$

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

Hyperparameter tuning was done during cross-validation to optimize F1-Score. F1-Score was chosen as the key measure because a balance between recall and precision is important for recurrence detection. The positive class (recurrence) is the minority class, so minimizing false negatives can allow for more samples to be predicted as the recurrence class, making recall important. However, although recurrence is the minority class and having enough sample size is important, false positives should also be minimized, so the model will not say patients had a recurrence when they did not, making precision important. After cross-validation and tuning, each model's performance was evaluated on the development set, and the best combination of sampling and classification methods was selected as the final model to evaluate on the held-out test set.

4.3.5.4. Diagnosis code filtering

For all patients, ICD-CM diagnosis codes were used to filter out patients with Stage 0 or Stage IV carcinomas and get recurrence dates. For patients with Stage 0 carcinomas, the patient was filtered out of the cohort if the carcinoma did not develop into invasive carcinoma. For patients with metastatic cancer, metastatic cancer was defined as a secondary neoplasm diagnosis made with the initial breast cancer diagnosis or a code for metastatic cancer. For patients without clinical notes (893 patients) or without notes with recurrence keyword mentions (391 patients), ICD-CM diagnosis codes were used to detect recurrence. Recurrence is defined as a cancer of the same type that has come back after a period of time to either where the original cancer occurred or a different location in the body. The period of time can be quantified as the time from the last follow-up after completion of cancer treatment to the point of a new diagnosis. From the PCORnet CDM tables, only diagnosis codes without descriptions were provided. The diagnosis codes were searched in the UMLS ICD-10-CM and ICD-9-CM vocabularies to obtain the code descriptions. From the code descriptions, diagnosis records were filtered down to codes for primary neoplasms, follow up appointments, and secondary neoplasms. Then, patients with recurrence were identified as those with secondary neoplasms after cancer treatment completion or patients with breast cancer that has recurred in the breast. Patients with distant recurrence or regional recurrence were more easily identified with secondary neoplasm codes, while patients with in-breast recurrence were more difficult to detect with codes. The codes for breast cancer indicate malignant neoplasms in an area of the breast, and in-breast recurrences are also denoted as malignant neoplasms in an area of the breast. Therefore, time between diagnoses was examined to ascertain whether patients had local recurrence or not. In addition, for patients that had clinical notes without recurrence

keyword mentions, their notes were manually examined and compared against diagnosis codes to ascertain their recurrence statuses.

4.3.6. Tumor receptor status extraction

4.3.6.1. Named entity recognition pipeline development

For patients with clinical notes, all notes were split into sentences, filtered down to those including tumor receptor status mentions, and concatenated together. All note types were used, and the two primary note types were progress notes and pathology reports. The tumor receptor status mentions were categorized as: ER+ (estrogen receptor positive), ER- (estrogen receptor negative), PR+ (progesterone receptor positive), PR- (progesterone receptor negative), ER_PR+ (hormone receptor positive), ER_PR- (hormone receptor negative), HER2+, HER2-, TNBC (triple negative), and TPBC (ER+, PR+, HER2+). Annotation guidelines were developed to manage the process and ensure reproducibility. The guidelines along with example note snippets are described in Appendix 1. The notes were split into sentences, and each sentence was annotated using Prodigy. The base model used for named entity annotation was “en_core_sci_md” from the scispaCy Python package (108) of spaCy models for biomedical and clinical text processing. The “en_core_sci_md” is a spaCy pipeline with 50,000 word vectors that was trained on biomedical literature from semantic scholar. The “en_core_sci_md” pipeline was used for tokenization to enable easier annotations.

Four annotation datasets were created. The training set contained notes from a random sample of 200 patients. The set of 200 notes contained 841 sentences, out of which, 697 sentences contained named entities. The development set contained notes from a random sample of 50 patients to evaluate training performance and gage whether more samples

needed to be annotated. The set of 50 notes contained 214 sentences, out of which, 171 sentences contained named entities. The training and development sets were annotated using Prodigy's "ner.manual" recipe, which allows users to manually annotate notes without any suggestions from the base model. An example annotation is shown in Figure 11.

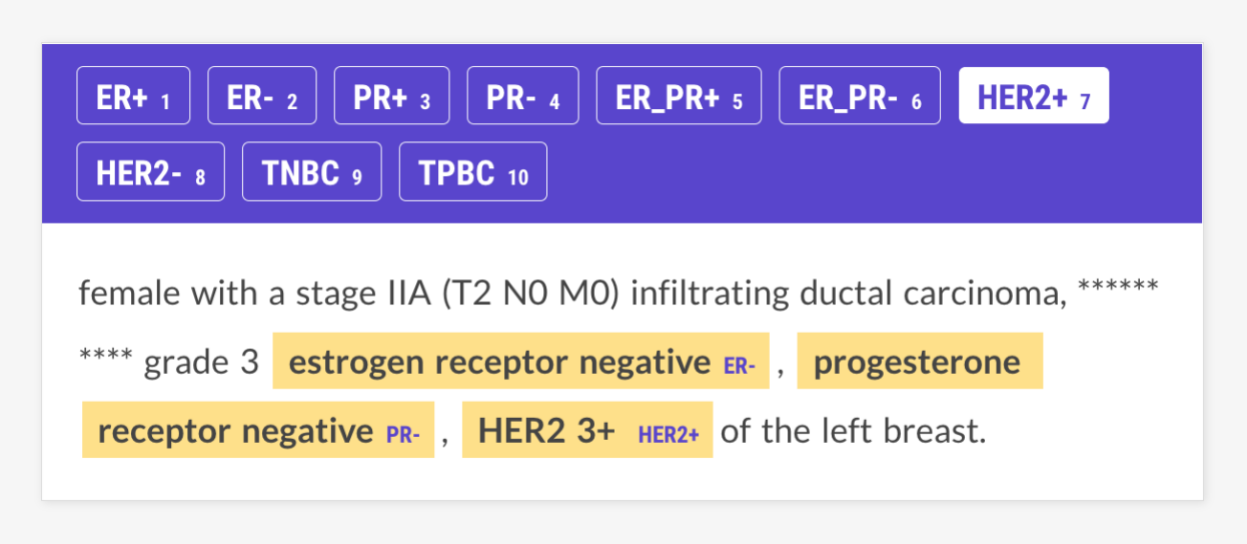


Figure 11. Example annotation using the Prodigy annotation system

Cross-validation was not used for training to prevent note sentence leakage. Since the name entity recognition task was done on each note sentence, using cross-validation could allow for sentences from one patient to be in training and testing folds. This could potentially bias the result because sentences from one patient note will contain similar patterns of writing. A model trained on the training set (200 patients) and evaluated on the development set (50 patients) was outputted as a spaCy model. The third dataset contained notes from a random sample of 180 patients (706 sentences) for model fine-tuning using Prodigy's "ner.correct" recipe. The "ner.correct" recipe shows the outputted model's predicted annotations and allows the user to correct or remove them. It updates the model in the loop with the corrected tokens.

4.3.6.2. Evaluation

A random sample of notes from 50 patients was annotated by a second annotator and used as a held-out test set for final evaluation. Cohen's kappa coefficient was used to measure inter-rater reliability (See 4.3.5.3 Measures of evaluation, p. 76).

Performance on the held-out test set was evaluated with three measures: precision, recall, and F1-Score (See 4.3.5.3 Measures of evaluation, p. 76). Overall accuracy was also used as a metric. The following labels were evaluated: ER+, ER-, PR+, PR-, ER_PR+, ER_PR-, HER2+, HER2-, TNBC, and TPBC. Precision, recall, and F1-Score were evaluated for the named entity recognition task overall and for each individual label.

The labels were also aggregated into groups (HR+/HER2-, HR+, HR-/HER2+, TNBC) and performance was evaluated at the patient level to understand how many patients would be correctly characterized by the model. Precision, recall, and F1-Score were used for evaluation, and the micro average was used to get the overall results because of the sample imbalance between groups.

4.4. **Results**

4.4.1. Study population

The tables extracted are shown in Table 7. Data for breast and oral cancer were extracted and provided together by NC TraCS. Breast cancer record counts in Table 7 were calculated after separating the datasets by cancer registry diagnosis. Providers were not designated for breast and oral cancer data separately, so the total number reflects all providers involved in patient care breast and oral cancer patients.

Table 7. Structured data summary for breast cancer data.

Table Name	Record Count
Clarity	

Clarity ID	3,944
Deaths	66
PCORnet CDM	
Condition	9,407
Death	0
Death Cause	0
Demographic	3,806
Diagnosis	2,803,612
Encounter	1,041,231
Lab Result CM	1,411,710
Address History	3,806
Med Admin	519,509
OBS Clin	335,882
OBS Gen	277
Prescribing	442,046
Procedures	427,358
Provider	196,813
Vital	789,939
NC DHHS Death Database	
Death	31

The NCCR cohort included 3997 patients with 9630 records, and the CDW-H cohort included 3984 patients with 9590 records (Figure 12). The thirteen patients from the NCCR that were not in the CDW-H were not included in the study cohort. The largest age groups in the cohort were 50-64 years (N=1536, 38.55%) and 65 years or more (N=1443, 36.22%) (Figure 13). The cohort was predominantly comprised of White patients (N=2836, 71.18%) and Black or African American patients (N=797, 20.01%) (Table 8). All patients were female.

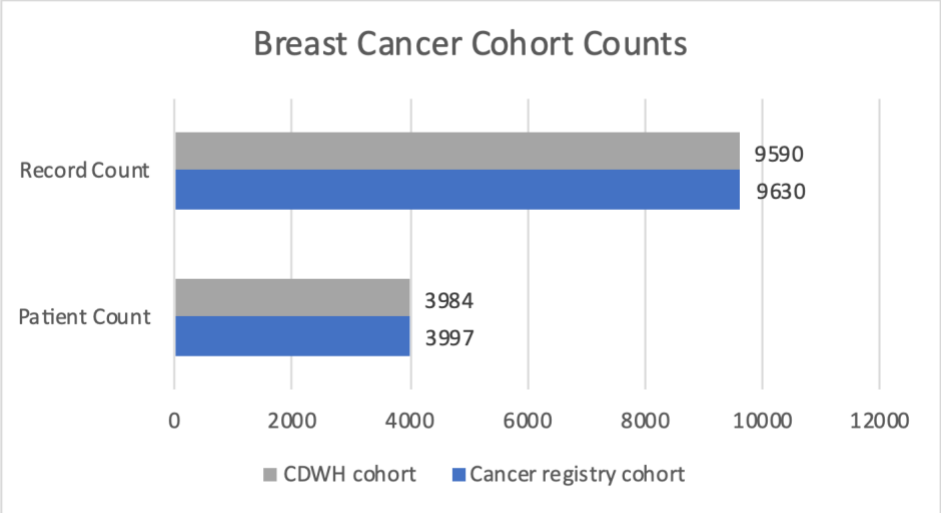


Figure 12. Breast cancer cohort distribution from CDW-H and NCCR.

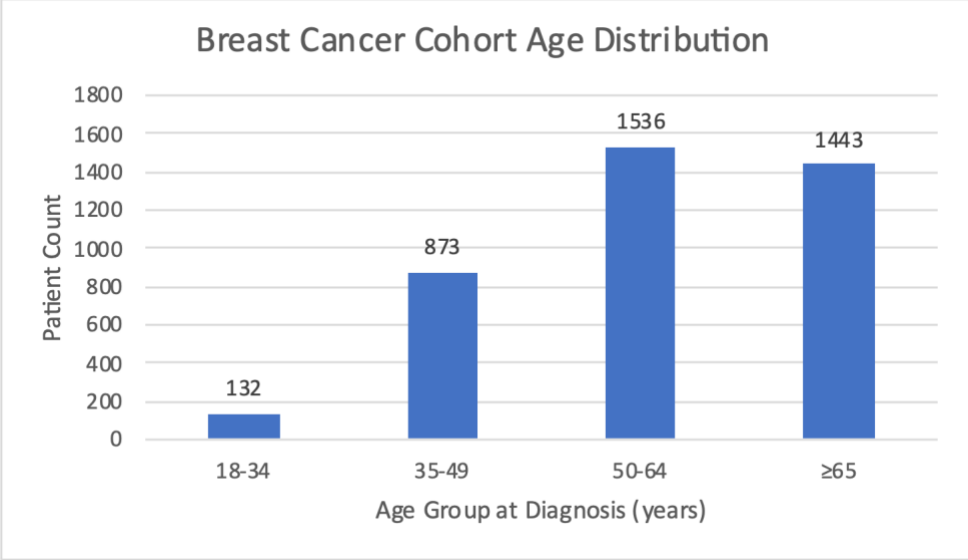


Figure 13. Age group distribution in the CDW-H breast cancer cohort.

Table 8. Racial distribution of patients in the CDW-H breast cancer cohort.

Race	CDW-H Patient Count (N=3984)
American Indian or Alaska Native	25 (0.63%)
Asian	72 (1.81%)
Black or African American	797 (20.01%)
Native Hawaiian or Other Pacific Islander	4 (0.10%)

No information	24 (0.60%)
Other	154 (3.87%)
Refuse to answer	5 (0.13%)
Unknown	62 (1.56%)
White	2836 (71.18%)
Missing	5 (0.13%)

4.4.2. Recurrence detection

4.4.2.1. Natural language processing pipeline

To assess annotation performance, a random sample of 50 patient notes was annotated by a graduate research assistant and a medical student. The inter-rater reliability was a Cohen's kappa statistic of 84.67%.

Two oversampling methods (SMOTE, ADASYN), three undersampling methods (random undersampling, near miss version 3 undersampling with two sampling strategies, Tomek Links), and one combination method (SMOTE Tomek) were evaluated with logistic regression, linear SVM, SVM, and random forest with 5-fold cross-validation on the training set (499 no recurrence, 98 recurrence). The performance in terms of, precision, recall, F1-Score, and ROC-AUC score are shown in Table 9. The baseline performance without any sampling methods had high precision between 94-100%, but recall ranged from 10-55% across algorithms. The oversampling methods seemingly overfit all machine learning classifiers, while the undersampling methods have mediocre cross-validation performance. The final logistic regression model uses L2 regularization with a stochastic average gradient solver and the maximum number of iterations was set to 10,000. Random forest was implemented with 100 trees using the Gini index to measure the quality of each split. Linear SVM was implemented with an L2 penalty and squared hinge loss function. SVM was implemented with C = 1.0 and a

scaled gamma value (1 / (number of features * feature variance)). Performance on the development set better reflects how the model would perform on unseen clinical notes.

Table 9. Recurrence detection cross-validation results with data sampling methods and classifiers using clinical note features.

Model Name	Precision	Recall	F1-Score	ROC-AUC
Baseline (no sampling)				
Logistic Regression	100.00%	20.42%	63.10%	92.39%
Linear SVM	98.18%	55.26%	83.05%	94.04%
SVM	100.00%	41.95%	76.63%	93.33%
Random Forest	94.17%	33.63%	71.50%	91.85%
SMOTE				
Logistic Regression	95.23%	92.59%	93.98%	98.92%
Linear SVM	98.01%	97.80%	97.90%	99.65%
SVM	99.40%	96.00%	97.69%	99.71%
Random Forest	98.19%	91.40%	94.76%	98.74%
ADASYN				
Logistic Regression	94.90%	93.58%	94.28%	98.54%
Linear SVM	97.60%	97.59%	97.59%	99.49%
SVM	99.39%	94.39%	96.87%	99.55%
Random Forest	97.82%	92.20%	94.99%	98.45%
Random Undersampling				
Logistic Regression	85.77%	78.63%	82.59%	91.53%
Linear SVM	90.66%	78.58%	85.10%	92.33%
SVM	91.72%	77.58%	85.10%	91.34%
Random Forest	88.05%	81.68%	85.18%	90.41%
Near Miss (resample majority)				
Logistic Regression	78.50%	78.68%	76.80%	86.28%
Linear SVM	84.71%	77.63%	81.05%	87.07%
SVM	84.79%	75.58%	79.95%	87.18%
Random Forest	84.53%	81.63%	83.12%	89.71%
Near Miss (resample all)				
Logistic Regression	81.76%	78.47%	80.56%	86.52%
Linear SVM	84.83%	72.32%	80.24%	86.25%
SVM	89.79%	75.37%	83.35%	86.39%
Random Forest	76.67%	69.37%	74.93%	86.82%
Tomek Links				

Logistic Regression	100.00%	19.42%	62.26%	92.53%
Linear SVM	100.00%	56.16%	83.74%	94.30%
SVM	100.00%	42.79%	77.09%	93.54%
Random Forest	95.00%	33.68%	71.56%	91.20%
SMOTE Tomek				
Logistic Regression	95.85%	93.78%	94.88%	98.95%
Linear SVM	98.02%	98.00%	97.99%	99.82%
SVM	99.60%	95.40%	97.47%	99.78%
Random Forest	97.73%	92.40%	95.09%	98.48%

Performance for each sampling and classification method combination was also evaluated on the development set, and the results are shown in Table 10. Three approaches resulted in comparably high F1-Scores: linear SVM with SMOTE oversampling (88.03%), logistic regression with ADASYN oversampling (86.61%), and linear SVM with SMOTE Tomek over/undersampling (86.11%). Using SMOTE and SMOTE Tomek with linear SVM had precision of 93.55% and 93.10% respectively and using ADASYN oversampling with logistic regression had a precision of 76.74%. However, using ADASYN oversampling with logistic regression resulted in a recall of 78.57%, which is higher than recalls of SMOTE (69.05%) and SMOTE Tomek (64.29%). Since the goal for performance in the task was to find a balance between precision and recall, logistic regression with ADASYN oversampling was selected as the final algorithm to be tested on the held-out test set.

Table 10. Recurrence detection development set results with data sampling methods and classifiers.

Model Name	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**
Baseline				
Logistic Regression	100.00%	38.10%	74.74%	90.12%
Linear SVM	92.00%	54.76%	81.98%	92.00%
SVM	95.00%	45.24%	77.99%	91.54%
Random Forest	100.00%	30.95%	70.48%	86.93%

SMOTE				
Logistic Regression	78.38%	69.05%	84.30%	90.25%
Linear SVM	93.55%	69.05%	88.03%	91.07%
SVM	95.00%	45.24%	77.99%	90.82%
Random Forest	88.00%	52.38%	80.26%	85.72%
ADASYN				
Logistic Regression	76.74%	78.57%	86.61%	90.90%
Linear SVM	93.55%	69.05%	88.03%	91.14%
SVM	95.00%	45.24%	77.99%	90.95%
Random Forest	76.00%	45.24%	75.11%	82.96%
Random Undersampling				
Logistic Regression	55.00%	78.57%	77.98%	88.43%
Linear SVM	66.00%	78.57%	82.79%	88.82%
SVM	64.00%	76.19%	81.47%	88.06%
Random Forest	46.03%	69.05%	71.87%	82.90%
Near Miss (resample majority)				
Logistic Regression	64.71%	78.57%	82.28%	90.02%
Linear SVM	67.39%	73.81%	82.18%	88.84%
SVM	67.39%	73.81%	82.18%	89.29%
Random Forest	57.14%	66.67%	76.63%	87.10%
Near Miss (resample all)				
Logistic Regression	64.71%	78.57%	82.28%	90.02%
Linear SVM	67.39%	73.81%	82.18%	88.83%
SVM	67.39%	73.81%	82.18%	89.29%
Random Forest	65.91%	69.05%	80.45%	89.82%
Tomek Links				
Logistic Regression	100.00%	38.10%	74.74%	90.01%
Linear SVM	95.83%	54.76%	82.62%	91.94%
SVM	95.00%	45.24%	77.99%	91.34%
Random Forest	100.00%	30.95%	70.48%	86.84%
SMOTE Tomek				
Logistic Regression	82.86%	69.05%	85.49%	90.81%
Linear SVM	93.10%	64.29%	86.11%	91.71%
SVM	95.00%	45.24%	77.99%	90.96%
Random Forest	86.36%	45.24%	76.80%	84.91%

Evaluating the pipeline with ADASYN oversampling and logistic regression on the held-out test set resulted in an overall accuracy of 91.00%, precision of 83.07%, recall of 85.24%, F1-Score of 84.10%, and ROC-AUC score of 94.51%. The confusion matrix for the result on the

held-out test set is shown in Figure 14. The number of false positives is slightly higher than the number of false negatives, but all recurrence predictions were manually reviewed after predicting on unlabeled examples to ensure the recurrence predictions were correct.

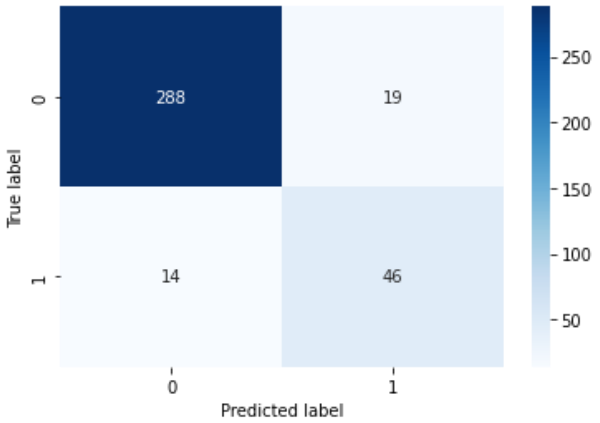


Figure 14. Confusion matrix displaying performance on the held-out test set from classifying recurrence with the logistic regression model that was trained on oversampled clinical notes

Manually reviewing the minority class predictions from the unlabeled set to verify performance was considered an easier task than manually reviewing the whole set. Predicting recurrence labels for the set of unlabeled notes (1479 patient notes) resulted in 1355 no recurrence predictions and 124 recurrence predictions. Of the 125 recurrence predictions, 33 predictions were incorrect. 7 of the predictions were new primaries, and 26 predictions were mentions of other recurrent conditions. The most common points of confusion were recurrent seromas and recurrent episodes of major depressive disorder. New primaries are difficult to distinguish from recurrences as the language surrounding new primaries is very similar to language surrounding recurrences, and clinicians will often write “new primary vs. recurrence” in the notes as something that needs to be clarified.

4.4.2.2. ICD-CM diagnosis code filtering

The 2700 patients with recurrence-related clinical notes and structured data, the 391 patients with clinical notes that did not have recurrence keywords and structured data, and the 893 patients with structured data only were filtered using ICDM-CM diagnosis codes to filter out patients with Stage 0/IV carcinomas. 333 patients had Stage 0 carcinoma, and 519 patients had Stage IV carcinoma. Recurrence dates were obtained for 772 patients, and 2878 patients did not have local or distant recurrence. The final cohort distribution was 2878 (78.85%) without recurrence and 772 (21.15%) with recurrence.

4.4.3. Tumor receptor status extraction

After training on 187 patient notes and evaluating on a set of 50 patient notes, the model had the following performance: precision of 90.34%, recall of 87.92%, and F1-Score of 89.12%. It also had an overall accuracy of 89.00%. The performance on each label is presented in Table 11. The model performed poorly in identifying mentions of HER2+ and HER2- statuses. Identifying HER2 status is important but confusing for a model because of the many variations in how each status can be written (See Appendix 1). Also, a clear point of confusion is the significance of dashes (-) and plus (+) signs. Take the example sentence, "Breast, right, core biopsy: -HER-2 FISH analysis: Positive (Her2\/***** ratio is 2.4 with an average HER2 copy number of 6.7)." In this sentence, “-HER-2 FISH” was identified as HER2- because of the dash before the string. “Her2\/***** ratio is 2.4” was also identified as HER2-, which is incorrect.

Table 11. Tumor receptor status extraction model performance per label: initial results

Label	Precision	Recall	F1-Score
ER+	94.59%	98.13%	96.33%
ER-	92.86%	86.67%	89.66
PR+	95.65%	95.65%	95.65%
PR-	100.00%	100.00%	100.00%
ER_PR+	94.44%	79.07%	86.08%

ER_PR-	100.00%	100.00%	100.00%
HER2+	53.33%	53.33%	53.33%
HER2-	78.72%	74.00%	76.29%
TNBC	100.00%	83.33%	90.91%
TPBC	100.00%	100.00%	100.00%
Overall	**90.34%**	**87.92%**	**89.12%**

The rest of the annotations were done by manually correcting model predictions on a set of 180 patient notes. The overall model performance after fine tuning was a precision of 88.5%, recall of 87.92%, and an F1-Score of 88.51%. The overall accuracy was also 89.00%. The results per label are shown in Table 12. The performance in detecting HER2+ and TPBC mentions dropped. There is only one example of TPBC in the development set, explaining why the performance dropped from 100% to 0% in comparison to initial results.

Table 12. Tumor receptor status extraction model performance per label: fine-tuning

Label	Precision	Recall	F1-Score
ER+	94.59%	98.13%	96.33%
ER-	100.00%	86.67%	92.86%
PR+	86.79%	100.00%	92.93%
PR-	100.00%	78.57%	88.00%
ER_PR+	94.59%	81.40%	87.50%
ER_PR-	100.00%	100.00%	100.00%
HER2+	46.67%	46.67%	46.67%
HER2-	80.85%	76.00%	78.35%
TNBC	100.00%	100.00%	100.00%
TPBC	0.00%	0.00%	0.00%
Overall	**89.12%**	**87.92%**	**89.12%**

The model was tested on the double-annotated held-out test set for final performance evaluation. The overall model performance was a precision of 90.40%, recall of 86.39%, and F1-Score of 88.35%. The overall accuracy was also 88.00%. The results per label are shown in Table 13.

Table 13. Tumor receptor status extraction model performance per label: held-out test set

Label	Precision	Recall	F1-Score
ER+	91.74%	94.34%	93.02%
ER-	100.00%	72.73%	84.21%
PR+	97.22%	98.59%	97.90%
PR-	92.86%	81.25%	86.67%
ER_PR+	96.36%	88.33%	92.17%
ER_PR-	-	-	-
HER2+	36.36%	80.00%	50.00%
HER2-	78.72%	78.72%	78.72%
TNBC	100.00%	35.29%	52.17%
TPBC	100.00%	20.00%	33.33%
Overall	**90.40%**	**86.39%**	**88.35%**

The recall for the HER2+ label increased almost two-fold and the precision dropped, leading to a slight increase in F1-Score. The precisions for triple negative and hormone receptor positive-HER2 positive labels are very high, and their recall scores are very low, indicating several false negatives. The ER_PR- label does not have scores because there were no hormone receptor negative mentions in the held-out test set, and the model did not incorrectly classify any entities as hormone receptor negative either. The model performs best on detecting estrogen and progesterone receptor status, and there is lower performance in detecting HER2 status. The model was trained to highlight HER2 scores in case they are needed to further stratify patients downstream. As such, a primary point of confusion for HER2 status evaluation is that if a score was mentioned next to a HER2 status, the score was included in the annotation span. This means that even if the model can detect a HER2 status, the prediction would be incorrect if it did not also detect the score. An example illustrating a prediction that is technically incorrect but correct in terms of the patient's tumor receptor status is depicted in Figure 15. The entity outlined in blue is marked as correct in evaluation, while the entity

outlined in black is marked as incorrect in evaluation because it does not encompass the score. Two examples of correct model predictions from the same patient are shown in Figure 16. The first example shows tumor receptor status being detected in a pathology report, while the second example is from a progress note.

Figure 15. Technically correct and incorrect tumor receptor status named entity recognition prediction examples

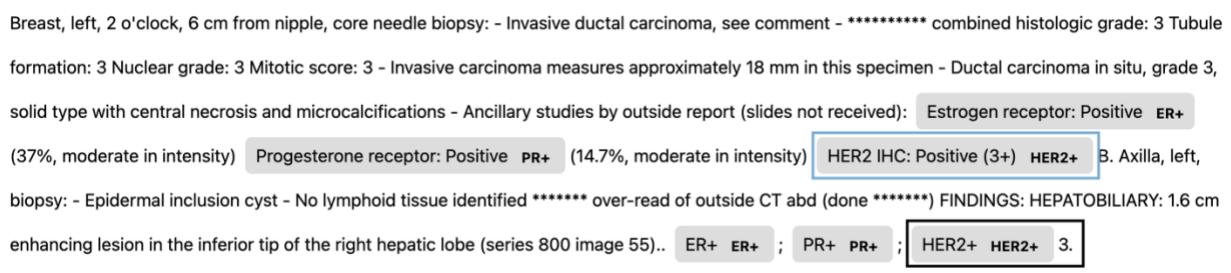
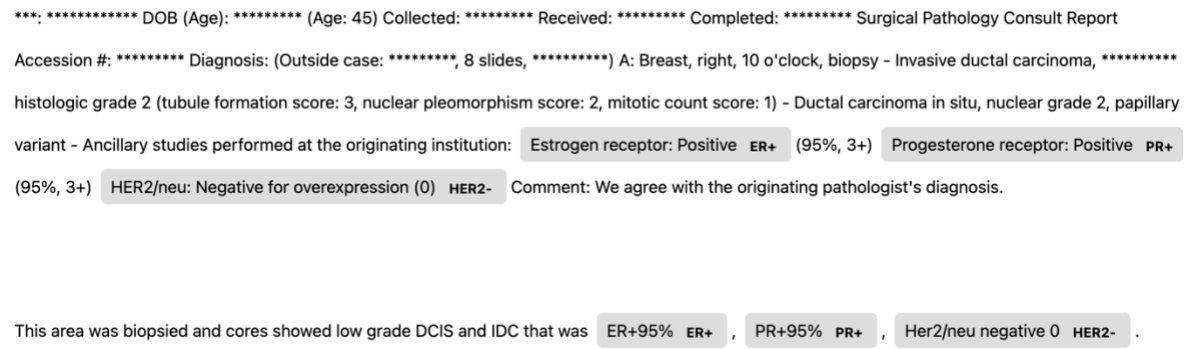


Figure 16. Correct tumor receptor status named entity recognition prediction examples



After aggregating the named entities in the held-out test set and the model predictions into molecular subtype groups, the performance was evaluated on the patient level. The results are shown in Table 14. The number of patients that fall into the criteria is shown in the "Support" column. The micro average was used to get the overall results because of the sample number imbalance.

Table 14. Tumor receptor status extraction: patient level performance

Tumor Receptor Status Groups	Precision	Recall	F1-Score	Support (Number of Patients)
HR+/HER2-	94.44%	97.14%	95.77%	35
HR+	95.24%	100.00%	97.56%	40
HR-/HER2+	0.00%	0.00%	0.00%	1
HR-/HER2-	66.67%	90.00%	72.73%	5
Micro Average	**92.86%**	**96.30%**	**94.55%**	**81**

HR: hormone receptor

4.5. **Discussion**

4.5.1. Recurrence detection

The best recurrence detection model achieved an overall accuracy of 91.00%, precision of 83.07%, recall of 85.24%, F1-Score of 84.10%, and ROC-AUC score of 94.51% on the held-out test set of clinical notes. From prior work, Savova et al (2014)(131) had the following performance: 59% precision, 92% recall, and 72% F1-Score. The best recurrence detection model developed in this study performed better than the prior model in terms of precision and F1-Score but not recall. More recent natural language processing models focused specifically on local or distant recurrence individually, so their performances were unable to be compared to those of the model in this study.

From a technical perspective, lower classification performance can be attributed both to a class imbalance between patients with and without breast cancer recurrence and to the limitations of clinical notes as a sole data source. Of the 2700 patients with recurrence keywords in their clinical notes, 291 (10.7%) patients were found to have recurrence. Pan et al (2017)(152) conducted a meta-analysis of 88 trials with ER-positive female breast cancer patients, encompassing 62,923 women. The study reported that the loco-regional recurrence risk from diagnosis to year 20 ranges from 7-12% based on pathological nodal status. They also

reported that the distant recurrence risk from diagnosis to year 20 ranges from 18-46% for Stage I tumors and 29-57% for Stage II tumors, with both ranges varying by nodal status. Using a hybrid approach with both unstructured and structured data was able to better capture the ranges presented in the literature. The final cohort in this study included 2878 (79%) patients without recurrence and 772 (21%) patients with recurrence.

With increasing rates of overall survival, having outcomes like recurrence-free survival or progression-free survival is critical to make advances in breast cancer-related clinical research. These outcomes are often manually abstracted and require large amounts of time and cognitive effort. Recurrence information was not included in data extracted from the NCCR, indicating a possible need for population health surveillance models to populate the registry. There has been a shift toward informatics pipelines to increase the efficiency of obtaining survival-based outcomes. From that perspective, hybrid approaches that use both clinical notes and structured EMR data are valuable because they mitigate the weaknesses of using each individually. Models with clinical notes may not be as portable to different institutions due to clinician writing patterns, while ICD-CM diagnosis codes are standard across institutions. Diagnosis codes lack context, which clinical notes provide. To be even more thorough, more data sources could be compared across institutions.

4.5.2. Tumor receptor status extraction

Unstructured clinical notes are rich sources of information that may not be contained in the structured EMR data. For breast cancer research, it is difficult to identify various cancer characteristics in the structured EMR data, including tumor receptor status. Tumor receptor status is important for patient prognosis and to determine therapeutic management of breast

cancer (153). In the PCORnet CDM tables, there are codes for estrogen receptor status but not for progesterone receptor status or HER2 status. Due to a lack of information in the structured data, extracting tumor receptor status from the clinical notes in a critical task.

Named entity recognition approaches to extract tumor receptor status are generally performed on one type of note like pathology reports. However, not every patient in the dataset has pathology reports, so other note types were also included in the analysis. This led to model confusion in differentiating between aspects in pathology reports and progress notes. In addition, pathology reports typically have standard formats that are filled in with numbers, but progress notes are mostly written by hand, which will allow for human errors like spelling mistakes. Using a word vector-based pre-trained model helps to mitigate this weakness, but as the pre-trained model was not trained on clinical notes, there is still a margin of error. Also, while the task is to identify tumor receptor status mentions, the purpose is to identify what molecular subtype of breast cancer a patient has. Tying tumor receptor status mentions to molecular subtype for a patient, the assumption would be that each mention would be in reference to the patient; however, there are sometimes mentions of family member's tumor receptor statuses in patient notes. Future research needs to consider how to disambiguate tumor receptor status for a patient and all other people mentioned in a clinical note.

4.5.3. Limitations

There are a few limitations in this work. Although recurrences typically happen within the first five years, many patients with recurrence in the dataset had previous breast cancer diagnoses before 2014 (outside dataset date range 04/2014 – 01/2021), which was abstracted from clinical notes. For patients who only had structured data, any pre-2014 diagnoses were

missed, allowing for error in the no recurrence predictions from ICD-CM code filtering. Another limitation is a lack of diversity in molecular subtype in the random sample selected for testing. There was only one patient with HER2-enriched breast cancer, making it difficult to evaluate performance for the category. Also, the tumor receptor status annotations were done by me and a PhD student in biological & biomedical sciences programs. Ideally, a sample of notes would be annotated by a clinician, but to mitigate this weakness, annotation guidelines were developed and used after discussion with a board-certified medical oncologist. All data used comes from a single institution, so there was no evaluation for generalizability.

REFERENCES

108. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: fast and robust models for biomedical natural language processing. arXiv preprint arXiv:190207669. 2019.

115. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011;12:2825-30.

118. Zhou S, Wang N, Wang L, Liu H, Zhang R. CancerBERT: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. J Am Med Inform Assoc. 2022.

120. Gorin SS, Haggstrom D, Han PK, Fairfield KM, Krebs P, Clauser SB. Cancer care coordination: a systematic review and meta-analysis of over 30 years of empirical studies. Annals of Behavioral Medicine. 2017;51(4):532-46.

121. Kamińska M, Ciszewski T, Łopacka-Szatan K, Miotła P, Starosławska E. Breast cancer risk factors. Prz Menopauzalny. 2015;14(3):196-202.

122. Kabel AM. Tumor markers of breast cancer: New prospectives. Journal of Oncological Sciences. 2017;3(1):5-11.

123. Venetis K, Crimini E, Sajjadi E, Corti C, Guerini-Rocco E, Viale G, et al. HER2 Low, Ultra-low, and Novel Complementary Biomarkers: Expanding the Spectrum of HER2 Positivity in Breast Cancer. Frontiers in Molecular Biosciences. 2022;9.

124. Network NCC. Breast Cancer (Version 3.2022) 2022 [Available from: https://www.nccn.org/professionals/physician_gls/pdf/breast.pdf.

125. Ritzwoller DP, Hassett MJ, Uno H, Cronin AM, Carroll NM, Hornbrook MC, et al. Development, Validation, and Dissemination of a Breast Cancer Recurrence Detection and Timing Informatics Algorithm. J Natl Cancer Inst. 2018;110(3):273-81.

126. Lambert P, Pitz M, Singh H, Decker K. Evaluation of algorithms using administrative health and structured electronic medical record data to determine breast and colorectal cancer recurrence in a Canadian province. BMC Cancer. 2021;21(1):763.

127. Jung H, Lu M, Quan ML, Cheung WY, Kong S, Lupichuk S, et al. New method for determining breast cancer recurrence-free survival using routinely collected real-world health data. BMC Cancer. 2022;22(1):281.

128. Izci H, Tambuyzer T, Tuand K, Depoorter V, Laenen A, Wildiers H, et al. A Systematic Review of Estimating Breast Cancer Recurrence at the Population Level With Administrative Data. JNCI: Journal of the National Cancer Institute. 2020;112(10):979-88.

129. Hassett MJ, Uno H, Cronin AM, Carroll NM, Hornbrook MC, Ritzwoller D. Detecting Lung and Colorectal Cancer Recurrence Using Structured Clinical/Administrative Data to Enable Outcomes Research and Population Health Management. Med Care. 2017;55(12):e88-e98.

130. Hothorn T, Hornik K, Zeileis A. ctree: Conditional inference trees. The comprehensive R archive network. 2015;8.

131. Carrell DS, Halgrim S, Tran D-T, Buist DSM, Chubak J, Chapman WW, et al. Using Natural Language Processing to Improve Efficiency of Manual Chart Abstraction in Research: The Case of Breast Cancer Recurrence. American Journal of Epidemiology. 2014;179(6):749-58.

132. Zeng Z, Espino S, Roy A, Li X, Khan SA, Clare SE, et al. Using natural language processing and machine learning to identify breast cancer local recurrence. BMC Bioinformatics. 2018;19(17):498.

133. Wang H, Li Y, Khan SA, Luo Y. Prediction of breast cancer distant recurrence using natural language processing and knowledge-guided convolutional neural network. Artif Intell Med. 2020;110:101977-.

134. Ling AY, Kurian AW, Caswell-Jin JL, Sledge GW, Jr, Shah NH, Tamang SR. Using natural language processing to construct a metastatic breast cancer cohort from linked cancer registry and electronic medical records data. JAMIA Open. 2019;2(4):528-37.

135. Banerjee I, Bozkurt S, Caswell-Jin JL, Kurian AW, Rubin DL. Natural Language Processing Approaches to Detect the Timeline of Metastatic Recurrence of Breast Cancer. JCO Clin Cancer Inform. 2019(3):1-12.

136. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. J Am Med Inform Assoc. 2013;20(e1):e147-e54.

137. Tamang SR, Hernandez-Boussard T, Ross EG, Gaskin G, Patel MI, Shah NH. Enhanced Quality Measurement Event Detection: An Application to Physician Reporting. EGEMS (Wash DC). 2017;5(1):5.

138. Hochheiser H, Castine M, Harris D, Savova G, Jacobson RS. An information model for computable cancer phenotypes. BMC Med Inform Decis Mak. 2016;16(1):121.

139. Yala A, Barzilay R, Salama L, Griffin M, Sollender G, Bardia A, et al. Using machine learning to parse breast pathology reports. Breast Cancer Res Treat. 2017;161(2):203-11.

140. Breitenstein MK, Liu H, Maxwell KN, Pathak J, Zhang R. Electronic Health Record Phenotypes for Precision Medicine: Perspectives and Caveats From Treatment of Breast Cancer at a Single Institution. Clin Transl Sci. 2018;11(1):85-92.

141. Savova GK, Tseytlin E, Finan S, Castine M, Miller T, Medvedeva O, et al. DeepPhe: A Natural Language Processing System for Extracting Cancer Phenotypes from Clinical Records. Cancer research. 2017;77(21):e115-e8.

142. Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. arXiv preprint arXiv:190605474. 2019.

143. Institute TNCTaCSNT. CDW-H Frequently Asked Questions [Available from: https://tracs.unc.edu/index.php/services/informatics-and-data-science/cdw-h/cdw-h-faq.

144. Institute TNCTaCSNT. EMERSE [Available from: https://tracs.unc.edu/index.php/services/informatics-and-data-science/emerse.

145. Norgeot B, Muenzen K, Peterson TA, Fan X, Glicksberg BS, Schenk G, et al. Protected Health Information filter (Philter): accurately and securely de-identifying free-text clinical notes. npj Digital Medicine. 2020;3(1):57.

146. Montani I, Honnibal M. Prodigy: A new annotation tool for radically efficient machine teaching. Artificial Intelligence to appear. 2018.

147. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002;16:321-57.

148. He H, Bai Y, Garcia EA, Li S, editors. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence); 2008: IEEE.

149. Mani I, Zhang I, editors. kNN approach to unbalanced data distributions: a case study involving information extraction. Proceedings of workshop on learning from imbalanced datasets; 2003: ICML.

150. Tomek I. Two modifications of CNN. IEEE Trans Systems, Man and Cybernetics. 1976;6:769-72.

151. Gianinazzi ME, Rueegg CS, Zimmerman K, Kuehni CE, Michel G, Group SPO. Intra-rater and inter-rater reliability of a medical record abstraction study on transition of care after childhood cancer. PLoS ONE. 2015;10(5):e0124290.

152. Pan H, Gray R, Braybrooke J, Davies C, Taylor C, McGale P, et al. 20-Year Risks of Breast-Cancer Recurrence after Stopping Endocrine Therapy at 5 Years. N Engl J Med. 2017;377(19):1836-46.

153. Pironet A, Poirel HA, Tambuyzer T, De Schutter H, van Walle L, Mattheijssens J, et al. Machine Learning-Based Extraction of Breast Cancer Receptor Status From Bilingual Free-Text Pathology Reports. Frontiers in Digital Health. 2021;3.

5. VALIDATING DRUG REPURPOSING CANDIDATES WITH RETROSPECTIVE CLINICAL ANALYSIS

5.1. **Introduction**

Drug repurposing, or identifying new uses for existing drugs, can reduce the time cost and time needed to put a drug to market. Drug repurposing candidates are FDA-approved drugs and safe for humans, so they do not need to go through Phase I clinical trials. In 2019, the Coronavirus Disease 2019 (COVID-19) global pandemic drove families across the world to stay constrained in their homes until vaccines and treatments for COVID-19 were developed. During this time, drug repurposing became a household term, and the research community responded quickly by adapting existing or developing new drug repurposing informatics pipelines. While many proposed candidates for COVID-19 were not effective in treating the disease, drug repurposing informatics methods were useful for hypothesis generation. Real world data (RWD), like EHR or insurance claims data, are valuable sources of data for candidate generation because they enable researchers to estimate drug effects in humans longitudinally. Much of the evidence that drove drug repurposing during the COVID-19 pandemic was in vitro efficacy. While preclinical in vitro analyses are quick, the results may not translate to efficacy in humans, which was observed in large observational trials with COVID-19 patients using hydroxychloroquine, an anti-malarial drug (154). A well-supported drug repurposing candidate would have support from various data sources like RWD, preclinical evidence, and if possible, clinical trials in similar conditions.

Breast cancer is the most common malignancy among women. With improved outcomes in recent years, patients with early-stage breast cancer have largely been treated for a cure. It is a heterogenous disease on the molecular level, with treatment guided by molecular subtypes of cancer cells. The purpose of this work is to validate drug repurposing candidates in breast cancer with EMR. Rather than selecting one drug repurposing candidate and simulating a clinical trial using EMR, as has been done in previous studies (5), this study proposes a higher level approach to finding drug signals. Non-cancer drug signals were evaluated by conducting a recurrence-free survival analysis using drug features and other covariates. A statistical approach, Cox proportional hazards model, was used as a baseline to compare to a similar study by Wu et al (2019)(155) which focused on discovering non-cancer drug effects on overall survival in patients of various cancer types. Two machine learning approaches, random survival forest and survival support vector machine (SVM), were also developed.

5.2. **Prior Work**

Over the past decades, there has been increasing implementation of 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 (5). 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 (15) to identify key literature associated with drug repurposing and validation in Chapter 2 (p. 7). 10 studies using electronic health records for either drug repurposing candidate prediction or validation were selected. One additional study that did not appear in the literature review

search results is also discussed here. Of the 11 studies, 5 used clinical records in validation and 6 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 15.

Table 15. EMR validation sample sizes in prior work

Study	**Sample size (in patients)**
EMR Use in Validation	
Khatri et al (2013)(6)	2,515
Xu et al (2015)(5)	42,165
Gayvert et al (2016)(22)	–
Gottlieb et al (2014)(21)	–
Xu et al (2018)(156)	–
EMR Use in Prediction	
Paik et al (2015)(157)	530,000
Koren et al (2018)(83)	30,000
Kuang et al (2016)(81) and (2016)(158)	64,515
Low et al (2017)(82)	9,945
Wu et al (2019)(155)	43,310 (first site); 98,366 (second site)

5.2.1. EMR data use in validation

In studies using clinical records for validation, the validation methods used included Cox proportional hazard analysis (5, 6, 22), other statistical analysis (156), and off-label use extraction (21). Of all the studies, Xu et al (2015)(5) 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 retrieved 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 (5).

Other studies using Cox proportional hazards models aimed to associate predicted drug use with treatment success (6, 22, 156). Khatri et al (2013)(6) 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)(22) 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)(156) and Gottlieb et al (2014)(21) did not provide detailed methodologies for their validation processes. Xu et al (2018)(156) provided background for patient record extraction, cohort selection, and stated t-test p-values along with derived conclusions. Gottlieb et al (2014)(21) extracted off-label uses from EMR but did not provide a methodology for the process.

5.2.2. EMR data use in drug candidate prediction

In studies using clinical records for drug candidate prediction, both statistical analysis methods (81, 155, 157, 158) and machine learning methods (82, 83) 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)(83) 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 to predict treatment success for potential anti-hypertensive agents. For cohort identification, Koren et al (2018)(83) only included patients that had at least two initial systolic and diastolic blood pressure values in a given timeframe. Low et al (2017)(82) 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% precision) (82). 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)(82) 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

(82). 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.

Wu et al (2019)(155) used a Cox proportional hazards model for prediction. They aimed to detect non-cancer drug effects from EMR at Vanderbilt University (43,310 patients) and externally validated their model at Mayo Clinic (98,366 patients). Their cohorts included patients with prostate cancer, breast cancer, lung cancer, colorectal cancer, and other unmentioned cancer types. Two clinicians compiled a list of 146 drugs by filtering out known antineoplastic drugs, drugs used to support cancer care, over the counter drugs, and short-term use drugs. The compiled set of drugs was used as the non-cancer drug features. In the model, other covariates included patient demographics, tumor type and stage, and ICD-9-CM diagnosis codes grouped into phenome-association study phenotypes. The covariates were screened with a univariate Cox model, and those with a p-value less than 0.3 were included in the predictive model. The study reported hazard ratios with 95% confidence intervals (CI) for a drug list, which was ranked based on false discovery rate (FDR)-adjusted p-value. Drugs that met the FDR-adjusted p-value < 0.1 cut-off were included in the ranked drug list. To provide supporting evidence for their top predictions, they performed an in-depth biomedical literature search and clinical trial search. 9 drugs were detected at both sites: rosuvastatin, simvastatin, amlodipine, tamsulosin, metformin, omeprazole, warfarin, lisinopril, and metoprolol. Simvastatin and amlodipine both had FDR-adjusted p-values < 0.001.

Three studies used variations of fixed effect models for prediction. Paik et al (2015)(157) 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)(81) 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)(81) 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 (81). 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)(158) 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)(81), the study did not include any details on cohort identification and EHR data extraction (158).

5.3. **Methods**

5.3.1. Subjects

The target population for this study was female breast cancer patients above the age of 18 years. All records from the CDWH for patients in the NCCR between April 4, 2014 and January 13, 2021 (approximately 7 years) were extracted, and patients designated as having breast cancer in the NCCR were included in the cohort.

5.3.2. Study design

A pipeline was created to validate drug candidate predictions in breast cancer. The pipeline consisted of three steps: data extraction and processing, algorithm development, and evaluation. All patient data for patients with a breast cancer diagnosis in the NCCR were extracted from the CDWH and NCCR. CDWH data included were unstructured clinical notes and structured data including PCORnet CDM tables and a clarity table. The task for the study is to predict recurrence as a binary outcome and as a time to event analysis with recurrence-free survival as the outcome. The primary purpose of this study is to examine the significant or important drug features in the classification.

5.3.3. Data preprocessing

The Unified Medical Language System (UMLS) Metathesaurus was used as the primary gateway for information from various vocabularies on procedure codes and medication details. UMLS is maintained by the National Library of Medicine, and the vocabularies accessed in this work are: Current Procedural Terminology (CPT), Healthcare Common Procedural Coding System (HCPCS), and International Classification of Diseases 9^{th} and 10^{th} revisions procedure codes (ICD-9-PCS, ICD-10-PCS) for procedures, and RxNorm for medications and indications.

Data were divided into five categories spanning across the treatment timeline: demographics (i.e., age at diagnosis, race), biopsies, other procedures (i.e., cancer-related surgery, radiation therapy), and all prescriptions classified as breast cancer drugs, other cancer drugs, and non-cancer drugs. Age at diagnosis was calculated by taking the difference between a patient's date of birth and first biopsy date. If a patient did not have a biopsy done at UNC, the age at diagnosis was marked as unknown. Race was extracted from the CDW-H clarity table. Biopsies were extracted after mapping CPT, HCPCS, ICD-10, and ICD-9 procedure codes to the UMLS Metathesaurus and getting code descriptions. Records including the terms "breast" and "biopsy" in the code description were extracted. Cancer-related surgeries and radiation therapy were extracted using keywords such as "mastectomy" or "radiation".

Prescribed medications were first separated into two groups: those with RxNorm concept unique identifiers (CUI) codes and those without codes. Data processing with the RxNorm CUI code group was done in three steps: (1) all medications were searched for in RxNorm by code and medication indications were pulled from the "may_treat" variable, (2) drugs that did not return any results after searching by code were searched by raw name to get indications, and (3) remaining medications with no results were exported to CSV for manual review. With the no code group, steps 2 and 3 were followed. The drug names that did not return any results were cleaned by hand to exclude dosage forms and expand names if they were abbreviated. The medications were then searched by cleaned name to obtain indications. Medications with pain indications were filtered out from the dataset. All medication names were then collapsed by excluding dosage forms with manual review to ensure that all drug features included in the model would be unique drugs, rather than string variations of the same

drug. Also, the UMLS Metathesaurus updated during the data wrangling process, and indication results were returned in different orders. Indications for each drug were reordered to remove string variant duplicates.

An overarching dataset was compiled with all data variables (Table 16) and sorted by patient and date to understand patient timelines. For visualizing patient timelines, breast imaging and Breast Imaging-Reporting and Data System (BI-RADS) scores were used as a marker for disease status. Breast cancer patient notes were filtered to those with a note type including the partial string “imag”, and regular expressions were used to query for notes with BI-RADS mentions in the note text. This dataset was then structured into a dataset suitable for modeling. Imaging scores were not incorporated in the dataset for modeling. Figure 17 shows a visualization of Jane Doe’s (pseudonym for randomly selected female patient) clinical record timeline.

Table 16. EMR drug repurposing dataset structure

DATE_TIME	ENCOUNTERID	PATID	LABEL	CATEGORY	VALUE
Date with or without time depending on the record	Encounter ID for structured data; Note ID for unstructured data; PRESCIRBINGID for medications	Patient ID	For procedures, this is a list with code and description (ex: [19081, breast biopsy]	Categories: • Demographics • Biopsy • Other procedures (surgery / radiation therapy) • Prescriptions (breast cancer drugs, other cancer drugs, non-cancer drugs)	The numerical value for the given row (ex: age in years)

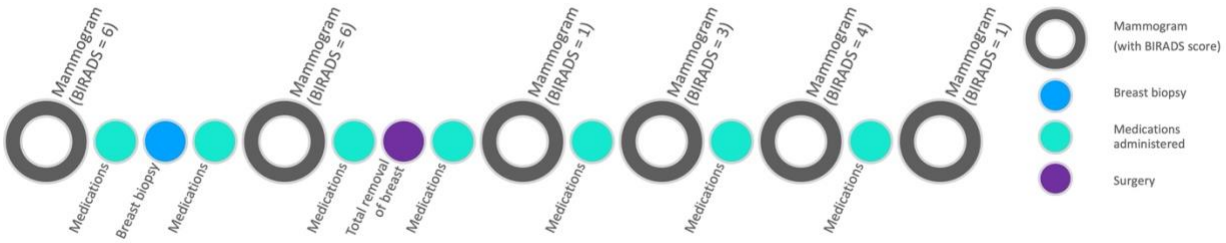


Figure 17. Jane Doe's clinical record timeline

5.3.4. Methods

5.3.4.1. Binary recurrence classification

Machine learning classifiers were used for a binary recurrence classification without time to event data. A machine learning approach was selected over other methods like co-occurrence statistics to account for feature interactions with other covariates. With a machine learning approach, models will be conditioned on feature interactions rather than individual feature relationships with the outcome variable. A random forest classifier and support vector machine were used to predict recurrence status without time to event for breast cancer patients with each row in the dataset corresponding to one patient using covariates such as age at diagnosis group, race, whether the patient had surgery, and whether the patient had radiation therapy. All prescriptions that the patient had received between the date of initial diagnosis and recurrence date were also included as features. The prescription features were binary, such that if a patient had received a prescription, the value for that patient row and prescription column would be 1. The features from the binary classification without time to event were not examined in detail. Two experiments were conducted per model: a baseline

experiment which included cancer-related drugs in the models and a filtered experiment which excluded cancer-related drugs.

5.3.4.2. Recurrence-free survival (time to event) analysis

A multivariate Cox proportional hazards model, random survival forest, and survival SVM were used to compare drug exposures to effect on recurrence-free survival (censored). Overall survival was not a good fit for the primary outcome of this study because the survival rate at UNC for patients who were in the EHR system between April 2014 and January 2021 was 97.5%. All models were implemented with the following features: age at diagnosis group, race, whether the patient had cancer-related surgery or not, whether the patient had radiation therapy or not, and all unique prescriptions the patient had received between the date of initial diagnosis and recurrence. The output variable was recurrence status and time to event. A variable screening was done to only include drugs that had been taken by at least 50 patients in the cohort. Two feature variations after screening were compared to ensure validity of the model: (1) all drug features and (2) excluding breast cancer drugs and drugs meant to treat effects of cancer treatment (e.g., radiation dermatitis). The expectation from the first feature variation would be to see that the standard course of treatment (i.e., tamoxifen) would be most significant.

5.3.4.3. Dataset division for training and testing

The Cox proportional hazards model was implemented with the whole dataset: 2878 (78.85%) patients survived recurrence-free, and 772 (21.15%) patients survived with recurrence. To train and test the machine learning classifiers, the dataset was split into 80% of

the data for training and 20% of the data for testing. 5-fold cross-validation was used for hyperparameter tuning.

5.3.5. Measures of evaluation

5.3.5.1. Binary recurrence classification

Four metrics were examined as measures of evaluation: F1-Score, precision (i.e., positive predictive value), recall (i.e., sensitivity), and ROC-AUC score. ROC-AUC score is the area under the receiver operating characteristic curve. A ROC curve plots the false positive rate against the true positive rate (i.e., recall), and the AUC represents the degree of separability between two classes.

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

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

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

$$False\ positive\ rate = \frac{FP}{TN + FP}$$

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

5.3.5.2. Recurrence-free survival (time to event) analysis

The censored concordance index (C-index) was used to evaluate the Cox proportional hazards model, and hazard ratios with 95% CIs and FDR-adjusted p-values were used to evaluate the features. Censored C-index was also used to evaluate the random survival forest and survival SVM. Feature importance from the random survival forest and survival SVM were

evaluated on the test set with permutation importance and SHapley Additive exPlanations (SHAP). Drug features were considered important if they had at least a 0.1% impact on the model. A Shapley value represents the average feature contribution in making a prediction (159). The premise behind using measures of explainability is to see which drug features are contributing significantly to the recurrence-free survival prediction.

5.3.6. Supporting evidence search

Top drug candidate predictions were searched in relation to breast cancer or any cancer in PubMed to provide supporting evidence to predictions. Studies with preclinical evidence and observational studies were separated and included as evidence. For meta-analysis search results, the manuscripts were examined to pull the original studies. The clinical trials database, ClinicalTrials.gov (160), was also searched with “Breast Cancer” and "Cancer” as the condition or disease and the drug name as the other term. The drug candidate predictions were also discussed with a board-certified medical oncologist to identify any off-label usage of the drugs for breast cancer treatment at UNC. Associated literature evidence and protocols for any off-label usage was also included as supporting evidence.

5.4. **Results**

3650 female patients with breast cancer and timelines of at least two weeks of encounters were included in this study. 2878 patients survived without local or distant recurrence of their breast cancers, and 772 patients had recurrence events. The age at diagnosis ranged from 20 years to 96 years with a median age of 60 years. There were 1141 drugs as starting features, excluding pain medications. Drugs taken by less than 50 patients and meant for short-term conditions (e.g., common cold) or non-therapeutic agents (e.g., imaging

agent) were removed from the dataset. After filtering, 151 drugs remained, including breast cancer therapies. After removing breast cancer therapies, 121 drugs remained.

5.4.1. Baseline results

5.4.1.1. Binary recurrence classification

The baseline experiment was to use all drugs prescribed to patients as features along with all covariates. The random forest classifier and support vector machine were implemented with class weights, weighting the minority (recurrence) class higher than the majority (no recurrence) class. The random forest classifier achieved performance of 80.4% ROC-AUC score, a precision of 75.7%, recall of 71.7%, and F1-Score of 73.3%. The support vector machine achieved a performance of 81.9% ROC-AUC score, precision of 71.9%, recall of 73.3%, and F1-Score of 72.5%.

5.4.1.2. Recurrence-free survival (time to event) analysis

The baseline experiment was to use all drug features, including breast cancer drugs, with the covariates to predict recurrence-free survival and identify drugs based on significance by FDR-adjusted p-value or feature importance. The Cox proportional hazards model had a C-index of 0.83. The drug features with negative coefficients and FDR-adjusted p-values less than 0.1 are shown in Table 17.

Table 17. Top important drugs associated with breast cancer recurrence-survival from Cox proportional hazards model (baseline model: including cancer-related drugs)

Drug Name	**RxNorm Indication**	**Hazard Ratio (95% CI)**	**FDR-adjusted p-value**
Tamoxifen	• Breast neoplasms • Gynecomastia • Precocious puberty • Pancreatic neoplasms	0.29 (0.21 to 0.39)	$p < 0.001$
Anastrozole	• Breast neoplasms	0.24 (0.16 to 0.34)	$p < 0.001$

Letrozole	• Breast neoplasms	0.37 (0.27 to 0.50)	p < 0.001
Ondansetron	• Vomiting	0.40 (0.29 to 0.54)	p < 0.001
Silver sulfadiazine	• Wound infection	0.51 (0.35 to 0.76)	0.0217
Albuterol sulfate	• Asthma • Bronchial spasm	0.57 (0.39 to 0.84)	0.0771
Ibandronate	• Postmenopausal osteoporosis	0.15 (0.04 to 0.60)	0.0938
Atorvastatin	• Coronary artery disease • Hyperlipoproteinemias • Hypertriglyceridemia • Hypercholesterolemia	0.57 (0.37 to 0.87)	0.0979

The random survival forest had a C-index of 0.82, and the top ten drug features were tamoxifen, letrozole, anastrozole, midazolam, pregabalin, celecoxib, docusate sodium, exemestane, silver sulfadiazine, and gabapentin. As expected, the most highly weighted drugs in the model are meant to treat breast cancer. Tamoxifen, letrozole and anastrozole are treatments taken after primary treatment to reduce breast cancer recurrence risk. The survival SVM had a C-index of 0.82 as well. The top ten drug features were tamoxifen, ondansetron, anastrozole, prochlorperazine maleate, letrozole, silver sulfadiazine, cyclophosphamide, carboplatin, ibandronate, and paclitaxel. The feature weights for drugs that both random survival forest and survival SVM found important are shown in Table 18. Features that were also found important using SHAP are marked with an asterisk (*) in the table.

Table 18. Top important drugs associated with breast cancer recurrence-survival from both random survival forest and survival SVM (baseline models: including cancer-related drugs)

Drug Name	**RxNorm Indication**	**Random Survival Forest Feature Weight**	**Survival SVM Feature Weight**
Tamoxifen*	• Breast neoplasms • Gynecomastia • Precocious puberty • Pancreatic neoplasms	0.0184 ± 0.0089	0.0295 ± 0.0168
Letrozole*	• Breast neoplasms	0.0181 ± 0.0093	0.0099 ± 0.0088

Anastrozole*	• Breast neoplasms	0.0177 ± 0.0102	0.0157 ± 0.0110
Midazolam	• Status epilepticus • Psychomotor agitation • Anxiety disorders	0.0132 ± 0.0153	0.0014 ± 0.0033
Celecoxib	• Adenomatous polyposis coli • Rheumatoid arthritis • Ankylosing spondylitis • Dysmenorrhea • Osteoarthritis	0.0033 ± 0.0057	0.0012 ± 0.0008
Silver sulfadiazine*	• Wound infection	0.0021 ± 0.0028	0.0073 ± 0.0068
Diclofenac*	• Inflammation • Rheumatoid arthritis • Ankylosing spondylitis • Photophobia • Dysmenorrhea • Juvenile arthritis • Osteoarthritis • Keratosis	0.0010 ± 0.0013	0.0014 ± 0.0052

5.4.2. Filtered model results

5.4.2.1. Binary recurrence classification

The random forest classifier and support vector machine were implemented after excluding cancer-related drug features. The random forest classifier achieved a performance of 74.8% ROC-AUC score, 68.4% precision, 66.4% recall, and 67.3% F1-Score. Support vector machine achieved a performance of 77.0% ROC-AUC score, 67.8% precision, 70.0% recall, and 68.7% F1-Score.

5.4.2.2. Recurrence-free survival (time to event) analysis

The multivariate Cox proportional hazards model was fit on the whole dataset without breast cancer therapies and had a C-index of 0.79. Seven drug features had negative coefficients with FDR-adjusted p-values less than 0.1, and six of them had FDR-adjusted p-

values less than 0.05. The top nine drug features were: midazolam, silver sulfadiazine, lorazepam, ephedrine, ibandronate, albuterol sulfate, heparin, docusate sodium, and atorvastatin. The drugs with their hazard ratios, 95% CIs, and FDR-adjusted p-values are shown in Table 19.

Table 19. Top non-cancer drugs with improved breast cancer recurrence-free survival from Cox proportional hazards model

Drug Name	**RxNorm Indication**	**Hazard Ratio (95% CI)**	**FDR-adjusted p-value**
Midazolam	• Status epilepticus • Psychomotor agitation • Anxiety disorders	0.42 (0.34 to 0.52)	$p < 0.001$
Silver sulfadiazine	• Wound infection	0.41 (0.28 to 0.60)	$p < 0.001$
Ephedrine	• Rhinitis • Orthostatic hypotension • Asthma • Bronchial spasm	0.61 (0.49 to 0.78)	0.0011
Ibandronate	• Postmenopausal osteoporosis	0.06 (0.02 to 0.24)	0.0017
Albuterol sulfate	• Asthma • Bronchial spasm	0.56 (0.39 to 0.81)	0.0326
Docusate sodium	• Constipation	0.72 (0.57 to 0.90)	0.0564
Atorvastatin	• Coronary artery disease • Hyperlipoproteinemias • Hypercholesterolemia	0.56 (0.37 to 0.85)	0.0675

After filtering out breast cancer drugs and drugs meant to treat cancer treatment effects, the random survival forest had a C-index of 0.78. The top ten drug features by way of permutation importance on the test set were midazolam, celecoxib, ephedrine, phenylephrine, albuterol sulfate, ibandronate, diclofenac, omeprazole, atorvastatin, pantoprazole. The survival SVM also had a C-index of 0.76. The top ten drug features were midazolam, ibandronate, celecoxib, denosumab, lorazepam, phenylephrine, heparin, ephedrine, scopolamine, and pentoxifylline. Features that both random survival forest and survival SVM found important are

shown with feature weights in Table 20. Features that were also found important using SHAP are marked with an asterisk (*) in the table.

Table 20. Top non-cancer drugs associated with breast cancer recurrence-survival from both random survival forest and survival SVM

Drug Name	**RxNorm Indication**	**Random Survival Forest Feature Weight**	**Survival SVM Feature Weight**
Midazolam*	• Status epilepticus • Psychomotor agitation • Anxiety disorders	0.0457 ± 0.0183	0.0429 ± 0.0148
Celecoxib*	• Adenomatous polyposis coli • Rheumatoid arthritis • Ankylosing spondylitis • Dysmenorrhea • Osteoarthritis	0.0202 ± 0.0108	0.0071 ± 0.0060
Ephedrine	• Rhinitis • Orthostatic hypotension • Asthma • Bronchial spasm	0.0105 ± 0.0191	0.0039 ± 0.0058
Albuterol sulfate*	• Asthma • Bronchial spasm	0.0031 ± 0.0043	0.0035 ± 0.0045
Ibandronate	Postmenopausal osteoporosis	0.0030 ± 0.0040	0.0136 ± 0.0172
Atorvastatin*	• Coronary artery disease • Hyperlipoproteinemias • Hypercholesterolemia	0.0019 ± 0.0038	0.0026 ± 0.0023
Famotidine*	• Dyspepsia • Heliobacter infections • Zollinger-Ellison syndrome • Urticaria • Gastroesophageal reflux • Duodenal ulcer • Heartburn • Peptic esophagitis • Stomach ulcer	0.0015 ± 0.0020	0.0036 ± 0.0059
Alendronate*	• Postmenopausal osteoporosis	0.0012 ± 0.0016	0.0020 ± 0.0067

	• Extramammary paget disease		
Heparin*	• Unstable angina • Pulmonary embolism • Postoperative complications • Thrombophlebitis • Thromboembolism • Myocardial infarction • Cerebral infarction • Coronary thrombosis	0.0011 ± 0.0025	0.0042 ± 0.0050

5.5. **Discussion**

A statistical approach and two machine learning approaches to survival analysis were taken to analyze drug exposure effects with respect to recurrence-free survival. A binary classification approach was also taken with two machine learning classifiers to understand performance without including time to event. Top non-cancer drugs from the time to event analyses were identified for each baseline model that included breast cancer drugs and each model after filtering breast cancer drugs from the dataset. Biomedical literature and ClinicalTrials.gov, the clinical trials database in the US, were searched to find supporting evidence for top drug candidate predictions.

Midazolam was a top drug feature in the models, and preclinical evidence has supported its use to treat lung cancer, neurogliomas, pancreatic adenocarcinoma, and hepatocellular carcinoma (161-164). One study by Lu et al (2021)(165) also supported the use of midazolam in inhibiting cell proliferation in breast cancer. Bisphosphonates, which are meant for osteoporosis, such as zoledronic acid and clodronate are being prescribed to breast cancer patients at UNC to prevent metastatic recurrence in the bone. Ibandronate, a bisphosphonate, was a top feature in the models, and there are previous studies that indicate it could be

effective for reducing breast cancer recurrence in the bone (166). However, recent clinical trials have presented that it does not improve disease-free survival or overall survival (167, 168).

Statins are another category of drugs that have been predicted in drug repurposing retrospective clinical analyses (82, 155), and atorvastatin was a top drug feature in this study. Recent preclinical evidence has shown that atorvastatin can trigger cancer cells to undergo necrosis (169). In addition, 58 clinical trials were found for atorvastatin in association with any cancer with 15 clinical trials examining its effects in breast cancer. Most studies are currently recruiting or have unknown status, so we were unable to further support or challenge the association. Similarly, for celecoxib, a top drug feature in random survival forest and survival SVM, 32 clinical trials were found investigating its use for different cohorts of breast cancer patients. A clinical trial found that it can induce COX-2 inhibition which in turn supports anti-tumor activity in primary breast cancer tissue (ClinicalTrials.gov Identifier: NCT01695226)(170). Another clinical trial found that celecoxib did not have an impact on disease-free survival for HER2- breast cancer patients (ClinicalTrials.gov Identifier: NCT02429427)(171). Future studies are warranted to understand whether the drug repurposing candidates predicted in this work will have an impact in breast cancer patients.

This study differs from others in terms of sample size and approach. In prior work, many researchers have examined cohorts with over 10,000 patients over time to conduct large-scale drug screenings. In addition, efforts in the past have largely been hypothesis-driven rather than data-driven, using clinical hypotheses or literature signals to pinpoint drugs and conduct retrospective clinical analyses for a few drugs. In this study, approximately 150 drugs were

examined in a recurrence-free survival analysis, using a high-level data-driven approach rather than a specific hypothesis-driven approach to examine drug effects in a breast cancer cohort.

5.5.1. Limitations

In comparison to other studies that take a high-level approach to retrospective clinical analysis for drug repurposing, the sample size used in this study is very small. Low et al (2017)(82) examined drug repurposing for breast cancer with overall survival as the primary outcome, and their sample size was 9,945 patients, which is 2.7 times larger than the sample size in this study. The limited sample size and imbalanced data led to decreased power, explaining the models' intermediate level of separation between the recurrence and recurrence-free classes. There was also no evaluation for generalizability as all patients came from one institution. To expand the sample size and have more power, future work will consider including patients from the legacy EHR system at UNC and patient data from other institutions in the Carolina Collaborative.

REFERENCES

5. Xu H, Aldrich MC, Chen Q, Liu H, Peterson NB, Dai Q, et al. Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality. J Am Med Inform Assoc. 2015;22(1):179-91.

6. Khatri P, Roedder S, Kimura N, De Vusser K, Morgan AA, Gong Y, et al. A common rejection module (CRM) for acute rejection across multiple organs identifies novel therapeutics for organ transplantation. J Exp Med. 2013;210(11):2205-21.

15. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS medicine. 2009;6(7):e1000097.

21. Gottlieb A, Altman RB. Integrating systems biology sources illuminates drug action. Clin Pharmacol Ther. 2014;95(6):663-9.

22. Gayvert KM, Dardenne E, Cheung C, Boland MR, Lorberbaum T, Wanjala J, et al. A computational drug repositioning approach for targeting oncogenic transcription factors. Cell Rep. 2016;15(11):2348-56.

81. Kuang Z, Thomson J, Caldwell M, Peissig P, Stewart R, Page D. Computational Drug Repositioning Using Continuous Self-Controlled Case Series. KDD. 2016;2016:491-500.

82. Low YS, Daugherty AC, Schroeder EA, Chen W, Seto T, Weber S, et al. Synergistic drug combinations from electronic health records and gene expression. J Am Med Inform Assoc. 2017;24(3):565-76.

83. Koren G, Nordon G, Radinsky K, Shalev V. Machine learning of big data in gaining insight into successful treatment of hypertension. Pharmacol Res Perspect. 2018;6(3):e00396.

154. Martinez MA. Lack of Effectiveness of Repurposed Drugs for COVID-19 Treatment. Front Immunol. 2021;12.

155. Wu Y, Warner JL, Wang L, Jiang M, Xu J, Chen Q, et al. Discovery of Noncancer Drug Effects on Survival in Electronic Health Records of Patients With Cancer: A New Paradigm for Drug Repurposing. JCO Clin Cancer Inform. 2019(3):1-9.

156. Xu C, Ai D, Shi D, Suo S, Chen X, Yan Y, et al. Accurate Drug Repositioning through Non-tissue-Specific Core Signatures from Cancer Transcriptomes. Cell Rep. 2018;25(2):523-35.e5.

157. Paik H, Chung A-Y, Park H-C, Park RW, Suk K, Kim J, et al. Repurpose terbutaline sulfate for amyotrophic lateral sclerosis using electronic medical records. Sci Rep. 2015;5:8580.

158. Kuang Z, Thomson J, Caldwell M, Peissig P, Stewart R, Page D. Baseline Regularization for Computational Drug Repositioning with Longitudinal Observational Data. IJCAI (U S). 2016;2016:2521-8.

159. Molnar C. Interpretable machine learning: Lulu. com; 2020.

160. Zarin DA, Tse T, Williams RJ, Califf RM, Ide NC. The ClinicalTrials. gov results database—update and key issues. New England Journal of Medicine. 2011;364(9):852-60.

161. Seo JA, Jeon HY, Kim M, Lee YJ, Han ET, Park WS, et al. Anti-metastatic effect of midazolam on melanoma B16F10 cells in the lungs of diabetic mice. Biochem Pharmacol. 2020;178:114052.

162. Oshima Y, Sano M, Kajiwara I, Ichimaru Y, Itaya T, Kuramochi T, et al. Midazolam exhibits antitumour and anti-inflammatory effects in a mouse model of pancreatic ductal adenocarcinoma. Br J Anaesth. 2022;128(4):679-90.

163. Qi Y, Yao X, Du X. Midazolam inhibits proliferation and accelerates apoptosis of hepatocellular carcinoma cells by elevating microRNA-124-3p and suppressing PIM-1. IUBMB Life. 2020;72(3):452-64.

164. Wang C, Datoo T, Zhao H, Wu L, Date A, Jiang C, et al. Midazolam and Dexmedetomidine Affect Neuroglioma and Lung Carcinoma Cell Biology In Vitro and In Vivo. Anesthesiology. 2018;129(5):1000-14.

165. Lu HL, Wu KC, Chen CW, Weng HK, Huang BM, Lin TY, et al. Anticancer Effects of Midazolam on Lung and Breast Cancers by Inhibiting Cell Proliferation and Epithelial-Mesenchymal Transition. Life (Basel). 2021;11(12).

166. Adjuvant bisphosphonate treatment in early breast cancer: meta-analyses of individual patient data from randomised trials. Lancet. 2015;386(10001):1353-61.

167. Livi L, Scotti V, Desideri I, Saieva C, Cecchini S, Francolini G, et al. Phase 2 placebo-controlled, single-blind trial to evaluate the impact of oral ibandronate on bone mineral density in osteopenic breast cancer patients receiving adjuvant aromatase inhibitors: 5-year results of the single-centre BONADIUV trial. Eur J Cancer. 2019;108:100-10.

168. Vliek SB, Noordhoek I, Meershoek-Klein Kranenbarg E, van Rossum AGJ, Dezentje VO, Jager A, et al. Daily Oral Ibandronate With Adjuvant Endocrine Therapy in Postmenopausal Women With Estrogen Receptor-Positive Breast Cancer (BOOG 2006-04): Randomized Phase III TEAM-IIB Trial. J Clin Oncol. 2022:Jco2100311.

169. Abolghasemi R, Ebrahimi-Barough S, Bahrami N, Ai J. Atorvastatin Inhibits Viability and Migration of MCF7 Breast Cancer Cells. Asian Pac J Cancer Prev. 2022;23(3):867-75.

170. Brandão RD, Veeck J, Van de Vijver KK, Lindsey P, de Vries B, van Elssen CH, et al. A randomised controlled phase II trial of pre-operative celecoxib treatment reveals anti-tumour transcriptional response in primary breast cancer. Breast Cancer Res. 2013;15(2):R29.

171. Coombes RC, Tovey H, Kilburn L, Mansi J, Palmieri C, Bartlett J, et al. Effect of Celecoxib vs Placebo as Adjuvant Therapy on Disease-Free Survival Among Patients With Breast Cancer: The REACT Randomized Clinical Trial. JAMA Oncol. 2021;7(9):1291-301.

6. CONCLUSION

This research had two aims: (1) to produce a computational phenotyping algorithm using electronic medical records, and (2) to build a pipeline for retrospective clinical record analysis to validate drug repurposing candidates. Under the first aim, there are two angles to computational phenotyping. The first angle examines using embedding models and foundational natural language processing methods to predict oral cancer risk with pathology notes. The second angle examines data extraction from clinical notes and structured EMR data on outcomes and tumor characteristics to facilitate breast cancer research. Under the second aim, a drug repurposing study examining drug effects in breast cancer with recurrence-free survival analysis was conducted.

In Chapter 3, the study aimed to determine the likelihood of a patient having oral cancer by using spaCy pipelines and embedding models (e.g., SciBERT) to analyze microscopic descriptions of oral samples from dental pathology notes. Mortality and morbidity rates of oral cancer decrease with early diagnosis, and risk stratification is an open research area. We presented an approach to predict whether oral pathology samples have high probability of being cancerous or not to aid pathologists in deciding to refer a patient to an oncologist. Researchers have tackled this problem with approaches like biomarker discovery, but there is a lack of NLP studies focusing on oral cancer and very few on head and neck cancers, providing an opportunity to explore NLP methods for oral cancer.

In this study, the developed approach presents a unique use case and examination of using natural language processing methods on to classify pathology reports. This work could potentially streamline the referral process for clear cases of cancer and provide clinical decision support for cases with clinical features that may not traditionally be associated with cancer. Further research in the project will consist of connecting patient dental records to clinical records to understand whether patients with pre-cancerous symptoms in dental records later developed oral cancer as well.

In Chapter 4, the work demonstrated the feasibility of phenotyping a breast cancer cohort by identifying breast cancer recurrence and extracting tumor receptor status mentions. A binary classification approach was used to identify breast cancer recurrence in clinical notes, and manual code-based filtering and chart review were done for patients without notes that mention recurrence or without notes at all. A named entity recognition approach with a pre-trained word vector pipeline was used to extract tumor receptor status mentions from clinical notes of all types (e.g., pathology reports, progress notes). This chapter focused on building a pipeline for data extraction to enable breast cancer research.

Computational phenotyping for cohort characterization and stratification is becoming increasingly important for researchers to produce findings that can be clinically relevant and applicable. There are significant amounts of time and effort devoted to manual chart abstraction by subject matter experts and researchers, which creates a large bottleneck for progress in clinical research. Structured data contains many kinds of data like patient demographics, medications, procedure, diagnoses, and encounters, but the structured data does not provide context or insight into a patient's lifestyle, family history, social history, and

more. In addition, often when data are presented like for a procedure, there are no details on the outcomes of the procedure in the structured data. All contextual information is contained in the clinical notes. The notes are also heterogeneous in that clinicians have writing patterns that change over time, depending on their training and personal styles. This makes clinical natural language processing a difficult and interesting task but also presents the problem of portability. The goal behind developing natural language processing algorithms is that they could be useful beyond the institutions they were developed in. In terms of portability, structured data is standardized, making it easier to evaluate in new settings. The unstructured data may have some standard note types such as progress notes, but many notes are unlabeled. Therefore, using structured data and clinical notes individually for clinical research can pose problems that a hybrid approach could overcome. In Chapter 4, a hybrid approach using clinical notes and structured EMR data for breast cancer recurrence detection was developed. Future directions consist of testing on data from external sites and other cancers to test for both portability and generalizability by condition.

For tumor receptor status extraction, the clinical notes are the only source of information, presenting an opportunity to develop clinical information extraction models. There have been many efforts to produce scalable clinical information extraction tools that can be used across sites in different locations; however, developing portable computational phenotyping tools based on clinical notes is still an active research area. Researchers are developing standards and protocols for tool development that could tackle the problem of portability, but the state of research is not at the point of tools working out of the box. Training and fine-tuning on local institutional data is necessary to achieve reasonable results. To extract

tumor receptor status for patients, a pre-trained word vector pipeline was fine-tuned on annotated clinical notes for named entity recognition. Future research will consist of developing the tumor receptor status model to identify minority named entities like HER2+ cases with better performance and stratify receptor status even further by detecting scores for each receptor type.

In Chapter 5, statistical and machine learning approaches to breast cancer drug repurposing with retrospective clinical analysis were presented using recurrence-free survival as the primary outcome. Models were developed and compared with different feature variations, and top predictions were examined in biomedical literature and the clinical trials database, providing two forms of supporting evidence. In retrospective clinical analysis, there are multiple barriers to conducting drug repurposing research. While EHRs contain large amounts of rich data, the data are not prepared for research, leading to significant efforts in preparing and wrangling data into a form that is ready for analysis. Therefore, the work in Chapter 4 was used to facilitate the drug repurposing work in Chapter 5. This study was unique to the area of drug repurposing in that data variables from unstructured clinical notes and structured EMR data were combined to compile the dataset. Also, the study had a small sample size in comparison to other prior studies but still produced drug repurposing candidates that were supported by biomedical literature and clinical trials. In terms of technical approach, this study is innovative in that it takes an explainable machine learning approach to drug repurposing EHR data, while using a statistical model for comparison. Past retrospective clinical analyses with machine learning methods have used logistic regression for propensity score matching or predicting survival as a binary variable. To the best of my knowledge, there is a lack

of studies in the literature investigating drug repurposing in EHR with a machine learning time to event analysis. A limitation to using explainable machine learning approaches is that with limited sample size and high dimensionality, it is difficult to find accurate separation between imbalanced classes. However, using different feature variations and incorporating supporting evidence for evaluation helped to mitigate this weakness. Future work consists of providing more technical validity for drug candidate predictions by expanding the dataset to include breast cancer data from other institutions. The primary goal of this work is to make a meaningful contribution to clinicians and patients; therefore, after providing technical validity, future research will consist of taking some drug candidate predictions to clinic for evaluation.

This work adds to the growing body of literature in computational phenotyping and EHR-based drug repurposing, and it also has many implications for clinical research. Clinicians are overloaded with information that will continue to grow. Humans have a cognitive bias, the availability heuristic, where decisions are made based on readily available information, indicating that if the information needed to make a decision is not readily available, it likely will not be included in the decision-making process. With growing data sizes, retrospective clinical analysis will require pipelines that can be reused and adapted to different tasks. The computational phenotyping pipeline developed in this research can be expanded upon and used as a starting point for future cancer research with retrospective EHR data. With the heavy cost and time burden of drug development, drug repurposing is an attractive alternative solution for hypothesis generation. Using EHR data for drug repurposing has the unique advantage of being able to observe a patient cohort over time and see drug effects on outcomes either from primary treatment or treatments for co-morbidities. While EHR data can

be a powerful data source, considerations for data security, patient privacy, and data preparation have made it difficult for researchers to fully exploit the data for drug repurposing. This study advances on past work in the field and marks a step forward for using retrospective clinical analysis for drug repurposing.

Appendix 1. Guidelines for Annotating Clinical Notes

Guidelines for annotating patient clinical notes as indicating breast cancer recurrence or not:

- Reasons for marking a patient as having breast cancer recurrence:
    - Clearly stated diagnosis
        - "recurrent breast cancer"
        - "breast cancer recurrence"
    - Recurrence based on patient history and current diagnosis
        - Having a diagnosis in one year, and having the same diagnosis in a following year with verbiage describing it as recurrence
        - "breast cancer [YEAR] local recurrence [YEAR]"
    - Distant recurrence
        - The patient had breast cancer and is receiving treatment for removal of a cancerous lesion in another part of the body.
- Reasons for marking a patient as not having breast cancer recurrence
    - Clearly stated lack of recurrence
        - "Without sx/sxs of recurrence"
        - "No evidence of recurrent disease"
    - Note only describing how to reduce risk of recurrence
    - Recurrence of other conditions (not breast cancer)
    - Past recurrent breast cancer pre-2014 (outside data range)
    - New primary

        - "The patient has a recurrence, which is actually a new primary in her left breast."

Guidelines for annotating patient clinical notes with mentions of tumor receptor status:

- Reasons for rejecting notes:
    - If a note does not contain any mention of the status of a patient anywhere in the note, it should be ignored. In other words, if there are no named entities to highlight, reject the note.
        - Example: "The patient estrogen receptor status is unspecified."
- Do not annotate reference ranges:
    - reference range: estrogen receptor and progesterone receptor: <1%=negative; =or>1% = positive her2/neu: 0, 1=negative for overexpression, 2=equivocal, 3=positive for overexpression
- How to annotate:
    - Schema organization:
        - Named entity labels are bold
        - Examples for how entities in a particular label may be written in the clinical notes are located under each bolded label. Annotation examples are underlined.
    - Types:
        - Estrogen receptor
            - **ER+**
                - estrogen receptor positive

- er positive
- er+
- estrogen-receptor-positive
- strongly positive estrogen receptor
- strongly positive er
- weakly positive estrogen receptor
- weakly positive er
- Examples directly from note: (portion underlined is how the annotation should be)
  - estrogen receptor 95% positive (3+)
  - grade 2 er+

- **ER-**
  - estrogen receptor negative
  - er negative
  - er-
  - estrogen-receptor-negative
  - Examples directly from note: (portion underlined is how the annotation should be)
    - estrogen receptor (ventana, clone sp1): interpretation: negative comp uter-assisted quantitative score: 0%

- It is not necessary to annotate the score for ER status.

- Progesterone receptor
  - **PR+**
    - progesterone receptor positive
    - pr positive
    - pr+
    - progesterone-receptor-positive
    - strongly positive progesterone receptor
    - strongly positive pr
    - weakly positive progesterone receptor
    - weakly positive pr
    - Examples directly from note: (portion underlined is how the annotation should be)
      - <u>progesterone receptor 80% positive</u> (2+)
  - **PR-**
    - progesterone receptor negative
    - pr negative
    - pr-
    - progesterone-receptor-negative
    - Examples directly from note: (portion underlined is how the annotation should be)

        - progesterone receptor (ventana, clone 1e2): interpretation: negative computer-assisted quantitative score: 0%
            - It is not necessary to annotate the score for PR status.

- HER2 (human epidermal growth factor receptor 2)
    - HER2 status can be tricky to denote. Always rely on the score if there is one.
    - **HER2-**
        - her2-
        - her2 negative
        - her2/neu negative
        - her2neu negative
        - her2/neu-
        - her2neu-
        - h2n negative
        - h2n-
        - her2 low
        - h2n low
        - her2/neu low
        - her2- (0)
        - her2 negative (0)

- her2/neu negative (0)
- her2neu negative (0)
- her2/neu- (0)
- her2neu- (0)
- h2n negative (0)
- h2n- (0)
- her2- (1)
- her2 negative (1)
- her2/neu negative (1)
- her2neu negative (1)
- her2/neu- (1)
- her2neu- (0)
- h2n negative (1)
- h2n- (1)
- her2 low (1+)
- her2/neu low (1+)
- her2neu low (1+)
- h2n low (1+)
- her2 equivocal
- her2 equivocal (2)
- her2/neu equivocal (2)
- her2neu equivocal (2)

    - h2n equivocal (2)
    - her2 low (2+)
    - her2/neu low (2+)
    - her2neu low (2+)
    - h2n low (2+)
    - her2 2+ fish not amplified
    - her2 2+ fish nonamplified
- **HER2+**
    - her2 positive
    - her2+
    - her2/neu+
    - h2n+
    - her2 2+ fish amplified
    - her2 positive (2+fish)
    - her2+ (2+fish)
    - her2/neu+ (2+fish)
    - her2neu+ (2+fish)
    - h2n+ (2+fish)
    - her2 positive (3)
    - her2+ (3)
    - her2/neu+ (3)
    - her2neu+ (3)

        - h2n+ (3)
        - Examples directly from note: (portion underlined is how the annotation should be)
            - her2/neu (ventana, clone 4b5, fda-approved): interpretation: positive for overexpression computer-assisted quantitative score: 3+
            - invasive portion receptors were er/pr negative, her2 3+.
- Combinations
    - Triple-negative
        - **TNBC**
            - tnbc
            - ER-/PR-/HER2-
            - estrogen receptor negative, progesterone receptor negative, her2 negative
            - Examples directly from note: (portion underlined is how the annotation should be)
                - patient reportedly with triple negative breast cancer
                - pt's history of tnbc
    - Triple-positive

        - This should only be used as a label if there are no spaces between positive tumor receptor status mentions.
        - **TPBC**
            - er/pr/her2+
            - er+/pr+/her2+
            - hr+her2+
- Hormone receptor status
    - **ER_PR+**
        - er/pr+
        - er/pr positive
    - **ER_PR-**
        - er/pr-
        - er/pr negative
        - Examples directly from note: (portion underlined is how the annotation should be)
            - invasive portion receptors were <u>er/pr negative</u>.

Appendix 2. Breast Cancer Patient Count per Drug

The number of breast cancer patients who have taken each drug are shown in the table below.

Drug Name	Patient Count
SODIUM CHLORIDE	2426
ONDANSETRON	2420
MIDAZOLAM	2314
CEFAZOLIN	1960
PHENYLEPHRINE	1490
EPHEDRINE	1423
DOCUSATE SODIUM	1278
PREGABALIN	1246
ONDANSETRON HCL	1218
HEPARIN	1061
LORAZEPAM	877
LETROZOLE	830
DIPHENHYDRAMINE	772
PEGFILGRASTIM	735
GLYCOPYRROLATE	701
GABAPENTIN	669
SULFAMETHOXAZOLE-TRIMETHOPRIM	652
FAMOTIDINE	651
ANASTROZOLE	644
POLYETHYLENE GLYCOL	606
TRIAMCINOLONE ACETONIDE	511
DOCETAXEL	455
ALBUTEROL SULFATE	451
NEOSTIGMINE METHYLSULFATE	450
SILVER SULFADIAZINE	408
OMEPRAZOLE	380
PREDNISONE	368
SCOPOLAMINE	367
PACLITAXEL	365
AZITHROMYCIN	357
PANTOPRAZOLE	354
ATORVASTATIN	350
ENOXAPARIN	349
EXEMESTANE	338
LEVOTHYROXINE	330

FLUTICASONE PROPIONATE	307
LEVOFLOXACIN	302
AMLODIPINE	299
HYDROCHLOROTHIAZIDE	295
VANCOMYCIN	251
CLINDAMYCIN HCL	249
CARBOPLATIN	240
TRASTUZUMAB	239
CIPROFLOXACIN	233
BENZONATATE	223
FUROSEMIDE	221
METRONIDAZOLE	220
NYSTATIN	219
ALBUMIN, HUMAN	217
LOSARTAN	213
METFORMIN	212
TRAZODONE	206
BACITRACIN ZINC	194
SENNOSIDES	190
DOXYCYCLINE HYCLATE	187
MAGNESIUM SULFATE	187
METOPROLOL TARTRATE	184
METOPROLOL SUCCINATE	182
VALACYCLOVIR	177
CLINDAMYCIN	172
INSULIN	172
CEFTRIAXONE	172
SERTRALINE	171
MAGNESIUM	170
ESCITALOPRAM	167
DEXTROSE	160
PERTUZUMAB	158
METHYLENE BLUE	158
VASOPRESSIN	154
LORATADINE	151
METHYLPREDNISOLONE	149
CEFEPIME	145
GENTAMICIN	145
ALENDRONATE	144

NITROGLYCERIN	144
VITAMIN D3	137
CITALOPRAM	137
ALPRAZOLAM	137
IBANDRONATE	134
CETIRIZINE	134
ZOLPIDEM	134
BUPROPION HCL	130
MOMETASONE	124
PEG 3350-ELECTROLYTES	121
MELOXICAM	121
CLONAZEPAM	119
METHYLPREDNISOLONE SODIUM SUCCINATE	116
TC-99M-MEDRONATE SODIUM	115
MELATONIN	115
PRAVASTATIN	108
HYDRALAZINE	105
CEFDINIR	105
DENOSUMAB	104
ALTEPLASE	104
OSELTAMIVIR	103
SODIUM BICARBONATE	102
ALUMINUM-MAG HYDROXIDE-SIMETHICONE	99
IRON	98
METOCLOPRAMIDE	96
DEXAMETHASONE SODIUM PHOSPHATE	95
DIPHENOXYLATE-ATROPINE	95
BISACODYL	95
INSULIN LISPRO	93
CALCIUM CARBONATE	93
SUCRALFATE	92
MONTELUKAST	92
CLOBETASOL	91
CEPHALEXIN	90
IPRATROPIUM BROMIDE	89
VITAMIN D2	88
MECLIZINE	86
CARVEDILOL	86
MIRTAZAPINE	86

PIPERACILLIN-TAZOBACTAM	82
SIMVASTATIN	79
MUPIROCIN	78
OLANZAPINE	72
NICOTINE	71
BALANCED SALT IRRIGATION SOLUTION	71
PENTOXIFYLLINE	71
SPIRONOLACTONE	69
LOPERAMIDE	65
CAPECITABINE	58
GLIPIZIDE	58
ZOLEDRONIC ACID	55
DICYCLOMINE	54
HYDROXYZINE	54
METHYLPREDNISOLONE ACETATE	54
CLOTRIMAZOLE	54
DIPHTH,PERTUSSIS(ACEL),TETANUS	52
TRANEXAMIC ACID	52
SIMETHICONE	52
ERYTHROMYCIN	51
BUDESONIDE-FORMOTEROL	50
DICLOFENAC	49
VITAMIN B12	48
FLUTICASONE-SALMETEROL	48
FOLIC ACID	47
NOREPINEPHRINE BITARTRATE	46
CYANOCOBALAMIN (VIT B-12)	46
ROSUVASTATIN	44
SUMATRIPTAN	44
BUSPIRONE	43
MULTIVITAMIN	43
TROPICAMIDE	42
DOXYCYCLINE MONOHYDRATE	42
PAPAVERINE	42
FERUMOXYTOL	41
CLOTRIMAZOLE-BETAMETHASONE	41
MOXIFLOXACIN	40
OFLOXACIN	40
HYDROCORTISONE ACETATE	39

TRETINOIN	38
TIOTROPIUM BROMIDE	38
THROMBIN (RECOMBINANT)	38
TAMSULOSIN	38
PROPRANOLOL	38
PROCHLORPERAZINE EDISYLATE	38
PROCHLORPERAZINE MALEATE	37
LATANOPROST	36
HYDROXYCHLOROQUINE	35
QUETIAPINE	35
OXYBUTYNIN CHLORIDE	34
TBO-FILGRASTIM	33
PAROXETINE	33
AZELASTINE	33
POTASSIUM PHOSPHATE	33
FILGRASTIM	32
LEUPROLIDE	32
MAGNESIUM CITRATE	32
OLOPATADINE	31
OMEPRAZOLE MAGNESIUM	31
HYDROCORTISONE	30
POTASSIUM PHOSPHATES-MBASIC AND DIBASIC	30
COLCHICINE	30
BUDESONIDE	30
FULVESTRANT	30
MUPIROCIN CALCIUM	29
FEXOFENADINE	29
CHOLECALCIFEROL (VITAMIN D3)	29
POLYMYXIN B SULFATE	28
VITAMIN E	28
POTASSIUM CHLORIDE	27
ASCORBIC ACID (VITAMIN C)	27
DICLOXACILLIN	27
THROMBIN (BOVINE)	27
AMPICILLIN-SULBACTAM	27
EMPAGLIFLOZIN	26
CARBOXYMETHYLCELLULOSE SODIUM	26
KETOCONAZOLE	26
LEVETIRACETAM	26

ALLOPURINOL	25
RALOXIFENE	25
DEXTROMETHORPHAN-GUAIFENESIN	24
MIRABEGRON	24
GLIMEPIRIDE	24
CALCIUM	24
INSULIN NPH ISOPHANE	24
EZETIMIBE	24
CARBACHOL	24
CONJUGATED ESTROGENS	23
ADO-TRASTUZUMAB EMTANSINE	23
SYNTHROID	23
APIXABAN	23
MEROPENEM	23
HALOPERIDOL LACTATE	22
ENALAPRIL MALEATE	22
SITAGLIPTIN	22
GOSERELIN	22
LISINOPRIL	22
TRIAMTERENE-HYDROCHLOROTHIAZIDE	22
LOSARTAN-HYDROCHLOROTHIAZIDE	22
ACYCLOVIR	22
SOLIFENACIN	21
ERTAPENEM	21
ROPINIROLE	21
CICLOPIROX	21
LACTULOSE	21
AMOXICILLIN	21
TOREMIFENE	20
OXYTOCIN	20
ATROPINE	20
LAMOTRIGINE	20
HYDROXYZINE HCL	20
SODIUM PHOSPHATE	19
CLINDAMYCIN PHOSPHATE	19
THIAMINE HCL (VITAMIN B1)	19
NICOTINE TRANSDERMAL PATCH	19
TERBINAFINE HCL	19
CLARITHROMYCIN	19

AMIODARONE	18
DONEPEZIL	18
BIMATOPROST	18
MAGNESIUM HYDROXIDE	18
FLUOROURACIL	18
AZTREONAM	18
THROMBIN (HUMAN)-FIBRINOGEN-APROTININ-CALCIUM	18
WHITE PETROLATUM-MINERAL OIL	17
FLUOCINONIDE	17
THYROID (PORK) TABLET	17
BECLOMETHASONE DIPROPIONATE	17
TOBRAMYCIN	17
HYDROXYZINE PAMOATE	17
PHENTERMINE	17
RIVAROXABAN	17
DORZOLAMIDE	17
RIFAXIMIN	17
VALSARTAN	16
RIZATRIPTAN	16
OXYMETAZOLINE	16
CIPROFLOXACIN-DEXAMETHASONE	16
ISOSORBIDE MONONITRATE	16
UREA	16
METHIMAZOLE	16
FLUOCINOLONE	16
LOVASTATIN	15
TIMOLOL MALEATE	15
LEVOCETIRIZINE	15
HYDROCORTISONE SODIUM SUCCINATE	15
HYLAN G-F 20	14
HYOSCYAMINE	14
NYSTATIN-TRIAMCINOLONE	14
HYDROGEN PEROXIDE	14
ESZOPICLONE	14
POLYMYXIN B SULFATE-TRIMETHOPRIM	14
SENNA	14
BETAMETHASONE DIPROPIONATE	14
RANITIDINE	14
FOSFOMYCIN TROMETHAMINE	13

NEOMYCIN	13
NIFEDIPINE	13
FLURBIPROFEN	13
INSULIN ASPART	13
PSYLLIUM	13
BRIMONIDINE	13
HYOSCYAMINE SULFATE	13
RISPERIDONE	13
CHOLESTYRAMINE	13
SODIUM,POTASSIUM,MAG SULFATES	13
DIAZEPAM	13
CEFUROXIME AXETIL	13
LINACLOTIDE	12
UDENYCA	12
EPINEPHRINE HCL	12
PROTAMINE	12
POVIDONE-IODINE	12
HYALURONIDASE	12
RAMIPRIL	12
NITROFURANTOIN MACROCRYSTAL	12
OCTREOTIDE ACETATE	12
VITAMIN B COMPLEX TABLET	12
MINOCYCLINE	12
PALBOCICLIB	12
NICOTINE GUM	11
ATOVAQUONE-PROGUANIL	11
GLUCAGON	11
AMPICILLIN	11
MICONAZOLE NITRATE	11
CANAGLIFLOZIN	11
LANSOPRAZOLE	11
PIOGLITAZONE	11
LEVALBUTEROL	11
PENICILLIN V POTASSIUM	11
ADENOSINE	10
NEOMYCIN-BACITRACIN-POLYMYXIN	10
BENZTROPINE	10
FLUOXETINE	10
BENZOCAINE	10

ROLAPITANT	10
CISPLATIN	10
TACROLIMUS	10
PSEUDOEPHEDRINE	10
ELIQUIS	10
ARIPIPRAZOLE	10
ARFORMOTEROL	10
OLMESARTAN	10
DEXAMETHASONE	10
PRAMIPEXOLE	10
INSULIN HUMAN	10
GLYCERIN	9
NALOXONE	9
BACITRACIN	9
HALOBETASOL PROPIONATE	9
DESVENLAFAXINE SUCCINATE	9
SODIUM POLYSTYRENE SULFONATE	9
CARBAMAZEPINE	9
LINZESS	9
AMOXICILLIN-POTASSIUM CLAVULANATE	9
DEXTROAMPHETAMINE-AMPHETAMINE	9
TEMAZEPAM	8
CHLORHEXIDINE GLUCONATE	8
TERCONAZOLE	8
XARELTO	8
MINERAL OIL-HYDROPHIL PETROLAT	8
AMMONIUM LACTATE	8
EXENATIDE	8
COLLAGENASE CLOSTRIDIUM HISTOLYTICUM	8
UMECLIDINIUM	8
ECONAZOLE	8
LINAGLIPTIN	8
IMIQUIMOD	8
LUBIPROSTONE	8
DESONIDE	8
RAMELTEON	8
MIDODRINE	8
HYDROQUINONE	8
LOTEPREDNOL ETABONATE	8

BENAZEPRIL	8
AROMASIN	8
CALCIPOTRIENE	7
BUMETANIDE	7
DIGOXIN	7
DILTIAZEM	7
MEMANTINE	7
BEVACIZUMAB	7
HALOPERIDOL	7
TOLTERODINE	7
TROSPIUM	7
VERAPAMIL	7
IRBESARTAN	7
IRON SUCROSE	7
KETOTIFEN	7
TORSEMIDE	7
FLUOROMETHOLONE	7
DAPAGLIFLOZIN	7
SODIUM FERRIC GLUCONATE COMPLEX	7
NEULASTA	7
DULAGLUTIDE	7
DIVALPROEX	7
FILGRASTIM-SNDZ	7
DRONABINOL	7
VALSARTAN-HYDROCHLOROTHIAZIDE	6
RISEDRONATE	6
PRASTERONE (DHEA)	6
PRAZOSIN	6
DEXLANSOPRAZOLE	6
FEBUXOSTAT	6
FAMCICLOVIR	6
DESVENLAFAXINE	6
VITAMIN K1	6
BENAZEPRIL-HYDROCHLOROTHIAZIDE	6
SEVELAMER CARBONATE	6
NEBIVOLOL	6
ESTRADIOL	6
AZELAIC ACID	6
ADVAIR	6

FENOFIBRATE NANOCRYSTALLIZED	6
MEDROXYPROGESTERONE	6
BISACODYL ENEMA	6
FILGRASTIM-AAFI	6
CANDESARTAN	6
CARBIDOPA-LEVODOPA	6
VILAZODONE	6
ZIPRASIDONE	6
ZINC SULFATE	6
METHYLPHENIDATE	6
DAPTOMYCIN	6
TRASTUZUMAB EMTANSINE	6
IRON DEXTRAN	6
ENALAPRIL-HYDROCHLOROTHIAZIDE	5
AMINOCAPROIC ACID	5
DAPSONE	5
SODIUM HYPOCHLORITE	5
OMEGA-3 ACID	5
EFLORNITHINE	5
TRIAMCINOLONE	5
UMECLIDINIUM-VILANTEROL	5
PRENATAL VITAMIN-CALCIUM-IRON-FOLIC ACID	5
NORETHINDRONE-ETHINYL ESTRADIOL	5
AMLODIPINE-BENAZEPRIL	5
AMLODIPINE-VALSARTAN	5
EPIRUBICIN	5
HYDROXYUREA	5
PROAIR HFA	5
PYRIDOSTIGMINE BROMIDE	5
LORCASERIN	5
NEOMYCIN-POLYMYXIN-HYDROCORTISONE	5
FLUTICASONE FUROATE-UMECLID-VILANT	5
CALCITRIOL	5
LACOSAMIDE	5
ACETIC ACID	5
MESALAMINE	5
DOBUTAMINE	5
PENICILLIN G BENZATHINE	5
TAMOXIFEN	5

BISOPROLOL-HYDROCHLOROTHIAZIDE	5
FERROUS GLUCONATE	5
MANNITOL	5
TELMISARTAN	5
SANTYL	5
VORTIOXETINE	5
OXACILLIN	5
RIFAMPIN	5
GUAIFENESIN	5
MYCOPHENOLATE MOFETIL	5
PERMETHRIN	4
LEFLUNOMIDE	4
IVERMECTIN	4
ZALEPLON	4
VINORELBINE	4
PILOCARPINE	4
SODIUM IODIDE	4
ONABOTULINUMTOXINA	4
TELMISARTAN-HYDROCHLOROTHIAZIDE	4
SOTALOL	4
SODIUM ACETATE	4
SPIRONOLACTONE-HYDROCHLOROTHIAZIDE	4
SULFACETAMIDE SODIUM	4
MINERAL OIL ORAL	4
SULFASALAZINE	4
SUPREP BOWEL PREP KIT	4
SEVELAMER HCL	4
TAZAROTENE	4
SALICYLIC ACID	4
VALPROATE SODIUM	4
RITUXIMAB	4
LUPRON DEPOT	4
MVI NO.4 WITH VIT K	4
TETRACYCLINE	4
PYRIDOXINE (VITAMIN B6)	4
NIVESTYM	4
TRELEGY ELLIPTA	4
PROGRAF	4
OMEGA 3-DHA-EPA-FISH OIL	4

JANUMET	4
ZONISAMIDE	4
AMILORIDE	4
BISOPROLOL FUMARATE	4
ANORO ELLIPTA	4
FELODIPINE	4
COLESEVELAM	4
DABIGATRAN ETEXILATE	4
FESOTERODINE	4
FEXOFENADINE-PSEUDOEPHEDRINE	4
FIDAXOMICIN	4
FLUDROCORTISONE	4
IODINE-POTASSIUM IODIDE	4
FLUOROMETHOLONE ACETATE	4
CALCIUM-VITAMIN D3	4
FLUTICASONE FUROATE	4
DESOXIMETASONE	4
CALCITONIN (SALMON)	4
DESLORATADINE	4
GEMFIBROZIL	4
DRONEDARONE	4
FERRIC SUBSULFATE	4
CEFUROXIME SODIUM	4
CEFTAZIDIME	4
ENTRESTO	4
ESOMEPRAZOLE MAGNESIUM	4
TETRAHYDROZOLINE	3
FLUOCINOLONE ACETONIDE	3
THYROID (PORK)	3
FLUTICASONE FUROATE-VILANTEROL	3
OCTREOTIDE	3
THYROID	3
BORIC ACID	3
METHYLERGONOVINE	3
NORGESTIMATE-ETHINYL ESTRADIOL	3
EPOETIN ALFA	3
METHOTREXATE SODIUM	3
DESMOPRESSIN	3
ELUXADOLINE	3

URSODIOL	3
METHENAMINE HIPPURATE	3
NORETHINDRONE ACETATE-ETHINYL ESTRADIOL	3
DESCOVY	3
GENTAMICIN-VANCOMYCIN	3
POTASSIUM ACETATE	3
BROMPHENIRAMINE-PSEUDOEPHEDRINE	3
POTASSIUM PHOSPHATE, MONOBASIC	3
DARBEPOETIN ALFA	3
TRULICITY	3
PHENOBARBITAL	3
DEXILANT	3
EFINACONAZOLE	3
NALTREXONE	3
TREPROSTINIL SODIUM	3
OXALIPLATIN	3
OXCARBAZEPINE	3
BENZOYL PEROXIDE	3
MUCINEX	3
BACITRACIN-POLYMYXIN B	3
TRIAZOLAM	3
PEG3350-SOD SULF-NACL-KCL-ASCORBATE-C	3
BIOTIN	3
TRASTUZUMAB-DKST	3
FENOFIBRATE	3
BISMUTH SUBSALICYLATE	3
NEOMYCIN-POLYMYXIN-DEXAMETHASONE	3
MINERAL OIL	3
TRAMETINIB	3
MINERAL OIL ENEMA	3
TRIMETHOPRIM	3
ATEZOLIZUMAB	3
BIVALIRUDIN	3
AMANTADINE HCL	3
EPINEPHRINE	3
GLUTAMINE	3
CALCIUM ACETATE	3
CANAGLIFLOZIN-METFORMIN	3
PROGESTERONE	3

SODIUM PICOSULFATE-MAGNESIUM OXIDE-CITRIC ACID	3
LORATADINE-PSEUDOEPHEDRINE	3
VITAMIN B2	3
LACTOBACILLUS ACIDOPHILUS	3
LITHIUM CARBONATE	3
VITAMIN C	3
SODIUM PHOSPHATES	3
CEFTAROLINE FOSAMIL	3
COLESTIPOL	3
HYPROMELLOSE	3
ACETAZOLAMIDE	3
CAPSAICIN	3
SILVER NITRATE	3
CIMETIDINE	3
ISOPROTERENOL	3
VIIBRYD	3
SACUBITRIL-VALSARTAN	3
MAGNESIUM CHLORIDE	3
ZOLMITRIPTAN	3
CYCLOSPORINE MODIFIED	3
LYRICA	3
CARBAMIDE PEROXIDE	3
PRIMIDONE	3
LURASIDONE	3
ITRACONAZOLE	3
CROMOLYN	3
ACYCLOVIR SODIUM	3
RHO(D) IMMUNE GLOBULIN	3
OTEZLA	2
CEFOXITIN	2
TERBUTALINE	2
ORLISTAT	2
DOLUTEGRAVIR	2
REPAGLINIDE	2
SITAGLIPTIN-METFORMIN	2
TERAZOSIN	2
CEFADROXIL	2
ELTROMBOPAG	2
BAMLANIVIMAB	2

SILDENAFIL	2
B-COMPLEX WITH VITAMIN C TABLET	2
CLEVIDIPINE	2
CLARITIN-D	2
AZELASTINE-FLUTICASONE	2
TENOFOVIR DISOPROXIL FUMARATE	2
AZATHIOPRINE	2
OLAPARIB	2
CILOSTAZOL	2
CHLORPHENIRAMINE	2
SACCHAROMYCES BOULARDII	2
SALMETEROL	2
TIROSINT	2
CETIRIZINE-PSEUDOEPHEDRINE	2
OXYBUTYNIN	2
PENCICLOVIR	2
REMDESIVIR	2
COSENTYX	2
PHENYTOIN SODIUM	2
CALTRATE	2
PHYTONADIONE (VITAMIN K1)	2
CRESTOR	2
CRISABOROLE	2
PIMECROLIMUS	2
SYMBICORT	2
PREMARIN	2
TIVICAY	2
SUVOREXANT	2
CYCLOPHOSPHAMIDE	2
POLYSACCHARIDE IRON COMPLEX	2
POSACONAZOLE	2
PREDNISOLONE SODIUM PHOSPHATE	2
PREDNISOLONE ACETATE	2
BOOSTRIX TDAP	2
DEXTRAN-HYPROMELLOSE	2
BETAMETHASONE, AUGMENTED	2
PERFOROMIST	2
DIPHENHYDRAMINE-ZINC ACETATE	2
PAROXETINE MESYLATE	2

CLOZAPINE	2
RALTEGRAVIR	2
DICLOFENAC-MISOPROSTOL	2
PEMBROLIZUMAB	2
CYPROHEPTADINE	2
RACEPINEPHRINE	2
RABEPRAZOLE	2
QUININE	2
SODIUM FLUORIDE	2
CORTISONE	2
SODIUM NITROPRUSSIDE	2
DEXTROMETHORPHAN HBR	2
PRUCALOPRIDE	2
NIZATIDINE	2
SAXAGLIPTIN	2
MEGESTROL	2
ACETYLCYSTEINE	2
MICAFUNGIN	2
MICARDIS	2
FLOVENT	2
FLECAINIDE	2
MILNACIPRAN	2
MINOXIDIL	2
MODAFINIL	2
MOMETASONE FUROATE	2
VITAMINS A,C,E-ZINC-COPPER	2
INDAPAMIDE	2
INFLIXIMAB	2
MULTAQ	2
EZETIMIBE-SIMVASTATIN	2
LEUCOVORIN CALCIUM	2
EVOLOCUMAB	2
ETOPOSIDE	2
TRULANCE	2
AFLIBERCEPT	2
MYCOPHENOLATE SODIUM	2
EPLERENONE	2
VENTOLIN HFA	2
GLYBURIDE	2

GLUCOPHAGE	2
VALTREX	2
GENTIAN VIOLET	2
ALCLOMETASONE	2
LIOTHYRONINE	2
GEMCITABINE	2
METHOTREXATE	2
VITAMIN D3-FOLIC ACID	2
VALPROIC ACID	2
IBRANCE	2
IMATINIB	2
ALBIGLUTIDE	2
VALGANCICLOVIR	2
FOSAMAX	2
ETHAMBUTOL	2
MOMETASONE-FORMOTEROL HFA	2
XELJANZ	2
ZARXIO	2
JARDIANCE	2
ATOMOXETINE	2
ACAMPROSATE	2
NIACIN	2
NERATINIB	2
NEORAL	2
ESOMEPRAZOLE SODIUM	2
ZINC OXIDE	2
MYFORTIC	2
ESTRACE	2
TRASTUZUMAB-HYALURONIDASE-OYSK	2
ARMOUR THYROID	2
VESICARE	2
NADOLOL	2
APREMILAST	2
ALREX	1
SORBITOL	1
SELEXIPAG	1
SELENIUM	1
ZIOPTAN	1
SELENIUM SULFIDE	1

SECUKINUMAB	1
ABILIFY	1
VITAMIN B6	1
ALEMTUZUMAB	1
ABEMACICLIB	1
ALMOTRIPTAN MALATE	1
SODIUM CITRATE-CITRIC ACID	1
CALCIUM-VITAMIN D2	1
VIT C-VIT E-COPPER-ZINC OX-LUTEIN	1
SOFOSBUVIR-VELPATASVIR	1
ACETYLCHOLINE CHLORIDE	1
VIT A-C-E-LUTEIN-MINERALS	1
VISMODEGIB	1
ALPELISIB	1
ABATACEPT	1
CERTOLIZUMAB PEGOL	1
CARAFATE	1
CEFPODOXIME	1
ZINC CHLORIDE	1
ZAFIRLUKAST	1
YUVAFEM	1
SODIUM CITRATE	1
ZINC ACETATE	1
WARFARIN	1
CEFUROXIME	1
VORICONAZOLE	1
CELEBREX	1
VOLTAREN	1
VIVELLE-DOT	1
SENEXON-S	1
CELECOXIB	1
SPIRIVA	1
VITAMIN-BIOTIN	1
YUPELRI	1
CARBINOXAMINE	1
CANDESARTAN-HYDROCHLOROTHIAZIDE	1
XIIDRA	1
XANAX	1
XELODA	1

ADCIRCA	1
SAXAGLIPTIN-METFORMIN	1
VERTEPORFIN	1
ATENOLOL	1
BEPOTASTINE BESILATE	1
TENOFOVIR ALAFENAMIDE	1
APRACLONIDINE	1
BENICAR	1
ARGATROBAN	1
ARIKAYCE	1
BENDAMUSTINE	1
TETANUS AND DIPHTHERIA TOX	1
TETANUS IMMUNE GLOBULIN	1
TRAVOPROST	1
TRASTUZUMAB-ANNS	1
ATACAND	1
THROMBIN (HUMAN)-FIBRIN-APROT-CA	1
AZILSARTAN MEDOXOMIL	1
AMPHOTERICIN B	1
THROMBIN(HM PLAS)-FIBRIN-APROT-CA	1
TORSEMIDE 100 MG TABLET	1
ATIVAN	1
THYROTROPIN ALFA	1
TIGECYCLINE	1
AYR SALINE NASAL GEL	1
TIMOLOL	1
TOLNAFTATE	1
AVATROMBOPAG	1
AUVI-Q	1
TITANIM DIOX-ZINC OXIDE-HOMOSALATE-OCTINOXATE-MERADIMATE	1
TOBRAMYCIN-DEXAMETHASONE	1
TOBRADEX	1
TELMISARTAN-AMLODIPINE	1
TEGRETOL	1
ALUMINUM CHLORIDE	1
AMIKACIN	1
VEMLIDY	1
CALCIPOTRIENE-BETAMETHASONE	1
VEDOLIZUMAB	1

CAFFEINE	1
ALUMINUM HYDROX-MAGNESIUM CARB	1
STUDY ONDANSETRON	1
BROVANA	1
SULFACETAMIDE SODIUM-SULFUR	1
AMBIEN	1
BROMOCRIPTINE	1
BREXPIPRAZOLE	1
AMBRISENTAN	1
BREO ELLIPTA	1
BORTEZOMIB	1
BEVESPI AEROSPHERE	1
AMINOPHYLLINE	1
SYNDROS	1
AMITIZA	1
AMLODIPINE-ATORVASTATIN	1
AMMONIUM LACTATE-SODIUM LACTATE-POTASSIUM LACTATE	1
AMOXAPINE	1
TADALAFIL	1
TALIMOGENE LAHERPAREPVEC	1
AMOXICILLIN-CLARITHROMYCIN-LANSOPRAZOLE	1
TRIPTORELIN PAMOATE	1
TAZORAC	1
TRIHEXYPHENIDYL	1
TRIFLUOPERAZINE	1
JAKAFI	1
CYCLOSPORINE	1
SARECYCLINE	1
METHENAMINE MAND-SOD BIPHOS	1
FLAVOXATE	1
MICONAZOLE	1
MEXILETINE	1
FLUCONAZOLE	1
FLUMAZENIL	1
METOLAZONE	1
FONDAPARINUX	1
FORMOTEROL FUMARATE	1
METHYLDOPA	1
FOSINOPRIL-HYDROCHLOROTHIAZIDE	1

MITOMYCIN	1
MESNEX	1
MESNA	1
FROVATRIPTAN	1
GANCICLOVIR	1
MECOBALAMIN-LEVOMEFOLATE CALCIUM-PYRIDOXAL PHOS	1
GENVOYA	1
GLIPIZIDE-METFORMIN	1
GLYBURIDE-METFORMIN	1
MAFENIDE	1
MIFEPRISTONE	1
MOEXIPRIL	1
GLYCINE DILUENT INTRAVENOUS SOLUTION	1
NATURE-THROID	1
EPTIFIBATIDE	1
ERGOTAMINE TARTRATE	1
NICARDIPINE	1
NEXIUM	1
ESMOLOL	1
NERLYNX	1
ESTRADIOL-NORETHINDRONE ACETATE	1
NEO-KIDNEY AUGMENT SELECTED RENAL CELLS	1
ESTROPIPATE	1
NATEGLINIDE	1
FEMARA	1
NASCOBAL	1
NARCAN	1
NALTREXONE HCL	1
ETANERCEPT	1
NALOXEGOL	1
MYRBETRIQ	1
MYCOPHENOLATE	1
ETHACRYNIC ACID	1
FACTOR VIIA	1
MACITENTAN	1
LULICONAZOLE	1
CETUXIMAB	1
IRBESARTAN-HYDROCHLOROTHIAZIDE	1
INDACATEROL	1

INGENOL MEBUTATE	1
INTRAROSA	1
LETAIRIS	1
LENALIDOMIDE	1
INVOKAMET	1
LATUDA	1
LASIX	1
LAPATINIB	1
LANOXIN	1
LEVOMEFOLATE CA-B6-MEB12-ALGAL OIL	1
IRON,CARBONYL-VITAMIN C	1
ISENTRESS HD	1
LABETALOL	1
KEVZARA	1
ISOSORBIDE DINITRATE	1
KETOPROFEN	1
JULUCA	1
JANUVIA	1
IVABRADINE	1
ABALOPARATIDE	1
LEVOMEFOLATE-ALGAL OIL	1
GLYCOPYRROLATE-FORMOTEROL	1
HYDROCORTISONE VALERATE	1
LOXAPINE SUCCINATE	1
LOVENOX	1
GOLIMUMAB	1
GRALISE	1
GRANIX	1
HORIZANT	1
HUMATE-P	1
HUMIRA	1
HYALURONIDASE, HUMAN RECOMBINAN	1
HYDROCORTISONE-IODOQUINOL	1
LEVOMILNACIPRAN	1
LISDEXAMFETAMINE	1
LIOTRIX	1
LINEZOLID	1
LIFITEGRAST	1
LIALDA	1

LEXAPRO	1
LEVOXYL	1
INCRUSE ELLIPTA	1
LEVONORGESTREL-ETHINYL ESTRADIOL	1
NICOTROL	1
EPOETIN ALFA-EPBX	1
NITRO-BID	1
PROBENECID	1
PYLERA	1
COREG	1
PSEUDOEPHEDRINE-GUAIFENESIN	1
PROTONIX	1
PROPYLENE GLYCOL	1
COUMADIN	1
PROPAFENONE	1
COZAAR	1
PROCHLORPERAZINE	1
PROBENECID-COLCHICINE	1
CONJUGATED ESTROGEN-MEDROXYPROGESTERONE	1
CRIZOTINIB	1
PRISTIQ	1
PREDNISOLONE	1
JANTOVEN	1
CYTARABINE	1
DALTEPARIN (PORCINE)	1
DAUNORUBICIN	1
POTASSIUM CITRATE	1
POTASSIUM CITRATE-CITRIC ACID	1
COPAXONE	1
COLCRYS	1
EPINASTINE	1
CLOBETASOL-EMOLLIENT	1
RUXOLITINIB	1
CHLORDIAZEPOXIDE	1
ROMIPLOSTIM	1
CHLORTHALIDONE	1
RIVASTIGMINE	1
CIPRO HC	1
CLEMASTINE	1

RIOCIGUAT	1
RILPIVIRINE	1
RIFABUTIN	1
COENZYME	1
RIBAVIRIN	1
REXULTI	1
REVEFENACIN	1
RESTASIS	1
REPATHA	1
RENOVA	1
REFRESH RELIEVA	1
RANOLAZINE	1
RANIBIZUMAB	1
DEMECLOCYCLINE	1
DESLORATADINE-PSEUDOEPHEDRINE	1
PLEGRIDY	1
ELETRIPTAN	1
DOXORUBICIN	1
DOXYLAMINE SUCCINATE	1
DULERA	1
DUPILUMAB	1
OMEGA3-DHA-EPA-OTHER OMEGA3S-FISH OIL	1
OMEGA-3 FATTY ACIDS-FISH OIL	1
DUPIXENT	1
OMALIZUMAB	1
OMADACYCLINE	1
OLMESARTAN-AMLODIPINE-HYDROCHLOROTHIAZIDE	1
PITAVASTATIN CALCIUM	1
ELVITEG-COB-EMTRICIT-TENOFO ALAFENAM	1
ELVITEG-COB-EMTRICIT-TENOFO DISOPRO	1
EMTRICITABINE-TENOFOVIR DISOPROXIL FUMARATE	1
ENBREL	1
NXSTAGE RFP	1
NP THYROID	1
NORETHINDRONE	1
ENTECAVIR	1
NORETHINDRONE ACETATE	1
OPSUMIT	1
DOXAZOSIN	1

OSPEMIFENE	1
DOPAMINE	1
PIROXICAM	1
PINDOLOL	1
DETROL LA	1
DEXMETHYLPHENIDATE	1
PHENOXYBENZAMINE	1
PHENOBARBITAL SODIUM	1
DEXRAZOXANE HCL	1
DEXTROAMPHETAMINE	1
PENICILLIN G SODIUM	1
PENICILLIN G POTASSIUM	1
DEXTROMETHORPHAN POLISTIREX	1
DIFLUCAN	1
DILANTIN	1
DIOVAN	1
PALIPERIDONE	1
DISULFIRAM	1
DOCOSANOL	1
DOFETILIDE	1
DOLUTEGRAVIR-RILPIVIRINE	1
ABACAVIR-LAMIVUDINE	1