2018-04-23

Using machine learning methods to improve chronic disease case definitions in primary care electronic medical records

Lethebe, Brendan Cord

UNIVERSITY OF CALGARY

Using machine learning methods to improve chronic disease case definitions in primary care electronic medical records

by

Brendan Cord Lethebe

A THESIS
SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
DEGREE OF MASTER OF SCIENCE

GRADUATE PROGRAM IN COMMUNITY HEALTH SCIENCES

CALGARY, ALBERTA
April, 2018

I Abstract

Background: Chronic disease surveillance at the primary care level is becoming more feasible with the increased use of electronic medical records (EMRs). However, the quality of surveillance information is directly dependent on the quality of the case definitions that identify the conditions of interest.

Purpose: To determine whether machine learning algorithms can produce chronic disease case definitions comparable to committee created case definitions in a primary care EMR setting.

Methods: A chart review was conducted for the presence of hypertension, diabetes, osteoarthritis, and depression in a cohort of 1920 patients from the Canadian Primary Care Sentinel Surveillance Network database. The results of this chart review were used as training data. The C5.0, Classification and Regression Tree, Chi-Squared Automated Interaction Detection decision trees, Forward Stepwise logistic regression, Least Absolute Shrinkage and Selection Operator penalized logistic regression were compared using 10-fold cross validation. Sensitivity, specificity, positive predictive value and negative predictive value were estimated and compared for the four chronic conditions of interest.

Results: Validity measures were similar across algorithms. For hypertension, sensitivity ranged between 93.1-96.7%, while specificity ranged from 88.8-93.2%. For diabetes, sensitivities ranged from 93.5-96.3% with specificities between 97.1-99.0%. For osteoarthritis, sensitivities ranged from 82.0-84.4% with specificities between 92.7-94.0%. For depression, sensitivities went from 81.4-88.3%, and specificities ranged from 93.4-94.9%. Compared with the committee-created case definitions, these metrics were equivalent or better using the machine learning method.

Conclusions: Machine learning algorithms produced accurate case definitions comparable to committee-created case definitions. It is possible to use machine learning techniques to develop high quality case definitions from EMR data.

II Preface

Chapter 2 is primarily a summary a machine learning algorithms that are commonly used for classification problems in the published literature. Several of these algorithms are not implemented in this project, but they are important to provide context for section 2.1, where we summarize the studies that have used machine learning for similar purposes using electronic medical record data. For readers that are experienced with machine learning, chapter 2 can be safely skipped.

Chapter 3.3 is a summary of previously published work by Williamson et. al in 2014. The results of this study represent committee-created case definition development, which are to be directly compared with the results in chapter 5. It is also important to report the committee-created case definitions, to be compared with the machine learning case definitions in terms of their complexity, and number of features included.

Chapter 5 has been submitted to the CMAJ Open and is currently under review.

Chapter 6 is a general discussion that expands upon the interpretation of Chapter 5.

III Acknowledgements

Thanks to the members of my supervisory committee: Tolulope Sajobi, Hude Quan, and Paul Ronksley. Your contributions were invaluable. A special consideration for my supervisor, Tyler Williamson. Thank you for taking a shot on me.

IV Dedication

For my mother, Ranger Randi Lethebe.

Table of Contents

List of Tables

List of Figures and Illustrations

V Abbreviations

- AIC - Akaike Information Criterion
- ANN - Artificial Neural Network
- AUC - Area Under the Curve
- AUROC - Area Under the Receiver Operator Characteristic
- CaRT - Classification and Regression Trees
- CHAID - CHi-squared Automated Interaction Detection
- CPCSSN - Canadian Primary Care Sentinel Surveillance Network
- DAD - Discharge Abstract Database
- EHR - Electronic Health Record
- EMR - Electronic Medical Record
- FN - False Negatives
- FP - False Positives
- HBA1C - Haemoglobin A1c or glycated haemoglobin test
- ICD - International Classification of Diseases
- KNN - K Nearest Neighbours
- LASSO - Least Absolute Shrinkage and Selection Operator
- NPV - Negative Predictive Value
- PPV - Positive Predictive Value
- RPART - Recursive PARTitioning
- SQL - Structured Query Language

- SVM - Support Vector Machine
- TN - True Negatives
- TP - True Positives

1 Background

1.1 Chronic Disease

Chronic diseases are especially taxing to the healthcare system and lives of Canadians, as they afflict a large portion of the population for long periods of time. This adversely affects the well-being of many, and constitutes a large portion of the cost to the healthcare system. It is estimated that among Canadians aged 55 to 74, the prevalence of multi-morbidity (the presence of 2 or more chronic conditions) is about 60% [1]. Other studies have reported that half of those suffering from chronic conditions have more than one chronic condition [2]. Medical costs for patients with chronic conditions are very high. A 2004 estimate suggests that the cost of chronic disease care in Canada constitutes 42% of the direct healthcare costs. Furthermore, diseases often have indirect costs as well, which are the costs incurred from loss of productivity [3]. Including cardiovascular disease, these conditions are estimated to be associated with approximately three quarters of all deaths in the country [4]. Additionally, those suffering from chronic diseases are more frequent users of emergency departments, and tend to be hospitalized more often than those without chronic conditions [5].

1.1.1 Chronic Disease Surveillance

Chronic disease surveillance is essential for understanding the burden of chronic disease in Canada, supporting healthcare planning, resource allocation, identifying populations of interest for intervention or research, and monitoring trends of disease over time [6]. High quality surveillance is achieved only with a large volume of quality data, and quality tools for identifying patients. Traditionally, these data sources were primarily comprised of physician reimbursement records. The emergence of the electronic medical record (EMR) has provided another source of data, and a quality set of case definitions are the necessary tools for the job.

1.2 Administrative Data

Administrative data is health data that is routinely collected in administering the healthcare system. This method of data collection is referred to as “passive” or “secondary”, as it is often collected for another primary purpose [7]. This type of data is collected at all levels of healthcare [8]. They tend to be simple entries that summarize the care given. These summaries often use coding

systems or ontologies to standardize the information, making it common across sites and regions. There are many coding systems used throughout the world to succinctly detail the care received by a patient. In many cases, these codes are submitted by a healthcare provider to a government or health insurance company for reimbursement. This process has earned the coding system the moniker "billing codes". This is not an entirely accurate name for the coding systems though, as many uses of these codes are not related to billing and reimbursement whatsoever.

One of the first studies to propose the use of administrative data as a source for research was in 1973 [9]. This study compared hospital admission and utilization rates in various communities in Vermont, depending on their proximity to hospitals. Since then, it has been highly debated whether administrative health data, which is viewed as a secondary source of data, is appropriate for research at the population level. It is considered secondary data since, when used for research purposes, it is being used for a purpose secondary to it's primary purpose. Here the administrative data is collected for internal documentation and billing purposes, and research is a secondary purpose. For this reason, some are skeptical of the merits of administrative data for population surveillance [10]. Another source of concern is the lack of variables available, as some doubt the accuracy or interpretation of billing and procedure codes [11, 12, 13]. Furthermore, using administrative data sources for surveillance has shown to be poor for conditions that are underdiagnosed, such as depression [14]. Some conditions that are mainly treated at the primary care level have also proven to be difficult to identify using administrative data.

On the other hand, administrative data is appealing to many others. Administrative data boasts the kind of population coverage that primary data collection cannot attain. This type of coverage allows for surveillance and study of cohorts with rare conditions with low prevalence that would be difficult to identify using other methods. Other studies have shown that billing and procedural codes are reliable and accurate [15, 16]. Administrative data can be accessed from many different sites, and doesn't require active enrollment of patients into studies. The time and cost-effective nature of administrative data have led many to use it for research purposes [17]. Due to the simplicity and similarity of administrative data sources, some success has been found by using administrative data sources from both Canada and the United States in the same study [18]. This shows the potential of administrative data for large-scale or even aggregated studies. Some organizations are developing data models and software that allow for the sharing of codes and methods across

research institutions and data sources [19].

1.2.1 Data in Inpatient Settings

The Discharge Abstracts Database (DAD) is a Canadian example of administrative data routinely collected in an inpatient setting. This type of data and others similar to it have been used by researchers to identify patients with certain conditions for quite some time [20]. When a patient is discharged from the hospital, the paper chart is sent to be "coded" and an administrative data entry is made corresponding to that visit. This is an invaluable tool for identifying patients that suffer from acute conditions that typically result in a hospitalization, such as stroke [21]. The DAD contains procedure codes and International Classification of Diseases (ICD) codes that summarize a patient's care and do not contain lab values, clinical notes, radiology/pathology reports, or clinical measures [22].

It can be seen that inpatient administrative data sources are useful for the identification of patients with conditions that are primarily treated in the inpatient setting. It is possible to identify chronic conditions that are typically treated in outpatient settings, such as hypertension, in inpatient administrative data with some degree of success [20]. This may not be true for all conditions though, as some conditions tend to be treated solely in outpatient settings. A potentially more thorough method of identifying chronic disease cases with administrative data is to use data from primary care, where the bulk of the care for chronic conditions actually occurs.

1.2.2 Administrative Data in Primary Care Settings

Traditionally, administrative data from outpatient data sources such as ambulatory care and family physician offices have been lacking. The bulk of the data that has come from the primary care level has been physician claims data, or billing codes along with some prescription medication information. In the primary care setting, patients come in for a wide variety of purposes and receive care with varying amounts of documentation on the paper chart, and coding without the assistance of trained "coders" in most cases. This has caused some to believe that traditional primary care administrative data is incomplete and unreliable. It is difficult to ascertain the quality of ambulatory billing records in terms of how they capture the wide array of unique interactions with the healthcare system. These problems have been partially addressed by the emergence of the electronic medical record (EMR).

1.3 Electronic Medical Records (EMR)

The electronic medical record (EMR) is an emerging tool for the collection of data from all levels of healthcare. The EMR aims to provide a comprehensive record of the health care system that can be accessed quickly and efficiently. This is a natural extension of the increased utilization of technology in all aspects of society. There is evidence that shows EMRs are helpful in increasing efficiency and reducing costs [23]. These findings are disputed elsewhere, where there are claims that the use of EMRs may slow down workflows by as much as 48 minutes per clinic day [24].

EMRs are important tools in quantifying the resource expenditure in the healthcare system, but they also increase the ability to share data. A computerized record can easily be accessed and shared by a clinician to a patient, or to a researcher. This versatility also makes an EMR a promising data source for surveillance and observational research purposes.

The EMR often contains administrative billing data, medications, encounter diagnoses, problem lists, laboratory values, demographics, and examination data. This differentiates the EMR from some of the more traditional administrative data sources that primarily include ICD-9 billing codes and CPT procedural codes. The EMR has the capacity to provide a much clearer picture of the care provided in all settings of the healthcare system. Another advantage of the EMR is that it relies less on the coding systems and ontologies that were so heavily used in other administrative data sources. The EMR has opened up the door for clinician free-text to be used in conjunction with coding systems to gain additional insight into a patient's care and condition.

One of the central complications of the current EMR landscape is that there is a wide variety of EMR systems in use throughout the Canadian healthcare system. Vendors sell EMR systems with vastly different data collection instruments, data formats, and structure of backend databases. Simple EMR systems may not include laboratory or examination data, while more complex systems may include a wide array of other information. This makes it challenging to provide population-level surveillance across multiple EMR systems, but there are methods in current use (and more being developed) that will standardize EMR data.

The effectiveness of EMR usage in various healthcare settings has been evaluated in some publications. Studies have been conducted to determine the economic benefit of EMR implementation, showing an increase in productivity [25]. These studies showed that institutions benefited in terms of resource management, cost control, and quality assessment. Also, a committee estimated that

the time saved by using the EMR will enable half of physicians to add one patient visit per day after two years of implementation [26].

In primary care settings, the implementation of EMR systems has shown mixed results. While there have been claims that EMR systems would have a positive impact on the quality of care, several quality indicator studies have been performed with mixed results. Ultimately, there is very little evidence to suggest that the use of an EMR is associated with an improvement in the quality of care [27, 28]. Another study has shown that EMRs can have a positive impact on record-keeping of medication prescriptions, leading to more efficient dispensing of medications, and a decrease in adverse drug events [29].

Some evidence has been shown that the benefit of an EMR system is also dependent on the reimbursement system in which a clinic operates. The return on investment seems to be lower for those in a fee-for-service setting and higher in capitation settings [30]. Fee-for-service is a payment model in which a physician gets reimbursed for each treatment provided. The capitation system, generally speaking, is when a physician is reimbursed a set amount for each patient assigned to them [31].

Some concern has been raised that the use of an EMR system at the primary care level can have an adverse impact on the patient-clinician relationship, but overall the benefits outweigh the negatives [32].

It has been shown that that EMR systems have high up-front costs, with an initial decrease in productivity due to documentation and training, but that cost-savings are realized over time. Once implemented, the electronic systems have shown to have process and structural benefits that increase overall productivity [33]. Although the literature contains critiques of the implementation of an EMR, it is clear that EMRs are here to stay.

Perhaps the biggest barrier to the efficacy of EMR systems is the lack of standardization. There are a variety of different EMR vendors that offer systems with a great deal of disparity. The key differences include the type and amount of data collected for each patient, the coding vocabulary, and the consistency of data quality checks. Also, at the site level, differences in practice may lead to a disparity in the quality of data. This makes it extremely difficult to provide chronic disease surveillance in a population when it is often difficult to compare one EMR system to another.

1.3.1 Differences between EMRs and Electronic Health Records (EHR)

There has been some ambiguity regarding the difference between the EMR and the electronic health record (EHR), partially due to the infancy of the research and implementation of electronic records in health. In some circles, these terms are used interchangeably. This is not appropriate and an important distinction should be made here.

An EMR is essentially a clinical data repository, operated by the organization in which it is implemented. This is where clinical, billing, demographic, medication and other data are stored, primarily for use by the institution operating it.

An EHR is a patient-oriented record. An EHR typically is only functional if EMR systems already exist, as it is comprised of data from multiple levels of healthcare. This is sometimes referred to as cross-institutional [34]. In many cases, EHRs are owned by the patient (or other stakeholder), and are used to describe the overall health of an individual patient, with data from multiple sources. Often a patient can even append information to their EHR.

With these definitions, it is clear that there is a difference between an EMR and an EHR. An EMR is primarily for use by the organization providing the care, and often contains information from just one site, especially at the primary care level [35]. The EHR, on the other hand, is a much broader entity, sometimes drawing from multiple levels of care and attempting to describe the overall health of a patient [36].

This study does not deal with EHR data and will focus entirely on building case definitions from extracted EMR data.

1.4 The Potential of the EMR for Surveillance

With the widespread adoption of the EMR in primary care in Canada, the opportunity for chronic disease surveillance has increased dramatically. A 2014 study estimated that 56% of physicians in Canada use an EMR [37]. Case identification before the advent of the EMR relied heavily on billing codes. The EMR provides the opportunity for case definition development using a wider array of information. Rather than simply using billing codes, case definitions can include physician free text, medication prescriptions, problem list entries, laboratory values, and more. By having access to more data, the case definitions used to identify those suffering from specific conditions can be more accurate, thereby improving disease surveillance.

There are many studies in which a case definition has been implemented in an EMR (in inpatient and outpatient settings) and the validity estimated [38, 39, 40, 41]. In order to estimate the validity of a case definition, a reference standard must first be established.

1.5 Case Definitions

A case definition is a set of criteria that defines a condition in a specific context. If a person's record meets that criteria, they are considered a case (for the condition of interest). Case definitions can vary in complexity and can be implemented in different settings. For example, consider case definitions for hypertension.

The Mayo Clinic definition for hypertension is "a common condition in which the long-term force of the blood against your artery walls is high enough that it may eventually cause health problems, such as heart disease"[42]. While that may be a comprehensive clinical definition, it would be very difficult to implement at a population level. Perhaps a more relevant definition of hypertension for identification would be a systolic blood pressure measurement over 140 mmHg, or a diastolic blood pressure above 90 mmHg. Another case definition may require two or more blood pressure measurements over 140 mmHg and 90 mmHg within 1 year. These specific and measureable criteria are much better suited to population surveillance in a database containing blood pressure measurements.

Case definitions have been applied to administrative data sources for many years and are primarily built on ICD billing codes [20, 43, 44]. As with all classification procedures, case definitions may not perfectly identify the intended patient group. Patients may not meet the case definition even if they are true cases, or patients who do not actually have the disease may falsely be assumed to be cases.

In the database setting, a case definition appears as a set of rules that result in a binary classification: disease positive or disease negative. Accordingly, the quality of any surveillance system in a population is directly related to the accuracy of the case definition (or definitions) used to identify the condition(s) of interest. A strong set of case definitions needs to be developed and validated for use in each particular data source to ensure that we make the most of the data sources available. The validation step requires the establishing of a reliable reference standard for a sample that represents a population.

1.6 Reference Standard

A reference standard is an indication of true disease status for a patient. A gold standard is a type of reference standard that is regarded as absolutely true for a disease [45]. Unfortunately, in health research reference standards are rarely considered as gold standards due to some ambiguity around the true status of disease. Since we are not able to establish a "gold standard" for disease status, we will have to rely on a reference standard that is perhaps the most common, the chart review.

A chart review is a reference standard used for administrative health data, including EMR data. This is a manual review of a patient record by a trained abstractor to determine disease status. These reviews are typically preceded by the development of a list of criteria in the chart that would define the disease. In an EMR setting, it is common for a manual review of the EMR data to be considered a chart review [46].

Another potential reference standard is the identification of a patient via some other administrative data source, like the DAD. This method of establishing a reference standard is subject to some scrutiny, as it uses an administrative data source to confirm another administrative data source. Others have compared case definitions to other previously validated case definitions to determine accuracy [47].

Yet another potential reference standard is the inclusion of a patient in a disease registry. For some chronic conditions, registries are created to track individuals throughout the healthcare system. Some have used the presence in a disease registry to confirm the diagnosis status of a case definition in an EMR. The quality of the reference standard, in this case, is solely dependent on the validity of the registry. Many registries have a low false positive rate, but it can be difficult to assess the false negative rate. This is because it is rare for a non-case to be included in a registry, and it can be quickly verified. False negatives, on the other hand, are much harder to identify in a registry, because it requires information about patients not included in the registry. Another popular method is to get laboratory confirmation for certain characteristics (ie. blood sugar in patients with diabetes) [48]. It should be emphasized that the decision on what to use as a reference standard is typically based on costs and convenience.

Some studies have used a chart review as the reference standard diagnosis by which to evaluate the validity of case definitions [49, 50, 41, 39, 51]. These include case definitions for rheumatoid arthritis, bipolar disorder, stroke/TIA, and eight case definitions that will be discussed in Chapter 3.

Others have compared the agreement in case identification methods between EMR data and more classical sources of administrative data, with varying amounts of success [52]. Another study used billing codes for a case definition, but used free-text as the reference standard [38].

1.7 Committee-Created Case Definitions

The traditional method of case definition development is via a committee of healthcare professionals. Essentially, a group of clinicians are brought together and through discussion and debate, determine a set of criteria that distinguish between cases and non-cases. This is a process that can take a long time to complete. Once a case definition is determined, it is evaluated against a reference standard to get an estimate of the accuracy of the case definition.

This model of case definition development is summarized in two to three steps. The first is the development of criteria using clinical knowledge, and creating a case definition. The second step is to evaluate the validity of the case definition against a reference standard. A third step can be incorporated, wherein adjustments are made to the case definition after reviewing the validity results.

We propose a new method of case definition development, wherein we reverse the order of procedure. Here we will start with the reference standard, which we will use as the labels for a statistical learning approach. These methods create a set of rules that maximize accuracy and can be interpreted as a case definition. Using this method, we can forego the tedious process of committee-creation, and use an algorithm that will select the features that most accurately distinguish between cases and non-cases. These algorithms are often referred to as machine learning algorithms.

1.8 Objectives

1. Determine if machine learning methods are capable of creating chronic disease case definitions in a primary care EMR that are comparable in terms of simplicity and validity measures with committee-created case definitions.

2. Compare different machine learning methods to establish which methods are best suited for case definition development.

2 Machine Learning

Machine learning is a system of statistical methods that are relatively new to the health research community. A common definition for machine learning is that it is a process through which a computer "learns" a model, without being explicitly programmed [53]. There is an important distinction here between classical regression-based methods and machine learning methods.

In the classical regression-based predictive statistics, there is some outcome that is desired to be predicted. Predictor variables are selected that may be able to aid in the prediction of the outcome and a model is constructed, often through some hierarchical structure. Then a machine performs calculations that give the regression coefficients or summary statistics based on this model. Machine learning does not simply crunch the numbers and spit out coefficients, rather, it makes decisions about the actual structure of the model; the predictors or features selected, the interactions between features, the complexity of the model and so on. This is a key distinction, as machine learning methods are being implemented in large data settings, where the complexity and scale of data and features are often too large for traditional regression methods.

Generally, machine learning models can be broadly categorized as two categories: supervised and unsupervised methods. Supervised methods are used when there is a known outcome, and a model is being trained to predict the outcome. This requires at least a sample of data that includes the outcome, so a model can be trained to accurately predict. Unsupervised learning, on the other hand, does not necessarily have an outcome. Often, unsupervised learning methods are used to identify clusters in the features space, differentiating groups or clusters of observations based on a number of features.

For this study, we are looking at binary classification; where a disease is either present or absent. This is a supervised machine learning approach, and there are a variety of machine learning methods, ranging from the simple to the complex. In order to develop case definitions, the interpretability of the case definitions is equally as important as the accuracy of the case definition.

Here we will briefly summarize the common machine learning methods that have been implemented in the published literature, as well as the algorithms that will be used in this study. A reader experienced in machine learning methods can safely skip this chapter.

2.0.1 Naïve Bayesian Classifier

The Naïve Bayesian Classifier is a simple concept that compares the probabilities of being in a given class given the values of various candidate features. It has been used extensively in text categorization problems [54]. This algorithm implements Bayes' Theorem [55], which is defined as follows:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(A \cap B)}{P(B)} \tag{1}$$

Where $\cap$ represents the intersection, or the joint operator. Essentially, the Naïve Bayesian Classifier computes the conditional probability, $P(A|B)$, of each class (disease or no disease) based on each covariate. It assumes independence of each covariate, and the estimated probability of disease is defined as follows [56]:

$$\hat{y} = \underset{k\varepsilon\{1,\ldots,K\}}{\arg\max}\ p(C_k)\prod_{i=1}^{n} p(x_i|C_k) \tag{2}$$

Where C represents a target class, the x_i represent the values of the predictor variables. The predicted class is the class that maximizes the conditional probability. It is clear that this classifier assumes conditional independence of all predictor variables [57], as the posterior probability is simply the product of all individual conditional probabilities. It has been shown that the Naïve Bayesian Classifier and other discriminant methods are comparable to logistic regression in small sample sizes, but the performance of logistic regression outperforms it with a larger sample size, and has a smaller asymptotic error [58].

In terms of case definition development, the Naïve Bayesian Classifier is not very useful. With all of the data available in an EMR, and with the clinician free-text tokenized into individual vectors, there is potential for hundreds and even thousands of variables to be considered. The Naïve Bayesian Classifier has no method of feature selection, so every variable is included in the model. This makes it a poor option for case definition creation, even if the predictions are reasonably accurate. The interpretability of the model is lacking, as each variable is still included.

This methodology may prove to be of some use in EMR classification though. As mentioned earlier, the number of variables, especially when including tokenized free-text, is quite large, and many computation-heavy algorithms will perform slowly with that many variables to consider. We

can use the simple Bayes' Theorem formula to calculate conditional probabilities for each variable, and thus eliminate uninteresting variables by only including the variables that meet the following criteria, where T is some threshold value.

$$|P(feature|\text{DISEASE}) - P(feature|\text{NO DISEASE})| > T. \tag{3}$$

This is a simple way to remove a portion of the variables that have no effect on predicting disease status, in order to increase efficiency of the algorithm. This becomes important in cases with large amounts of data, where complex computation problems may extend beyond the limits of hardware. Others have used odds ratios, information gains (to be discussed in a later section), exponentiated probability differences, and more for this same purpose [59].

Similar to the Naïve Bayesian Classifier, Bayesian Networks that do not require the independence of variables can be used. These methods are acceptable for prediction, but are typically not interpretable as case definitions, and will not be used for the case definition development.

2.0.2 K-Nearest Neighbours

The K-Nearest Neighbours method is a simple, heuristic approach to a classification problem. New cases are predicted by finding the most common class in the most similar K observations [60]. This method is most useful in settings where predictors are measured, as it relies heavily on distance measures. Additionally, weights can be added to observations or predictor variables to improve prediction [61]. It can still be used in cases with binary predictors, but it acts very similar to the Naïve Bayesian Classifier in this setting.

The K-nearest neighbours method will not be used for case definition development, as the resulting model is not interpretable, especially in a setting with many potential features. This method does not reduce the feature space at all.

2.0.3 Logistic Regression

Regular logistic regression is not generally considered to be a machine learning algorithm. This is a classical method, where a model is constructed *a priori*. However, there are machine learning approaches that implement logistic regression models. Logistic regression is a generalized linear model wherein the outcome variable is binary. Multinomial and proportional odds models are available for categorical or ordinal outcomes but for our purposes, we will only look at the binary

outcome (disease or no disease). Logistic regression models the logit function as a linear combination of the predictor variables [62]. The logit function is as follows, where π is the probability of the outcome:

$$logit(\pi) = \log(\frac{\pi}{1-\pi}) \tag{4}$$

We use the logit function because it has favorable properties for regression in a binary response setting. A regression model is typically constructed as the linear combination of some number of predictor variables. By definition, this assumes that the outcome being modelled can truly be modelled in a linear fashion. This is problematic in the case of a binary outcome when we are modelling the probability, π, since π can only be defined on the interval [0,1]. If one tried to model π using a linear combination of predictor variables, it is possible to produce estimated probabilities that fall outside of the possible range for π. The odds, $\frac{\pi}{1-\pi}$, are defined from [0,∞), and are thus more favorable for modelling, but still have problems near 0. The natural logarithm of the odds (also called the logit), $\log(\frac{\pi}{1-\pi})$, is defined from $(-\infty, \infty)$ and is widely used because of this property. Since the log-odds are defined for all possible values, we can construct a linear model that does not have boundary-problems.

The regression equation is used as follows:

$$log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n \tag{5}$$

Which can be rewritten as:

$$\pi = \frac{e^{\beta_0+\beta_1 X_1+...+\beta_n X_n}}{1+e^{\beta_0+\beta_1 X_1+...+\beta_n X_n}} \tag{6}$$

Where the X's are the values of the predictor variables. Logistic regression uses the principles of maximum likelihood to estimate the β coefficients. This is done by maximizing the likelihood function [63]:

$$L(\beta|y) = \prod_{i=0}^{N} \binom{n_i}{y_i} \pi^{y_i} (1-\pi)^{n_i - y_i} \tag{7}$$

Here π is defined as the probability of "success" for an event, y refers to the number of successes in n independent trials. This equation is maximized by differentiating with respect to β and

finding the maxima. The prior equation (0.6) defines π using the β values. The result is a set of β coefficients that are used to model the log of the odds, which can further be decomposed into estimates of probabilities. Computers, however, do not calculate the coefficients in this manner. Statistical software must use an iterative approach to approximate the maxima of the likelihood function. These software use matrix calculations to calculate the maximum likelihood and regression coefficients.

It can be easily seen that logistic regression is a very useful tool in terms of classification, but logistic regression alone is of very little use in automated case definition development. There are no mechanisms for feature selection, and in the case of EMR data with tokenized free-text, the sheer number of variables make logistic regression a poor choice. Logistic regression assigns a coefficient to each predictor variable included in the model. The methods that will be used in this analysis take into account thousands of candidate features. A vector of thousands of coefficients is not particularly useful when trying to implement a case definition that makes intuitive sense for clinical purposes.

One of the issues with logistic regression is multi-collinearity, wherein multiple predictor variables that are correlated are assigned weights that are not especially indicative of their individual ability to detect disease status.

There are methods of using logistic regression for automated case definition development in settings with many predictor variables. It is important to note that the results of a logistic regression are not interpretable as a case definition without a few liberties being taken. The logistic regression will produce linear combinations of the log of the odds. It is common practice to take these log-odds estimates and finding optimal cut-points to determine the predicted class. In order to convert these results into a case definition, however, simplifying assumptions will be required. For example, one might consider positive coefficients to be inclusion criteria, and negative coefficients to be exclusion criteria. This is the approach that was taken in this project. The two most common methods of feature selection using a logistic regression method are forward stepwise or penalized logistic regression.

2.0.4 Forward Stepwise Logistic Regression

When there are a large number of variables to be considered in a logistic regression, some have opted to use a method called stepwise logistic regression to aid in variable selection [64, 65]. It has previously been used for case definition development [66]. Stepwise logistic regression is typically used in one of two ways, forward or backward. In the forward approach, the model is built from the empty model, and complexity is added step-by-step. The backwards approach assumes the most complex model, and features are removed step by step. This study will solely use forward stepwise logistic regression, as there are potentially thousands of features. This is where you start with the base model, including only an intercept:

$$log(\frac{\pi}{1-\pi}) = \beta_0 \tag{8}$$

From this base model, we add variables one by one, maximizing some measure of model "quality". The most popular metric of this is called the Aikake Information Criterion (AIC), and is defined as such:

$$AIC = 2k - 2log(L) \tag{9}$$

Another measure that is used is the Bayesian Information Criterion (BIC):

$$BIC = -2log(L) + klog(n) \tag{10}$$

Here k represents the number of features included in the model, n represents the number of data points, and L represents maximum likelihood estimate[67, 68]. The model with the minimum AIC is selected as the final model, and variables are chosen by the order of the magnitude of AIC change. Using this method provides an automated way of selecting the features that will go into the final model, thus selecting the features that will comprise the case definition for the desired condition.

Stepwise methods have been widely criticized by statisticians. It has been shown that accuracy measures (R^2 values in particular) are biased in subset selection approaches [69]. Others have shown evidence that predictions using forward stepwise methods fare no better than models that include all candidate features [70].

This does not mean that stepwise methods are invalid in the case of EMR case definition development. Large data sources with many covariates require data mining techniques that almost always run the risk of overfitting. Stepwise methods are no different. This method will be explored in the case definition development.

2.0.5 LASSO Logistic Regression

Instead of using stepwise methods, another variable selection strategy is to use a penalty called a LASSO (Least Absolute Shrinkage and Selection Operator). This method shrinks some coefficients, but more importantly, it sets coefficients to 0. It can be used as an automated way to exclude extraneous variables from an otherwise massive logistic regression equation. Essentially, we are applying the following constraint to a regular logistic regression.

$$log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n \text{ SUBJECT TO } \sum_j |\beta_j| < t \tag{11}$$

Where t is a tuning parameter or constant. This tuning parameter forces the model to include only the variables that contribute the most to the outcome in the sample. It is widely held that imposing such a limit introduces some amount of bias, but is useful in improving predictive accuracy and simplicity.

Dr. Tibshirani is credited with the development of the LASSO regression, as well as the ridge regression [71]. The ridge penalty is very similar to the LASSO penalty, and is defined as such:

$$\sum_j \beta_j^2 < t \tag{12}$$

The LASSO penalty is sometimes referred to as the L_1 penalty and the ridge penalty is referred to as the L_2 penalty. Methods have been developed to blend these methods together using a tuning parameter, and this method is often called Elastic-Net regression [72]. This method of penalization has also been implemented in other modelling situations, such as time-to-event, or survival analyses [73]. We will not be using ridge or Elastic-Net regression for variable selection in case definition development, as we will restrict this study to LASSO logistic regression, as it is the favorable method for removing features and creating concise feature lists, compared to the ridge regression.

It can be seen that the bounds of the L_1 penalty are maximized when one of β_1 or β_2 are set to 0. The limit of the ridge penalty, on the other hand, do not have this same property. The L_1 penalty uses an absolute value to restrict the β coefficients whereas the L_2 penalty uses the sum of squares. For the purposes of variable selection, the L_1 penalty is preferable to the L_2 penalty, as the L_1 penalty favours the setting of β coefficients equal to zero. The L_1 penalty achieves the largest combined β coefficient estimates by setting the variables of lesser importance equal to zero. Figure 1 shows this relationship.

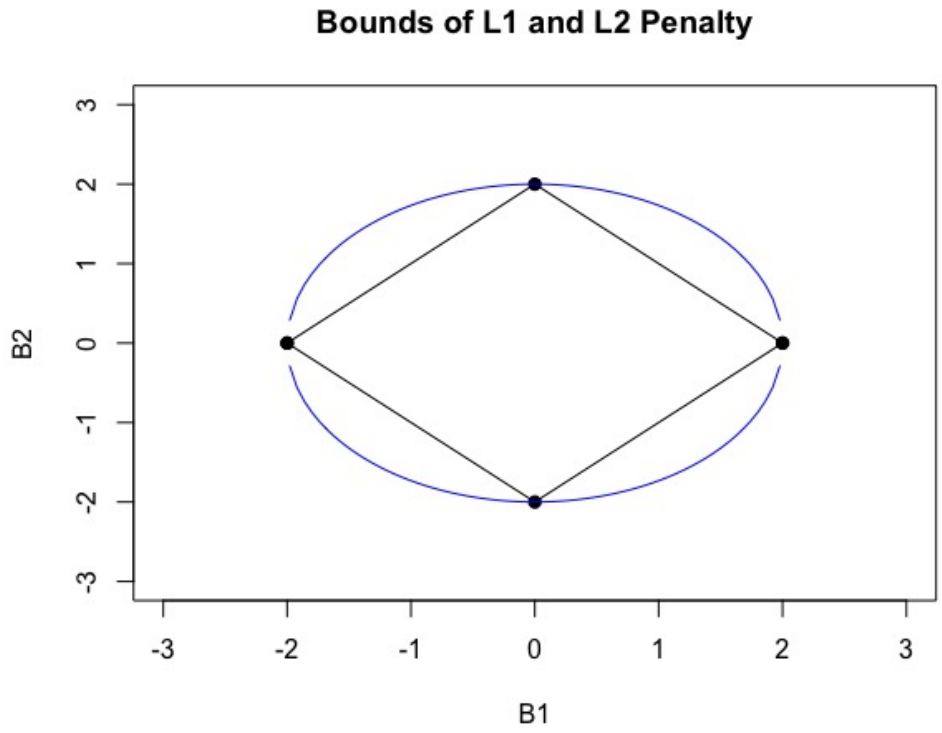

Figure 1: Bounds of β coefficients when t=2. The black curve is the L_1 penalty and the blue curve is the L_2 penalty.

When using the L_1 penalty, the model is created at uniform cut-points of the t parameter. Typically, the final parameter value is selected as the value at which the accuracy is maximized in a cross-validation setting (discussed in a later section).

2.0.6 Decision Tree Algorithms

Another classification method that should be considered for chronic disease case definition development is that of decision tree classifiers. There are several types of decision tree methodologies, but they all share the same basic principles. A decision tree is a system of rules and conditions that can be visualized as a "tree" [74]. See Figure 2 for an example of a decision tree.

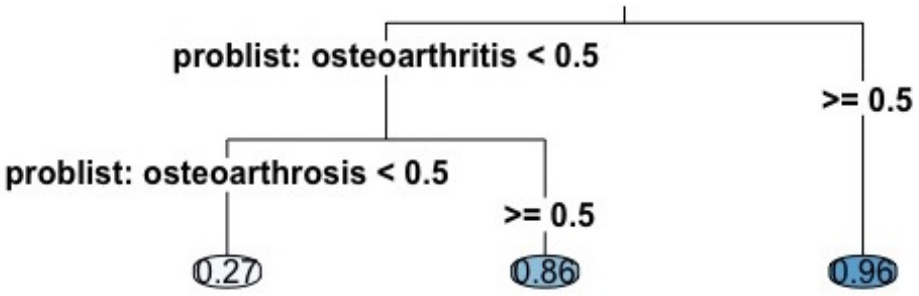

Figure 2: An example of a decision tree to show the proportion of patients diagnosed with osteoarthritis. The top of the decision tree represents all the patients in the sample. Note that each end-node of the tree indicates a conditional proportion, so the total sum of all nodes need not be 1.

In this example, of those that had the word "osteoarthritis" occur in their primary care problem list, 96% truly had osteoarthritis. For the remaining patients, 86% of those that had the word "osteoarthrosis" in their problem list truly had osteoarthritis, and 27% that had neither word in the problem list truly had osteoarthritis.

A decision tree is composed of a set of splits (typically binary splits) on different variables. The end-nodes, sometimes called leaf-nodes, are the endpoints where a class is specified. These decision trees can be decomposed into a set of rules that can be interpreted as a case definition. One of the main advantages of using a decision tree algorithm as opposed to a logistic regression method of variable selection is that the decision tree algorithm is capable of accounting for the association between variables very easily.

The case definitions developed by the aforementioned logistic regression methods result in overly simplified definitions in some cases. The final case definition consists of a list of inclusion

and exclusion criteria determining the disease status. Decision tree algorithms are different in that they formulate binary splits based on a feature, and then make subsequent partitions based on the previous splits. This allows for the combination of variables and exclusion criteria in the different branches of the tree.

As mentioned, there are several different algorithms for creating decision tree classifiers. The main difference between each of these is the metric that is used as a "purity metric". This is a measure of how well a split in the data differentiates between those with and without the disease. In each algorithm, all possible splits are considered, over every variable, and the one that maximizes the purity index is chosen. This process is repeated until there are no more splits that improve the purity index. Below is a description of three of these algorithms.

2.0.6.1 Classification and Regression Trees / Recursive Partitioning

One of the more popular decision tree algorithms is called Classification and Regression Trees (CART), developed by Brieman, Friedman, and others [75]. These ideas were expounded upon and renamed Recursive PARTitioning (RPART) by Therneau, Atkinson, and others [76, 77]. This methodology is not only capable of building classification trees, but is also able to build regression trees. Regression trees are required when the outcome is measured.

For classification purposes, the purity index that is typically used is called the Gini Index [78], which is defined as such:

$$I_G(p) = \sum_i \sum_{j \neq i} p_i p_j = \sum_i \sum_j p_i p_j - \sum_i p_i^2 = 1 - \sum_i p_i^2 \tag{13}$$

Where p_x is the probability of belonging to class x in a node of the decision tree. The Gini index is calculated for every possible split in the predictor variables, and the one with the largest improvement in purity (or the biggest decrease in the Gini index) is chosen. This is repeated until there are no remaining splits that improve the index. This is where the pruning of the tree takes place. There are several pruning algorithms that are used, the most popular of which is called cost complexity pruning. This is done by finding the minimum of the following metric, for all possible subtrees of the unpruned tree:

$$f(t) = \frac{err(prune(T,t),S) - err(T,S)}{|leaves(T)| - |leaves(prune(T,t))|} \tag{14}$$

Where err(T,S) is the error rate of the unpruned tree, prune(T,t) is the pruned decision tree, and leaves(T) is the number of leaves in the unpruned tree. The tree that minimizes this value, up to some threshold (sometimes called the complexity parameter) is selected as the new tree. This process is repeated until there are no more pruning steps that fall under the specified complexity parameter.

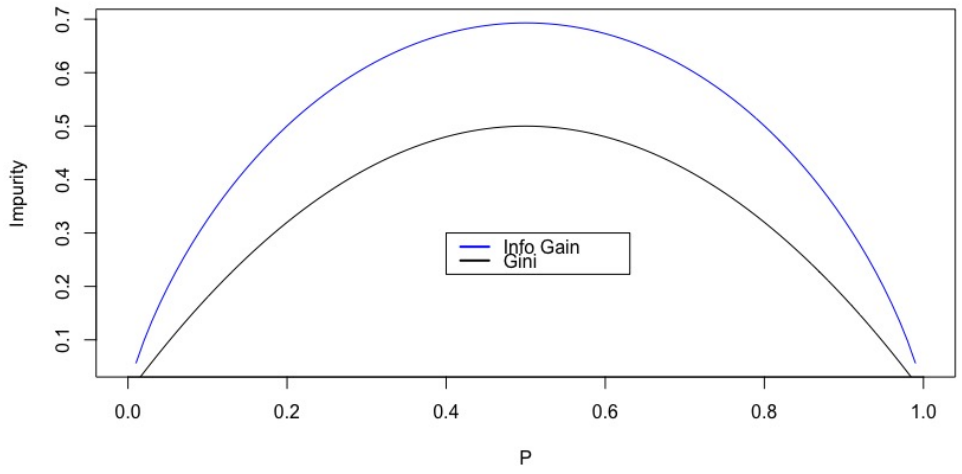

Figure 3: A comparison of the Information Gain and Gini impurity measures.

2.0.6.2 C4.5 and C5.0 Decision Trees

The C4.5 and C5.0 algorithms were developed by Ross Quinlan [79]. C5.0 is an updated version of the C4.5 algorithm [80] [81]. Both of these algorithms develop trees using an impurity index called information gain. This metric uses entropy theory and is calculated as such:

$$gain(A) = -\sum_{j=1}^{k} p_i \times log_2(p_i) \tag{15}$$

Where p_i represents the proportion of cases for predictor i given a split on the predictor A. Visual comparisons of the Gini Impurity index and Information Gain is seen in Figure 3. The C4.5

and C5.0 algorithms take this calculation one step further and calculate the gain ratio, which is defined as follows:

$$GainRatio(A) = \frac{gain(A)}{split(A)} \tag{16}$$

Where split is defined as:

$$split_A(D) = -\sum_{j=1}^{v} \frac{D_j}{D} \times log_2(\frac{D_j}{D}) \tag{17}$$

Where D refers to the number of observations at the various levels of the potential split. D_j refers to the number of observations at a level of the potential split and D is the total number of patients pre-split. Using split-info in the calculation of the gain ratio takes into account the number and size of each branch of the tree. Otherwise, there is a bias towards multi-valued attributes. It has been shown that gain ratio is typically preferable to information gain alone [82].

The C5.0 algorithm also offers a variety of features that might be useful for the development of case definitions. Most of these features were added on to the preexisting C5.0 algorithm.

Like the CART algorithm, a complexity parameter that limits the amount of pruning can be used. This parameter is represented as a value between 0 and 1, where 0 represents no pruning, and 1 represents the most pruning. The C5.0 algorithm also includes an option where one can specify the minimum number of observations that can be in a node of the tree. This is a method of avoiding overfitting.

The algorithm also includes an option to include a cost matrix, wherein one can specify the penalty associated with misclassification of true positives and true negatives. This is a useful tool in that it allows for the development of case definitions with higher sensitivity or specificity respectively, and it is important when dealing with rare conditions.

Another feature that the C5.0 algorithm has implemented is a process called winnowing. Winnowing is another method used for feature selection that is useful in high-dimensional settings. When there are a large number of predictors, it is often useful to perform a feature selection step before creating a decision tree, often for computing efficiency purposes. This step is done using a simple measure of a predictor's relationship with the outcome. An example of this was seen earlier with the following method:

$$|P(feature|\text{DISEASE}) - P(feature|\text{NO DISEASE})| > T. \quad (18)$$

The winnowing applies a similar method to drop variables and decrease the computational load required by the C5.0 algorithm. These methods have been shown to have a positive impact in some cases and a negative impact in others.

2.0.6.3 CHAID Decision Tree

The Chi-squared Automatic Interaction Detector (CHAID) is another decision tree algorithm that determines splits using the significance of a chi-squared test statistic. As opposed to CaRT methods, CHAID is designed solely for classification problems and is not to be used for regression. CHAID is also designed for using categorical features exclusively. As part of the algorithm, optimal collapsing of non-binary categorical features based on cross tabulations is done to ensure binary splits.

The decision to split in CHAID is based on the p-value of the χ^2 test [83, 84]. The split with the lowest p-value when cross tabulated with the outcome is selected at each node. A threshold α value is specified to limit the number of splits to only those that have a p-value less than α.

Note that it is possible to create a split in a CHAID decision tree that does not indicate a difference in outcome category. For example, it is possible to have a binary split in which both end-nodes of the tree indicate the same disease status. This is because the χ^2 test detects a difference between proportions and does not require those proportions to be on opposite sides of the proportion 0.5. Both the CaRT and C5.0 methodologies (Gini index and information criterion) will also include these redundancies but these algorithms contain pruning steps that will generally remove them.

2.0.7 Random Forest Classifier

An extension to the decision tree classifiers is the random forest algorithm [53, 85]. This is a popular method that is used often in classification problems. A random forest model is called an ensemble tree method. This means that multiple trees are constructed and the predicted class is decided by the majority vote from each of the trees.

There are several methods that can be used to construct an ensemble of decision trees. For example, the bagging method is when random samples of the training data are taken without re-

placement, and a decision tree developed for each sample [86]. Another example is that of random split selection, wherein the split is selected at random from the K best possible splits [87]. The most common method involves creating random subsets of variables and training multiple trees with different subsets of candidate features [88]. This ensures that the classification is not being dominated by a few features. This method has been shown to improve predictive accuracy in many cases, when compared to a single decision tree. It has been suggested that due to the Law of Large Numbers, random forest models do not over-fit the training data [89].

Another ensemble tree method is called boosting. There are various decision tree boosting approaches, but they follow the same blueprint. A common method is called AdaBoost, which generates an ensemble of trees. The first decision tree is generated, and those cases that are misclassified by the first tree in the training set are given a higher weight in the subsequent tree [90]. The resulting ensemble of trees is then used to predict the outcome class. This method has been shown to improve predictions in some cases [91].

The problem with the random forest and other ensemble tree approaches in the context of case definition development is that there is no way to interpret these methods into human readable case definitions. It is common to create 500+ individual trees in a random forest approach, and there is no way to decompose those into rule sets, given that the classification is based on majority vote. We will not use the random forest method for case definition development, but it will be an important area of future research to quantify the differences in accuracy between these complex methods of classification and the interpretable methods.

2.0.8 Support Vector Machines

A powerful machine learning tool that has been used in many settings is the Support Vector Machine (SVM) [92]. Visually, this method attempts to draw a line in the feature space that best separates cases from non-cases [93]. Rather than doing this using parametric assumptions of linearity or polynomial regressions, it uses various "kernel" functions. Different implementations of SVMs use different functions, some of the simpler implementations acting similar to a Loess smoother or a moving average algorithm. Others, like figure 4, use a non-linear kernel to generate a linear combination of features that geometrically distinguish between class A and class B.

SVM's, like the random forest models are considered "black-box" algorithms, meaning they

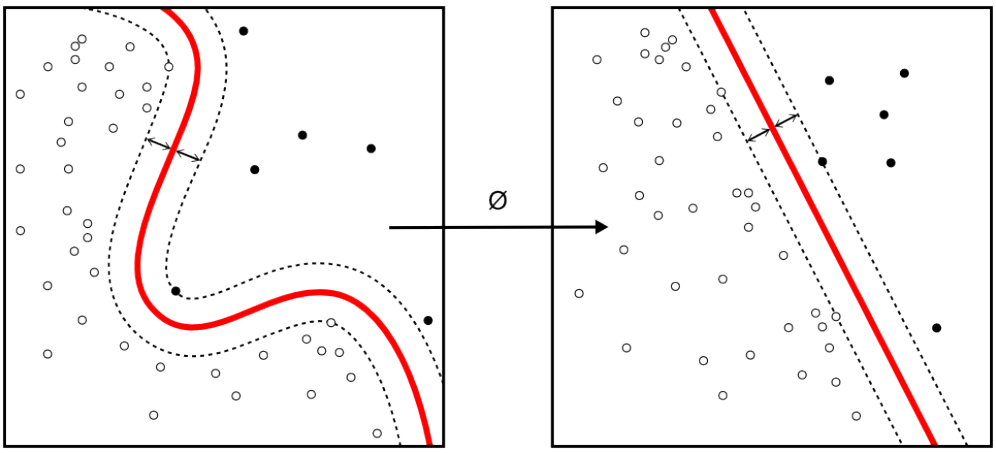

Figure 4: A two-dimensional support vector machine illustration

are not directly interpretable. This is because the coefficients may have a linear representation, similar to a logistic regression, but the linear coefficients need to be transformed through the kernel function. This makes it a poor choice for case definition development as a set of rules cannot be derived from the model, and therefore will not be implemented in this project.

2.0.9 Artificial Neural Networks

Artificial Neural Networks (ANN) are generally considered to be the most complicated [94]. These models were inspired by the cells and networks that naturally occur in the brain. In the classification setting, the algorithm generates a number of "nodes", where a classification algorithm or prediction method is implemented on a transformation of the data. Some implementations of an ANN are as simple as a parallel series of logistic regressions, each including a different polynomial transformation of the features and outcome. Deep Learning is an implementation of neural networks, and is very popular in today's machine learning landscape [95].

The most important step of the ANN is called back propagation. This is a method that recursively assigns weights in an automated fashion to each of the nodes that were developed for classification. Those nodes that predict best are assigned higher weights, and nodes that provide little predictive benefit are assigned low weights.

This method, although often regarded as the most accurate and flexible, does not offer a human interpretable case definition, and will not be used for this project.

2.0.10 Tuning Parameters

It should be noted that each machine learning algorithm has a number of tuning parameters that need to be set. These parameters allows users to customize the model as they see fit. Common tuning parameters include a loss matrix, allowing for a preference structure to exist. This means a model can be trained that will focus on achieving a higher sensitivity, at the cost of specificity, or vice versa. This is of particular importance in rare disease cases, where the algorithm may tend to favour specificity much more than sensitivity. Other parameters control the amount of complexity in the models, or the number of features. Decisions on tuning parameter selection must be supported by some level of evidence, requiring a method of comparing different tuning parameter values for model fitting.

The machine learning algorithms that will be implemented in this study will be the C5.0 decision tree, the CaRT/rPart decision tree, the CHAID decision tree, LASSO-penalized logistic regression, and Forward stepwise logistic regression.

2.1 Application of Machine Learning Methods in EMRs and EHRs

Thus far, the implementation of machine learning in the EMR space is confined primarily to outcomes prediction, clinical decision support, or an attempt to synthesize clinical free-text [96]. These methods are not generally used for disease surveillance, rather, they are typically used to identify high risk patients for which a clinical intervention can be implemented. The ideal machine learning algorithms selected for different studies are determined mainly by the purpose of the study. When predictive accuracy is the most important factor, and interpretation is of lesser import, more complicated "black box" algorithms can be used. When the desired product is an interpretable model, a set of rules, or an attempt to expose features and relationships between variables that contribute to prediction, the more concise and interpretable methods are much preferred over the "black box" algorithms. Here we will discuss the problems for which machine learning algorithms have been implemented in EMR and EHR systems.

2.1.1 Outcomes Prediction

As stated earlier, a large portion of the published studies that adopt machine learning for EMR data analysis focuses on outcomes prediction or forecasting. Mani et al. used Naïve Bayes, logistic

regression, K-nearest neighbours, CaRT decision trees, random forest, and SVMs to forecast type 2 diabetes in a non-diabetic population using electronic medical record data [97]. The primary measure of model accuracy compared here was the Area Under the Receiver Operating Curve (AUROC). In this implementation, the random forest algorithm achieved the best AUROC.

Another study from Cooper et al, published in 1997, used logistic regression, K-nearest neighbours, decision tree algorithms, Bayesian Networks including Naïve Bayes, and ANNs to predict mortality in pneumonia patients [98]. This study showed very little differences between algorithms in predictive accuracy, and concluded that when predictive accuracy is similar, the simplicity of the final model should be used to select the optimal algorithm.

Similarly, Wolfe et al. used logistic regression, classification trees, and ANNs to predict a composite outcome of mortality and intensive care unit admission following traumatic events [99]. This attempt was unsuccessful, as they were unable to meet pre-specified benchmarks for predictive accuracy in this implementation.

A 2014 study published by Mani et al. used SVMs, Naïve Bayes, Tree-Augmented Naïve Bayes, K-Nearest Neighbours, CaRT decision trees, random forest, logistic regression, and Lazy Bayesian classifiers in an attempt to detect the late-onset of neonatal sepsis in infants for whom a blood culture test has been ordered on suspicion of sepsis [100]. Here the SVM-based feature selection methods contributed most to predictive accuracy. The final reported accuracy achieved was an AUROC of 0.78.

Palaniappan et al. implemented various decision trees, Naïve Bayesian, and ANN algorithms to predict the diagnosis of heart disease obtained from the Cleveland Heart Disease database [101]. Here the Naïve Bayesian algorithm was labelled as the method that fit the desired criteria best, although a consideration was taken for the decision tree algorithm, as it is the most interpretable outcome.

A much earlier study, published in 1997, from Shankle et al. used CaRT and stepwise logistic regression methods to detect early stages of dementia, using mostly survey data [102]. The interesting finding from this study is that a single feature was determined to be the most predictive, and the only feature included in the final models. This shows how machine learning algorithms can, in some settings, massively simplify prediction problems.

Kawaler et al. used Naïve Bayes, K-Nearest Neighbours, SVMs, the C4.5 decision tree, and

random forest algorithms to predict patients with a venothromboembolism event in the 90 days following a hospitalization [103]. The conclusion of this study was that the Naïve Bayesian, random forest, and SVM methods achieved the best predictive accuracy, measured with a balance of sensitivity and positive predictive value.

A study published in 2014 documented the use of SVMs to aid in the prediction of survival probabilities at 6 months, 1 year, and 2 years following cancer diagnosis, taking into account the type of cancer diagnosis [104]. These methods showed a large improvement in predictive accuracy at 6 months compared with a clinician panel (AUROC of 0.87 to 0.79, respectively). However, this difference was not observed for the two-year survival time (AUROC of 0.76 vs 0.75, respectively).

2.1.2 Phenotyping and Case Definitions

Another common use of machine learning methods in the EMR space is phenotyping. Phenotypes are remarkably similar to case definitions, as they often aim to distinguish between those that have a condition and those that do not have a condition. However, phenotypes do not need to be directly interpretable as rule sets and can therefore use "black-box" machine learning methods. Case definitions, on the other hand, need to be interpretable.

In 1990, a study from Baxt et al. was published that used an ANN to phenotype acute myocardial infarction in patients presenting into the emergency department [105]. This method achieved a sensitivity of 92%, and a specificity of 96%. This was deemed to be "substantially better" than the performance of case definitions developed by a committee of physicians and other analytical approaches. Similarly, Pakhomov et al. used a Naïve Bayesian algorithm with a bag-of-words approach to identify patients with heart failure from the clinical text in an EMR [106].

In 2012, Carroll et al. published a study that used a LASSO logistic regression method to create a phenotype for rheumatoid arthritis [38]. The reference standard used to define true cases was the presence of an ICD-9 code. This model was trained separately across 3 different EMR systems and showed portability across systems.

It is commonly desired to have a phenotype or case definition for the smoking status of a patient. Many EMR systems have a classification system that attempts to capture this information, but this is often captured in free-text entries. Thus, natural language processing efforts have been made to classify free text as indicative of current smokers, past smokers, and non-smokers. An

organization called Informatics for Ingegrating Biology and the Bedside (i2b2) has established data models, software, and methods for large-scale surveillance. A challenge was put out in an i2b2 workshop that led to several groups attempting the smoking status classification problem [107]. Savova et al. used SVMs along with a simple bag-of-words approach, along with a consideration of negatory terms to classify smoking status. The final reported F1-score values was 85.57 [108]. Clark et al. used SVM's and achieved scores between 85 and 90 [109]. Aramaki et al. used K-Nearest Neighbours combined with a natural language tagging algorithm to achive an F1-Score of 0.88 [110]. Carrero et al. used boosted decision trees to achieve a score of 0.75 [111]. Other methods included ANNs, C4.5, and Naïve Bayesian Classifiers.

Another i2b2 challenge workshop was published by Uzuner et al. in 2010 where natural language processing and machine learning techniques were used to extract important medication information from free-text entries. This included medication names, dosages, route of administration, duration, and more [112].

Wang et al. used advanced Natural Language Processing algorithms to detect coronary angiogram results and ovarian cancer diagnoses in the General Practice Research Database [113]. The strongest submission had an overall F1-score of 85.7.

Chen et al. published a phenotyping study in 2013 where SVMs were used to identify patients with rheumatoid arthritis, colorectal cancer, and venous thromboembolism [114]. The data used here was "annotated" cohorts, where presence or absence of disease was already established via chart review. They also incorporated a semi-supervised machine learning technique called active learning. The active learning models performed better than the traditional passive learning algorithms. AUC values were strong for rheumatoid arthritis, boasting values between 0.82 and 0.9. The predictions for colorectal cancer were also strong with values ranging between 0.81 and 0.93. Venous thromboembolism, on the other hand, resulted in very poor predictions with values ranging between 0.27 and 0.44

Another large phenotyping study was published in 2013 by Srimani et al., where a comparison was made between boosted decision trees, logistic regression, random forest, Naïve Bayesian, and more to create phenotypes for various conditions in multiple datasets [115]. The interesting conclusion of this study is that in some EMR systems and conditions, complicated "ensemble" machine learning methods are not needed, and simpler, more interpretable case definitions are

preferred. Liang et al. used Deep Learning, SVMs, and decision tree methods for phenotyping several conditions in 2014 [116]. The main result of this study was showing that there are some conditions and settings where a Deep Learning algorithm can provide higher accuracy than the "shallower" methods like decision trees and SVMs.

A 2014 study from Peissig et al. used machine learning algorithms to create phenotypes for acute myocardial infarction, acute liver injury, atrial fibrillation, cataracts, congestive heart failure, dementia, type 2 diabetes, diabetic retinopathy, and deep vein thrombosis [117]. The algorithms used here are from the WEKA suite of machine learning algorithms [118]. Here a branch of algorithms called inductive logic programming outperformed the selected decision tree algorithms.

Lin et al. used ridge logistic regression to develop a phenotype to identify patients with rheumatoid arthritis along with methotrexate-induced liver transaminase abnormalities [119]. The resulting model achieved statistically significant improvements in predictive accuracy from the baseline system.

2.1.3 Other Application of Machine Learning to EMRs

A handful of other studies have been published that implement machine learning methods to EMR data for purposes that cannot be classified as outcome prediction or phenotyping/case definition development.

A 2011 study from Boxwala et al. used logistic regression and SVMs to detect "suspicious access" to EMR data [120]. Zhang et al. published a study in the same year using a Naïve Bayesian approach to identify the roles of users as they access EMR data [121]. By implementing these methods, suspicious access can be flagged and brought to the attention of the appropriate stakeholders. These are important areas, as they show how machine learning can be used to monitor EMR data access and improve data security.

Another important usage of machine learning methods in an EMR setting is in the deidentification process. When administrative data is used for research or other purposes, it needs to be stripped of all identifiable features, to protect the identity of the individuals represented in the data. The advent of the EMR, and especially the inclusion of clinician free-text, allowed further opportunity for identifiable features to be available in the data. A 2007 study from Szarvas et al. used the C4.5 decision tree algorithm to extract identifiable features from an EMR dataset, for the purpose

of meeting the guidelines of the Health Information Portability and Accountability Act [122]. Two very similar studies from Neamatullah et al. and Patric et al. were published in 2008 and 2010 used SVMs and other natural language methods to extract identifiers [123, 124].

Other implementations use machine learning to parse the clinician texts to improve the interpretability of the text itself. A strong example of that is a 2012 study that used Naïve Bayes, C4.5, and SVMs to identify sentences and fragments of text that are negatory or speculative [125]. For example, it is important to note that the text "not hypertension" or "maybe hypertension" are not necessarily the same as "hypertension" without that qualifier. Another study, published in 2011, used SVMs to detect abbreviations and shorthand in clinician free text [126]. This is an important consideration, as the medical community uses a variety of shorthand's that could blur the meaning of certain phrases or words across different physicians or areas. Yet another study, published in 2009 by Uzuner et al. used advanced natural language and machine learning methods to identify free-text diagnoses as present, absent, uncertain, or misassigned [127].

Pakhomov et al. used a Naïve Bayesian classifier and bag-of-words to assign billing codes to free-text diagnoses in 2006 [128]. A similar study from the same author was published in 2008 and SVMs to summarize the findings of foot examinations in clinician free-text [129].

An intriguing implementation of unsupervised machine learning methods in EMR data was published in 2013 by Lasko et al. Here, they used clustering methods to identify subgroups of patients with longitudinal serum uric acid measurements [130]. Interestingly, this unlabeled cluster method was able to achieve an AUC of 0.97 in distinguishing between uric acid signatures of patients with gout vs patients with acute leukemia.

2.1.4 Gaps in the Machine Learning Literature

The use of machine learning algorithms for disease phenotyping is not a new concept. There are a handful of studies that have done this in the past. What makes this study unique is that we will be using only interpretable machine learning methods. As mentioned earlier, there are several studies that have generated disease phenotypes using ANNs, SVMs, and random forest algorithms. While these methods have shown promising results in distinguishing between cases and non-cases, human readable methods are not commonly used. This is a necessary step to bridge the gap between the classic method of case definition development and the automated methods that could be used in the

future. Furthermore, it has been shown in these previous studies that phenotypes, case definitions, and prediction models may be portable across EMR systems, but that accuracy is maximized when models are trained in each setting. Therefore, it is necessary to create case definitions that are specific to Canadian datasets.

Here we will implement machine learning algorithms to create case definitions for chronic diseases. In order to compare the committee method of case definition development with the machine learning method, these methods must be implemented in an EMR system that has had previous case definition work performed. There are a few large EMR networks in Canada that are viable candidates.

3 The Canadian Primary Care EMR Landscape

3.1 CPCSSN

The Canadian Primary Care Sentinel Surveillance Network (CPCSSN) is an organization that collects EMR data for the purposes of chronic disease surveillance and clinical research, practice audit and feedback, and health services and policy research in Canada. It currently includes 11 different regional EMR networks across Canada, including over 1180 sentinels (primary care physicians) covering more than 1.7 million Canadian patients across 8 provinces/territories [136]. CPCSSN has received funding for the surveillance of eight chronic diseases: hypertension, diabetes, osteoarthritis, depression, dementia, COPD, epilepsy, and Parkinson's Disease from the Public Health Agency of Canada.

CPCSSN collects de-identified data from each of the 11 networks pertaining to each of the eight previously specified conditions, at regular intervals. These data are then cleaned and stored in a standard data model. The need for cleaning arises from some of the issues with the raw EMR data. These include, but are not limited to: identifiable features, misspellings, missing data, inconsistent data, and standardization issues [137].

A common problem with many data sources is that those who generate and record the data often do not directly benefit from their efforts. CPCSSN provides quarterly reports and statistics to the physicians, thus allowing the creators of the data to also be end-users of the data. Also, some work has been done to improve quality improvement projects in the CPCSSN network [138]. This feedback system also provides an incentive to have more rigorous data discipline. This means that the data collected is more reliable and accurate, as those who are producing the data actually use the results of their work. By improving data discipline, we reduce the amount of data cleaning that needs to be done.

Several studies have been performed using CPCSSN data including quantifying obesity rates [139], representativeness assessments of the CPCSSN population [140], hypertension management and prevalence [141], and more. These are all examples of EMR data being leveraged to provide surveillance across EMR systems, thanks to the common data model CPCSSN developed.

The cleaned CPCSSN data is stored in a relational database, with a unique, anonymized patient ID attached to it. Each entry also contains a field for the site and network ID, signifying which

clinic and which EMR system that entry fell under. Table 1 shows the basic structure of the CPCSSN data.

Table	Description
Allergy Intolerance	Allergies or reactions to substances, and resultant medications
Billing	ICD-9 codes submitted for reimbursement
Encounter	Each primary care visit, including reason for visit
Encounter Diagnosis	An entry for each diagnosis for each encounter
Exam	Measurements including height, weight, blood pressure, etc.
Family History	Family history of notable conditions
Health Condition	Listing of conditions the clinician saw fit to include in the problem list
Lab	Laboratory results for various tests (eg. HbA1c, Fasting Glucose, etc.)
Medication	ATC codes and other information describing therapies
Network	Description of each of the 11 EMR Networks
Patient	Unique ID for each patient, including sex, birth year and birth month
Patient Demographic	Patient information including ethnicity, language, etc.
Referral	Referrals to specialty clinics
Risk Factor	Listing of risk factors (e.g. alcohol, smoking, etc.)
Site	Description of Sites
Vaccine	List of Vaccinations received

Table 1: Brief description of each of the CPCSSN tables.

The result is a standardized set of EMR data from multiple networks that represents a large number of Canadian's primary care information. This is a good setting for the implementation and testing of case definitions for chronic disease.

Case definitions for each of the eight key conditions have been developed and validated using samples from CPCSSN data [41, 142, 143].

3.2 Validation of the Current CPCSSN Case Definitions

In 2014, a study by Williamson et al was published that estimated accuracy measures for commitee-created case definitions. [41]. This study established a reference standard for a large sample of patients, determining the presence or absence of eight conditions. After a thorough deliberation process by a team of physicians and clinical professionals, the case definitions were created. The validation sample and reference standard derived from this study will be used as the training observations for the machine learning methods. By comparing validity estimates of committee-created case definitions with machine learning developed case definitions on the exact same sample, we can make a fair assessment of the quality of the new case definitions.

An age-stratified random sample of 1920 patients was taken such that 85% of the population was over the age of 60. This was done to ensure that there were large enough number of cases for each condition. Note that this will have an effect on our estimates of predictive value and accuracy, as the sample prevalence of these conditions has been falsely inflated. However, this was deemed a necessary step, as chart reviews are costly. The true disease status of these patients was determined by a chart review by a trained abstractor. Table 2 has the validation results.

Disease	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
Hypertension	84.9 (82.6-87.1)	93.5 (92.0-95.1)	92.9 (91.2-94.6)	86.0 (83.9-88.2)
Diabetes	95.6 (93.4-97.9)	97.1 (96.3-97.9)	87.0 (83.5-90.5)	99.1 (98.6-99.6)
Depression	81.1 (77.2-85.0)	94.8 (93.7-95.9)	79.6 (75.7-83.6)	95.2 (94.1-96.3)
COPD	82.1 (76.0-88.2)	97.3 (96.5-98.0)	72.1 (65.4-78.8)	98.4 (97.9-99.0)
Osteoarthritis	77.8 (74.5-81.1)	94.9 (93.8-96.1)	87.7 (84.9-90.5)	90.2 (88.7-91.8)
Dementia	96.8 (93.3-100.0)	98.1 (97.5-98.7)	72.8 (65.0-80.6)	99.8 (99.6-100.0)
Epilepsy	98.6 (96.6-100.0)	98.7 (98.2-99.2)	85.6 (80.2-91.1)	99.9 (99.7-100.0)
Parkinsonism	98.8 (96.4-100.0)	99.0 (98.6-99.5)	82.0 (74.5-89.5)	99.9 (99.8-100.0)

Table 2: Validity estimates for the 8 CPCSSN case definitions including 95% confidence intervals.

These committee-created case definitions achieved very strong validity results. This statement was made regarding the expectations of validity:

Because acceptable limits for individual metrics need to be suited to the question of interest, we considered all measures above 70% acceptable, with any falling into the 70% to 80% range meriting additional investigation. (Williamson et al 2014)

The case definitions themselves are somewhat complicated, including long lists of medications, some exclusion criteria. Table 3 through 10 contain the committee-created case definitions. Table 9 contains the case definition for Parkinson's disease, which is primarily identified through the ICD-9 code 332, but also includes a list of four medications. However, an exclusion statement states that if a patient is only identified via the medications, they are to be excluded. It can be seen that this makes the medication criteria completely unnecessary in the case definition, as there are no situations in which the medication criteria determine disease status.

It is clear that these case definitions met the criteria for the acceptable limits, and these case definitions have since been used throughout the CPCSSN network.

Billing	Problem List	Medication	Lab Result
Two occurrences of the following code within two years: 250 - Diabetes, mellitus	Any occurrence of the following code: 250 - Diabetes, mellitus	ATC Code A10BF01, A10BB01, A10BB09, A10BB12, A10AB01, A10AC01, A10AD01, A10AE01, A10AB05, A10AD05, A10AE05, A10AE04, A10AB04, A10AD04, A10BA02, A10BD03, A10BH01, A10BB03, A10AC03 The following diagnoses, if met, make the medication criteria alone insufficient: 256.4 - Polycystic Ovarian Syndrome 648.8 - Gestational Diabetes 249 - Secondary (chemical induced) diabetes 790.29 - Hyperglycemia NOS 775.1 - Neonatal diabetes mellitus	1) Any HbA1c ≥ 7 2) Two occurrences within one year of: Fasting Glucose ≥ 7

Table 3: The current CPCSSN diabetes case definition. A patient is considered diabetic if they meet the criteria from any of the columns.

Billing	Problem List	Medication
Minimum two occurrences of the following codes within two years: 1) 401 - Essential hypertension 2) 402 - Hypertensive heart disease 3) 403 - Hypertensive chronic kidney disease 4) 404 - Hypertensive heart and chronic kidney disease 5) 405 - Secondary hypertension	Any occurrence of the following codes: 1) 401 - Essential hypertension 2) 402 - Hypertensive heart disease 3) 403 - Hypertensive chronic kidney disease 4) 404 - Hypertensive heart and chronic kidney disease 5) 405 - Secondary hypertension	ATC Code C07AB04, C09XA02, C03DB01 C08CA01, C07AB03, C07CB03 C09AA07, C09AA01, C07AG02 C03BA04, C09AA08, C09AA02 C09BA02, C09CA02, C09DA02 C08CA02, C09DA02, C08CA02 C09AA09, C03AA03, C03EA01 C03BA11, C09CA04, C09DA04 C09AA03, C09BA03, C09DA01 C02LB01, C03BA08, C09CA07 C07AA06, C09AA10, C03DB02 C09CA03, C08DA01 The following diagnoses, if met, make the medication criteria alone insufficient: 346 - Migraines 428 - CHF 410 - Myocardial Infarction 412 - Myocardial Infarction 250 - Diabetes 427 - Cardiac Arrhythmia 333.1 - Tremor 456.0 - Esophageal Varices 456.1 - Esophageal Varices

Table 4: The current CPCSSN hypertension case definition. A patient is considered hypertensive if they meet the criteria from any of the columns.

Billing	Problem List	Medication
Age $\geq$ 35 for any of the below criteria		
Any occurrence of the following codes: 1) 491.2 - Chronic bronchitis 2) 492 - Emphysema 3)496 - Chronic airway obstruction not elsewhere classified	Any occurrence of the following codes: 1) 491.2 - Chronic bronchitis 2) 492 - Emphysema 3)496 - Chronic airway obstruction not elsewhere classified	ATC Code R03BB04 R01AX03 R03BB01 R03AK04 The following diagnoses, if met, make the medication criteria alone insufficient: 493 - Asthma

Table 5: The current CPCSSN COPD case definition. A patient is considered COPD case positive if they meet the criteria from any of the columns.

Billing	Problem List	Medication
Any occurrence of the following codes: 1) 296 - Episodic mood disorders 2) 311 - Depressive disorder not elsewhere classified	Any occurrence of the following codes: 1) 296 - Episodic mood disorders 2) 311 - Depressive disorder not elsewhere classified	ATC Code N06CA01 N06AB04 N06AB10 N06AB03 N06AB08 N06AX11 N06AG02 N06AB06 N06AF04 The following diagnoses, if met, make the medication criteria alone insufficient: 300 - Anxiety disorders

Table 6: The current CPCSSN depression case definition. A patient is considered a depression case if they meet the criteria from any of the columns.

Billing	Problem List
Any occurrence of the following codes: 1) 715 - Osteoarthritis and allied disorders 2) 721 - Spondylosis and allied disorders	Any occurrence of the following codes: 1) 715 - Osteoarthritis and allied disorders 2) 721 - Spondylosis and allied disorders

Table 7: The current CPCSSN osteoarthritis case definition. A patient is considered an osteoarthritis case if they meet the criteria from any of the columns.

Billing	Problem List	Medication	Encounter Diagnosis
Include: 345.* - Epilepsy and recurrent seizures 780.3 - Convulsions Exclude: 345.2 - Petit mal status (without AED) 345.3 - Grand mal status (without AED) 780.3 - Convulsions (without AED) * AED = Antiepileptic drug	Include: 345.* - Epilepsy and recurrent seizures 780.3 - Convulsions Exclude: 345.2 - Petit mal status (without AED) 345.3 - Grand mal status (without AED) 780.3 - Convulsions (without AED) * AED = Antiepileptic drug	ATC Code N03AF01, N05BA09, N03AE01 N03AX12, N03AX09, N03AX14 N03AF02, N03AA02, N03AB02 N03AX16, N03AA03, N03AX11 N03AG01, N03AG04, N03AX18 N03AD01, N03AX10, N03AF03 Exclusion: Exclude any case identified only through medication criteria	Include: 345.* - Epilepsy and recurrent seizures 780.3 - Convulsions Exclude: 345.2 - Petit mal status (without AED) 345.3 - Grand mal status (without AED) 780.3 - Convulsions (without AED) * AED = Antiepileptic drug

Table 8: The current CPCSSN epilepsy case definition. A patient is considered an epilepsy case if they meet the criteria from any of the columns.

Billing	Problem List	Medication	Encounter Diagnosis
Include: 332.* - Parkinson's disease	Include: 332.* - Parkinson's disease	ATC Code N04BD02 N04BD01 N04BA01 N04BA03 Exclusion: Exclude any case identified only through medication criteria	Include: 332.* - Parkinson's disease

Table 9: The current CPCSSN Parkinson's disease case definition. A patient is considered a Parkinson's case if they meet the criteria from any of the columns.

Billing	Problem List	Medication	Encounter Diagnosis
Include: 290.* - Dementias psychosis 331.* - Other cerebral degenerations 294.1 - Dementia in conditions classified 294.8 - Other persistent mental disorders 797.* - Senility without mention of psychosis 438.* - Late effects of cerebrovascular disease Then Exclude: 797.* (without med) 438.*(without med) 290.8 - Other specified senile psychotic conditions 290.9 - Unspecified senile psychotic condition 331.3 - Communicating hydrocephalus 331.4 - Obstructive hydrocephalus 331.5 - Idiopathic normal pressure hydrocephalus (INPH) 331.81 - Reye's syndrome	Include: 290.* 331.* 294.1 294.8 797.* 438.* Then Exclude: 797.* 438.* 290.8 290.9 331.3 331.4 331.5 331.81	ATC Code N06DX01 N06DA03 N06DA04 N06DA02	Include: 290.* 331.* 294.1 294.8 797.* 438.* Then Exclude: 797.* 438.* 290.8 290.9 331.3 331.4 331.5 331.81

Table 10: The current CPCSSN dementia case definition. A patient is considered a dementia case if they meet the criteria from any of the columns.

4 Methodology

Here we will outline the new method of case definition development, which should be considered a candidate to replace or supplement the committee-created case definition process. This method involves 5 steps:

1. Obtaining the data and chart review
2. Generating model features
3. Selecting the optimal tuning parameters (Bootstrap Phase)
4. 10-fold cross validation
5. Final case definition interpretation

4.1 Obtaining Data and Chart Review

In order to create case definitions using machine learning methods, a sample with a reliable reference standard must be established. It is important that the sample contains both cases and non-cases of the disease, and that the prevalence of the condition in the sample is generalizable to the population. The reference standard that will be used in this study is a manual chart review of the EMR by trained abstractors.

In order to assess the validity of a case definition against an EMR, we must calculate a few metrics.

4.1.1 Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value, and Accuracy

In the binary case of disease-status, where both the case definition and the reference standard result in a positive or negative indication of disease status, we can form a 2x2 table to represent the agreement and disagreement of the two measures. An example of this is table 11.

Here we use the symbols TP, FP, FN, and TN to represent the counts for each category in the table. TP stands for the "true positives". The true positives are those patients that were correctly classified as having the disease, as per the gold standard and the case definition. FP represents the "false positives", which are those whom the case definition classified as having the disease, but in truth do not have the disease. FN represents the "false negatives", which are those whom the case

	Reference Standard +	Reference Standard -
Test +	TP	FP
Test -	FN	TN

Table 11: A 2x2 classification table.

definition classified as not having the disease, but do in fact have the disease. TN represents the "true negatives", or those that were correctly classified as not having the condition or disease. In most cases, these four values are used to estimate four different measures of accuracy: sensitivity, specificity, positive predictive value, and negative predictive value.

Sensitivity is defined as the proportion of those that do in fact have the disease that are properly classified [144]. Sensitivity is sometimes referred to as recall when used outside of epidemiology and medicine. If the sensitivity of a case definition is 90% then we would expect about 90 out of every 100 patients with the disease to be correctly classified as having the disease by the case definition. Sensitivity is represented as follows.

$$Sensitivity = \frac{TP}{TP+FN} \tag{19}$$

Specificity is defined as the proportion of those that do not have the disease that are correctly classified. If the specificity of a case definition is 95%, we would expect about 95 out of every 100 patients without the disease to be correctly classified as not having the disease by the case definition. Specificity is represented as follows:

$$Specificity = \frac{TN}{TN+FP} \tag{20}$$

Positive predictive value (PPV) is defined as the proportion of those classified as having the disease that do in fact have the disease [145]. PPV is sometimes referred to as precision. If the PPV of a case definition were 80%, we would expect about 80 out of every 100 patients who were classified as having the disease to be correctly classified as having the disease. PPV is represented as follows:

$$PPV = \frac{TP}{TP+FP} \tag{21}$$

Negative predictive value (NPV) is defined as the proportion of those classified as not having the

disease that truly do not have the disease. If the NPV of a case definition is 85%, we would expect about 85 out of every 100 patients who were classified as not having the disease to be correctly classified as not having the disease. NPV is represented as follows:

$$NPV = \frac{TN}{TN + FN} \tag{22}$$

Finally, accuracy is simply defined as the proportion of classifications that were correct:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{23}$$

Some will refer to the complement of accuracy (1-Accuracy) as the misclassification rate.

4.1.2 F1-Score

The F1-score is the harmonic average of sensitivity and PPV [145]. This is defined as:

$$F1 = \frac{Sensitivity * PPV}{Sensitivity + PPV} \tag{24}$$

Here the sensitivity is a measure of accuracy for those that are true cases, but the positive predictive value includes non-cases as well. This measure is often better suited for conditions with low prevalence compared with the misclassification rate.

4.1.3 G-Mean

The G-mean, similar to the F1-score is a function of the sensitivity and the PPV. Rather than the harmonic mean, the G-mean is the geometric mean:

$$F1 = \sqrt{Sensitivity * PPV} \tag{25}$$

It can be seen that these two measures are very similar in definition, and often result in the same set of tuning parameters and case definitions.

4.1.4 Naïve Mean

This is a little-used metric that has gone by different names. It will be called the Naïve mean in this study. This is the arithmetic mean of the sensitivity and specificity.

$$Naïve\ Mean = \frac{sensitivity + specificity}{2} \tag{26}$$

This weighs sensitivity and specificity equally, regardless of prevalence. The downside of this metric is that it has no incentive to yield strong predictive values.

4.1.5 Prevalence and Validity

It is important to consider how prevalence can influence the validity measures of a case definition. Prevalence, as discussed earlier, is the proportion of a population with a disease at a specific point in time or over some period of time. When we reexamine the definitions of PPV and NPV, it can be shown that they are dependent on disease prevalence. PPV uses both the true positive and false positive counts. The true positives are those with the disease and the false positives are those without. By using both values in the calculation of PPV, it is clear that both PPV and NPV are directly dependent on the prevalence of the condition [146]. This dependence can also be seen by rearranging the PPV and NPV formula as such:

$$PPV = \frac{sensitivity * d}{sensitivity * d + (1 - specificity)(1 - d)} \tag{27}$$

$$NPV = \frac{specificity * (1 - d)}{specificity * (1 - d) + (1 - sensitivity) * d} \tag{28}$$

Where d is equal to the disease prevalence [147]. It is widely held that as prevalence increases so does the PPV, and NPV decreases [148].

Sensitivity is calculated with the true positives and the false negatives, both of which truly do have the disease. Specificity is calculated with the true negatives and the false positives, both of which do not have the disease. This has led to the common stance that sensitivity and specificity are more or less invariant to prevalence [149].

The issue of prevalence becomes important when evaluating the accuracy of a case definition. If the sample has an artificially high or low prevalence (due to some type of bias or mechanism in the selection process) the validity estimates may be inaccurate in the real-world application of the case definition. Therefore, it is important to validate case definitions in settings that resemble the true population. It has been shown that many administrative data validation studies have reported PPV and NPV inappropriately, as the prevalence of the validation cohort did not represent the true population [150].

4.2 Generating the Model Features

In order to create a classification model, one must create the data frame that holds each of the candidate features. There is a false sentiment with machine learning that the algorithms do all the work. This is not true; the user must specify the candidate features and perform data manipulation and transformation in order to fit a machine learning model.

For simplicity, we have created binary features only. This data frame is comprised of a single row for each unique patient in the sample. The features are added as columns to this data frame. They are set to 0 if a patient does not meet the criteria and 1 if they do.

4.2.1 ICD Codes

In the CPCSSN data model, ICD codes can be found in the Billing, Encounter Diagnosis, and Health Condition tables. The codes from the Billing table are the classic administrative data offerings, being the codes that were submitted for reimbursement. The Encounter Diagnosis and Health Condition ICD codes are assigned via CPCSSN's cleaning algorithms. This is in an attempt to summarize the free-text that is put in the encounter diagnosis and problem list tables into a more familiar format and language. Not every entry in the Encounter Diagnosis and Problem List tables are assigned an ICD code, as there is sometimes ambiguity in the text. However, the free-text is also available, and is included in the analysis.

Every unique ICD code that appeared in the sample was used as a feature. This includes truncated codes, such that there are unique features for ICD code "250.*" and "250.0". In this example, the "250.*" feature includes the "250.0" cases as well. This method allows for the distinction between whole-number codes, and specific decimal indications.

Furthermore, separate features identifying any instance of a code, two instances within one year, and two instances within two years were selected. Temporal cosiderations have been included case definitions in the past, so it is included here. The features were also stratified by which table they come from, so there is a feature for ICD code 401 in the billing record and an independent record for code 401 in the encounter diagnosis table.

Note that this method of feature generation requires no clinical knowledge or specialization to the condition of interest. Every unique ICD code is included.

4.2.2 ATC Codes

The feature generation of the ATC codes is similar to the billing codes. The Medication table is the only source of such information. Quantity, dose strength, and route were not included as features, as there is not a standard or consistent notation for these fields in the CPCSSN data model.

Every unique ATC code was included as a unique feature, including temporal stratification of two occurrences within one year and two occurrences within two years. Truncation of the ATC codes were conducted such that families and classes of drugs were also captured, and not just the active ingredient.

4.2.3 Laboratory Values

CPCSSN collects lab data for the test and measurements seen in table 12.

Glucose Tolerance
Fasting Glucose
Hemoglobin
Total Cholesterol
Triglycerides
Serum Creatinine
Thyroid Stimulating Hormone
Albumin Creatinine Ratio
HDL
HbA1c
LDL
Microalbumin
eGFR/GFR
International Normalized Ratio

Table 12: The laboratory information that is recorded in the CPCSSN database

Features were created for each of the above tests. A range of plausible values was determined. For example, for the HbA1c test, the result is stored as a percentage. Plausible values are in the 3-11% range, so a systematic number of cut-points were calculated. A feature for an HbA1c score

greater than 7.0, 7.1, 7.2, etc. was generated, including features for two occurrences within one year, and two occurrences within two years.

4.2.4 Referrals

Referral information in CPCSSN is primarily stored as a SNOMED CT code. For example, the code "183519002" refers to a referral to a cardiology service. A feature was generated for each unique referral code. There are many referrals that were not given a SNOMED CT code, but clinician text was assigned. The information from the text was also be used.

4.2.5 Free Text

Another asset that the EMR brings is the inclusion of physician free-text. Invaluable information that was previously unavailable in administrative data sources is now readily available. This also brings complications though. Automated recognition and handling of free text is a field of study commonly referred to as Natural Language Processing (NLP). It is being widely used in combination with machine learning for all sorts of classification and prediction purposes [151]. This is a broad area of study, but for this analysis, a tokenization or "bag-of-words" approach will be implemented.

Tokenization is a very simple concept: partition a sentence into individual words. Rather than the computer storing the entry as, "Essential Hypertension - r/o secondary causes," we can store the sentence as "Essential," "Hypertension," "r/o," "secondary," and "causes." This tokenization will allow a computer to see that the word hypertension appears in the entries of the patients with a reference standard indication of hypertension. Sometimes this process is referred to as "bag of words" because the structure of the sentence is not preserved, just the words within it [152].

In terms of our classification problem wherein we hope to estimate y $\varepsilon\{0,1\}$ with x $\varepsilon\{x_1, x_2, ..., x_n\}$, this means creating a binary or count variable for each word that appears in the physician free-text. Most text is case-sensitive, so a common practice in NLP methods is to convert all characters into lower case, remove all symbols, and sometimes remove all numbers as well. This is a simple way of standardizing the text.

Basic tokenization methods fail to take into account negative indication terms that indicate the absence of a disease. An important distinction should be made for terms that include a negative descriptive statement; for example, "no diabetes" or "borderline hypertension". Clearly, we do

not want to include these statements as indications of disease, but the basic tokenization or bag of words method is unable to properly classify the statement.

The common way to deal with this problem is to expand the bag-of-words method to include two or three-word combinations. This will allow us to observe the cases in which a negative indication term is applied to a condition. Obviously, using this approach is more computationally expensive, as we have now increased the number of tokenization's since we will still include each word on its own as well. In large datasets, this can lead to a large number of variables being created. Some classifiers with limited feature selection tools may be unable to sift through that many variables.

One solution to this problem is to only look for negative indication terms around certain key words. If you are creating a case definition for osteoarthritis, you can specify that the tokenization include the word "osteoarthritis" by itself as a variable, with the word preceding it as another variable (e.g. "no osteoarthritis" or "query osteoarthritis"), or even with two words preceding it as another variable (e.g. "query for osteoarthritis"). This method is instrumental in dealing with negative indication terms. A similar method is to create a list of potentially negative indication terms (e.g. "no", "query", "maybe", etc.) and discount the words following it.

For this study, free-text features were generated using a bag-of-words approach. Free text was obtained from the fields shown in table 13.

CPCSSN Table	Field
Encounter	reason orig
Encounter Diagnosis	diagnosistext orig diagnosistext calc
Health Condition	diagnosistext orig diagnosistext calc
Medication	name orig name calc
Referral	name orig name calc

Table 13: The tables and fields from which text was used in feature generation

This generates a feature for each unique word that appears in these fields. Each of these fields is treated separately, so there can be differentiation between the source of the text. This is important as there may be a significant difference between the words appearing in some fields compared to others. For example, it may be more indicative of true disease status to have the word "hypertension" appear in the problem list as opposed to the encounter diagnosis field.

Further text feature generation was performed using wildcard searches for keywords, phrases, and truncations. For example, the wildcard search for "hypert*" indicates a feature that includes the text "hypertension" and "hypertensive", words that would not be deemed the same by the bag-of-words approach. Similarly, groupings of words can be put together to improve the prediction of a feature. For example, when trying to distinguish between type 1 and type 2 diabetes, we generated a feature that combines "type 1", "t1dm", and "insulin dependent diabetes". Also, features were generated for keywords across all text fields.

4.3 Validation Methods

When a machine learning algorithm is implemented, there are steps that should be taken to maximize accuracy and get fair estimates of validity. The two most common methods for these steps are called bootstraps [153] and cross-validation. Both of these methods are used to split the available data into training and test frames. The training frame is set of observations used to train the model, and the test frame is an independent set of observations where the trained model can be evaluated in terms of its predictive accuracy.

4.3.1 Bootstrap Validation

The bootstrap takes a random sampling approach to the splitting of the available data [154, 155]. If there are N observations in a sample, the bootstrap will randomly sample some number X of those observations with replacement. The most common value of X is equal to N. For example, if we have 1000 observations, the bootstrap would randomly select 1000 observations with replacement, meaning a single observation can be included multiple times. This constitutes the training data. The test data is comprised of those observations that were not selected for the training data.

It is important to note that by randomly sampling observations with replacement for the training set, we are using repeated observations. It is common for observations to be included in the

training set several times over. The test set, on the other hand, does not contain any duplicate observations. This means that although we are training our model on a set containing duplicated entries, the models are evaluated using a set completely unique from the training set, with no duplicates. Simulations have shown that using these methods provides reasonable estimates of the sampling distribution of statistics [156].

The model is then trained on the training set, and evaluated on the test set. It is generally suggested that the process should be repeated between 1000 and 2000 times [157] These steps are repeated many times over, in order to get estimates of validity and variability for the method. This method is commonly used in model assessment, as well as getting estimates of variance for non-parametric or small sample statistics [158].

4.3.2 K-fold Cross Validation

K-fold cross validation takes a different approach than the bootstrapping method. Rather than taking random samples with replacement, this method generates K "folds" of the data [159]. The most common value of K is 10, which has been shown through simulations to have desirable properties [160].

With 10-fold cross validation, the available observations are randomly distributed into 10 equally-sized parts, or folds. Nine of these folds are used for training, and the remaining fold is used for testing. This process is repeated 10 times, such that each fold has been tested on once, and trained on nine times. The result is 10 different estimates of validity, allowing for an assessment of the variability of the estimates.

4.3.3 Bootstrap vs Cross Validation

In this study, the bootstrapping method will be used for selecting optimal tuning parameters, and 10-fold cross validation will be used for determining the final estimates of sensitivity, specificity, positive predictive value, negative predictive value, and accuracy.

The bootstrapping method is ideal for setting the tuning parameters because they allow the algorithm to train on the exact sample size that is observed. If there are N observations in a sample, the bootstrapped method trains on N observations each time. This is important when it comes to selecting parameters that shape the complexity or size of the final case definition. Also, the bootstrap allows for more than 10 iterations of the algorithm. The more iterations at this

step, the better, as it allows for the differentiation of tuning parameters with more clarity and less variation.

The 10-fold cross validation approach will be used for the final assessments of accuracy because it gives a more accurate estimate of confidence intervals than the bootstrap method. It is widely held that the cross-fold method yields estimates with less bias, but more variability than the bootstrap method [161]. It is vital that the final estimates be conducted on N observations. More than N observations will give more narrow confidence intervals than we have the sample size to estimate. The 10-fold cross validation is well suited to this problem, as we test on each observation once, and once alone. The bootstrap does not ensure that we are testing on each observation.

4.4 Loss Functions

Every algorithm or model is fit to minimize or maximize some loss function. For example, linear regression minimizes the sum of the squared residuals. Logistic regression maximizes the likelihood function. For the machine learning algorithms being used, there are several candidate metrics that can be minimized or maximized depending on the goals of the case definition. For example, a condition with high prevalence may be well suited by minimizing the misclassification rate. On the other hand, a rare condition with a low prevalence may benefit from maximizing the positive predictive value, to ensure that those identified are true cases, or maximize the F1-score, to ensure a higher rate of sensitivity.

4.4.1 Selecting Optimal Tuning Parameters

The first step in selecting optimal tuning parameters is to select the metric that should be minimized/maximized. Here we have minimized the misclassification rate, as the conditions considered were not rare. This means that we will ultimately generate one case definitions per algorithm for each condition.

A bootstrapped validation method was implemented to select the tuning parameters for each of the aforementioned metrics. Random sampling with replacement of the available observations was performed to form the training set. The remaining observations were used as the test set. For each combination of tuning parameters, the bootstrap was performed 30 times. Once all the iterations were performed, the values were plotted using a Loess smoother, and the maximum/minimum

value was selected as the optimal tuning parameters for that particular metric. Here we will discuss the various tuning parameters for each of the algorithms.

4.4.1.1 Tuning Parameters - C5.0 Decision Tree

The C5.0 decision tree has a number of potential tuning parameters. The most important is the ability to input a loss matrix, allowing the user to create decision trees that specialize in sensitivity or specificity, regardless of the prevalence of the disease. The matrix specifies a ratio between the weight of a false positive versus a false negative. Ratios from 1:10 thru 10:1 are included as tuning parameters.

The C5.0 decision tree also has a parameter called "CF", standing for confidence factor. This is a value between 0 and 1, which determines the required amount of "confidence", or a measure of the required change in information gain to make a split. Essentially, this is a continuous measure of how complex, or how many splits the decision tree should have. Values closer to 0 will be larger, more inclusive trees, whereas the values closer to 1 will be smaller trees, with fewer splits. We will iterate through values 0.01 through 0.8 in the bootstrapping phase.

There are a few other parameters that might be of use as tuning parameters. Winnowing is a feature available to the C5.0 algorithm that "winnows" out features that provide no predictive value. This feature often makes no difference in the final decision tree but is useful in reducing compute time when there are many features. C5.0 also includes a feature that decomposes the decision tree into a set of rules. This is useful for determining the final case definition, but makes little difference in the splits. Another feature is called "minCases", which specifies the minimum required number of observations at a node in the decision tree for a further split to be made. If this parameter is set too low, overfitting and unnecessary splitting may occur. This value will be set at 10, and remain unchanged.

4.4.1.2 Tuning Parameters - CaRT/rPart

Similar to the C5.0 algorithm, the Recursive PARTitioning (RPART) implementation of the CaRT algorithm includes the input of a loss matrix, and ratios 1:10 thru 10:1 were implemented. It also has parameter called "cp", which stands for complexity parameter, which functions the same

as the "CF" parameter in the C5.0 algorithm. CaRT also includes an analog to the "minCases" feature. An additional feature that is included here is called "maxdepth" which limits the total height of the tree. This will not be limited in the CaRT iterations. Note that the CHAID algorithm has a similar feature.

4.4.1.3 Tuning Parameters - Loss Matrices

The C5.0 and CaRT implementations both include a feature that allows a user to specify the weights of false positives and false negatives. This is a critical feature, as it allows these algorithms to be very versatile when attempting to minimize or maximize different metrics.

Take, for example, a condition with a prevalence of 10%. A machine learning algorithm minimizing the misclassification rate will favor algorithms with high specificity, but will not have an incentive to have high sensitivity. In this sample, if we were to assume that no patients had the condition, we would be 90% correct, with a misclassification rate of 10%. This does not provide the algorithm much incentive to improve sensitivity if it is at the expense of specificity. The naïve mean, however, of the sample if all observations are assumed as non-cases is the average of 0% (sensitivity) and 100% (specificity), which is 50%. Clearly there are metrics are much improved with an increase in sensitivity.

If a decision tree algorithm with a 10% prevalence is fit with a loss matrix that weighs false positives 10 times higher than false negatives, the splits made by the algorithm will be selected with a roughly equal balance of sensitivity and specificity. It should follow that implementations with a 1:1 loss matrix will maximize the misclassification rate, and a 1:10 loss matrix will maximize the naïve mean.

4.4.1.4 Tuning Parameters - CHAID

The CHAID algorithm does not include a loss matrix, which inhibits its usefulness. There are other parameters that can be adjusted. The first is the threshold for the p-values. As stated earlier, the CHAID algorithm determines splits via the p-value of a Chi-squared test. As this threshold is adjusted, the number of splits increases or decreases accordingly. Iterations from 0.001 to 0.2 were performed. Another parameter is the maximum height of the tree, similar to the "maxdepth" feature

in the rPart implementation. Iterations of this variable are performed from 1-7 in the bootstrapping phase.

4.4.1.5 Tuning Parameters - Forward Stepwise Logistic Regression

Forward stepwise logistic regression has only one parameter that is adjusted. This is the parameter that specifies the penalty added to the AIC calculation for an additional variable being added to the model. As features are added, the AIC increases with each new feature. At a certain point, the addition of another variable is outweighed by this penalty parameter, and no further variables are added to the model. Iterations are performed across all plausible values of this parameter.

4.4.1.6 Tuning Parameters - LASSO Logistic Regression

LASSO regression, similar to forward stepwise logistic regression, just has one parameter that is to be adjusted. This parameter is the penalty or limitation applied to the total sum of the absolute values of the regression coefficients. When this penalty is low, more features are included in the model. When this penalty is high, very few features are included. Iterations will be performed across all values of this parameter.

The result of the bootstrapping phase is a set of parameters for each of the machine learning algorithms, minimizing one of the metrics previously described, that can now be evaluated for predictive accuracy.

4.5 Validation and Interpretation

Ten-fold cross validation was used to get final estimates of validity for the case definitions. This splits the data into 10 unique folds, nine of which are used for training, and one used for testing. This is repeated 10 times, so that each observation is used for testing once. The result is a 2x2 table that can be used to calculate final estimates. This was performed four times for each algorithm, once for each of the metrics used in the tuning parameter selection phase.

Once the validity estimates are established, the final case definitions are created by training the model on all the available observations. The decision tree case definitions are easily interpretable.

The final tree can be logically decomposed into a set of rules. In fact, the C5.0 algorithm has an option that constructs sets of rules rather than trees.

The logistic regression algorithms require a larger amount of interpretation. These models result in linear combinations of log odds estimates of coefficients. These are not directly interpretable as case definitions. In order to decompose these results to a case definition, all features with positive coefficients are added to an inclusion list, and features with negative coefficients are added to an exclusion list. A patient is assumed to be a non-case unless they meet more inclusion criteria than exclusion criteria.

5 Manuscript 1 - Comparing feature selection methods for the development of common chronic disease case definitions in a primary care electronic medical record database

5.1 Background

With the widespread adoption of electronic medical records (EMRs), the quantity and types of data routinely collected at the primary care level has expanded beyond simple coding systems. These systems now include information such as prescription medications, referral information, laboratory values, examination values, physician free text describing health conditions and encounter diagnoses. Increasingly, these data is being used for health services research, practice audit and feedback, and chronic disease surveillance [162, 163, 30, 164]. However, these activities require valid and robust case definitions to identify cohorts of patients with certain conditions. A case definition is a set of criteria in a patient's EMR that indicates disease status. The Canadian Primary Care Sentinel Surveillance Network (CPCSSN) is an organization that collects the EMR data for over 1.7 million Canadians from many EMR vendors and transforms them into a common data model. CPCSSN has developed case definitions for eight chronic conditions including hypertension, diabetes, osteoarthritis, depression, dementia, epilepsy, COPD, and Parkinson's disease (Williamson et al) [41]. Each case definition was developed by a committee of physicians, epidemiologists, biostatisticians and database experts before validation. This approach to case definition development relies on the committee's clinical experience to determine the data features or patterns that are indicative of the condition of interest before the definition is operationalized, and is then compared to a reference standard.

The purpose of this work is to introduce a new approach to developing and validating case definitions using machine learning methods, which removes the requirement of first developing a consensus-based definition candidate for the validation exercise. We propose a method in which the results of the reference standard are used as the training data for machine learning classifiers that can be interpreted as case definitions, specialized to maximize predictive accuracy. To date, few studies have applied machine learning techniques to create case definitions or decision rules [165, 106, 166, 101]. A few studies have used decision trees to create rule sets or models in administrative data settings [167, 168, 169]. Most of the published machine learning EMR studies

explore predictive modelling rather than case definition development [170, 97, 171]. Predictive modelling excercises do not require a human readable set of rules, and tend to use complex "black-box" machine algorithms methods. This proposed method is likely to improve efficiency and reduce the costs associated with developing new case definitions. This new method is well suited to applications in EMR databases, where there are many candidate features.

5.2 Methods

5.2.1 Data Source

CPCSSN is Canada's largest and only pan-Canadian primary care EMR database [136]. CPCSSN is a federation of 11 primary care practice-based research networks from 8 provinces and territories in Canada. CPCSSN extracts data on over 1.7 million Canadians from 12 different EMR systems and aggregates the data into a standardized format using various cleaning algorithms and transformations [172, 173].

This project uses data collected by Williamson et al. in 2014 as the reference standard [41]. For that study, the electronic chart of 1920 CPCSSN patients were reviewed for the presence or absence of hypertension, osteoarthritis, diabetes, and depression, and four other conditions in an age-stratified random sample. A chart review was deemed to be the optimal reference standard for accuracy and cost. The size of this sample was selected based on target confidence interval widths for proportions. The extraction was performed on June 30th, 2012. The reference standard was established independently of case definitions. Sampling was conducted such that 90% of the sample was over the age of 60. Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated for each of the committee-created case definitions. This was done to ensure that there would be a large enough number of chronic disease cases and ensure coverage for each condition. This strategy falsely inflated the prevalence of chronic disease in the sample, which has minimal effect on the estimates of sensitivity and specificity, but likely overestimates the PPV and underestimate the NPV, as the predictive values are related to the prevalence of the condition. This project attempts to create case definitions for hypertension, diabetes, osteoarthritis, and depression.

We had access to the entire CPCSSN record for each of these 1920 patients from the June 30, 2012 extraction. The data was received by permission from CPCSSN after the study was reviewed

by the University of Calgary Conjoint Health Research Ethics Board.

5.2.2 Data Features

Features were selected from the CPCSSN billing, encounter diagnosis, problem list, medications, laboratory, and examination tables. From the billing, encounter diagnosis, and problem list table, indicator columns for each unique ICD-9 code that appeared in any of the 1920 patient's EMR was created, with separate indicators for codes appearing twice within one year and twice within two years. From the CPCSSN medications table, each unique Anatomical Therapeutic Chemical code that appeared in any of the 1920 patient's medication entries was identified. From the CPCSSN laboratory table, binary indicators for the occurrence of HBA1C and fasting glucose tests with scores over multiple thresholds (e.g. HbA1C > 7) were created. An indicator for each word that appears in the physician free-text, as well as disease-specific keyword searches were included. This text appears in the problem list, encounter diagnosis, encounter reason, and medication reason fields.

5.2.3 Feature Selection Algorithms

There are many machine learning classification algorithms, but not all are suitable for this problem. To create a case definition, a method that can be interpreted as a series of rules is preferred. We focused on feature selection algorithms that were capable of creating simple rule sets. We compared 3 decision tree classifiers and two methods of feature selection for logistic regression.

The 3 decision tree classifiers are Ross Quinlan's C5.0 classifier [80, 81], Terry Therneau's implementation of CaRT methodology in R [76, 77], and Kass' Chi-square automatic interaction detection (CHAID) [83, 84]. The C5.0 method decides splits using an information gain ratio. Earlier methods simply used an information gain, but the ratio component was implemented to create more concise trees, as it increases the incentive to make more equal splits. Therneau's CaRT decision tree method uses a purity measure called the Gini Index to determine optimal splits. The CHAID method determines splits via the p-value of a Chi-Squared test. These differences in purity measures mean they often yield different decision rules on the same training sets. The final class assignment of a decision tree is determined by the majority class of the terminal node.

These were compared with Forward Stepwise Logistic Regression [68, 65] and LASSO penalized logistic regression [174, 71]; both employ methods of feature selection that can be interpreted

as an inclusion/exclusion list. The Forward Stepwise method selects features by finding the feature that makes the biggest improvement in a measure called the Akaike Information Criterion, which is based on the likelihood function from the regression. The LASSO method assigns a restriction or "penalty" on the total sum of the coefficients. This restriction forces the model to only include the variables most associated with the outcome. These methods were selected due to their interpretability and popularity in the machine learning classification literature. For forward stepwise and LASSO logistic regression, case definitions were derived from a list of inclusion and exclusion variables. Variables with positive coefficients were added to the inclusion list and variables with negative coefficients were added to the exclusion list. If a patient's EMR indicates more entries on the inclusion list, they are considered a case. This allows a user to create a human readable set of rules rather than a linear combination of regression coefficients and determing class using a cutpoint in the log-odds estimate, which is not human readable.

5.2.4 Tuning Parameter Selection

Each of the feature selection algorithms has tuning parameters that limit the complexity of the model. The decision tree algorithms have a pruning parameter that limits the number of splits made. The LASSO algorithm has a parameter that limits the magnitude of coefficients. The Forward Stepwise method includes a parameter that adds a higher penalty for each new feature included in the model. In order to find the optimal tuning parameter values, a bootstrap technique was used. This involved the repeated random sampling with replacement of the 1920 observations, where not all observations are included in the training of the model. The observations upon which the model were not trained was used as the test data. The training set always contained 1920 observations, and the test set contained an average of 707 unique observations. This was repeated 30 times for each parameter value, and the value that minimized the mean misclassification rate was selected as the optimal tuning parameter.

5.2.5 Statistical Analysis

The machine learning algorithms are trained using the chart review as the reference standard. The committee-created case definitions were also compared against the reference standard. In order to get unbiased estimates, sensitivity, specificity, PPV, and NPV were reported using 10-fold cross validation along with 95% confidence intervals. 10-fold cross validation is a technique in which

the sample is split into 10 distinct and equal parts. The predictive model is trained on 9 of the 10 parts, and tested on the remaining part. There is some evidence that the optimal number of folds is 10, and that it performs better than a bootstrap for determining validity [160]. The only patient's excluded for missing data reasons were those that were unable to get a reference standard diagnsosis. The final reported case definitions was trained using all 1920 observations. All analyses were conducted using R Statistical Software version 3.3.1 [175].

5.3 Results

Table 14 shows the patient characteristics, including the prevalence of each of the 8 conditions. The prevalence of each chronic condition was high given that 85.4% of the study cohort was older than 60 years of age. Table 15 describes the 10-fold cross validation results for each of the algorithms across all 4 conditions. For hypertension, sensitivity ranged between 93.1-96.7%, while specificity ranged from 88.8-93.2%. For diabetes, sensitivities ranged from 93.5-96.3% with specificities between 97.1-99.0%. For osteoarthritis, sensitivities ranged from 82.0-84.4% with specificities between 92.7-94.0%. For depression, sensitivities went from 81.4-88.3%, and specificities ranged from 93.4-94.9%. A visual comparison of the validity estimates for the committee-created case definitions and the C5.0 estimates is seen in Figure 5. When comparing these values with the committee-created case definitions, we see that the machine learning methods yield comparable validity measures in each category, and are significantly better in a few conditions. In hypertension, sensitivity and NPV are significantly higher with every of the automated methods. In diabetes, specificity and PPV are significantly higher using C5.0 and CaRT methods. The machine learning case definitions more concise than the committee created case definitions. The CPCSSN Committee Created case definitions are available on www.cpcssn.ca.

Patient Characteristics	Percent
Gender (% Male)	44.5
Age $\geq$ 60 years	85.4
Disease Prevalence:	
Diabetes	16.8
Hypertension	50.1
Depression	20.2
Osteoarthritis	31.6
COPD	7.9
Dementia	4.9
Epilepsy	7.3
Parkinsonism	4.3
Number of Chronic Conditions (Of 8 CPCSSN conditions):	
0	22.7
1	34.6
2	26.1
3+	16.6

Table 14: Patient Characteristics

Condition	Algorithm	Sensitivity	Specificity	PPV	NPV
Hypertension	Committee	84.9 (82.6,87.1)	93.5 (92.0,95.1)	92.9 (91.2,94.6)	86.0 (83.9,88.2)
	C5.0	94.8 (93.1,96.1)	92.2 (90.2,93.7)	92.4 (90.5,93.9)	94.6 (92.9,96.0)
	CaRT	93.1 (91.3,94.6)	93.2 (91.4,94.7)	93.2 (91.4,94.7)	93.1 (91.3,94.6)
	CHAID	95.4 (93.8,96.6)	92.0 (90.0,93.6)	92.2 (90.3,93.8)	95.2 (93.6,96.5)
	FS	96.0 (94.5,97.1)	90.4 (88.3,92.1)	90.9 (88.9,92.6)	95.8 (94.2,97.0)
	LASSO	96.7 (95.3,97.7)	88.8 (86.6,90.7)	89.6 (87.6,91.4)	96.4 (94.9,97.5)
Diabetes	Committee	95.6 (93.4,97.9)	97.1 (96.3,97.9)	87.0 (83.5,90.5)	99.1 (98.6,99.6)
	C5.0	93.5 (90.0,95.8)	99.0 (98.3,99.4)	94.9 (91.7,97.0)	98.7 (98.0,99.2)
	CaRT	94.7 (91.5,96.8)	98.9 (98.3,99.4)	94.7 (91.5,96.8)	98.9 (98.3,99.4)
	CHAID	96.0 (93.0,97.7)	98.4 (97.7,99.0)	92.5 (89.0,96.0)	99.2 (98.6,99.5)
	FS	96.3 (93.4,98.0)	98.2 (97.4,98.8)	91.7 (88.1,94.3)	99.2 (98.6,99.6)
	LASSO	95.3 (92.3,97.3)	97.4 (96.5,98.1)	88.2 (84.2,91.3)	99.0 (98.5,99.5)
Osteo	Committee	77.8(74.5,81.1)	94.9 (93.8,96.1)	87.7 (84.9,90.5)	90.2 (88.7,91.8)
	C5.0	82.0 (78.7,85.0)	93.8 (92.3,95.0)	86.0 (82.9,88.7)	91.8 (90.2,93.2)
	CaRT	82.4 (79.1,85.3)	93.4 (91.9,94.7)	85.3 (82.1,88.0)	91.9 (90.3,93.3)
	CHAID	82.0 (78.7,85.0)	94.0 (92.6,95.2)	86.5 (83.3,89.1)	91.9 (90.2,93.2)
	FS	82.0 (78.7,85.0)	94.0 (92.6,95.2)	86.5 (83.3,89.1)	91.9 (90.2,93.2)
	LASSO	84.4 (81.2,87.1)	92.7 (91.1,94.0)	84.2 (81.0,87.0)	92.7 (91.1,94.0)
Depression	Committee	80.9 (76.9,84.9)	94.7(93.6,95.9)	79.2 (75.1,83.2)	95.2 (94.2,96.3)
	C5.0	84.2 (80.1,87.6)	94.9 (93.6,95.9)	80.7 (76.4,84.3)	96.0 (94.8,96.9)
	CaRT	87.1 (83.2,90.2)	93.4 (92.0,94.6)	76.9 (72.6,80.7)	96.6 (95.5,97.5)
	CHAID	88.3 (84.6,91.3)	93.7 (92.4,94.9)	78.0 (73.8,81.8)	97.0 (95.9,97.8)
	FS	81.4 (77.0,85.0)	94.8 (93.5,95.8)	79.7 (75.3,83.5)	95.3 (94.0,96.3)
	LASSO	81.4 (77.0,85.0)	94.0 (92.6,95.1)	77.3 (72.9,81.3)	95.2 (94.0,96.2)

Table 15: 10-fold cross validation estimates of sensitivity, specificity, PPV, and NPV. Shown are estimates for each of the 5 feature selection algorithms, and the reported validity estimates from the committee created case definitions reported by Williamson et al. 95% confidence intervals are recorded in brackets.

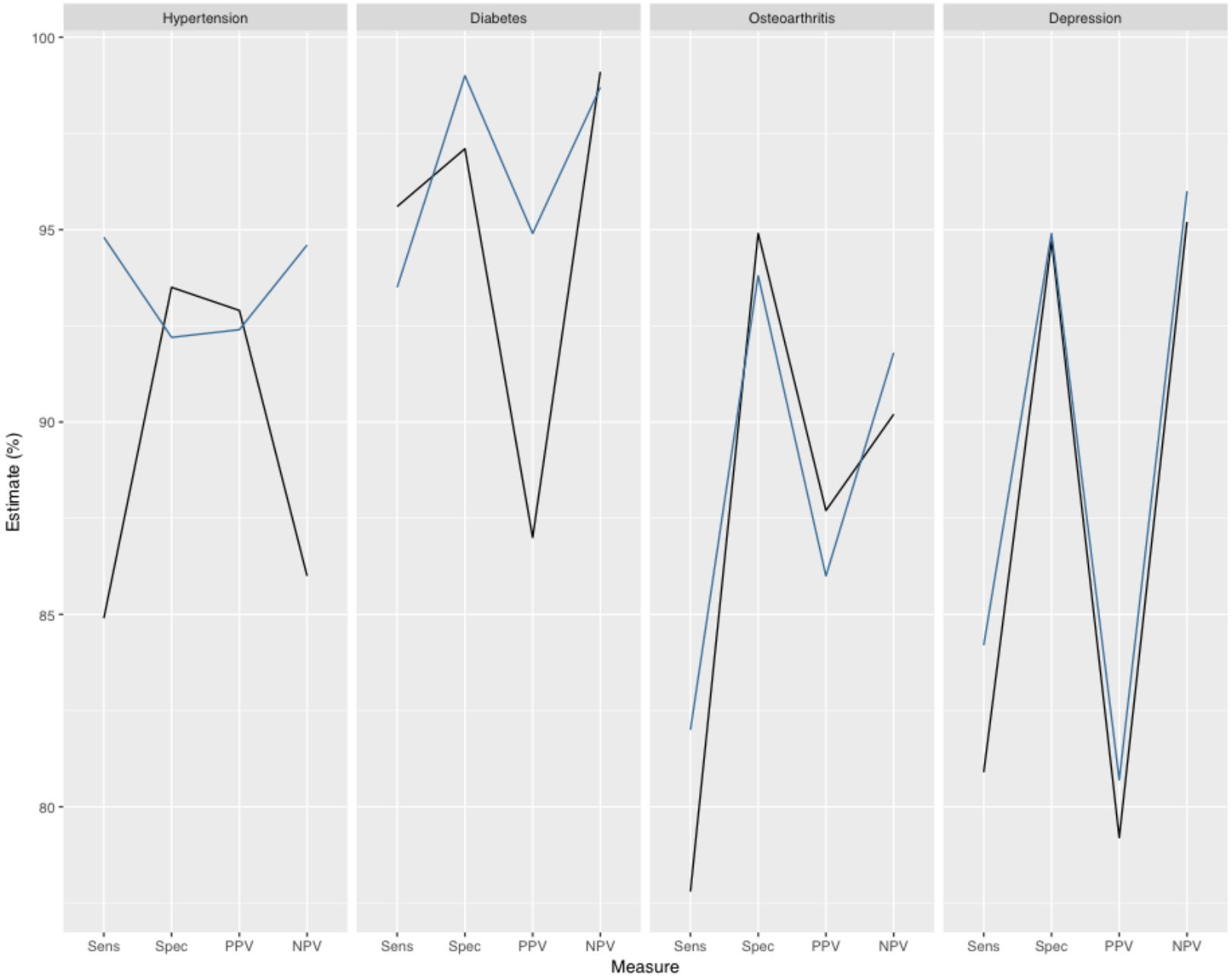

Figure 5: A visual comparison of the validity estimates from the committee-created method (black) and the C5.0 decision tree case definitions (blue).

Condition	Algorithm	Definition
Hypertension	C5.0	Encounter Code 401, Problem list code 401
	CaRT	Encounter code 401, Problem list code 401, ATC code C08CA01(Amlodipine)
	CHAID	Encounter code 401, Problem list code 401
	FS	Encounter code 401, Problem list code 401, Billing code 401 twice in 2 years
	LASSO	Billing code 401, Encounter text “hypertension”, Encounter code 401, Problem list code 401
Diabetes	C5.0	Encounter code 250 twice in 2 years, Problem list code 250
	CaRT	Encounter code 250 twice in 2 years, Problem list code 250 + billing text “acute”, Problem list code 250 + not ATC code N02BA01 (Aspirin)
	CHAID	Encounter code 250, Problem list code 250
	FS	Encounter code 250, Problem list code 250
	LASSO	Billing code 250, Billing text “mellitus”, Encounter code 250, Encounter code 250 2 in 2 years
Osteo	C5.0	Encounter code 715, Problem list code 715
	CaRT	Encounter code 715, Problem list code 715
	CHAID	Encounter code 715, Problem list code 715
	FS	Encounter code 715, Problem list code 715
	LASSO	Billing code 715, Encounter code 715, Problem list code 715
Depression	C5.0	Encounter code 311, Problem list code 311, Encounter code 300.09, Billing code 298 twice in 1 year, Problem list text ”depressive”, ATC Code N06AB10, problem list list code 298, encounter text ”affective” [eg. bipolar affective, seasonal affective], med N06AB04
	CaRT	Encounter code 311, Problem list code 311, ATC Code N06AB04, ATC Code N06AX16, ATC Code N06AB10, Problem list code 298.0, Billing code 296 twice in 1 year
	CHAID	Encounter code 311, Problem list code 311, ATC Code N06AB04, ATC Code N06AB10, Encounter code 298.0, Billing code 296 twice in 1 year
	FS	Encounter code 311, Problem list code 311, ATC Code N06AB04
	LASSO	ATC Code N06AB04, Billing text ”depressive”, Encounter code 311, Problem list code 311

Table 16: The final case definitions for each of the feature selection methods. The commas signify OR statements, the +’s signify AND statements.

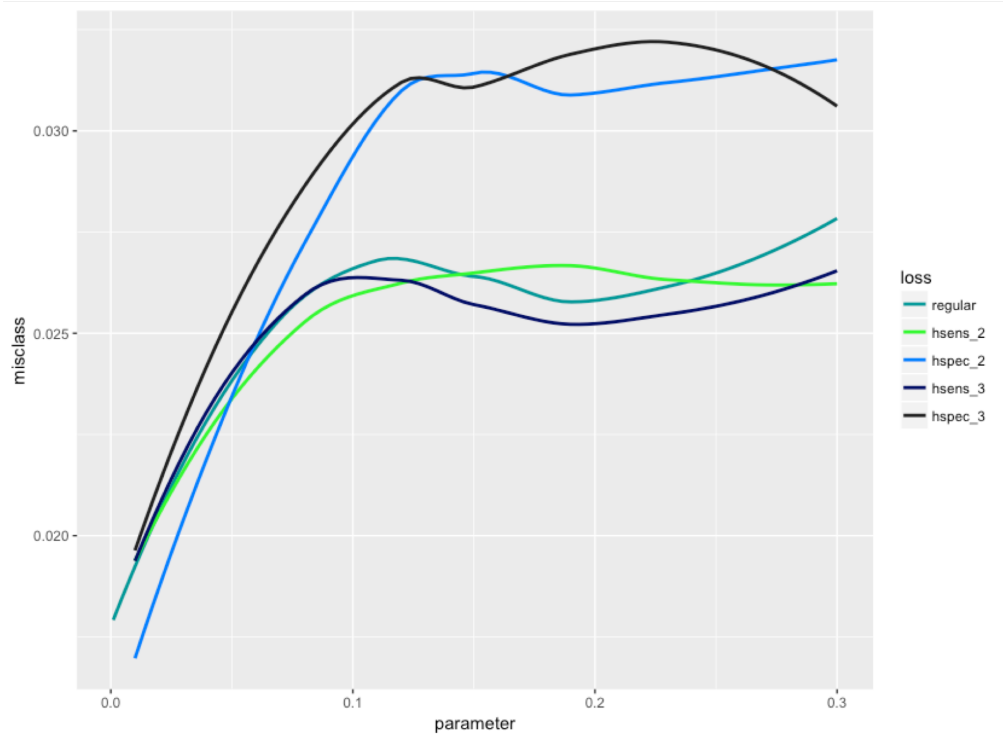

Figure 6: The diabetes bootstrap plot of the CaRT algorithm for tuning parameter selection.

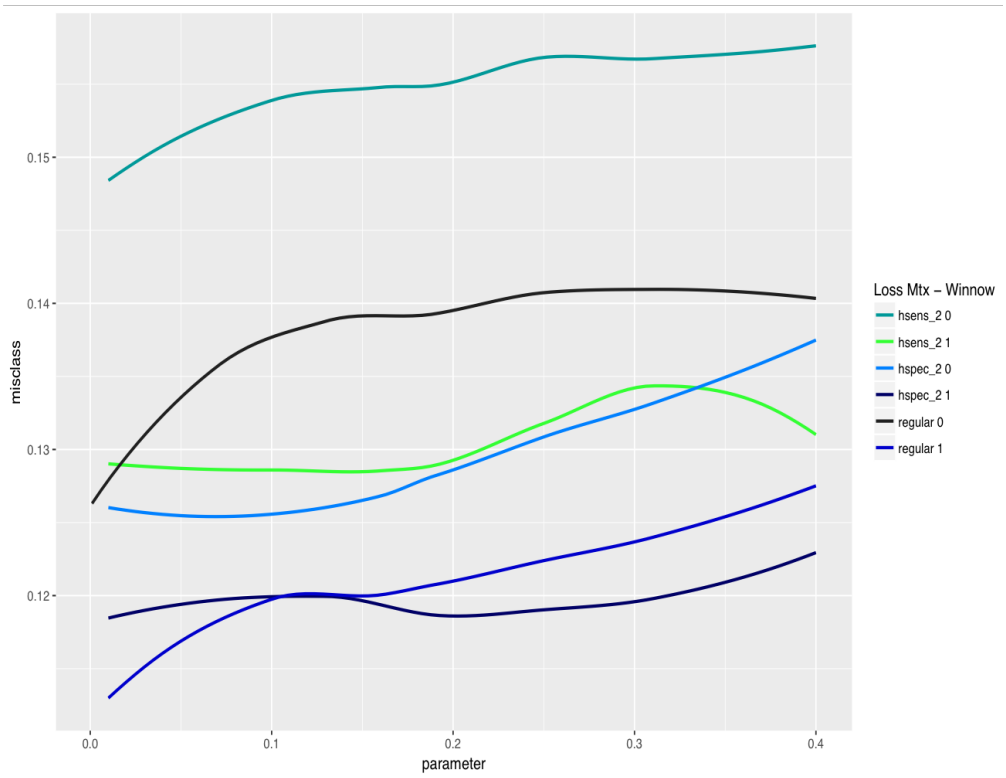

Figure 7: The osteoarthritis bootstrap plot of the C5.0 algorithm for tuning parameter selection.

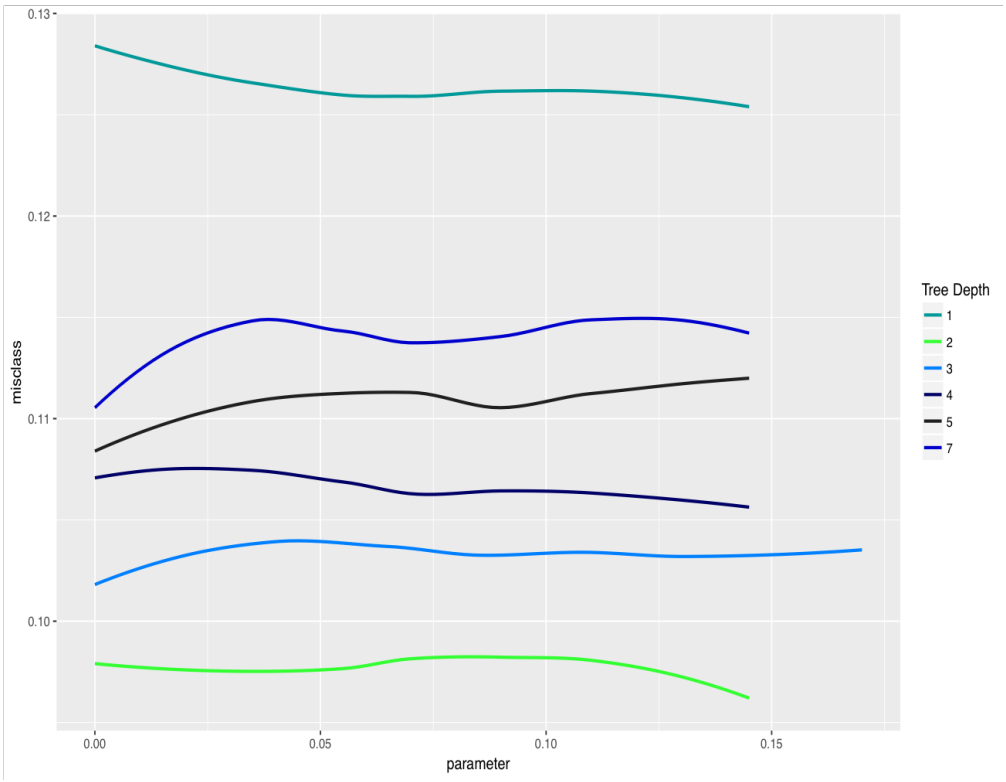

Figure 8: The osteoarthritis bootstrap plot of the CHAID algorithm for tuning parameter selection.

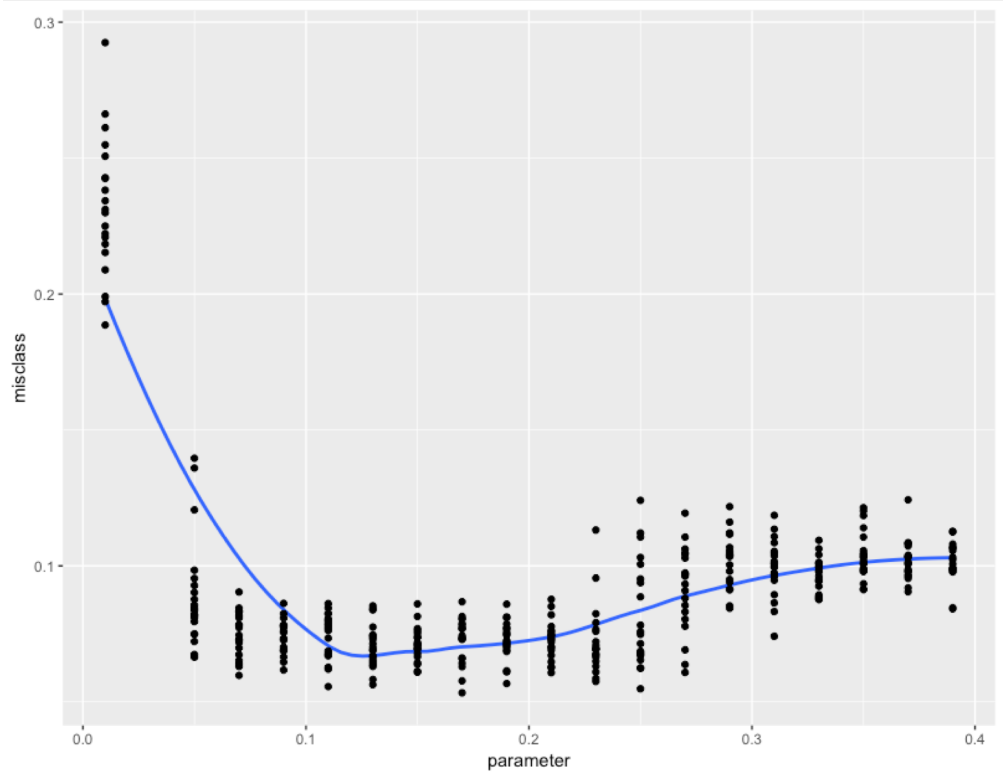

Figure 9: The hypertension bootstrap plot of the LASSO algorithm for tuning parameter selection.

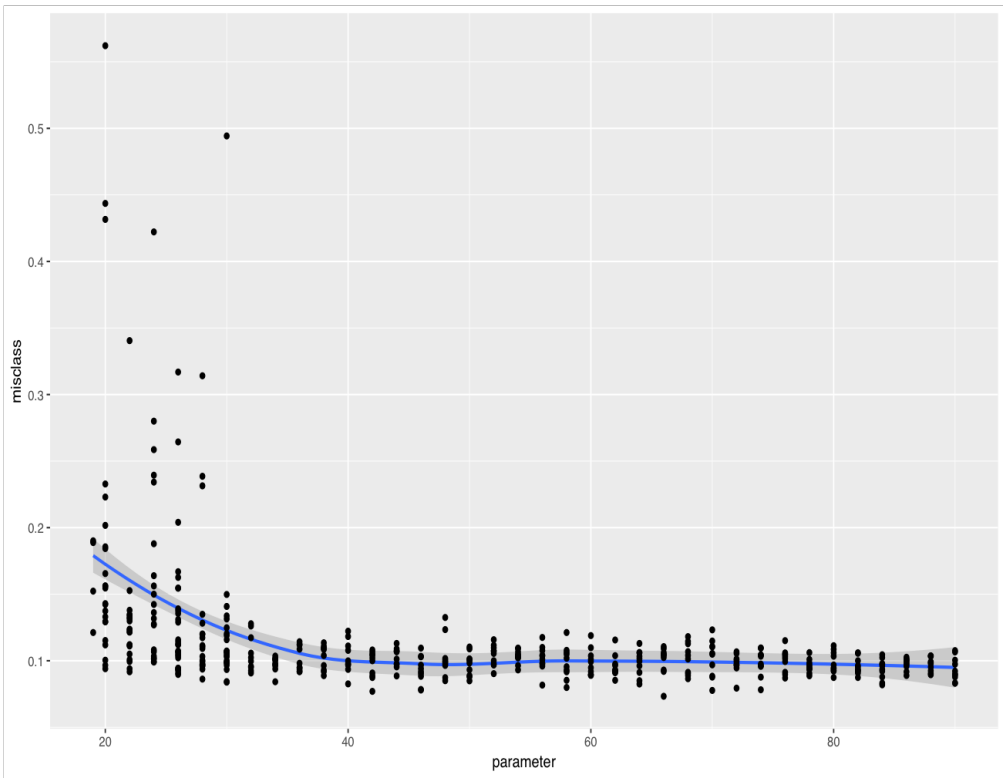

Figure 10: The osteoarthritis bootstrap plot of the Forward Stepwise algorithm for tuning parameter selection.

Generally, the various algorithms were similar in their validity measures. However, this was not the case for the depression case definitions. Both of the logistic regression methods resulted in lower sensitivity measures than the tree-based methods. This is an example of how some conditions may require a decision tree structure rather than a simple inclusion/exclusion list. There is little evidence to suggest that any one of the decision tree algorithms is better suited for case definition development from these results.

There was notable agreement in terms of the features selected for the final case definitions. Most case definitions included problem list and encounter diagnosis ICD codes. There is often direct agreement between the algorithms; for example, 4 of the 5 osteoarthritis case definitions are identical. More diversity exists between the depression case definitions. No two methods selected the exact same features. In other conditions, there are small differences and sometimes even logical inconsistencies in some of the definitions. For example, the CaRT case definition for diabetes includes the word “acute” and the absence of a prescription for aspirin. This emphasizes the need

to employ clinical knowledge as these minor rules are likely artifacts of the random sample, and not clinically relevant.

5.4 Interpretation

We have shown that equally valid case definitions can be created using machine learning methods as compared to committee created case definitions. Once a reliable reference standard is established for a sample, it is possible to create new case definitions that utilize all data that is extracted from the EMR. Furthermore, the validity measures from the 10-fold cross validation show a distinction between the logistic regression methods and the tree methods for the depression case definition. There appears to be parity between the decision tree methods in terms of validity measures, but notable differences in the features selected for the final case definitions.

The depression case definitions show the most obvious differences between algorithms. Unlike diabetes and hypertension, depression is often a more difficult condition to define using administrative data. This is due to the sometimes-subjective view of the condition itself, as well as the granularity of the ICD codes for the condition. A systematic review of hospital administrative data case definitions for depression found sensitivities ranging from 28-35% [176]. The ICD-9 codes ranging from 290 to 319 are all mental disorders, several of which could be defined as depression. The variability in coding makes feature selection less consistent across algorithms. This differs from the diabetes, hypertension, and osteoarthritis case definitions, which are typically well represented by a single code (250, 401, and 715 respectively). The C5.0 algorithm seems to favor specificity compared to the other decision tree algorithms, which have slightly higher sensitivities. Note that an increase in the specificity value results in more correct classifications than an equal increase in sensitivity, as there are only 386 depression cases and 1529 non-cases (20.2% prevalence). A key strength of this study is that we are able to get estimates of specificity and NPV, which many other machine learning publications lack, as they only report sensitivity and PPV.

It is important to consider the tradeoff between the accuracy and interpretability of a case definition. When considering accuracy, there may be cases where sensitivity may be valued over specificity (screening purposes) and vice versa. On the other hand, slight increases in the validity measures may not outweigh the loss of interpretability. In these examples, we are not able to make a clear distinction as to which algorithm is best suited for case definition development. At

this point, finding the optimal case definition using machine learning methods may require the implementation of several methods, and a judgement being made based on clinical and statistical considerations. Further validation of these methods on other conditions and in other settings may resolve this ambiguity.

Our study should be interpreted in light of its limitations. First, our sample population is older than the general population - which was intentional to ensure a higher prevalence of disease. Further validation needs to be performed to ensure that the validity estimates reported for these case definitions are generalizable to other ages. Second, there is always reason to believe that the reference standard is imperfect. A chart review for the presence or absence of these four conditions can be subjective and difficult to assess with the data used for review. This problem is a reality of any chart review exercise. Notwithstanding these limitations, this study draws from a relatively large sample size and was randomly sampled. Further, this study was draws on EMR data from across Canada, increasing the generalizability of these findings.

5.5 Conclusion

Going forward, it will be important to assess the abilities of machine learning algorithms to develop case definitions in rare conditions, such as dementia and epilepsy. For rare conditions, minimizing the misclassification rate may render case definitions with very high specificity but poor sensitivity. Since the vast majority of the reference set will not have the condition, the algorithms may favor case definitions that are very exclusive. Here there is very little room for improvement in the misclassification rate and the algorithm may be unable to improve sensitivity so other measures should be considered. For example, sensitivity, specificity, the F1-score, the G-mean, or weighted accuracy could be maximized instead.

Case definitions created using machine learning methods are as valid and generally involve fewer criteria than definitions developed by a committee of healthcare professionals. The decision tree algorithms performed best among the machine learning methods considered. This new method presents a simpler, and in our opinion more efficient, way of developing case definitions for EMR databases.

6 Discussion

The key finding from this study is that machine learning methods are useful for creating chronic disease case definitions that are comparable with committee-created case definitions. In fact, in some cases, these case definitions outperform the committee created definitions. The case definitions are also simple and concise, when compared to the committee-created definitions. Most notably, there are no long lists of medications. These methods allow for timely generation of high quality case definitions to improve chronic disease surveillance. The results of this study also warrant a further discussion about which algorithms are best suited for this process. More work needs to be done to identify the ideal loss functions to optimize in different settings.

6.1 Comparison of Decision Tree Methods and Logistic Regression Methods

Perhaps the most obvious distinction between the machine learning algorithms can be made between the logistic regression methods and the decision tree methods. It is clear from the 10-fold cross validation results that the logistic regression feature selection algorithms were consistently outperformed by their decision tree counterparts. This is due to two main factors: the lack of useful tuning parameters available to these algorithms, and the inability of the algorithms to explore feature interactions.

The fact that the logistic methods only offer one tuning parameter limit the versatility of the feature selection. This is most prominently illustrated by comparing these methods in conditions where the prevalence is low. In Chapter 5, the only metric that was considered was the misclassification rate. This is the metric that a logistic regression method is naturally inclined to maximize. For all conditions, when the misclassification rate is minimized, the logistic regression perform roughly equally to the decision tree method. The differences become apparent when minimizing other metrics, especially for rare conditions. Further work will be performed on conditions with a lower prevalence, warranting the maximization of other metrics, like the G-mean, F1-score, and Naïve Mean.

Thus far, the only tuning parameter optimized in the bootstrapping step for the LASSO regression is the penalty applied to the sum of the absolute value of the coefficient estimates. There are other options available to the LASSO regression that could become useful when optimizing other

metrics. As with many other regression methods that use likelihood estimates, it is possible to apply weights to observations. This method has been used in linear regression applications to help meet distributional assumptions, most notably the assumption of constant variance. We can repurpose this method to apply higher weights to the cases or non-cases, resulting in higher sensitivity or specificity case definitions. This effect is similar to that of the loss matrices in the decision tree algorithms.

Another limitation of the logistic regression methods is that the algorithm is unable to examine the relationship between combinations of features. The only algorithms that this implementation of the logistic methods can produce are a list of inclusion and exclusion features. This only allows for "or" statements with no opportunity for combinations of features, or "and" statements. For conditions that can be explained simply, using just a few features with no interactions, this is not a problem. For more complicated conditions, however, this is a major drawback of the logistic regression method. Interestingly, no exclusion criteria were selected in any of the final reported case definitions.

Yet another limitation of the logistic regression methods is that the feature selection is sometimes redundant. For example, the LASSO-generated case definition for diabetes includes a feature representing the presence of code 250, and the presence of code 250 twice in 2 years. This is a clear redundancy, as anyone who has been assigned the code 250 twice in two years must also meet the criteria for receiving the code 250. Here the likelihood estimate of the LASSO logistic regression included both features because it identified a difference in the estimate of the log-odds between those that were assigned code 250 twice within two years, and those who were assigned code 250 at all. The decision tree algorithms, on the other hand, will not select features that have this redundancy. This is a clear example where the decision tree algorithms are better suited for case definition development.

6.1.1 An Interesting Case Study

There are situations in which the LASSO and forward stepwise logistic regression methods might be preferable to the decision tree algorithms. Take, for example, the concept of frailty. Frailty is a state, often comprised of several conditions. There are many combinations of conditions and indications that could lead to a patient being deemed as frail. Attempting to use a decision

tree classifier in this setting would be inappropriate, as the decision tree algorithms are unable to account for the many independent conditions and features that need not be conditioned upon other conditions.

Take the Charlson Comorbidity Index for example [177]. This is a measure that attempts to quantify the 10-year mortality risk using a scoring system comprised of several conditions. As a patient is diagnosed with additional conditions, their score increases according to the weight of that condition. A decision tree classifier is unable to assess variables independently in this way. A logistic regression method for feature selection, on the other hand, is much better suited for the development of an index of this type, as it is able to generate lists of inclusion criteria, adjusting for other conditions, but not conditioning each new condition on the previously included conditions.

6.2 Overfitting - Artifacts Observed in the Case Definitions

There are several examples in these analyses of the machine learning feature selection methods choosing features that likely have no place being used in a case definition. For example, the CaRT definition for diabetes includes the prescription of Aspirin. The C5.0 definition for depression selected the free-text term "affective" from the encounter text. This feature is relevant for terms like "bipolar affective disorder" but may not be relevant for indications like "seasonal affective disorder". This feature may not be generalizable outside of this specific sample, as the patients who were assigned the text "seasonal affective disorder" also happened to be suffering from depression. This will not be the case with all patients.

A very interesting example of overfitting has been observed in other work, attempting to create a case definition that distinguishes between type 1 and type 2 diabetes. The C5.0 algorithm maximizing the naïve mean of sensitivity and specificity selected the word "LMC" as being predictive of diabetes status. "LMC" refers to a specialty clinic operating in Calgary that treats type 1 diabetic patients in southern Alberta. There is a perfectly fine clinical reason why this feature was selected for inclusion in the case definition: Those that are referred to the LMC clinic are type 1 diabetes cases. This is where we can distinguish between two types of overfitting. The first type is where the machine learning algorithm overfits the available observations and selects a feature that is an artifact. An artifact is an unwanted feature that has nothing to do with true disease status. The second type of overfitting can occur at the sampling level. In this case, the model has not

overfit the training data, rather the sample itself is overfit, and the observations within cannot be assumed relevant to the general public. This emphasizes the importance of tailoring the sample to the population you wish to make inferences about.

6.3 Customization of Case Definitions

Thus far, we have created case definitions that perform well compared with committee-created case definitions. Future work will need to be done on low-prevalence conditions, and conditions that are not easily defined. For example, the distinction between type 1 and type 2 diabetes is not easily made in ICD-9. It is widely believed that the way to differentiate the type 1 from the type 2 cases in administrative cases is through medications. Type 1 diabetes patients tend to be prescribed insulin, whereas type 2 diabetes patients get other glucose-lowering drugs. The free-text may also provide a crucial role. Further exploration of which loss functions are optimal for different types of conditions should be considered.

Perhaps the most powerful advantage of using machine learning methods for case definition development is flexibility. It is possible to create multiple case definitions that maximize different metrics. For example, a case definition with a high positive predictive value is useful in different circumstances than a case definition that achieves a balance in sensitivity and specificity. A high positive predictive value is achieved when specificities are high, as there are very few false positives. It is entirely possible in a rare condition to have a case definition that has very poor sensitivity, but strong positive predictive value. While this case definition would not be useful for getting an estimate of prevalence, we do know that of those identified, we are confident they are true cases. When identifying a cohort of patients for a study, it is important to be confident that all cases are true cases.

Take, for example, a rare condition like Parkinsonism. It is estimated that in industrialized countries, the prevalence of Parkinson's disease is about 0.3% [178]. A machine learning algorithm optimized for accuracy will likely generate a case definition with extremely high specificities, and will not assign much weight to sensitivity, as there are so few cases. This case definition, with extremely high specificity and lackluster sensitivity will likely have a high PPV. This means we are confident that those cases identified by the case definition are true positives, and not false positives. If a cohort of patients with Parkinson's was to be identified for some observational

study, it is critically important that every patient in the Parkinon's cohort actually has Parkinson's disease. This case definition is well-suited for this situation, even though its sensitivity is low, as PPV is the most important metric.

Alternatively, a case definition that maximizes the F1-score, G-mean, or Naïve mean achieves more of a balance between sensitivity and specificity, potentially at the expense of predictive values. A case definition that has moderate sensitivity and specificity, but poor predictive values is of little use for cohort development, but is much better equipped to illustrate the prevalence of disease, as it is much more accurate at a population level. This aids in the assessment of the burden of disease.

The Parkinson's example described earlier can illustrate this concept as well. While a case definition with low sensitivity but high PPV may be useful in identifying a cohort of patients we know to be Parkinson's cases, it is very poor at ascertaining an estimate of prevalence. A case definition with low sensitivity is doing a poor job of identifying every patient with Parkinson's. Implementing this case definition will very much underestimate the prevalence of Parkinsonism in a population. In order to get a valid estimate of how many cases there are in a population, a higher sensitivity is desired. A case definition with a balance of sensitivity and specificity is more appropriate for this purpose than the high PPV definition.

The machine learning methods offer a new opportunity for specialized case definitions. The previous method of case definition development, using committees, did not develop specialized case definitions for specific purposes. As these methods develop and are implemented for more conditions, the surveillance and identification of patients will be vastly improved and made more efficient.

6.4 Temporal Considerations

One of the greatest strengths of the EMR for chronic disease surveillance is the longitudinal nature of the data. As time goes on, an EMR may follow a patient for many years, collecting valuable information. This also presents a challenge that needs to be addressed in case definition and cohort development work. There are not clear answers in the literature that provide evidence for look-back windows in case definitions. There are many case definitions developed in the past that have used look-back periods of two years, some more, and some less. Perhaps the optimal look-back

period is condition-specific and should be explored independently for each condition.

The CPCSSN data used for this analysis includes billing data dating back to 2002. Most records, however, have occurred in just the past two years. There are a few methods that could be considered going forward. One option is to assign a higher weight to the more current diagnoses and observations compared with the older information. Another option is to specify the look-back period. For example, only look at records that have occurred in the last two years, one year, or six months. The efficacy of this approach would likely change depending on the condition. Some conditions may be more permanent than others, like Parkinsonism as compared with hypertension. An indication of Parkinsonism from three years ago is likely more relevant than an indication of hypertension from three years ago. Typically, when a patient is diagnosed with Parkinson's, they suffer from the condition for the rest of their life. A patient diagnosed with hypertension, on the other hand, may be prescribed anti-hypertensive drugs, or undergo a change in lifestyle that eases the condition, and may even rid them of the condition. These sorts of considerations should be included in the case definition development process going forward.

There are a few methods of performing a machine learning analysis that could incorporate this temporal consideration. One method is to include tags in the feature generation step, that specifies features as being indicative of code X appearing in the billing table in the year leading up to the reference standard being established. Thus far, the feature selection methods in this project include tags for multiple codes being observed within a one or two-year time period, but this does not require the observation to be in the most recent two years. If a patient was assigned multiple ICD codes for diabetes in 2010, they were considered to meet the criteria for two codes in one year, even though the reference standard was established in 2014. Therefore, a new set of features could be generated that only query the most recent years.

Another method of including the temporal information would be to restrict the entire dataset to the most recent window. For example, if the reference standard was established in 2014, complete the entire analysis on data solely recorded from 2012 to 2014, or 2013 to 2014. The validity estimates of each repeat analysis can be compared, and the optimal lookback window determined for each condition. Either of these two methods could potentially improve the validity of the generated case definitions.

6.5 Feature Considerations

There are a number of other features that could be included in case definition development. The laboratory data is likely being underutilized in the current analysis. The features generated for Chapter 5 were somewhat rudimentary. Cut-points were set and binary features were generated for any occurrence, any two occurrences within 1 year, and any 2 occurrences within 2 years. There are likely more informative ways to include this information. Temporal trends in laboratory values could be used, average values over given time periods, maximum or minimum values, etc. This should be an area of future research. Future CPCSSN data cuts will also include more laboratory information than currently available, which could also include valuable predictive information.

Another step that could be taken is to include more information from the examination tables. Currently, only blood pressure measurements are used from this table. Perhaps some of the available information, like BMI, weight, etc. would be useful in case definition development for conditions like diabetes.

Further development of the Natural Language Processing steps can also improve the case definitions. The bag-of-words approach used here has proven to be effective, but there is likely room for improvement using more complicated and powerful methods to incorporate physician free-text.

6.6 Reference Standards

The analysis used an age-stratified random sample and a subsequent chart review to obtain the sample. The age-stratification was done to ensure that the sample had a sufficient number of cases of chronic disease, as the older population tends to have higher rates of chronic disease. This false inflation of the prevalence will have some effects on the estimates of PPV, NPV, and accuracy, as these numbers are dependent on the sample prevalence. That being said, these numbers should be generalizable to the elderly CPCSSN population. While these estimates may be somewhat biased, the comparison between the machine learning algorithms and the committee-created case definitions is still valid, as both estimates used the exact same sample.

Work is currently being done by a few organizations to produce reliable reference standards without the costly acquisition of chart review data [46]. These methods are in their infancy, and have yet to be validated in most cases, but are a promising area of future research. Essentially, they require a high PPV definition for a condition. This results in a cohort of patients that are known

to be true cases are identified using a high PPV case definition. These values are used cases in a probability estimation exercise, involving a classifier such as a logistic regression using a large number of generated features. It is then possible to get a predicted probability of being a true case from this classifier. A cut-point in the probability estimate can then be specified and we can use the resulting binary state as a reference standard. The advantage to this method is that we can generate a reference standard on a much larger sample, without the need of a chart review. The reliability of the reference standard is questionable in this scenario though, and further validation work needs to be done to determine when and where this method is appropriate. Currently, it appears that using chart reviews is the best source of a reliable reference standard. Alternatively, if a disease registry exists, the presence of a patient in a registry is also a valuable reference standard. Unfortunately, these registries are not available for every condition, or in every healthcare setting.

6.7 Comparison with More Powerful Machine Learning Algorithms

The machine learning methods used in these analyses were selected based on their ability to be interpreted as a case definition. Decision tree algorithms are especially well-suited for this process. There are far more powerful and complicated machine learning algorithms that can be used for classification purposes, that are not as interpretable. A future area of study would be to compare the validity results of the simple machine learning methods used here as compared with the more complicated methods such as the random forest model, Support Vector Machines, or Artificial Neural Networks. If classification is much improved using these complicated algorithms, it would be an indication that there are complicated relationships and inter-dependencies of the features that cannot be captured using a single decision tree. This is an important area of future research that will provide much insight into the nature of case definition development.

Interpretability is a key part of case definitions as they exist in the current EMR landscape. A "black box" algorithm does not provide the same amount of inference that the simpler methods do. While these methods often have a method of determining which features are most predictive of disease status, they cannot be decomposed into a single set of rules. The decision as to which is more important: accuracy, or interpretability depends largely on the nature of the problem. In a purely predictive setting (e.g. predicting 30-day readmission rates in a tertiary care setting), the "black box" algorithms may be preferred, but the interpretable methods are better suited for case

definition development.

Interpretable models are also much easier to implement as a case definition, compared with the more complex methods. If an ensemble tree method, such as a random forest Model were used to create a classifier, the validity measures will likely be similar or more accurate than the single decision trees that we have developed. The issue with the random forest model, is that it is far more difficult to use the random forest model to classify new patients as case positive or case negative. A random forest model will typically create 500+ unique decision trees, each with varying depths, and a majority vote of the many trees determine case status. The implementation of this model requires every single feature used in any of the 500+ trees. This makes the use of this model very difficult.

Most EMR data lives in a relational database that is read with a Structured Query Language (SQL). A SQL statement is a human readable sentence that selects a portion of the data living in the relational database. The ideal case definition can be decomposed into human readable rules, and those rules can be decomposed into a SQL statement. This makes it very easy to implement a case definition and quickly identify patients as case positive or case negative. If we were to use the random forest model, there would be no way to implement the classifier in this way. In fact, the only way to interpret the random forest model is to perform all the feature generation steps that were performed in this analysis. This requires the generation of thousands of features, for every unique code, word, and lab value that exists in the database. Only after these computationally expensive steps are taken can the random forest model be implemented to identify patients as cases or non-cases. This is because every single feature that is included in any of the 500+ decision trees must be generated. This makes it very inconvenient to use a non-interpretable method like random forest, Support Vector Machines, Artificial Neural Networks, K-Nearest Neighbours, Logistic regressions using predicted probabilities, or even Naïve Bayesian classifiers for case definition development. The interpretable methods are much better suited for the implementation of case definitions.

6.8 Implementation of this Method

In spite of the limitations and shortcomings of this study, it has been shown that machine learning methods are capable of creating valid case definitions. This is evidence that the current method of

case definition development should be replaced or supplemented with machine learning components. Of course, clinical knowledge is very much required to ensure that the case definitions are clinically relevant and logical. This advanced knowledge of each condition will also be required to tailor the feature generation step as well.

The current landscape of the EMR in Canadian healthcare is anything but standardized. There are many EMR vendors, formats, data quality standards, and practices across the country. It is important that case definitions be developed in the setting for which they are to be used. The case definitions developed here are currently only generalizable or reliable for the CPCSSN network, as there is no validation of these definitions in any other setting. The major advantage of this method is that the development of case definitions for new settings and EMRs is now much improved and hastened. Although there is some evidence that common data models will solve at least a portion of the diversity problem in the EMR landscape, this standardization is still some time away. The opportunity for surveillance is here now, and this is a method of making surveillance possible in most any EMR setting.

6.9 What This Work Has Done to Further the Literature

Up to this point, the majority of studies that have used machine learning methods in an EMR setting have been for prediction and phenotype work. The resulting models are generally not human readable, or interpretable as a set of rules. This work has shown that machine learning methods can be leveraged to create a set of interpretable rules structured in a similar way as traditional case definitions. This method can potentially bridge the gap between those that value interpretability and clinical inference, and the "black box" machine learning methods, where interpretability and direct inference are unclear at best.

This work has also shown that the current method of committee-generated case definitions can be replaced or supplemented with a machine learning approach. Studies like this should also lead to further consideration for optimal loss functions, alternate machine learning algorithms, and other conditions that have been difficult to define using other methods.

6.10 Objectives and Conclusion

There were two objectives identified at in Chapter 2.

1. Determine if machine learning methods are capable of creating chronic disease case definitions in a primary care EMR that are comparable in terms of simplicity and validity measures with committee-created case definitions.

2. Compare different machine learning methods to establish which methods are best suited for case definition development.

For objective 1, we have proven that machine learning methods can perform as well, and even improve upon committee-created case definitions in the EMR setting. In some cases, the case definitions even outperform the existing case definitions in terms of both validity and simplicity. This is an exciting revelation, as it provides evidence for a shift in the process behind case definition development. This will not only improve the quality of disease surveillance, but it also improves the efficiency.

For objective 2, we have shown that decision tree algorithms are superior to feature selection methods for logistic regression in terms of case definition development. That being said, there is little evidence to suggest that there is any one algorithm that is best for this purpose. However, it is vitally important that those algorithms have a tuning parameter that allows for the explicit assignment of weights to false positives and false negatives. This allows for the greatest amount of customization and accuracy.

By achieving these three objectives, we have taken a major step towards the original goal: utilizing the massive potential of the EMR for disease surveillance. This research will aid in the allocation of resources, the management of chronic disease, and enable novel research to be performed that will change policy and patient's lives.

Bibliography

[1] Martin Fortin, Lise Lapointe, Catherine Hudon, and Alain Vanasse. Multimorbidity is common to family practice: is it commonly researched? *Canadian Family Physician*, 51(2):244–245, 2005.

[2] Catherine Hoffman, Dorothy Rice, and Hai-Yen Sung. Persons with chronic conditions: their prevalence and costs. *Jama*, 276(18):1473–1479, 1996.

[3] Stephen J Boccuzzi. Indirect health care costs. *Cardiovascular health care economics*, pages 63–79, 2003.

[4] Michael Mirolla. *The cost of chronic disease in Canada*. GPI Atlantic, 2004.

[5] Deborah T Vinton, Roberta Capp, Sean P Rooks, Jean T Abbott, and Adit A Ginde. Frequent users of us emergency departments: characteristics and opportunities for intervention. *Emerg Med J*, pages emermed–2013, 2014.

[6] *HPSC*, (accessed: November 2017). http://www.hpsc.ie.

[7] Patricia Huston and C David Naylor. Health services research: reporting on studies using secondary data sources. *CMAJ: Canadian Medical Association Journal*, 155(12):1697, 1996.

[8] Suzanne M Cadarette and Lindsay Wong. An introduction to health care administrative data. *The Canadian journal of hospital pharmacy*, 68(3):232, 2015.

[9] John Wennberg and Alan Gittelsohn. Small area variations in health care delivery: a population-based health information system can guide planning and regulatory decision-making. *Science*, 182(4117):1102–1108, 1973.

[10] Emilie K Johnson and Caleb P Nelson. Utility and pitfalls in the use of administrative databases for outcomes assessment. *The Journal of urology*, 190(1):17, 2013.

[11] Carl van Walraven, Carol Bennett, and Alan J Forster. Administrative database research infrequently used validated diagnostic or procedural codes. *Journal of clinical epidemiology*, 64(10):1054–1059, 2011.

[12] Graeme F Woodworth, Clinton J Baird, Giannina Garces-Ambrossi, James Tonascia, and Rafael J Tamargo. Inaccuracy of the administrative database: comparative analysis of two databases for the diagnosis and treatment of intracranial aneurysms. *Neurosurgery*, 65(2):251–257, 2009.

[13] HA Khwaja, H Syed, and DW Cranston. Coding errors: a comparative analysis of hospital and prospectively collected departmental data. *BJU international*, 89(3):178–180, 2002.

[14] Eliseo J Pérez-Stable, Jeanne Miranda, Ricardo F Muñoz, and Yu-Wen Ying. Depression in medical outpatients: underrecognition and misdiagnosis. *Archives of Internal Medicine*, 150(5):1083–1088, 1990.

[15] Michelle J Semins, Bruce J Trock, and Brian R Matlaga. Validity of administrative coding in identifying patients with upper urinary tract calculi. *The Journal of urology*, 184(1):190–192, 2010.

[16] Jane S Saczynski, Susan E Andrade, Leslie R Harrold, Jennifer Tjia, Sarah L Cutrona, Katherine S Dodd, Robert J Goldberg, and Jerry H Gurwitz. A systematic review of validated methods for identifying heart failure using administrative data. *Pharmacoepidemiology and drug safety*, 21(S1):129–140, 2012.

[17] Noralou P Roos, Charlyn Black, Norman Frohlich, Carolyn DeCoster, Marsha Cohen, Douglas J Tataryn, Cameron A Mustard, Leslie L Roos, Fred Toll, Keumhee C Carriere, et al. Population health and health care use: An information system for policy makers. *The Milbank Quarterly*, pages 3–31, 1996.

[18] Geoffrey M Anderson, Kevin Grumbach, Harold S Luft, Leslie L Roos, Cameron Mustard, and Robert Brook. Use of coronary artery bypass surgery in the united states and canada: influence of age and income. *Jama*, 269(13):1661–1666, 1993.

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

[20] Hude Quan, Nadia Khan, Brenda R Hemmelgarn, Karen Tu, Guanmin Chen, Norm Campbell, Michael D Hill, William A Ghali, Finlay A McAlister, et al. Validation of a case definition to define hypertension using administrative data. *Hypertension*, 54(6):1423–1428, 2009.

[21] Susan E Andrade, Leslie R Harrold, Jennifer Tjia, Sarah L Cutrona, Jane S Saczynski, Katherine S Dodd, Robert J Goldberg, and Jerry H Gurwitz. A systematic review of validated methods for identifying cerebrovascular accident or transient ischemic attack using administrative data. *Pharmacoepidemiology and drug safety*, 21(S1):100–128, 2012.

[22] Beth A Virnig and Marshall McBean. Administrative data for public health surveillance and planning. *Annual review of public health*, 22(1):213–230, 2001.

[23] Dwight C Evans, W Paul Nichol, and Jonathan B Perlin. Effect of the implementation of an enterprise-wide electronic health record on productivity in the veterans health administration. *Health Economics, Policy and Law*, 1(2):163–169, 2006.

[24] Clement J McDonald, Fiona M Callaghan, Arlene Weissman, Rebecca M Goodwin, Mallika Mundkur, and Thomson Kuhn. Use of internist's free time by ambulatory care electronic medical record systems. *JAMA internal medicine*, 174(11):1860–1863, 2014.

[25] Astrid M van Ginneken. The computerized patient record: balancing effort and benefit. *International journal of medical informatics*, 65(2):97–119, 2002.

[26] Kevin Renner. Cost-justifying electronic medical records. *Healthcare Financial Management*, 50(10):63–68, 1996.

[27] Jeffrey A Linder, Jun Ma, David W Bates, Blackford Middleton, and Randall S Stafford. Electronic health record use and the quality of ambulatory care in the united states. *Archives of internal medicine*, 167(13):1400–1405, 2007.

[28] Jane Metzger, Emily Welebob, David W Bates, Stuart Lipsitz, and David C Classen. Mixed results in the safety performance of computerized physician order entry. *Health Affairs*, 29(4):655–663, 2010.

[29] David W Bates, Joseph Studer, Cheryl A Reilly, Elizabeth L Cureton, Cynthia D Spurr, and Gilad J Kuperman. Evaluating the impact of a computerized ambulatory record. In *Proceedings of the AMIA Symposium*, page 964. American Medical Informatics Association, 2000.

[30] David W Bates, Mark Ebell, Edward Gotlieb, John Zapp, and HC Mullins. A proposal for electronic medical records in us primary care. *Journal of the American Medical Informatics Association*, 10(1):1–10, 2003.

[31] Kenneth Arrow, Alan Auerbach, John Bertko, Shannon Brownlee, Lawrence P Casalino, Jim Cooper, Francis J Crosson, Alain Enthoven, Elizabeth Falcone, Robert C Feldman, et al. Toward a 21st-century health care system: recommendations for health care reform. *Annals of Internal Medicine*, 150(7):493–495, 2009.

[32] Ruth Stashefsky Margalit, Debra Roter, Mary Ann Dunevant, Susan Larson, and Shmuel Reis. Electronic medical record use and physician–patient communication: an observational study of israeli primary care encounters. *Patient education and counseling*, 61(1):134–141, 2006.

[33] Jayna M Holroyd-Leduc, Diane Lorenzetti, Sharon E Straus, Lindsay Sykes, and Hude Quan. The impact of the electronic medical record on structure, process, and outcomes within primary care: a systematic review of the evidence. *Journal of the American Medical Informatics Association*, 18(6):732–737, 2011.

[34] Tracy D Gunter and Nicolas P Terry. The emergence of national electronic health record architectures in the united states and australia: models, costs, and questions. *Journal of medical Internet research*, 7(1), 2005.

[35] Jamie L Habib and Drug Benefit Trends. Ehrs, meaningful use, and a model emr. *Drug Benefit Trends*, 22, 2010.

[36] Dave Garets and Mike Davis. Electronic medical records vs. electronic health records: yes, there is a difference. *Policy white paper. Chicago, HIMSS Analytics*, pages 1–14, 2006.

[37] Guy Paré, Louis Raymond, Ana Ortiz de Guinea, Placide Poba-Nzaou, Marie-Claude Trudel, Josianne Marsan, and Thomas Micheneau. Barriers to organizational adoption of emr systems in family physician practices: a mixed-methods study in canada. *International journal of medical informatics*, 83(8):548–558, 2014.

[38] Robert J Carroll, Will K Thompson, Anne E Eyler, Arthur M Mandelin, Tianxi Cai, Raquel M Zink, Jennifer A Pacheco, Chad S Boomershine, Thomas A Lasko, Hua Xu, et al. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *Journal of the American Medical Informatics Association*, 19(e1):e162–e169, 2012.

[39] Victor M Castro, Jessica Minnier, Shawn N Murphy, Isaac Kohane, Susanne E Churchill, Vivian Gainer, Tianxi Cai, Alison G Hoffnagle, Yael Dai, Stefanie Block, et al. Validation of electronic health record phenotyping of bipolar disorder cases and controls. *American Journal of Psychiatry*, 172(4):363–372, 2015.

[40] Leanne Clifford, Amandeep Singh, Gregory A Wilson, Pearl Toy, Ognjen Gajic, Michael Malinchoc, Vitaly Herasevich, Jyotishman Pathak, and Daryl J Kor. Electronic health record surveillance algorithms facilitate the detection of transfusion-related pulmonary complications. *Transfusion*, 53(6):1205–1216, 2013.

[41] Tyler Williamson, Michael E Green, Richard Birtwhistle, Shahriar Khan, Stephanie Garies, Sabrina T Wong, Nandini Natarajan, Donna Manca, and Neil Drummond. Validating the 8 cpcssn case definitions for chronic disease surveillance in a primary care database of electronic health records. *The Annals of Family Medicine*, 12(4):367–372, 2014.

[42] *High Blood Pressure (Hypertension)*, 2016 (accessed: November 2017). http://www.mayoclinic.org/diseases-conditions.

[43] Karen Tu, Norman RC Campbell, Zhong-Liang Chen, Karen J Cauch-Dudek, and Finlay A McAlister. Accuracy of administrative databases in identifying patients with hypertension. *Open medicine*, 1(1):18, 2007.

[44] J Reneé Robinson, T Kue Young, Leslie L Roos, and Dale E Gelskey. Estimating the burden of disease: comparing administrative data and self-reports. *Medical care*, 35(9):932–947, 1997.

[45] AWS Rutjes, JB Reitsma, A Coomarasamy, KS Khan, and PMM Bossuyt. Evaluation of diagnostic tests when there is no gold standard. a review of methods. *HEALTH TECHNOLOGY ASSESSMENT-SOUTHAMPTON-*, 11(50), 2007.

[46] Tyler Williamson, Rebecca C Miyagishima, Janeen D Derochie, and Neil Drummond. Manual review of electronic medical records as a reference standard for case definition development: a validation study. *CMAJ open*, 5(4):E830–E833, 2017.

[47] Marcello Tonelli, Natasha Wiebe, Martin Fortin, Bruce Guthrie, Brenda R Hemmelgarn, Matthew T James, Scott W Klarenbach, Richard Lewanczuk, Braden J Manns, Paul Ronksley, et al. Methods for identifying 30 chronic conditions: application to administrative data. *BMC medical informatics and decision making*, 15(1):31, 2015.

[48] Danielle A Southern, Barbara Roberts, Alun Edwards, Stafford Dean, Peter Norton, Lawrence W Svenson, Erik Larsen, Peter Sargious, David CW Lau, and William A Ghali. Validity of administrative data claim-based methods for identifying individuals with diabetes at a population level. *Canadian Journal of Public Health/Revue Canadienne de Sante'e Publique*, pages 61–64, 2010.

[49] Reza Alaghehbandan, Don MacDonald, Brendan Barrett, Kayla Collins, and Yue Chen. Using administrative databases in the surveillance of depressive disorders—case definitions. *Population health management*, 15(6):372–380, 2012.

[50] Debra A Butt, Karen Tu, Jacqueline Young, Diane Green, Myra Wang, Noah Ivers, Liisa Jaakkimainen, Robert Lam, and Mark Guttman. A validation study of administrative data algorithms to identify patients with parkinsonism with prevalence and incidence trends. *Neuroepidemiology*, 43(1):28–37, 2014.

[51] Karen Tu, Myra Wang, Jacqueline Young, Diane Green, Noah M Ivers, Debra Butt, Liisa Jaakkimainen, and Moira K Kapral. Validity of administrative data for identifying patients

who have had a stroke or transient ischemic attack using emrald as a reference standard. *Canadian Journal of Cardiology*, 29(11):1388–1394, 2013.

[52] Andrea L Benin, Grace Vitkauskas, Elizabeth Thornquist, Eugene D Shapiro, John Concato, Mihaela Aslan, and Harlan M Krumholz. Validity of using an electronic medical record for assessing quality of care in an outpatient setting. *Medical care*, 43(7):691–698, 2005.

[53] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[54] Stuart Russell, Peter Norvig, and Artificial Intelligence. A modern approach. *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs*, 25:27, 1995.

[55] Maurice Kendall and Alan Stuart. The advanced theory of statistics. vol. 1: Distribution theory. *London: Griffin, 1977, 4th ed.*, 1977.

[56] Igor Kononenko. Id3, sequential bayes, naive bayes and bayesian neural networks. In *Proc. 4th European Working Session on Learning*, pages 91–98, 1989.

[57] Irina Rish. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46. IBM, 2001.

[58] Andrew Y Ng and Michael I Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems*, pages 841–848, 2002.

[59] Marko Grobelnik. Feature selection for unbalanced class distribution and naive bayes. ICML, 1999.

[60] Naomi S Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

[61] Richard J Samworth et al. Optimal weighted nearest neighbour classifiers. *The Annals of Statistics*, 40(5):2733–2763, 2012.

[62] David A Freedman. *Statistical models: theory and practice*. cambridge university press, 2009.

[63] Nuri H Salem Badi. Properties of the maximum likelihood estimates and bias reduction for logistic regression model. *Open Access Library Journal*, 4(05):1, 2017.

[64] Ross M Stolzenberg. Multiple regression analysis. *Handbook of data analysis*, 165:208, 2004.

[65] Ronald R Hocking. A biometrics invited paper. the analysis and selection of variables in linear regression. *Biometrics*, 32(1):1–49, 1976.

[66] Andrew Carr, HIV Lipodystrophy Case Definition Study Group, et al. An objective case definition of lipodystrophy in hiv-infected adults: a case-control study. *The Lancet*, 361(9359):726–735, 2003.

[67] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

[68] Ernst Wit, Edwin van den Heuvel, and Jan-Willem Romeijn. 'all models are wrong...': an introduction to model uncertainty. *Statistica Neerlandica*, 66(3):217–236, 2012.

[69] Alvin C Rencher and Fu Ceayong Pun. Inflation of r2 in best subset regression. *Technometrics*, 22(1):49–53, 1980.

[70] Ellen B Roecker. Prediction error and its estimation for subset-selected models. *Technometrics*, 33(4):459–468, 1991.

[71] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[72] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[73] Robert Tibshirani. The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395, 1997.

[74] Lior Rokach and Oded Maimon. *Data mining with decision trees: theory and applications*. World Scientific, 2008.

[75] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

[76] Terry M Therneau, Elizabeth J Atkinson, et al. An introduction to recursive partitioning using the rpart routines. Technical report, Technical report Mayo Foundation, 1997.

[77] Terry Therneau, Beth Atkinson, and Brian Ripley. *rpart: Recursive Partitioning and Regression Trees*, 2017. R package version 4.1-11.

[78] C Gini. Concentration and dependency ratios. *Rivista di politica economica*, 87:769–792, 1997.

[79] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[80] J Ross Quinlan. *C4.5: programs for machine learning*. Elsevier, 2014.

[81] Max Kuhn, Steve Weston, Nathan Coulter, and Mark Culp. C code for C5.0 by R. Quinlan. *C50: C5.0 Decision Trees and Rule-Based Models*, 2015. R package version 0.1.0-24.

[82] Earl Harris. Information gain versus gain ratio: A study of split method biases. In *ISAIM*, 2002.

[83] Gordon V Kass. An exploratory technique for investigating large quantities of categorical data. *Applied statistics*, pages 119–127, 1980.

[84] The FoRt Student Project Team. *CHAID: CHi-squared Automated Interaction Detection*, 2015. R package version 0.1-2.

[85] Tin Kam Ho. Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, volume 1, pages 278–282. IEEE, 1995.

[86] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[87] Thomas G Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, 40(2):139–157, 2000.

[88] Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8):832–844, 1998.

[89] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[90] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[91] Balázs Kégl. The return of adaboost. mh: multi-class hamming trees. *arXiv preprint arXiv:1312.6086*, 2013.

[92] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[93] William H Press, Saul A Teukolsky, William T Vetterling, and BP Flannery. Section 16.5. support vector machines. *Numerical recipes: the art of scientific computing*, 2007.

[94] Paul John Werbos. Beyond regression: New tools for prediction and analysis in the behavioral sciences. *Doctoral Dissertation, Applied Mathematics, Harvard University, MA*, 1974.

[95] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[96] William H Crown. Potential application of machine learning in health outcomes research and some statistical cautions. *Value in Health*, 18(2):137–140, 2015.

[97] Subramani Mani, Yukun Chen, Tom Elasy, Warren Clayton, and Joshua Denny. Type 2 diabetes risk forecasting from emr data using machine learning. In *AMIA annual symposium proceedings*, volume 2012, page 606. American Medical Informatics Association, 2012.

[98] Gregory F Cooper, Constantin F Aliferis, Richard Ambrosino, John Aronis, Bruce G Buchanan, Richard Caruana, Michael J Fine, Clark Glymour, Geoffrey Gordon, Barbara H Hanusa, et al. An evaluation of machine-learning methods for predicting pneumonia mortality. *Artificial intelligence in medicine*, 9(2):107–138, 1997.

[99] Rory Wolfe, Dean P McKenzie, James Black, Pam Simpson, Belinda J Gabbe, and Peter A Cameron. Models developed by three techniques did not achieve acceptable prediction of binary trauma outcomes. *Journal of clinical epidemiology*, 59(1):26–35, 2006.

[100] Subramani Mani, Asli Ozdas, Constantin Aliferis, Huseyin Atakan Varol, Qingxia Chen, Randy Carnevale, Yukun Chen, Joann Romano-Keeler, Hui Nian, and Jörn-Hendrik Weitkamp. Medical decision support using machine learning for early detection of late-onset neonatal sepsis. *Journal of the American Medical Informatics Association*, 21(2):326–336, 2014.

[101] Sellappan Palaniappan and Rafiah Awang. Intelligent heart disease prediction system using data mining techniques. In *Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on*, pages 108–115. IEEE, 2008.

[102] William Rodman Shankle, Subramani Mani, Michael J Pazzani, and Padhraic Smyth. Detecting very early stages of dementia from normal aging with machine learning methods. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 71–85. Springer, 1997.

[103] Emily Kawaler, Alexander Cobian, Peggy Peissig, Deanna Cross, Steve Yale, and Mark Craven. Learning to predict post-hospitalization vte risk from ehr data. In *AMIA annual symposium proceedings*, volume 2012, page 436. American Medical Informatics Association, 2012.

[104] Sunil Gupta, Truyen Tran, Wei Luo, Dinh Phung, Richard Lee Kennedy, Adam Broad, David Campbell, David Kipp, Madhu Singh, Mustafa Khasraw, et al. Machine-learning prediction of cancer survival: a retrospective study using electronic administrative records and a cancer registry. *BMJ open*, 4(3):e004007, 2014.

[105] William G Baxt. Use of an artificial neural network for data analysis in clinical decision-making: the diagnosis of acute coronary occlusion. *Neural computation*, 2(4):480–489, 1990.

[106] Serguei Pakhomov, Susan A Weston, Steven J Jacobsen, Christopher G Chute, Ryan Mever-

den, Véronique L Roger, et al. Electronic medical records for clinical research: application to the identification of heart failure. *Am J Manag Care*, 13(6 Part 1):281–288, 2007.

[107] Özlem Uzuner, Ira Goldstein, Yuan Luo, and Isaac Kohane. Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association*, 15(1):14–24, 2008.

[108] Guergana K Savova, Philip V Ogren, Patrick H Duffy, James D Buntrock, and Christopher G Chute. Mayo clinic nlp system for patient smoking status identification. *Journal of the American Medical Informatics Association*, 15(1):25–28, 2008.

[109] Cheryl Clark, Kathleen Good, Lesley Jezierny, Melissa Macpherson, Brian Wilson, and Urszula Chajewska. Identifying smokers with a medical extraction system. *Journal of the American Medical Informatics Association*, 15(1):36–39, 2008.

[110] Eiji Aramaki, Takeshi Imai, Kengo Miyo, and Kazuhiko Ohe. Patient status classification by using rule based sentence extraction and bm25 knn-based classifier. In *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*, 2006.

[111] FM Carrero, JM Gómez Hidalgo, E Puertas, M Maña, and J Mata. Quick prototyping of high performance text classifiers. In *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*, 2006.

[112] Özlem Uzuner, Imre Solti, and Eithon Cadag. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518, 2010.

[113] Zhuoran Wang, Anoop D Shah, A Rosemary Tate, Spiros Denaxas, John Shawe-Taylor, and Harry Hemingway. Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning. *PLoS One*, 7(1):e30412, 2012.

[114] Yukun Chen, Robert J Carroll, Eugenia R McPeek Hinz, Anushi Shah, Anne E Eyler, Joshua C Denny, and Hua Xu. Applying active learning to high-throughput phenotyping

algorithms for electronic health records data. *Journal of the American Medical Informatics Association*, 20(e2):e253–e259, 2013.

[115] PK Srimani and Manjula Sanjay Koti. Medical diagnosis using ensemble classifiers-a novel machine-learning approach. *Journal of Advanced Computing*, 1:9–27, 2013.

[116] Znaonui Liang, Gang Zhang, Jimmy Xiangji Huang, and Qmming Vivian Hu. Deep learning for healthcare decision making with emrs. In *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*, pages 556–559. IEEE, 2014.

[117] Peggy L Peissig, Vitor Santos Costa, Michael D Caldwell, Carla Rottscheit, Richard L Berg, Eneida A Mendonca, and David Page. Relational machine learning for electronic health record-driven phenotyping. *Journal of biomedical informatics*, 52:260–270, 2014.

[118] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[119] Chen Lin, Elizabeth W Karlson, Dmitriy Dligach, Monica P Ramirez, Timothy A Miller, Huan Mo, Natalie S Braggs, Andrew Cagan, Vivian Gainer, Joshua C Denny, et al. Automatic identification of methotrexate-induced liver toxicity in patients with rheumatoid arthritis from the electronic medical record. *Journal of the American Medical Informatics Association*, 22(e1):e151–e161, 2014.

[120] Aziz A Boxwala, Jihoon Kim, Janice M Grillo, and Lucila Ohno-Machado. Using statistical and machine learning to help institutions detect suspicious access to electronic health records. *Journal of the American Medical Informatics Association*, 18(4):498–505, 2011.

[121] Wen Zhang, Carl A Gunter, David Liebovitz, Jian Tian, and Bradley Malin. Role prediction using electronic medical record system audits. In *AMIA Annual Symposium Proceedings*, volume 2011, page 858. American Medical Informatics Association, 2011.

[122] György Szarvas, Richárd Farkas, and Róbert Busa-Fekete. State-of-the-art anonymization of medical records using an iterative machine learning framework. *Journal of the American Medical Informatics Association*, 14(5):574–580, 2007.

[123] Ishna Neamatullah, Margaret M Douglass, H Lehman Li-wei, Andrew Reisner, Mauricio Villarroel, William J Long, Peter Szolovits, George B Moody, Roger G Mark, and Gari D Clifford. Automated de-identification of free-text medical records. *BMC medical informatics and decision making*, 8(1):32, 2008.

[124] Jon Patrick and Min Li. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *Journal of the American Medical Informatics Association*, 17(5):524–527, 2010.

[125] Noa P Cruz Díaz, Manuel J Maña López, Jacinto Mata Vázquez, and Victoria Pachón Álvarez. A machine-learning approach to negation and speculation detection in clinical texts. *Journal of the Association for Information Science and Technology*, 63(7):1398–1410, 2012.

[126] Yonghui Wu, S Trent Rosenbloom, Joshua C Denny, Randolph A Miller, Subramani Mani, Dario A Giuse, and Hua Xu. Detecting abbreviations in discharge summaries using machine learning methods. In *AMIA Annual Symposium Proceedings*, volume 2011, page 1541. American Medical Informatics Association, 2011.

[127] Özlem Uzuner, Xiaoran Zhang, and Tawanda Sibanda. Machine learning and rule-based approaches to assertion classification. *Journal of the American Medical Informatics Association*, 16(1):109–115, 2009.

[128] Serguei VS Pakhomov, James D Buntrock, and Christopher G Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association*, 13(5):516–525, 2006.

[129] Serguei VS Pakhomov, Penny L Hanson, Susan S Bjornsen, and Steven A Smith. Automatic classification of foot examination findings using clinical notes and machine learning. *Journal of the American Medical Informatics Association*, 15(2):198–202, 2008.

[130] Thomas A Lasko, Joshua C Denny, and Mia A Levy. Computational phenotype discovery

using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PloS one*, 8(6):e66341, 2013.

[131] *EMRALD*, (accessed: November 2017). http://www.ices.on.ca/Research/Research-programs/Primary-Care-and-Population-Health/EMRALD.

[132] Karen Tu, Myra Wang, R Liisa Jaakkimainen, Debra Butt, Noah M Ivers, Jacqueline Young, Diane Green, and Nathalie Jetté. Assessing the validity of using administrative data to identify patients with epilepsy. *Epilepsia*, 55(2):335–343, 2014.

[133] SE Schultz, DM Rothwell, Z Chen, and K Tu. Identifying cases of congestive heart failure from administrative data: a validation study using primary care patient records. *Chronic diseases and injuries in Canada*, 33(3), 2013.

[134] Jessica Widdifield, Noah M Ivers, Jacqueline Young, Diane Green, Liisa Jaakkimainen, Debra A Butt, Paul O'Connor, Simon Hollands, and Karen Tu. Development and validation of an administrative data algorithm to estimate the disease burden and epidemiology of multiple sclerosis in ontario, canada. *Multiple Sclerosis Journal*, 21(8):1045–1054, 2015.

[135] Jessica Widdifield, Claire Bombardier, Sasha Bernatsky, J Michael Paterson, Diane Green, Jacqueline Young, Noah Ivers, Debra A Butt, R Liisa Jaakkimainen, J Carter Thorne, et al. An administrative data validation study of the accuracy of algorithms for identifying rheumatoid arthritis: the influence of the reference standard on algorithm performance. *BMC musculoskeletal disorders*, 15(1):216, 2014.

[136] *CPCSSN*, (accessed: November 2017). http://cpcssn.ca/.

[137] Richard Birtwhistle, Karim Keshavjee, Anita Lambert-Lanning, Marshall Godwin, Michelle Greiver, Donna Manca, and Claudia Lagacé. Building a pan-canadian primary care sentinel surveillance network: initial development and moving forward. *The Journal of the American Board of Family Medicine*, 22(4):412–422, 2009.

[138] Michelle Greiver, Neil Drummond, Richard Birtwhistle, John Queenan, Anita Lambert-Lanning, and Dave Jackson. Using emrs to fuel quality improvement. *Canadian Family Physician*, 61(1):92–92, 2015.

[139] Alanna V Rigobon, Richard Birtwhistle, Shahriar Khan, David Barber, Suzanne Biro, Rachael Morkem, Ian Janssen, and Tyler Williamson. Adult obesity prevalence in primary care users: an exploration using canadian primary care sentinel surveillance network (cpcssn) data. *Can J Public Health*, 106(5):283–289, 2015.

[140] John A Queenan, Tyler Williamson, Shahriar Khan, Neil Drummond, Stephanie Garies, Rachael Morkem, and Richard Birtwhistle. Representativeness of patients and providers in the canadian primary care sentinel surveillance network: a cross-sectional study. *CMAJ open*, 4(1):E28, 2016.

[141] Marshall Godwin, Tyler Williamson, Shahriar Khan, Janusz Kaczorowski, Shabnam Asghari, Rachel Morkem, Martin Dawes, and Richard Birtwhistle. Prevalence and management of hypertension in primary care practices with electronic medical records: a report from the canadian primary care sentinel surveillance network. *CMAJ open*, 3(1):E76, 2015.

[142] Nathan Coleman, Gayle Halas, William Peeler, Natalie Casaclang, Tyler Williamson, and Alan Katz. From patient care to research: a validation study examining the factors contributing to data quality in a primary care electronic medical record database. *BMC family practice*, 16(1):11, 2015.

[143] Amjed Kadhim-Saleh, Michael Green, Tyler Williamson, Duncan Hunter, and Richard Birtwhistle. Validation of the diagnostic algorithms for 5 chronic conditions in the canadian primary care sentinel surveillance network (cpcssn): a kingston practice-based research network (pbrn) report. *The Journal of the American Board of Family Medicine*, 26(2):159–167, 2013.

[144] Douglas G Altman and J Martin Bland. Diagnostic tests. 1: Sensitivity and specificity. *BMJ: British Medical Journal*, 308(6943):1552, 1994.

[145] David Martin Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. 2011.

[146] Mariska MG Leeflang, Anne WS Rutjes, Johannes B Reitsma, Lotty Hooft, and Patrick MM

Bossuyt. Variation of a test's sensitivity and specificity with disease prevalence. *Canadian Medical Association Journal*, 185(11):E537–E544, 2013.

[147] H Kelly, A Bull, P Russo, and ES McBryde. Estimating sensitivity and specificity from positive predictive value, negative predictive value and prevalence: application to surveillance systems for hospital-acquired infections. *Journal of Hospital Infection*, 69(2):164–168, 2008.

[148] Rajul Parikh, Annie Mathai, Shefali Parikh, G Chandra Sekhar, and Ravi Thomas. Understanding and using sensitivity, specificity and predictive values. *Indian journal of ophthalmology*, 56(1):45, 2008.

[149] R Brian Haynes. *Clinical epidemiology: how to do clinical practice research*. Lippincott williams & wilkins, 2012.

[150] Eric I Benchimol, Douglas G Manuel, Teresa To, Anne M Griffiths, Linda Rabeneck, and Astrid Guttmann. Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. *Journal of clinical epidemiology*, 64(8):821–829, 2011.

[151] Thorsten Joachims. *Learning to classify text using support vector machines: Methods, theory and algorithms*. Kluwer Academic Publishers, 2002.

[152] Yin Zhang, Rong Jin, and Zhi-Hua Zhou. Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1-4):43–52, 2010.

[153] Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer, 1992.

[154] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[155] Bradley Efron et al. Second thoughts on the bootstrap. *Statistical Science*, 18(2):135–140, 2003.

[156] Hal Varian. Bootstrap tutorial. *Mathematica Journal*, 9(4):768–775, 2005.

[157] James Carpenter and John Bithell. Bootstrap confidence intervals: when, which, what? a practical guide for medical statisticians. *Statistics in medicine*, 19(9):1141–1164, 2000.

[158] Herman J Adèr. *Advising on research methods: A consultant's companion*. Johannes van Kessel Publishing., 2008.

[159] Seymour Geisser. *Predictive inference*, volume 55. CRC press, 1993.

[160] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Stanford, CA, 1995.

[161] Yoshua Bengio and Yves Grandvalet. Bias in estimating the variance of k-fold cross-validation. In *Statistical modeling and analysis for complex data problems*, pages 75–95. Springer, 2005.

[162] Stephen B Thacker, Donna F Stroup, Richard B Rothenberg, and Ross C Brownson. Public health surveillance for chronic conditions: a scientific basis for decisions. *Statistics in medicine*, 14(5-7):629–641, 1995.

[163] Kenneth E Powell, Robert A Diseker, Rodney J Presley, Dennis Tolsma, Stic Harris, Kristen J Mertz, Kevin Viel, Doyt L Conn, and William McClellan. Administrative data as a tool for arthritis surveillance: estimating prevalence and utilization of services. *Journal of Public Health Management and Practice*, 9(4):291–298, 2003.

[164] Michael Klompas, Jason McVetta, Ross Lazarus, Emma Eggleston, Gillian Haney, Benjamin A Kruskal, W Katherine Yih, Patricia Daly, Paul Oppedisano, Brianne Beagan, et al. Integrating clinical practice and public health surveillance using electronic medical record systems. *American journal of preventive medicine*, 42(6):S154–S162, 2012.

[165] Tim Van den Bulcke, Paul Vanden Broucke, Viviane Van Hoof, Kristien Wouters, Seppe Vanden Broucke, Geert Smits, Elke Smits, Sam Proesmans, Toon Van Genechten, and François Eyskens. Data mining methods for classification of medium-chain acyl-coa dehydrogenase deficiency (mcadd) using non-derivatized tandem ms neonatal screening data. *Journal of biomedical informatics*, 44(2):319–325, 2011.

[166] Igor Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, 23(1):89–109, 2001.

[167] Juned Siddique, Gregory W Ruhnke, Andrea Flores, Micah T Prochaska, Elizabeth Paesch, David O Meltzer, and Chad T Whelan. Applying classification trees to hospital administrative data to identify patients with lower gastrointestinal bleeding. *PloS one*, 10(9):e0138987, 2015.

[168] Shang-Ming Zhou, Fabiola Fernandez-Gutierrez, Jonathan Kennedy, Roxanne Cooksey, Mark Atkinson, Spiros Denaxas, Stefan Siebert, William G Dixon, Terence W O'Neill, Ernest Choy, et al. Defining disease phenotypes in primary care electronic health records by a machine learning approach: a case study in identifying rheumatoid arthritis. *PloS one*, 11(5):e0154515, 2016.

[169] Robert J Carroll, Anne E Eyler, and Joshua C Denny. Naive electronic health record phenotype identification for rheumatoid arthritis. In *AMIA annual symposium proceedings*, volume 2011, page 189. American Medical Informatics Association, 2011.

[170] Jionglin Wu, Jason Roy, and Walter F Stewart. Prediction modeling using ehr data: challenges, strategies, and a comparison of machine learning approaches. *Medical care*, 48(6):S106–S113, 2010.

[171] Mohammed Khalilia, Sounak Chakraborty, and Mihail Popescu. Predicting disease risks from highly imbalanced data using random forest. *BMC medical informatics and decision making*, 11(1):51, 2011.

[172] R Birtwhistle, K Keshavjee, and A Lambert-Lanning. Building a pan-canadian primary care sentinel surveillance network: Initial development and moving forward. *Family Medicine*, 22:412–22, 2009.

[173] Richard Birtwhistle and Tyler Williamson. Primary care electronic medical records: a new data source for research in canada. *Canadian Medical Association Journal*, 187(4):239–240, 2015.

[174] J Friedman, T Hastie, and R Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[175] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[176] Kirsten M Fiest, Nathalie Jette, Hude Quan, Christine St Germaine-Smith, Amy Metcalfe, Scott B Patten, and Cynthia A Beck. Systematic review and assessment of validated case definitions for depression in administrative data. *BMC psychiatry*, 14(1):289, 2014.

[177] Mary Charlson, Ted P Szatrowski, Janey Peterson, and Jeffrey Gold. Validation of a combined comorbidity index. *Journal of clinical epidemiology*, 47(11):1245–1251, 1994.

[178] Robert L Nussbaum and Christopher E Ellis. Alzheimer's disease and parkinson's disease. *New England Journal of Medicine*, 348(14):1356–1364, 2003.