Using Abstraction to Overcome Problems of
Sparsity, Irregularity, and Asynchrony
in Structured Medical Data

By

Jacob Paul VanHouten

Dissertation
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
in

Biomedical Informatics
August, 2016
Nashville, Tennessee

Approved:

Thomas A. Lasko, M.D., Ph.D.

Christopher J. Fonnesbeck, Ph.D.

Katherine E. Hartmann, M.D., Ph.D.

Michael E. Matheny, M.D.

Nancy M. Lorenzi, Ph.D.

For my families

Especially Dale, DeeAnn, Sam, and Mike

Because you believed in me

ACKNOWLEDGEMENTS

This work was funded in part by the National Institutes of Health (Grant T32 GM07347) and the National Library of Medicine (Grant T15 LM007450).

I would like to thank my dissertation committee for their guidance throughout this work and throughout my education at Vanderbilt. Tom Lasko, my primary advisor, has shepherded me as well as challenged me throughout our five years working together, and I could not have asked for a better mentor to help me direct my interests in machine learning and medicine. I'm certain that whoever his second graduate student is will think likewise. Chris Fonnesbeck, who also served as my primary advisor in my MS degree in Biostatistics, has been a wealth of knowledge and support through both degree programs. Michael Matheny was the first person to encourage me to pursue the additional MS degree in Biostatistics, and I am glad I listened to him. Kathy Hartmann has been my graduate school liaison to the MSTP, as well as a fantastic boss when I assisted teaching her first year Causal Inference class. And Nancy, the self-proclaimed "pusher" of the group, has helped immensely in keeping me on track to finish my degrees at the appropriate times.

I would also like to thank others who have given me invaluable assistance and support. I would like to thank Cindy Gadd, Kevin Johnson, and Mark Frisse for stimulating discussions which helped shape my interests during my PhD training. I would like to thanks Jacek Bajor for his assistance with learning many of the more arcane computational tips and tricks, as well as for being a fantastic officemate.

I would also like to express my gratitude to the Departments of Biomedical Informatics, Biostatistics, and the Medical Scientist Training Program. I appreciate the support of the students, faculty, and staff, and am honored to call you all colleagues. I would not have made it through this training without you.

I want to say thank you to my family, especially my parents Dale and DeeAnn, as well as Granni and Pawpaw, my second parents, for the incredible love and support you showed me growing up. I also want to say thank you to my parents-in-law, Mike and Gina Bright, who have welcomed me completely into their close family circle as well.

My final thanks are to my amazing wife Courtney. Thank you for standing by me during these many years of work, stress, confusion, failure, and finally success. All of this, while still pursuing your own academic career. I could not ask for a better partner.

TABLE OF CONTENTS

Page

APPENDICES

LIST OF TABLES

Table Page

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Several challenges make analyzing health data with current machine learning methods difficult. Among these challenges are that the data are: largely and not randomly missing; collected irregularly in response to irregular clinic visits; and asynchronously collected at different visits.

In this dissertation, I explore the utility of modeling clinical data using various representations and whether they can be used to overcome the problems of sparsity, irregularity, and asynchrony from health data. I accomplish this through two means. First, I will perform a quantitative analysis of how data representation complexity of non-specific laboratory elements affects the discriminative performance of binary classifier models for highly specific procedural and demographic outcomes. I hypothesize that the representation that allows models to most effectively use non-specific information distributed throughout the medical record laboratory results will provide the best discrimination, calibration and confidence. Second, I explore the use of longitudinal, continuous data representations to query against particular combinations of laboratory results. I hypothesize that these experiments will demonstrate the potential value of this method for identifying rare phenotypes associated with unique clinical findings.

Non-technical summary

In biomedical research, one major focus is on identifying as-yet-unknown associations between clinical findings, diseases, outcomes, and successful treatments. For instance, it is desirable for a doctor to know that a patient with a particular genetic marker will have a less favorable reaction to a drug than another patient, so that they can potentially prescribe a drug that would work better for the patient.

Traditional ways that medical researchers approach uncovering these associations are randomized controlled trials and cohort studies. In both research designs, care is taken to ensure that the data about the study participants are correct, complete, and collected at the appropriate time designated by the study. Discoveries made using these approaches are considered reliable, but come with the increased cost of assuring the integrity of the data.

More recently, electronic health records have allowed medical researchers to explore associations between findings and diseases using information that is recorded as a byproduct of regular clinical care. Unlike trials and cohort studies, medical records data allow for analysis of larger populations over longer times, and this benefit can lead to discoveries which may have not been possible using more traditional methods. However, the data collected for patient care is significantly less curated than trial or cohort study data, and characteristics of the data make them more difficult to use for discovering new associations. Several methods of addressing the problems caused by these characteristics have been described. My work in this dissertation explores how these different methods affect researchers' ability to use electronic health data for identifying patterns and associations.

For example, the choice of how to represent the data that is extracted from the medical record may determine the performance of computerized methods used to discover relationships, even if the method and the relationships are the same across data representations. In order to provide some insight into the effects of the choice of data representation, I selected a specific computerized learning method and applied that method to several problems. In each problem, the goal was to distinguish between records of patients with a particular outcome of interest, such as a gall bladder surgery or a hip replacement, and those without, using only laboratory data found in the medical record. I show that more complex data representations do not necessarily lead to improved model performance. Chapter III contains the details on these experiments.

I also explored using continuous representations of laboratory data. With these representations, it was possible to look for associations, even for events that did not occur at the same time. I showed that continuous data representations could be used to explore which diagnostic codes are associated with particular laboratory findings. The details of these experiments are found in Chapter IV.

CHAPTER II

BACKGROUND

An electronic health record (EHR) is a computerized version of a patient's medical history over time. It contains many data elements related to a patient's medical care (Table 2.1)[1].

Table 2.1. Data domains contained within electronic health records.

Administrative and billing data
Patient Demographics
Progress notes
Vital signs
Medical histories
Diagnoses
Medications
Immunization dates
Allergies
Radiology images
Lab and test results

EHRs have improved clinical practice in terms of day-to-day record keeping. For example, EHRs are capable of simplifying the tasks of accessing, retrieving, and analyzing clinical information; electronic lookup reduces the need to sift through paper records. Rapid copying of electronic records allows for data to be easily shared with the patients, their families and all the members of the care team. Additionally, EHRs alleviate problems of illegibility that can arise during recording by hand.

However, many of potential benefits of EHR lie in their ability to enhance or enable more complex capabilities [2]. Here, I discuss three such capabilities: computerized physician order entry (CPOE) [3,4], clinical decision support (CDS) [5,6], and health information exchange (HIE) [7].

CPOE allows providers to use a computer to prescribe medications and place orders for laboratory and radiology tests, rather than filling out paper forms. Like EHR, structured CPOE also reduces errors arising from legibility or faulty interpretation of free text orders. Electronic input helps to reduce medical errors due to illegible writing or ambiguous units for ordered medications. As most CPOE systems interact directly with the EHR, this also removes the additional step of recording orders into the medical record, which reduces erroneous information.

A CDS system aids the provider in making medical decisions such as ordering tests or prescribing medications. Such systems can function by providing information and guidance to the provider during their decision process, allowing the treatment for an individual patient to be informed by evidence accumulated across many different studies [8]. This support may be given in the form of a clickable hyperlink, which could display current clinical guidelines for the management of a patient's disease. Alternatively, the support may be more active, opening a window and asking the provider if they really meant the information that they entered. The information delivered via CDS may also be patient-specific; allergy information or genetic markers of drug metabolism stored in the medical record can be shared with providers at the point of decision making in order to avoid potential hypersensitivity reactions or drug over- or under-dosing. Systems which combine CPOE and CDS systems have been shown to have moderate to high effects on doctors ordering the correct treatment, and some small effects on patient mortality [9].

HIE allows for efficient sharing of information between different clinical organizations. This is critical for improved care, as few patients receive all of their medical treatment at one institution. HIE may decrease overall costs to the system by reducing unnecessary repeat testing and inappropriate admissions [10,11]. HIE can also decrease wait time for physicians who need clinical records from another institution [12]. Traditionally, such records needed to be faxed, causing delays in decision making or clinical care. Even so, HIE does not ensure that the data are standardized and compatible between institutions, and mapping to formal ontologies may be necessary in order for systems to operate on the data that is exchanged [13].

Powerful in their own rights, the combination of CPOE, CDS, and HIE working together can further improve patient care. For instance, while ordering errors can be significantly reduced by the use of CPOE only, this effect is greatly increased when combined with CDS that alerts physicians to potentially better alternatives based on the orders entered. Moreover, displaying information that could alter a physician's decision based on patient history would greatly benefit from access to clinical records outside individual medical systems; HIE can allow CDS to use this information.

Secondary data usage

In addition to these direct operational benefits, EHRs can advance our understanding of health, medicine and medical care through "secondary use" of clinical data [14]. Especially compared to traditional methods of medical research, such as randomized controlled trials and cohort studies, the use of EHRs compares favorably in terms of cost, patient heterogeneity and representativeness, and length of records [15].

Aside from the upfront cost of EHR implementation and the necessary upkeep of the system, the additional cost of extracting and utilizing clinical data for research is minimal, as these data are collected during routine clinical care; the largest remaining cost is that of data cleaning [16]. Compare this to the cost of data collection in a clinical trial or cohort study, where additional workers must be hired to rigorously collect information about the participants. While the completeness of data from trials or cohorts may be superior to that of data collected during the processes of clinical care, the larger number of data elements captured, the larger sample size, and the relative cost of EHRs make them an appealing resource for research and discovery [17,18].

EHRs typically do not have strict criteria for the inclusion of patients into the record, except that the patient receives medical care; this is in contrast to randomized controlled trials and cohort studies, which often exclude patients that do not have desired characteristics. As a result, EHRs typically contain more data on populations that are underrepresented in trials and studies, such as the elderly, patients with multimorbidity, and patients of racial minority background [19].

EHRs are longitudinal by nature, and this characteristic lends these data to long-term outcomes research beyond what is feasible in a trial or cohort setting. This allows researchers to ask significantly more questions of the data, including identification of late-term effects of interventions that a shorter clinical trial may not be able to detect. Perhaps most powerfully, the discoveries made from secondary use of EHRs can directly feed back into the clinical environment. The benefits of CPOE, CDS, and HIE systems rely on current clinical guidelines and information in order for providers to continually improve care. A virtuous cycle of research findings leading to improved clinical care which spurs further research is the basis for the idea of a learning health system that facilitates quality improvement, clinical research and other data-driven approaches to improving health [20].

Learning from EHRs

Using large data sets such as EHRs to identify patterns and relationships requires methods that allow researchers to efficiently analyze large amounts of data. Ideally, such analysis should be performed efficiently, automatically, and make use of as much data as possible.

Statistical and machine learning approaches (sometimes collectively termed data science) are algorithmic methods for modeling complex data sets in order to learn and recognize patterns [21]. These approaches have gained widespread use in recent years, and this popularity has been driven by advancements in computational methods as well as the explosion of widely available large data sets. Data science techniques have been used in such varied tasks as image analysis, voice recognition, spam filtering, and many more, including medical diagnosis [22,23], prognosis [24] and phenotyping [25].

There are two main branches of learning algorithms: supervised and unsupervised [26]. Generally, supervised learning is concerned with learning relationships between data elements based on at least a subset of labeled data (output variables). Examples of such tasks could be predicting a patient's diastolic blood pressure given their systolic blood pressure, or classifying patients as having diabetes or not. These tasks could be performed using typical statistical approaches, such as linear or logistic regression, or using machine learning techniques such as random forests [27], support vector machines [28], or artificial neural network classifiers [29].

Unsupervised learning, on the other hand, deals largely with extracting underling structure from the data in the absence of clear labels. The input variables could be similar to those used in supervised learning, but instead of dividing instances into different classes, unsupervised learning tries to identify relationships and structure between the input variables. Examples of this type of learning in include clustering [30], dimensionality reduction [31], and signal separation techniques [32].

Clinical data can be useful for either supervised or unsupervised learning for discovering clinical associations. Here, I discuss the major types of data found in EHRs and some examples where they been used to learn patterns and identify associations in a clinical context.

Images

Medical images such as x-rays, MRIs, blood smear images, and microbiology slide preparations, are important components of EHRs. From these images, physicians can determine a patient's likely diagnosis and expected prognosis. Traditionally, such images have been reviewed by radiologists and pathologists, and the interpretations have been entered into the medical record for review by other healthcare professionals. While these summaries do provide high level interpretability of the image findings, they contain only partial information.

Recently, data science techniques have been applied to medical image analysis [33]. In this context, the features directly produced by the imaging technique can be identified via learning algorithms, labeled as having outcomes of interest, and directly used for pattern recognition. Example applications of such approaches include classification of different ultrasound heart views [34] , analysis of peripheral blood smears [35], automatic diagnosis of diabetic retinopathy from ophthalmology images [36], and detection of lung and colorectal cancers from thoracic imaging [37].

Free text forms

A significant portion of EHR data is stored as free text, or fields in which a provider can type whatever description or commentary about the patient's medical history they choose. Allowing such descriptive entries can be beneficial, in that subtle impressions about a patient's state can be flexibly recorded. Examples of such free text fields include patient history and physical, clinical progress notes, laboratory or radiology reports, and discharge summaries.

Natural language processing (NLP) is an approach for automatically parsing free text and converting it to meaningful representations [38–43]. These representations can then be used as substrates for machine learning, and have been successfully used to surveil for postoperative complications [44], identify the presence of chronic and acute diseases [45],and assign appropriate ICD-9 billing codes to radiology reports [46].

Structured Data

Unlike free text data, which can include almost any information, structured data has specific limitations on how and where it can be recorded. For instance, the data pertinent to a patient diagnosis might be recorded as the patient's name or medical record number, the diagnosis code assigned to that person, and the timestamp for when the code was assigned. Laboratory measurements could include the name of the laboratory test, the results of the test, whether the results were normal or abnormal, and a timestamp of the event. Structured data forms include diagnosis and billing codes, laboratory results, and tick boxes which indicate the presence of a finding or procedure.

While structured medical data does not allow the expressiveness of free text entries, the semantic homogeneity with which elements are recorded makes structured data more interpretable. Structured data have less ambiguous meanings for the same field than free text; for instance, two glucose measurements recorded in mg/dL mean the same thing, even if the actual values are different. This quality can allow for simpler aggregation of multiple patient records, as it can be assumed that a structured field for one patient has approximately the same interpretation for other patients. Structured medical data has been used to identify records with acute coronary syndrome [47], acute kidney injury[48], and myriad other conditions.

In this dissertation, I used only structured data; namely, laboratory results and billing codes. Laboratory results are added to a patient’s record as tests are ordered and returned, and these inform the healthcare team about the physiologic state of the patient. Billing codes are assigned to a patient’s medical record during interactions with the healthcare system, and they are typically used to indicate diagnoses that related to the medical trajectory of the patient. Historically, these codes have been coded using the International Classification of Diseases, Ninth Revision[49].

Challenges to learning from medical data

Despite these and many other examples of successful learning from clinical data, the task of extracting meaning from medical records remains difficult. Overcoming these difficulties is necessary for improving our ability to use clinical data for research and discovery. In this section I describe some of the specific challenges to learning from clinical data, as well as examples of how previous work has addressed these issues. While each data type in clinical records has its own specific considerations, I focus here on challenges that are common to most clinical data types, and specifically on ways they have been handled when using structured medical data as I have in this dissertation.

Error and Uncertainty

Within clinical data, there are numerous sources of unmodeled variation, also colloquially termed noise, and all can affect the outcomes of analyses if not accounted for. One particular example of potential data errors in a clinical setting is the uncertainty associated with measurements [50]. While laboratory measurements are largely accurate, there is still uncertainty in their values. In the best case, this might not affect the analysis at all; in the worst case however, the uncertainty could lead to false discoveries [50].

Another source of noise arises from misreporting information into the record, or even omitting important information entirely. Such errors could arise from recording correct information into the wrong patient's chart, copying and pasting a previous clinical note without appropriately updating the information, or simply forgetting to chart a clinical event [51]. As with the uncertainty associated with measurements, this can lead researchers to erroneous conclusions.

By far the most common method of addressing noise in biomedical models is to ignore it. This is an understandable approach; even though the laboratory measurements are an imperfect proxy for the underlying physiology of the patient, they are still the most likely value for the true state of the lab given the information available. However, this can still lead to errors as described above[52].

Simple data cleaning can go a long way in reducing errors found in medical records. Sometimes, this can be as simple as converting a result recorded in one unit of measurement to another, or recognizing that the recorded value is not biologically compatible with the clinical history. However, this can be problematic in some cases; without more information, there are many instances where a person's recorded weight would be a reasonable value, whether the intended units were kilograms or pounds. Determining the intended value for such a measurement can be difficult [53].

Some biomedical models address noise by modeling the uncertainty around the point estimates provided by the observed values. For instance, Gaussian process regression can be used to interpolate noisy observations while accounting for uncertainty [54,55], as can multiple imputation [56].

While the research in this dissertation does not directly address the issue of clinical data noise, the work in Chapter III can potentially be used to address missing and miscoded information. Experiments in both Chapters III and IV are designed with consideration of potential sources of error in the data.

Sparsity

Medical data is sparse, both in terms of time and in terms of recorded information. Across most patients' lifetimes, overwhelmingly more time is spent outside a clinical setting than in one. If Accordingly, if one imagines a patient's life as a timeline, the majority of data are not recorded in an EHR. As such, clinicians and researchers are left with only the limited view of the patient's risk factors and experiences that is recorded in their medical chart. Furthermore, many clinical systems do not communicate information about patients with other systems effectively, leading to missing data through failure to communicate. Sometimes this problem can be overcome by including information from associated registries or the Center for Medicare and Medicaid Services claims data, but such data validation is not available for all patients [57,58].

Even within the context of the clinical encounter, the data recorded are sparse. Of the thousands of possible measurements, procedures, and diagnosis codes available to physicians, only a small fraction are recorded in a patient's chart at any given visit. Part of this is intentional and largely positive; it makes little sense for a clinician to order a chest x-ray on every patient who comes in for an annual checkup. Additionally, the decisions about which data are recorded in a chart are driven by the actual practice of medicine, meaning that the data in the record are not missing at random [59,60].

In the statistics literature, the mechanism of data missingness has typically been described in terms of three categories: missing at random (MAR), missing completely at random (MCAR), and missing not at random (MNAR)[61]. Data that are missing at random mean that any differences between missing and observed values can be explained entirely by the observed data. Missing completely at random data go further, such that there are no systematic differences between missing values and observed values. MAR and MCAR data require fewer assumptions for valid inference. On the other end of the spectrum, missing not at random data are influenced to be missing by the values of the missing data. Returning to EHRs, data that are not collected by the physician because of their belief that the results will not be helpful for treatment decisions are clearly MNAR, which will complicate data analysis [62].

One method that has been used to control the effect of observation sparsity in medical studies has been to use only records that have at least a certain amount of data. This can be as extreme as only including records for patients that have entries for all of the variables of interest; a complete-case analysis [61]. However, given that the data are not missing at random, this sometimes leads to biased sampling and non-representativeness of populations. For instance, it was found that high risk surgical patients had over five times as much data as lower risk patients [63]. Using the amount of data as a decision tool for which records should be included in the model would over-represent the sicker populations.

Another method used to handle missing data is imputation[61], or setting a missing value to some reasonable guess. The simplest form of imputation is to just replace missing values with the population mean or median, which is a naïve estimate. However, if a significant proportion of data is missing, this can break dependencies between the variables of interest and cause models built on the imputed data to perform poorly [64].

The missing values can also be replaced conditionally on the non-missing observations for a record. In other words, given that the observed variables took the observed values, what is the most likely result for the missing value? Like naïve imputation, single imputation methods like this are subject to potential biases and can lead to incorrect conclusions [64]. Extending from single imputation, it is possible to produce several possible imputations, and average the results for inference. This has been demonstrated to be more robust than single imputation [56]. Multiple imputation has been shown to be effective at handling high proportions of missing data [65,66].

In my work, I address missing values and sparsity two different ways. In Chapter III, I explore the effect of different data representations on model performance. Where my input variables are counts of events, I do not have to impute; zero events are valid entries for this approach. I elected to replace missing values for laboratory results with the population mean. In Chapter IV, I demonstrate a method of handling missing values by interpolating between observed values, and setting values beyond the first and last values to the record-specific median.

Irregularity

Medical data are entered into patient charts as they are needed for clinical care. As a result, there is no standard frequency at which entries are made. It is likely that a patient will not see a medical provider for months or years, and then develop an illness which will require multiple clinical encounters over a short period of time.

Some researchers avoid irregularity, using only regularly sampled data such as is collected in intensive care units, fetal monitoring or continuous echocardiograms [67,68]. However, for the majority of medical data, this is not plausible.

Separating time into discrete bins is an approach that manages irregularity [69]. In the extreme case, a bin may be as large as the entire patient record; in terms of a binary event, this is equivalent to an indicator of whether the event occurred or not. From a data perspective, binning solves the problem of irregularity but induces another challenge: determining the resolution at which the data should be recorded and encoded for processing by a learning algorithm. If the data are captured at too low a resolution (much less often than the observed data points), then the information contained in the encounters associated with the acute event are significantly compressed. If on the other hand the data are captured at too high a resolution (much more often than the observations), then many of the entries would have missing values, and the data now exhibits the problems of sparsity described above. Additionally, inferences and predictions may be very sensitive to the choice of bin thresholds.

In this dissertation, I address data irregularity by using bins as well. In Chapter III, I represent the clinical data using bins at several different resolutions, and exploring the effect of these data representations on model performance. Within each bin, a summary measure such as the total count or mean result substitute for all of the values that fall within that time period. In Chapter IV, I transform the data into longitudinal functions. After this transformation, there is an interpolated estimate for every division at any arbitrary resolution.

Asynchrony

While irregularity is a property of the sampling rate of any individual variable, asynchrony is about a property of the relationship between sampled variables. As mentioned previously, not all variables are recorded for a patient at each of their visits; what is included in the chart is largely determined by clinical need. However, if a researcher wanted to look for associations between two related entities, such as hypertension and insomnia, they would want to be able to look at whether one affects the other. Yet, it is hard to determine such an effect if the variables are not observed at the same time. This leads to the question: "How temporally close is *close enough* to say that two things happened at the same time?".

As with irregularity, binning has been the main method of addressing asynchrony. Once the data are binned appropriately, say into discrete years, it is a matter of determining if two events happened within the same bin. Another more flexible approach is the sliding window, in which a specific bin width is designated, but the window is translated down the timeline. If any two events ever fall within the sliding window, they can be considered to have occurred close together. In any of these approaches, the challenge still remains determining what level of temporal relatedness is most appropriate.

The methods I used to manage irregularity also extend to managing asynchrony. In Chapter III, I binned all results into the same specific time bins, and results for different laboratory measurements that occupy the same relative time bin are assumed to have occurred at approximately the same times. In Chapter IV, the continuous longitudinal transformations allow all of the laboratory measurements able to be binned at any resolution. As a result, any arbitrary cross section of a record contains an interpolated estimate of all of the laboratory values of interest.

CHAPTER III

USING DATA ABSTRACTION MODELS OF NON-SPECIFIC LABORATORY RESULTS FOR CLASSIFICATION TASKS

Introduction

Computational approaches to phenotype identification often limit themselves to a small number of highly specific, expert-engineered features when defining phenotypes of interest [47,48,70]. This is in contrast to physicians, who generally use all available medical data when making diagnosis and treatment decisions, even if only through the use of heuristics. While the decision to include only strongly predictive features does provide computational and time savings, it also limits the sensitivity and specificity of the phenotype identification process. Exploring methods that allow computational approaches for phenotype discovery to make use of a larger portion of medical data elements is an essential step on the path to data-driven precision medicine [25,55,71–73].

An important source of such medical data are electronic health records (EHR)[14]. In addition to serving as a record of a patient's clinical care, EHR data may allow researchers to improve detection of patient conditions, procedures, or outcomes in situations where administrative coding is missing, or miscoded [57,74].

In this work, I distinguish between specific and non-specific evidence for an outcome of interest. For example, findings in the medical record that are specific for diabetes mellitus may include an elevated glucose result, the presence of metformin within a patient's medication list, or an ICD-9 code 250. In contrast, non-specific information may have either a known or unknown relationship with the outcome of interest, and is likely also associated with many other outcomes. The findings of coronary artery disease, increased serum creatinine, and medication orders for the antihypertensive drug Lisinopril are associated with, but not specific for, diabetes[75].

In aggregate however, such non-specific information may be useful in indicating the presence or absence of an outcome of interest. If the highly-specific indicators of a condition are missing or miscoded, as is common in EHR data, using non-specific information may allow for high fidelity labeling of cases and controls. Even when the outcome of interest is known with high confidence, inclusion of these other data elements could allow researchers to more precisely define distinct subpopulations of patients that may be of interest.

While much research has focused on the use of specific, expert-engineered features [55,67,70,76], comparatively little has explored the use of non-specific predictors in phenotype identification tasks. Where non-specific features have been included in models, their performance has often surpassed that of similar models with only expert-selected features. For instance, including the most common diagnoses, medications, and other information from the EHR improved detection of as-yet-undiagnosed diabetes over conventional risk models which used only BMI, smoking status, hypertensive status, gender, and age [77]. A natural language processing model identified clinical concepts mentioned in electronic medical records, which were then used to train adaptive elastic net penalized regression models with AUCs of 0.951 and 0.929 for identifying rheumatoid arthritis and coronary artery disease, respectively [78]. Sparse tensor factorization of unselected ICD-9 diagnosis codes and Healthcare Common Procedure Coding System procedure codes produced interpretable, concise phenotypes [79]. Joint probabilistic graphical models of free-text notes, medication orders, diagnosis codes, and laboratory tests identified phenotypes with higher normalized pointwise mutual information than models derived with Latent Dirichlet Allocation [80]. Topographical modeling of patients using high-dimensional genetic data, laboratory results, medications and vital signs allowed identification of subtypes of type II diabetes mellitus [81].

While these studies made use of non-specific predictors in building their models, none to my knowledge have assessed the discriminative ability of non-specific information without including selected, highly-specific, expert-generated features in the model, or which data representations best allow models to make use of this non-specific information. In this work, I explore the effect of different data representations on model performance, recognizing that greater representation complexity can come at a higher cost in computational resources and research effort. I also quantify the discriminative power of non-specific information distributed among laboratory test results. I accomplish this by applying a standard classification algorithm to several different binary classification tasks. I hypothesize that the representations that allow models to most effectively use non-specific information distributed throughout the medical record laboratory results will provide models with the best discrimination, calibration and confidence.

Background

In these experiments, I explored the effect of modeling clinical data using several different representations on model performance by building random forest classifiers for several demographic and surgical outcomes. I quantitatively evaluated the model performance using area under the receiver operating characteristic curve, a standard measure of discrimination, as well as the logarithmic scoring rule, which is a measure of model calibration, discrimination and confidence. Below, I provide background on the random forest classifier and these two measures of model performance.

Random forests

The random forest algorithm is a machine learning technique that has been used extensively in recent years for classification and regression problems [82]. While simple to parameterize, random forests often perform near the top of classification tasks compared other machine learning approaches [83,84]. Here, I provide background to the random forest and some intuition regarding its performance.

In order to understand random forests, it is appropriate to first understand classification trees, sometimes referred to as decision trees [85]. Such trees are simple representations of a greedy process for classifying instances in a data set. Classification trees are related to regression trees, except that the predicted outcome of a classification tree is a nominal class, while the predicted output of a regression tree is a real number.

The typical approach to learning the structure of a classification tree is to create recursive binary splits of the data set of interest. At each split, a single variable and threshold is selected; this is typically the variable and threshold that most reduce the heterogeneity of the data after the split is performed. Recursive splitting is continued in this way until a user-specified rule is achieved, such as a minimum accuracy or a maximum number of instances per terminal node in the tree. However, decision trees are typically poor classifiers and strongly dependent on the training data [86], which is in part why they have fallen out of favor for learning tasks.

A random forest classifier is an ensemble of classification trees, but with sources of randomness injected into their creation. This randomness decreases the correlation between individual trees, improving the strength of the overall forest classifier. Unlike classification or regression trees, individual trees in a random forest do not have access to the entire data set. For each tree in the random forest, the data are sampled with replacement to create a new training set. In addition to the randomly selected training set, each tree is only allowed access to a subset of the available input features when determining the optimal binary split.

As with typical classification trees, this recursive splitting continues until a specified rule is achieved. Often, this rule is that all instances from the data set be classified into distinct terminal nodes. The predictions for each instance in the data are then made on a per-tree basis and averaged over the total number of trees [27].

Random forests have many desirable properties that make them amenable to widespread use in machine learning. They typically scale well with the size of a training data set. They are more robust to output noise than some other machine learning approaches [27].

Tasks with many input features with weak predictive power can be efficiently used by random forests [27]. Random forests have the ability to learn non-linear combinations of weakly predictive variables to provide classifications with generally favorable error rates. Learning these non-linear relationships is automatic for the random forest algorithm, unlike regression approaches where any interactions or non-linearities of interest must be specified for inclusion in the model [27].

One practical benefit of random forests is that they provide an internal estimate of generalization error without the need of a separate test set. This is a result of the sampling that occurs when selecting a separate training set for each tree in the forest. On average, approximately 36% of the data are excluded from the new training set when sampling with replacement; these excluded instances are termed "out of bag" samples. When estimating the performance of a random forest classifier, the error achieved in classifying these out of bag instances approximates what would be found by classifying data from a separate test set.

Furthermore, this out of bag set can be used to determine the relative importance of individual predictors in the random forest. The effect on classification error of adding noise to each of the input variables can be quantified, revealing which variables are most important for accurate classification.

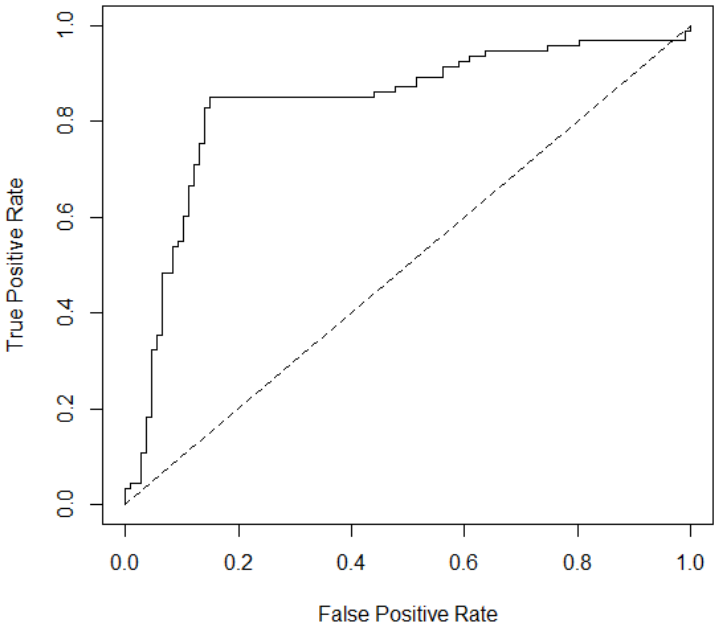
The properties of random forests are not all desirable. Compared to parametric models, random forests and other nonparametric methods typically run slower and require more parameters to be learned, despite their more relaxed assumptions. In many cases, the decision to use a parametric or nonparametric model will depend largely on how confident one is in the underlying distribution of their data [87].

While random forests perform well on tasks with some level of class imbalance, extreme imbalance can affect their performance. Several approaches have been proposed to improve performance on imbalanced data [88,89]. For my work, I selected an approach based on random balanced sampling of the majority and minority classes, which has been shown to improve discrimination performance[88]. Instead of simple sampling with replacement from the original set for each tree, sampling is performed with replacement from the instances in the minority class, then from the majority class to produce the same number. As a result, each tree is trained on a one-to-one ratio of cases to controls. This approach improves classification performance, even when still considering only out of bag performance.

Receiver operating characteristic curves

For machine learning tasks that produce probabilistic estimates, one of the most widely used tools of analysis is the receiver operating characteristic (ROC) curve, which is used to assess model discrimination, or the ability to separate positive and negative instances [90]. Visually, ROC curves provide a representation of the trade-off between true positive rate and false positive rate for a particular binary classification task over the entire range[90]. As the probability for correctly for detecting the outcome of interest increases, the likelihood of wrongly determining that an instance has the outcome of interest is non-decreasing. An example of an ROC curve is shown below.

Figure 3.1. An example ROC plot. The dashed line represents 50% accuracy. The solid line represents the ROC curve at each of the potential setting of false positive and true positive rates. Better ROC curves will approach the upper left-hand corner of the graph, which is a 100% true positive rate and a 0% false positive rate.



While the visual interplay of true positive rate and false positive rate at various thresholds represented by the ROC curve may be of interest to some researchers, the area under the curve (AUC) is the most often used numerical representation of test discrimination. Though several interpretations of AUC have been offered, one common conceptual explanation of AUC is that it is equal to the probability that the test will produce a higher predicted value for a randomly chosen positive example than for a randomly chosen negative example.

One benefit of AUC over simply quantifying how many instances the model classifies correctly is that AUC is insensitive to the prevalence of the outcome. For demonstration, imagine a scenario where there are many examples without the outcome of interest and relatively few with the outcome. In terms of raw accuracy, simply assigning all examples to the majority class with probability 1 would provide a high estimate of accuracy, but AUC would be close to 0.5, or random guessing. Using AUC instead of raw accuracy, on the other hand, provides an estimate of the classifier's ability to discriminate between positive and negative examples independent of the balance of the two classes.

AUC measures are also insensitive to miscalibration of the output probability estimate. Since the AUC measures the probability that a randomly chosen positive example has a score higher than a randomly chosen negative example, only the rank ordering of the estimates determines the AUC, not the actual predictions. Even if all the probability estimates are grossly wrong, they will produce the same AUC measure if their order remains unchanged. This also makes AUCs resistant to high- or low-probability outliers.

Logarithmic scoring rule

Another measure of model performance that is less frequently used in biomedical informatics is the logarithmic scoring rule. The logarithmic scoring rule is calculated as $s = \frac{1}{N}\sum s_i = \frac{1}{N}\sum \ln(r_i)$, where r_i is the probability assigned to the correct label for instance i. In the case of a binary $y \in [0,1]$ classifier predicting $\in [0,1]$, this can be simplified to $r = p^y(1-p)^{1-y}$[91].

This scoring rule is statistically strictly proper, meaning that the performance of a model measured by the rule is optimized uniquely when the classifier accurately predicts the true probabilities of the outcomes. There are three characteristics of the model that can be improved in order to increase the logarithmic scoring rule: calibration, discrimination, and confidence. Calibration is the agreement between the predicted probability of the outcome of interest and the true probability of that outcome. Confidence, in this sense, is how near to certainty the model predicts correct classifications. In other words, a classifier will have a higher (better) logarithmic scoring rule when the calibration, discrimination, or confidence (or any combination of the three) of the prediction is improved. This characteristic can make identifying which component is responsible for any improvement difficult, but the logarithmic scoring rule is still useful for comparing performance on a particular task even without this capability.

The logarithmic scoring rule ranges from -∞ to 0, where a score of zero is equivalent to assigning the correct class with probability 1, and $-\infty$ is equivalent to assigning the incorrect class with probability 1. To provide intuition for this result, the score can be converted back to the probability of predicting the correct class through exponentiation of the scoring rule. A classifier with a logarithmic scoring rule of -0.08 is equivalent to predicting the positive class probability $e^{-0.08}$ = 0.923 for all positive examples and 1 - 0.923 = 0.077 for all negative examples.

Methods

I trained random forest classifiers for eleven different classification problems with outcomes that were easy to extract from administrative data and that I believed would be essentially noiseless (Table 3.1, Appendix A). I selected these specific classification problems because they could be posed as binary classification tasks, and because these classification tasks using only laboratory data represented varying degrees of expected difficulty. For race, I simplified the model to predict white versus non-white. I assumed sex, race and CPT codes to perfectly indicate the presence of each outcome of interest - with acknowledged data limitations discussed in the background section - against which I could compare the predictive power of non-specific laboratory tests. In addition, I trained two additional models on what I expected to be very difficult conditional problems: 1) given that the patient received either a kidney or liver transplant or both, did the patient receive a kidney transplant, and 2) given that the patient received a hip or knee replacement or both, did they receive a hip replacement? I hypothesized that these conditional questions would be difficult problems because there would be significant overlap in the variable importance between the two questions. Although the prevalence of some procedures in this sample is lower than one percent, these numbers are in line with literature findings [92].

Table 3.1. Study population characteristics. The data set is highly imbalanced for many of the outcomes.

Outcome	Number (Proportion) with finding
Sex	152538 (46.87%) Male
Race	263849 (81.07%) White
Splenectomy	879 (00.27%)
Cholecystectomy	2843 (00.87%)
Pancreatectomy	557 (00.17%)
Appendectomy	1148 (00.35%)
Hemorrhoid Surgery	441 (00.14%)
Kidney Transplant	877 (00.27%)
Liver Transplant	1525 (00.47%)
Hip Replacement	2471 (00.76%)
Knee Replacement	2969 (00.91%)

I used data from the Vanderbilt University Medical Center Synthetic Derivative, the deidentified mirror of the hospital's electronic medical record used for research purposes [93]. This resource contains data on over 2.5 million patients going back as far as twenty years. After obtaining IRB consent, I selected the 150 most commonly recorded laboratory tests as potential model features; these account for roughly 95% of all laboratory results recorded in the Vanderbilt record. Of these, I excluded seven because they were not laboratory measurements (medication dose, IV start time, patient location, schedule, the provider who performed a specific test, the user's screenname, date for microbiology plate). I limited my study sample to the most recent eight years of data per record. I also required that individual records have results for at least 10 of the remaining 143 laboratory tests, at least one test for which there were three or more recordings, and no missing data for sex or race. This left a final study population of 325,461 records for training and testing.

I standardized the records by subtracting the population mean and dividing by the standard deviation for each laboratory test. I transformed the data into eight increasingly complex data representations for each patient record and classification task. These were 1) binary, or whether the test was ever ordered, 2) total counts of orders made for the test over the eight-year period, 3) counts per year for each of the eight years, 4) cumulative counts by year, 5) mean of all results in the eight-year span, 6) quintiles of all results in the eight-year span as defined by the sample population, and 7) a combination of order counts and result means (Table 3.2).

Table 3.2. Example representations of clinical data. Binary, counts, and means representations compress the data for a single laboratory into one number. Counts and cumulative counts incorporate a longitudinal component, and quintiles approximate the distribution of the record's laboratory results compared to the rest of the population.

	Glucose	Na	Cl	TRPI
Binary	[1]	[1]	[1]	[1]
Counts	[20]	[5]	[5]	[1]
Counts/yr.	[0, 2, 0, 1, 4, 5, 4, 4]	[0, 0, 0, 0, 1, 2, 1, 1]	[0, 0, 0, 0, 1, 2, 1, 1]	[0, 0, 0, 0, 0, 1, 0, 0]
Cumulative	[0, 2, 2, 3, 7, 12, 16, 20]	[0, 0, 0, 0, 1, 3, 4, 5]	[0, 0, 0, 0, 1, 3, 4, 5]	[0, 0, 0, 0, 0, 1, 1, 1]
Mean	[-0.10]	[0.32]	[-0.42]	[0.35]
Quintiles	[2, 5, 8, 5, 0]	[0, 0, 3, 2, 0]	[0, 3, 1, 1, 0]	[0, 0, 0, 0, 1]
Combo	[(20, -0.10)]	[(5, 0.32)]	[(5, -0.42)]	[(1, 0.35)]

I built random forest classifiers for each combination of representation and task, totaling 91 models. Given the high imbalance in my data for some of the classification tasks, I down-sampled the majority class to the same number of minority class examples, both sampled with replacement. This resulted in a one-to-one ratio of cases to controls for each decision tree in the forest. I optimized each forest's parameters to the specific task and representation for which it was trained.

For each task and representation, I report three measures of performance: AUC; the logarithmic scoring rule; and the average runtime per task. AUC and logarithmic scoring rule were computed only on out-of-bag samples.

Calculations were performed in the R statistical environment using packages downloaded from the Comprehensive R Archive Network (CRAN)[83,94–96]. This work was performed on a Linux server with 64 GenuineIntel 6 processors and 500GB of RAM. Random forests were built on 25 CPUs running in parallel, but the same configuration was used for all tasks and representations.

Results

The AUCs of the models ranged from 0.664 to 0.996 (Figure 3.2, Table 3.3). The easiest problem, on average, was detection of kidney transplant, while the hardest was the determination of whether a joint replacement patient received surgery on their hip or their knee; however, the performance of the classifier for identifying race using only the binary representation of laboratory data performed the worst overall. The models built using more complex data representations tended to have longer runtimes. The logarithmic scoring rules also showed varying levels of performance, ranging from -0.781 to -0.135 and largely tending to agree with the results from evaluating the AUCs (Table 3.4).

Table 3.3. Area under the curve (AUC) and average runtime for thirteen classification tasks and seven data representations. The highest performing representation for each task is bolded. Numbers in parentheses next to the outcomes are the dimensions of the input space.

Outcome	Binary (k=143)	Counts (k =143)	Means (k =143)	Quintiles (k= 715)	Year Bins (k =1144)	Cumulative (k =1144)	Combo (k = 286)
Sex	0.745 [0.743 , 0.746]	0.781 [0.779 , 0.782]	0.895 [0.894 , 0.896]	0.894 [0.892 , 0.895]	0.765 [0.763 , 0.766]	0.768 [0.766 , 0.77]	**0.902 [0.901 , 0.903]**
Race (white v all)	0.664 [0.662 , 0.667]	0.683 [0.680 , 0.685]	0.804 [0.802 , 0.806]	0.789 [0.787 , 0.790]	0.679 [0.677 , 0.681]	0.682 [0.680 , 0.685]	**0.805 [0.803 , 0.806]**
Splenectomy	0.913 [0.900 , 0.925]	0.927 [0.915 , 0.939]	0.934 [0.922 , 0.945]	0.937 [0.926 , 0.948]	0.920 [0.907 , 0.932]	0.922 [0.910 , 0.935]	**0.940 [0.929 , 0.951]**
Cholecystectomy	0.837 [0.828 , 0.846]	0.853 [0.844 , 0.862]	0.843 [0.834 , 0.852]	0.845 [0.836 , 0.854]	0.846 [0.837 , 0.855]	0.849 [0.840 , 0.858]	**0.858 [0.850 , 0.867]**
Pancreatectomy	0.937 [0.923 , 0.951]	0.948 [0.936 , 0.961]	0.937 [0.923 , 0.951]	0.946 [0.933 , 0.959]	0.938 [0.924 , 0.952]	0.943 [0.929 , 0.956]	**0.949 [0.936 , 0.962]**
Appendectomy	0.790 [0.775 , 0.806]	0.799 [0.784 , 0.815]	0.829 [0.814 , 0.844]	0.835 [0.820 , 0.849]	0.802 [0.786 , 0.817]	0.795 [0.780 , 0.811]	**0.840 [0.826 , 0.854]**
Hemorrhoid Surgery	0.754 [0.728 , 0.781]	0.755 [0.729 , 0.782]	0.759 [0.733 , 0.786]	0.762 [0.736 , 0.788]	0.765 [0.739 , 0.791]	0.766 [0.740 , 0.792]	**0.777 [0.751 , 0.803]**
Kidney Transplant	0.995 [0.991 , 0.998]	0.994 [0.990 , 0.998]	**0.996 [0.993 , 0.999]**	0.995 [0.991 , 0.998]	0.992 [0.988 , 0.996]	0.993 [0.989 , 0.997]	**0.996 [0.993 , 0.999]**
Liver Transplant	0.974 [0.968 , 0.979]	**0.977 [0.971 , 0.982]**	0.972 [0.967 , 0.978]	0.976 [0.970 , 0.981]	0.975 [0.969 , 0.98]	0.975 [0.970 , 0.981]	0.976 [0.971 , 0.982]
Kidney v Liver	0.960 [0.950 , 0.969]	0.982 [0.976 , 0.988]	0.978 [0.971 , 0.984]	0.981 [0.974 , 0.987]	0.975 [0.968 , 0.983]	0.981 [0.974 , 0.987]	**0.984 [0.978 , 0.990]**
Hip Replacement	0.930 [0.923 , 0.937]	0.960 [0.955 , 0.966]	0.950 [0.944 , 0.956]	0.958 [0.952 , 0.963]	0.955 [0.949 , 0.961]	0.957 [0.952 , 0.963]	**0.962 [0.956 , 0.967]**
Knee Replacement	0.948 [0.943 , 0.954]	**0.977 [0.973 , 0.981]**	0.967 [0.963 , 0.972]	0.974 [0.970 , 0.978]	0.972 [0.968 , 0.976]	0.975 [0.971 , 0.979]	**0.977 [0.973 , 0.981]**
Hip v Knee	0.666 [0.651 , 0.680]	0.748 [0.735 , 0.761]	0.719 [0.705 , 0.733]	0.749 [0.735 , 0.762]	0.732 [0.718 , 0.745]	0.737 [0.724 , 0.751]	**0.750 [0.737 , 0.763]**
Avg Runtime per Task (s)	180	183	184	683	1261	1982	225

Figure 3.2. Area under the ROC curve for thirteen outcomes and seven data representations. Lines connect results using the same representation. For the tasks of classifying race and sex, notice that the models using representations which do not include information about the laboratory result values perform significantly worse than models which make use of test values.

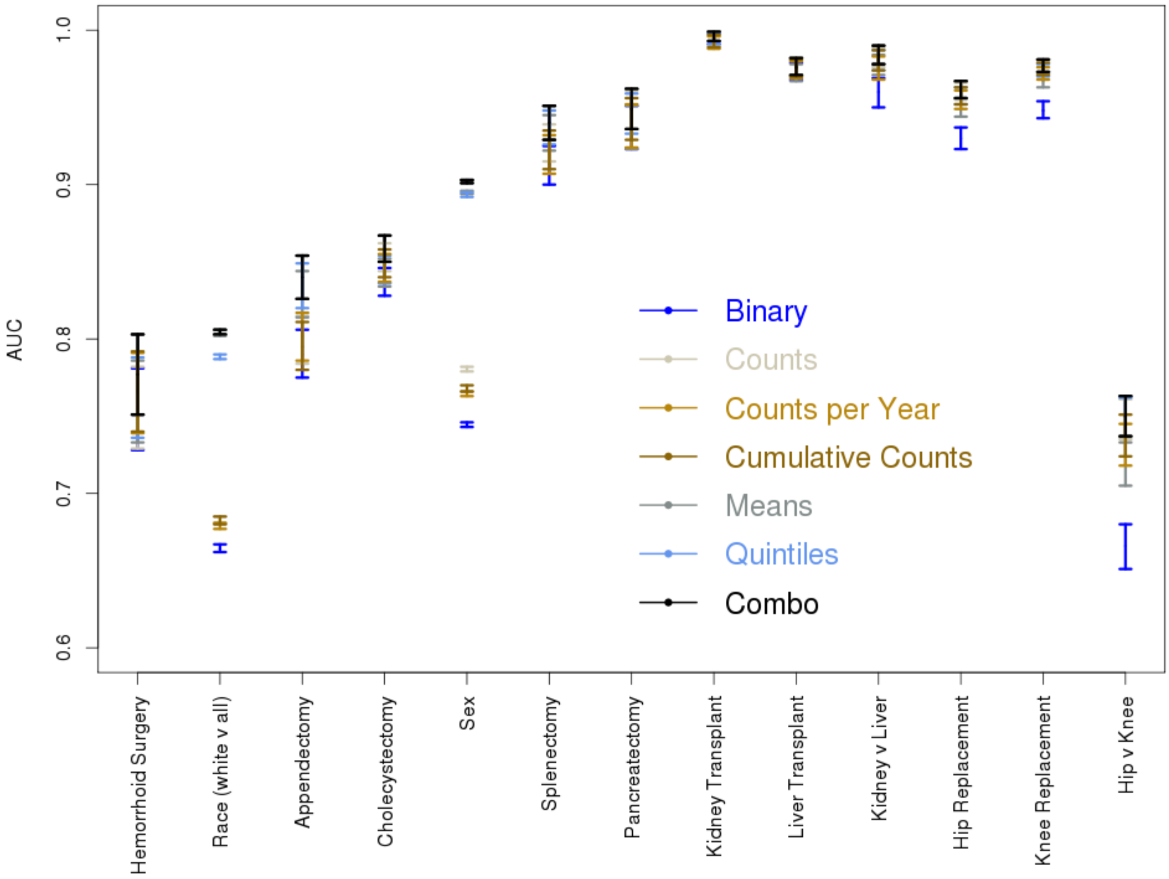


Table 3.4. Logistic scoring rule results for thirteen outcomes and seven data representations. The best performing representation for each task is bolded.

Outcome	Binary	Counts	Means	Quintiles	Year Bins	Cumulative	Combo
Sex	-0.718	-0.559	-0.436	-0.450	-0.580	-0.575	**-0.427**
Race (white v all)	-0.657	-0.564	**-0.486**	-0.500	-0.575	-0.575	**-0.486**
Splenectomy	-0.301	-0.271	-0.257	-0.262	-0.285	-0.286	**-0.248**
Cholecystectomy	-0.392	-0.359	-0.373	-0.381	-0.374	-0.373	**-0.354**
Pancreatectomy	-0.265	-0.238	-0.259	-0.256	-0.260	-0.255	**-0.232**
Appendectomy	-0.437	-0.404	-0.383	-0.390	-0.405	-0.413	**-0.379**
Hemorrhoid Surgery	-0.487	-0.429	-0.431	-0.420	**-0.423**	-0.424	-0.424
Kidney Transplant	-0.084	-0.070	-0.096	-0.070	-0.093	-0.085	**-0.064**
Liver Transplant	-0.165	**-0.135**	-0.176	-0.149	-0.156	-0.143	-0.137
Kidney v Liver	-0.280	-0.187	-0.237	-0.214	-0.242	-0.200	**-0.182**
Hip Replacement	-0.332	**-0.217**	-0.269	-0.247	-0.250	-0.247	-0.220
Knee Replacement	-0.291	**-0.154**	-0.219	-0.193	-0.198	-0.184	-0.160
Hip v Knee	-0.781	-0.597	-0.617	-0.596	-0.615	-0.612	**-0.589**

Discussion

I demonstrated the benefit of using non-specific laboratory results as input features to random forest classifiers predicting demographic and surgical labels. Using only low-specificity laboratory values, I achieved good discriminative prediction accuracy. This performance did not require the use of expert-derived features; nor did it require much data processing to achieve, as models built using lower complexity representations often performed as well as more complex ones.

Most often, models containing only the concatenation of mean test results and counts of orders performed the best on each task, with close to the minimum compute time. In other words, using result means and counts of laboratory orders alone was an efficient way to encode test results. While I used only random forest classifiers to explore the effect of including non-specific variables in various data representations, I expect that my results will extend to other classification algorithms, at least those that are as effective as random forests in extracting complex nonlinear relationships between input variables.

The calculated logarithmic scoring rules largely reaffirm the AUC rankings of the data representations. While it is impossible to separate whether the performance is due to model calibration, model discrimination, or model confidence, it is generally true that the models that performed the best in terms of AUC also performed the best in terms of logarithmic scoring rule.

The most important variables as determined by the random forests were not always the same among different data abstraction models within a specific task. For example, while the presence of an order for urine squamous epithelia or thyroid stimulating hormone (both of which are tests performed more often on women) was highly discriminative of sex in the binary representation, the mean results of creatinine, hemoglobin and mean corpuscular hemoglobin concentration were the most predictive features in the mean result value representation. An ordered test for the level of the anti-rejection drug tacrolimus was important for identifying records with a liver transplant; however, the total number of counts for liver function tests was more predictive in the abstraction model which contained count data.

For some tasks using the combination mean-count data abstraction models, the count of a particular laboratory result is more important than the mean value (Figure 3.3), and vice versa (Figure 3.4). For instance, the number of times a laboratory for lipase was ordered was the most important variable for determining the cholecystectomy status of a patient record, while the mean value of lipase proved to be more important for identifying records with appendectomy. The full suite of variable importance plots is included in Appendix B.

Figure 3.3. Variable importance plot for classifier predicting cholecystectomy using a combination representation consisting of counts and means of laboratory results.

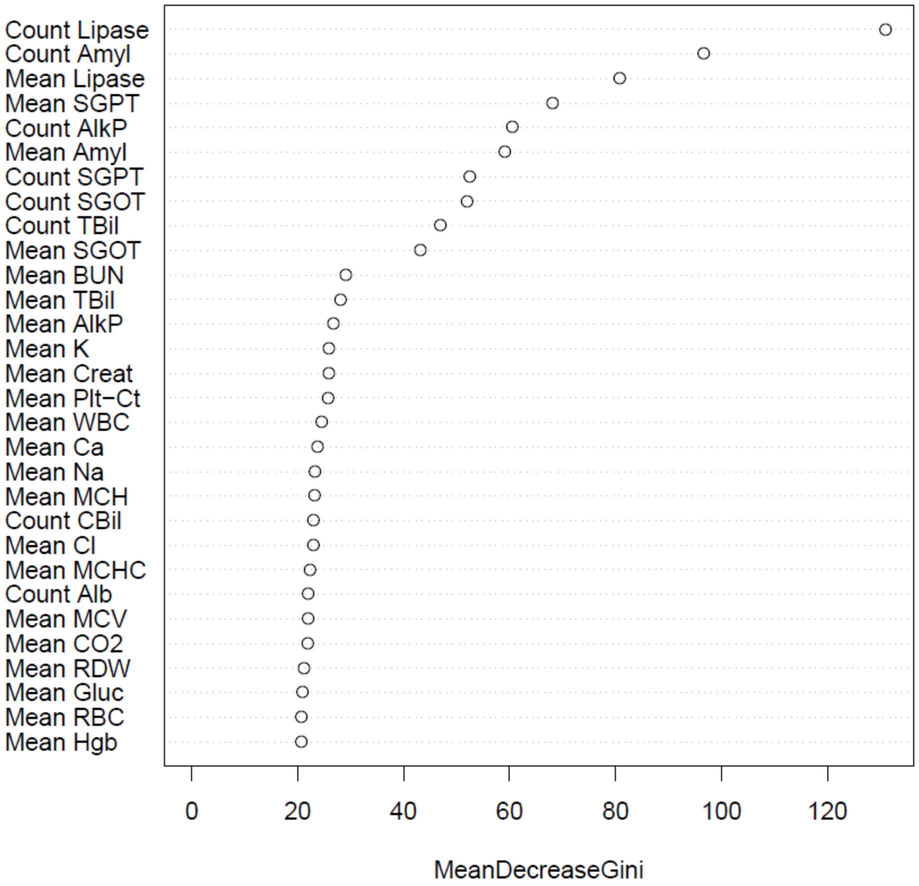
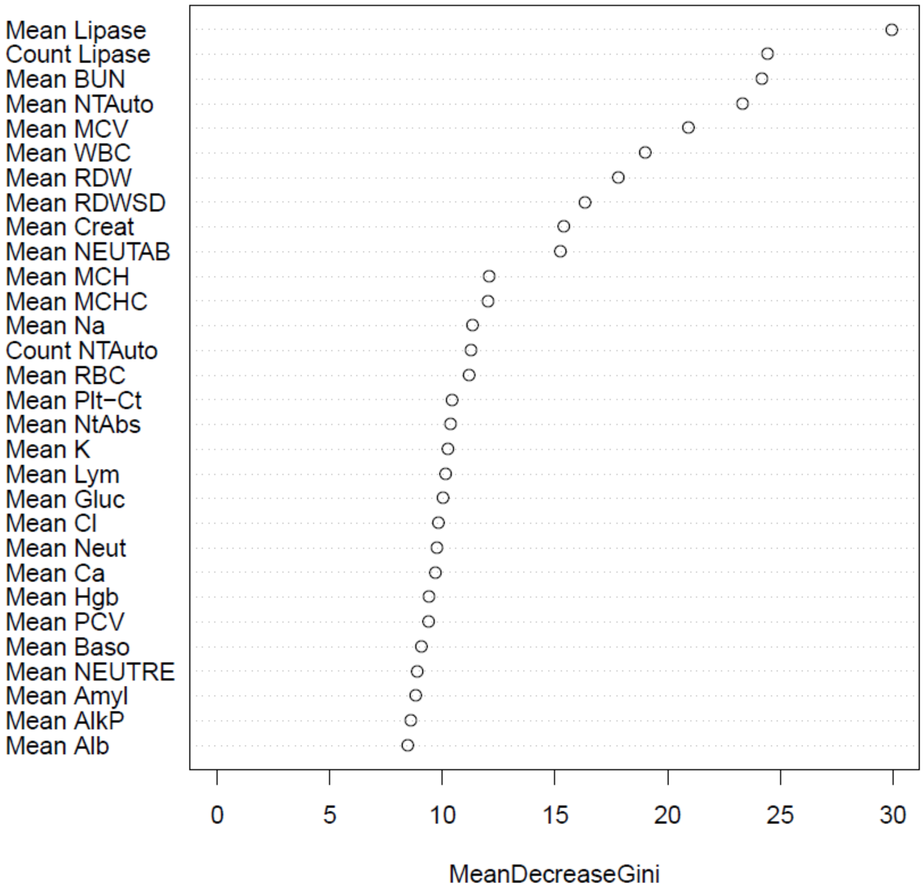


Figure 3.4. Variable importance plot for classifier predicting appendectomy using a combination representation consisting of counts and means of laboratory results.



Binary data representations for sex and race predictions performed significantly worse than other representations for these tasks. These representations contained no information about the values of laboratory results, only the fact that they were ordered. This suggests that orders do not depend on sex or race, but the results themselves do. Certainly, some differences are to be expected because some diseases are more common in men than women or in minorities than white patients. These small differences may be what the random forest is using to differentiate patients on the basis of race or sex.

Different data abstraction models caused the classifiers to focus on different variables. Representations that rely on counts of orders may be identifying features of clinical practice related to a particular outcome of interest; physicians will have specific patterns of ordering laboratory tests. It does not necessarily follow, however, that these clinical patterns are representative of the underlying physiology of the patient [97]. For instance, orders for blood levels of tacrolimus may have little to do with a patient's physiology, and more to do with making sure that the levels of tacrolimus remain in the therapeutic range; just counting the orders for this test would give a strong indication that the patient is in fact a transplant recipient. Representations that included information about the laboratory results, on the other hand, were likely picking up both information about the physical state of the patient and information about the practice pattern of the physician, through which the physiologic state can be altered.

While both the tasks of identifying patients with kidney transplant and patients with liver transplant separately were apparently simple tasks, it was surprising that the conditional task of determining which transplant had occurred given that it was one of these two was itself also a fairly simple proposition. For each task individually, either the presence or the results of tacrolimus level tests were discriminating features. But because both transplants require the use of tacrolimus, this variable was not as important when differentiating between procedures. Biological analytes related to the disease processes underlying the need for transplant were more important; for kidney transplant, these were such kidney-related entities as creatinine, blood urea nitrogen and phosphate, while for liver transplant the pertinent variables were alanine aminotransferase, aspartate aminotransferase, and alkaline phosphatase.

In the tasks of identifying records with knee replacement and with hip replacement, performance was good at even the lowest-complexity binary representation. However, when trying to identify which type of joint surgery had occurred in the record, performance dropped.

The variables that the random forests found to be most important for each individual task were very similar and non-specific: clotting test results such as prothrombin time, partial thromboplastin time, and international normalized ratio. When attempting to identify which type of joint surgery had occurred, these non-specific markers were no longer as useful, and the lack of any other strong predictors did not allow for high accuracy on that conditional classification problem.

To better understand the performance of the predictive models, I examined examples of records that were misclassified with high confidence, i.e., the predicted probability of the label was high but wrong. I believed this might provide some clues as to what was driving misclassifications. Interestingly, nineteen of the twenty records with the highest predicted probability of having a kidney transplant but labeled as a control turned out to be correctly classified by the algorithm, and misclassified by the CPT codes used as a gold standard label. This finding demonstrates that relying on high specificity markers of phenotypes is not without risks, as noise in that single value can corrupt the ability to identify records with the finding of interest. However, this was partially ameliorated by using the non-specific, diffuse information spread throughout the laboratory test results.

The timing of orders for laboratories appears to be less important than whether the order was placed at all. Counts per year and year-over-year cumulative counts only performed as well as total counts, not better. In the case of the count abstraction model, this may be due to the nature of the random forest and how variables are selected for inclusion in each tree. If the sum count of orders is the most information-dense representation, then a random forest classifier would need to select many individual variables from a representation of counts binned by year to encode the same data. As evidence for this hypothesis, the most common pattern of variable importance in counts by year and cumulative counts by year representations was that the most recent entry of counts per year and the most recent entry of year-over-year cumulative counts (which is equal to the sum of all counts over the eight years) were selected as most important. The cumulative abstraction model allowed the random forest classifiers access to total counts of laboratory orders, as well as intermediate counts. That the classifiers chose not to use these intermediate results is evidence that the distribution of counts over time was much less important than the total number of counts.

There are some limitations of this study. As with any research using EHR data, errors may have affected the performance of my classifiers. Extreme physiologic outliers of results and missing or miscoded entries were neither adjusted nor excluded. While this may have decreased accuracy of some models, the effect is likely negligible given the sheer volume of data.

While this work provides proof of concept that unselected, non-specific evidence from an EHR can be used to identify patients with specific conditions, future work in this area could make use of more data types to provide improved pattern recognition and discovery. Incorporating features medication orders, demographic information, and the output of natural language processing will likely improve the performance of such approaches.

CHAPTER IV

QUERYING DISEASES AGAINST EXACT LABORATORY COMBINATIONS USING CONTINUOUS DATA ABSTRACTION MODELS

Introduction

The irregular and asynchronous nature of medical data present challenges for using health records to identify relationships between clinical findings and the complex phenotypes with which patients may present[55]. As mentioned in Chapter II, information is entered into the patient chart as needed for clinical care, meaning there is no regular frequency at which data is recorded. Additionally, the choice of which data elements are collected is largely based on clinical decisions; with only a small subset of potential events measured simultaneously, determining which diseases are present at the same time as particular findings remains difficult. Addressing these issues requires significant decision making on the part of the medical researcher, such as how to handle missing data and how densely to bin the data for analysis. These decisions can have a large impact on the algorithm's performance [98].

In this work, I begin to investigate the utility of modeling clinical data as a means of addressing current limitations to using this data as substrates for statistical and machine learning algorithms. Specifically, I explore the use of inferred longitudinal functions of laboratory data and PheWAS [70] diagnosis codes for the purpose of querying of diagnosis codes against exact values for specific sets of laboratory results, or target, via correlations between a similarity metric between records and the target. After demonstrating face validity of this approach through univariate correlational analysis, I also show that accurately predicting the similarity score from a linear combination of diagnosis codes is achievable through linear regression.

Background

In these experiments I modeled laboratory and billing code data using two different interpolation techniques. I also explored methods of identifying associations between laboratory findings and PheWAS codes for diseases. Below, I provide some background on the data models and association measures explored.

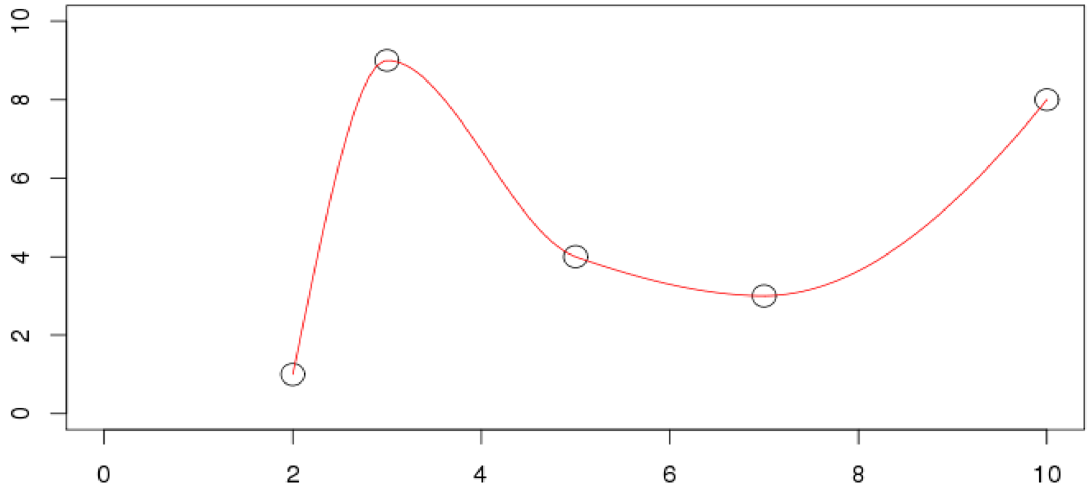
Data models

I modeled the laboratory results and billing code data as being generated from continuous functions. I used two interpolation techniques, piecewise cubic Hermite interpolation polynomials and continuous intensity curves, to generate estimates of these underlying functions given the observed data.

Piecewise cubic Hermite interpolation polynomials

While several methods of interpolation are in wide use [99,100], I chose to use piecewise cubic Hermite interpolation polynomials (pchip), a shape-preserving, smooth interpolation where the slope is calculated such that the values of the function do not locally overshoot the known function values [99]. Figure 4.1 shows a sample pchip interpolation.

Figure 4.1. Piecewise cubic Hermite interpolation polynomial applied to example data.



Continuous intensity curves

While pchip can efficiently interpolate functions with real-valued dimensions, transforming events which are either present or absent is a different task. Gaussian processes can be used to infer the intensity function of a sequence of events, but this is computationally demanding and time-intensive [101]. A faster alternative uses an approximation based on *k*-nearest neighbor density estimation, which I use in this work[102].

Similarity score

Attempting to find associations between specific combinations of laboratory findings requires a method to compare two entities, each possibly containing multiple values, into a single numeric summary. I selected the measure $s(x, x') = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|d_i|+1}$, where $d_i(x, x') = |x_i - x_i'|$, where i indicates an entry in x and a corresponding entry in x'. Subsequently, $s_i \in [0,1]$; records that perfectly match the target are assigned a similarity score of 1, and records that are perfect mismatches (meaning $d_i = \pm\infty$) are assigned a score of 0. Applying this method to laboratory values that have been standardized, as mine have been, gives the additional interpretation that a score of 0.5 (a half-match to the target) is equivalent to a record where $\sum d_i = 1$ (one standard deviation off).

When using this similarity measure for targets with multiple laboratory values, it is noteworthy that there are many different ways to achieve the same similarity result. For instance, a record that is within one standard deviation of the target in both laboratory results of a two-lab target would get a similarity score of 0.5. This is the same score that would be achieved by a record that is a perfect match on one of the laboratory values and a perfect mismatch on the other. This result stems from the summation occurring outside the fraction when calculating the similarity score.

Measures of association

In these experiments, I chose to explore two methods of assessing association between variables. I used correlation to measure the strength of univariate association between similarity measures and intensity of PheWAS codes [73]. In addition, I built penalized linear regression models to explore these associations while adjusting for associations between the similarity measures and other PheWAS codes [103].

Correlation

Correlation is a standard statistical tool for measuring the strength of association between two variables. One popular way of calculating correlation is Spearman's ρ[104]. To calculate Spearman's ρ, each instance in the data set is ranked from lowest to highest value. Spearman's ρ is then calculated by $\rho = \frac{cov(r_X,r_Y)}{\sigma_{r_X}\sigma_{r_Y}}$, where σ_{r_X} and $\sigma_{r_Y}$are the standard deviations for the rank variables and $cov(r_X, r_Y)$ is the covariance of the rank variables. Spearman's $\rho \in [-1,1]$; the closer the value of ρ is to ±1, the stronger the association between the two variables. If one of the variables tends to increase as the other increases, Spearman's ρ will be positive; if one variable decreases as the other increases, ρ will be negative. A Spearman correlation of 0 means that there is no relationship between the two variables

Unlike Pearson's product moment correlation (another measure of association) [105], Spearman's ρ is able to identify non-linear relationships because it uses the ranked values of the variables instead of the raw data. Furthermore, Spearman's ρ is more resistant to extreme values of variables, as the influence of any instance is limited to the value of its rank [106].

Linear regression

While correlation coefficients are an appropriate method of measuring the association between two variables, they do not adjust for other associated variables. To this end, regression models may be better suited.

Linear regression is a widely known and used approach for predicting one outcome from several simultaneously observed input variables. Linear regression predicts outcome y from input variables X using $\hat{y} = \hat{\beta}_0 + x_1\hat{\beta}_1 + \cdots + x_m\hat{\beta}_m$, where the $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \dots, \hat{\beta}_m)$ are estimated coefficients for each input variable that minimizes some measure of error between the predicted outcome $\hat{y}$ and the observed outcome y of interest. Optimizing the fit of the regression model is equivalent to solving the problem $\min_{\beta_0,\beta} \frac{1}{N} \sum_{i=1}^{N} w_i \, l(y_i, \beta_0 + \beta^T x_i)$.

However, standard linear regression techniques such as ordinary least squares do not perform well in terms of generalizability beyond the training set [103]. They can also fail to provide simple and interpretable models [107]. Penalized regression is a means of improving model performance and interpretation [103].

There are two main flavors of regression penalties: the L_2, or ridge regression penalty, and the L_1, or lasso penalty. The ridge regression penalty applies an L_2 bound to a regression model, which serves to continuously shrink the coefficients by placing a penalty on the sum of squared coefficients [103]. As a result of this coefficient shrinkage, which shrinks the variance of the estimates, the regression model tends to achieve better performance than unpenalized regression. Furthermore, variables with similar effect sizes retain penalized coefficients of similar magnitudes. However, ridge regression does not remove any of the coefficients; in a complex model with many coefficients, interpretation can be challenging.

The lasso penalizes a regression model by imposing the L_1 penalty on the sum of the absolute value of the regression coefficients [108]. It is a continuous shrinkage method, like ridge regression, but it also allows for the coefficients of the model to be driven to zero if the penalty is high enough. As a result, the lasso can be used for automatic feature selection through effectively setting the regression coefficients for irrelevant variables to zero. However, the lasso has its own set of caveats: if there are several highly correlated variables in the model, the lasso tends to select only one of the variables and remove all the others.

The elastic net is a regression model that is a weighted average of the lasso and ridge penalties [103]. This regression modeling strategy allows for automatic feature selection through lasso's sparsity induction, but does not have the limitation that only one variable out of several correlated features be kept. The elastic net fits a generalized model via penalized maximum likelihood, solving the problem $\min_{\beta_0,\beta} \frac{1}{N}\sum_{i=1}^{N} w_i\, l(y_i, \beta_0 + \beta^T x_i) + \lambda[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1]$, where $l(y, \eta)$ is the negative log-likelihood for observation $i.$ Notice this is similar to the objective function for standard linear regression, except that there are now two terms representing the ridge $||\beta||_2^2$ and lasso $||\beta||_1$ penalties. The mixing parameter $\alpha \in [0,1]$ controls the ratio of lasso to ridge penalty for a given model; α=1 is a pure lasso penalty, and α= 0 is pure ridge regression.

Typically, building an elastic net involves tuning λ, typically via cross-validation, to determine the optimum penalty for minimizing mean squared error [109]. When describing the model, it is common to report the model coefficients that are maintained at the largest λ where the cross-validated mean squared error is within one standard error of the minimum cross-validated mean squared error the λ_{1se}. Practically, this represents selecting a model that is essentially indistinguishable from the best-performing model in terms of mean squared error, while decreasing the risk that the model overfits to the data. I follow this approach in my experiments.

Methods

In my experiments, I used abstraction models of clinical data to determine univariate association measures and build regression models over values from the models of the data, instead of over the data itself.

I began with the same cohort of 325,461 records used in Chapter III. Members of the lab generated smooth interpolations by applying pchip to the standardized laboratory values at a resolution of 1000 total points over the eight-year period, or roughly one interpolated value every three days. I extrapolated values for each laboratory result outside of the first and last recorded value using the record-specific median. If a record did not have an instance of a particular laboratory test, I used the population mean for the entire length of the record.

I generated continuous intensity curve representations of ICD9 diagnosis codes represented at the highest level PheWAS diagnosis codes used in prior studies [70]. If a record did not have three or more entries for a particular PheWAS code, no curve was generated for that code-record combination. The intensity function was inferred for each highest-level PheWAS code for the most recent eight years of each patient record, with the initial years containing zero events if the record is shorter than eight years These intensities were computed with one point per day resolution, and then reduced by max pooling to 1000 points over eight years. As a result, the intensity curves and the continuous lab value interpolations were aligned to cover the same eight-year period per patient.

In order avoid handling collinearity within records while still using data from as many records as possible given computational constraints, I selected one cross section of laboratory results and PheWAS codes from each record. I selected this cross section uniformly and randomly from the section of curves between the first and last PheWAS code for each record. We excluded records for which there were no PheWAS codes, leaving 288,966 records from which we sampled cross sections to perform the association analysis.

Testing the approach

To explore whether using the data models would allow identification of known associations, I identified clinical targets with strong relationships based on clinical knowledge and expert recommendation. Using these target laboratory values, I calculated the similarity measure for each record and measured the correlation between these similarity scores and the intensity values for each of the high level PheWAS codes. I queried against single laboratory targets with strong known associations, as well as multiple distinct values for single laboratory targets where the value was known to determine the associated phenotypes. Based on early experiments, I selected a correlation threshold of 0.1 above which the majority of associations appeared correct. However, for some queries, no associations were correlated above 0.1. In my results, I report at least the top three correlated PheWAS diagnosis codes, as well as all PheWAS codes with correlations above 0.1. To assess the face validity of the resultant correlations, I employed clinical knowledge and non-exhaustive searches of the medical literature.

To investigate whether my method could identify clinical guidelines as well as biological associations, I turned my attention to measured blood levels of tacrolimus and cyclosporine, two commonly-measured anti-rejection medications given to transplant patients. I calculated correlation coefficients between the major transplant types and several levels of these drugs. Specific organ transplant surgeries should be more highly correlated with the laboratory values of tacrolimus and cyclosporine when the blood levels of these drugs are in the therapeutic range. Most patients who are taking these drugs will have had a transplant, and I hypothesized that I would be able to reasonably identify the transplanted organ based on the blood levels of these drugs. Even so, one difficulty for this task is that after most transplants, patients tend to continuously decrease their doses of these immunosuppressants, which can lead to a very wide therapeutic target when not considering time since operation [110,111].

To explore whether the associations identified would be affected by simultaneously considering other associations, I built penalized regression models using elastic net, predicting the similarity scores using the available PheWAS codes. After an initial grid search to optimize the α parameter, I determined that the relationship between α and the estimation error achieved by the models was very gradual. I therefore elected to use $\alpha = 0.5$ for my mixing parameter. The models were trained using 10-fold cross validation to determine the optimal penalty setting. Models were built in the R programming environment using the package glmnet [103,112].

For the penalized regression models, I evaluated the results qualitatively to see if the remaining regression terms had overlap with the PheWAS codes identified as being the most highly correlated with the similarity to the target value. I also quantitatively assessed the fit of the models by mean squared error, calculated as the average MSE over the ten cross-validation folds at the λ_{1se}.

Exploring the Data

To explore the data in our population, I targeted multiple laboratory results simultaneously to identify correlated findings by including additional results with the single laboratory targets. My hypothesis was that additional data elements would induce a new set of correlations between PheWAS codes and lab targets, and that some of these may be unexpected and novel. For any such query, I required that the record have data for at least one of the target laboratory measurements, but did not require that the record have a PheWAS code. I considered the absence of a PheWAS code informative in terms of diagnoses assigned to the record; while the absence of a laboratory result is informative of clinical practice and decision making, the absence of a PheWAS code does not theoretically contribute to information regarding the similarity of a record to the target of interest.

For these experiments, I chose to look at two use cases. First, I explored how the method would handle an abnormally high blood glucose measurement in the context of a normal hemoglobin A1C. Hemoglobin A1C is a measure of long-term glucose control, so a normal value would imply that the patient in question would have consistently had well-controlled blood glucose levels, despite the fact that their current glucose level is very high [113].

I also explored whether combinations of low packed cell volume (PCV), red cell distribution width (RDW), and mean corpuscular volume (MCV) could be used to identify known and novel relationships with different classes of anemia. This is a clinically relevant question, as RDW and MCV are often used in combination to classify anemias and to suggest potential etiologies [114,115].

Results

These results are in no way exhaustive of the potential findings, but instead should be considered as examples of the types of queries that could be asked of electronic health data using this approach.

Testing the approach

Extreme values for single laboratory values with known univariate associations were detected by my method (Table 4.1). Known associations with values of mean corpuscular volume from very low to very high were also detected by my method (Table 4.2).

I also demonstrate graphically the relationship between the laboratory results for tacrolimus and cyclosporine, two anti-rejection medications that must be tested in transplanted patients, and the main transplants associated with these drug levels (Figure 4.2).

Penalized regression predicting these single laboratory targets produced models with variable numbers of non-zero coefficients (Table 4.3, Appendices C, D). Of note, several of the models found suitable fits in terms of mean squared error by setting all coefficients to zero, equivalent to estimating the population mean similarity score for all instances. Regardless of the number of coefficients retained by the model, the fit as determined by the cross-validated mean squared error (Table 4.3).

Table 4.1. Top correlated PheWAS codes for selected single-laboratory targets. MSE = cross-validated mean squared error.

Analyte	Target (Normal)	PheWAS Code Description	Correlation
Glucose	450 mg/dL	Diabetes mellitus	0.3573
	(70-100)	Hypertension	0.1309
		Ischemic heart disease	0.1301
Creatinine	5.9 mg/dL	Renal failure	0.3290
	(0.70-1.50)	Hypertension	0.2875
		Ischemic heart disease	0.2559
		Disorders of lipoid metabolism	0.2152
		Congestive heart failure, nonhypertensive	0.1782
		Diabetes mellitus	0.1628
		Disorders of the kidney & ureters	0.1463
		Cardiac dysrhythmias	0.1441
		Gout and other crystal arthropathies	0.1331
		Cardiac conduction disorders	0.1255
		Cancer of kidney and urinary organs	0.1177
		Nonspecific chest pain	0.1175
		Cardiomyopathy	0.1165
		Kidney replaced by transplant	0.1109
		Hyperplasia of prostate	0.1073
		Heart valve disorders	0.1047

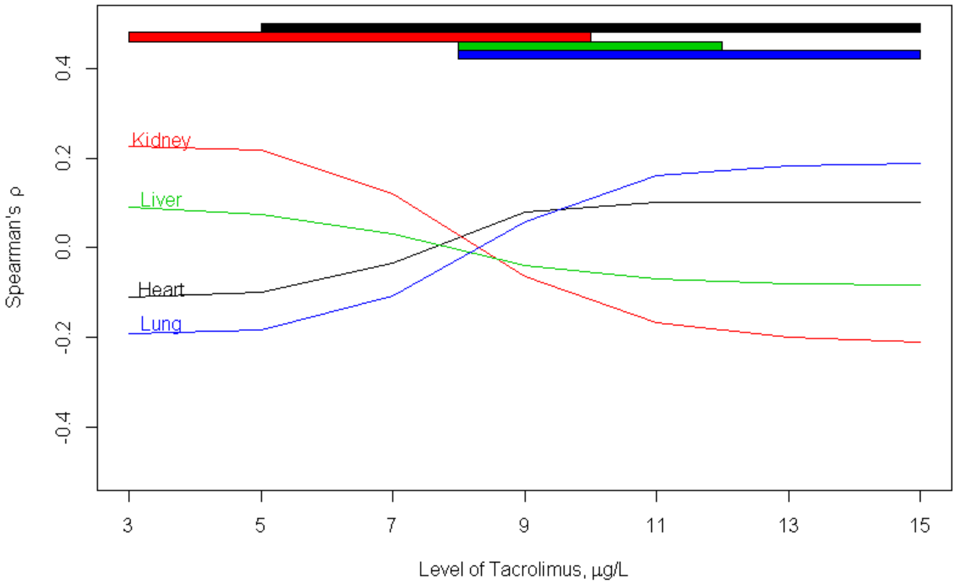
Troponin I	0.8 ng/mL	Renal failure	0.1481
	(<=0.03)	Congestive heart failure, nonhypertensive	0.1350
		Respiratory failure; insufficiency; arrest	0.1321
		Pleurisy	0.1137
		Cardiomegaly	0.1098
		Fluid, electrolyte, & acid-base balance disorders	0.1051
		Cardiac dysrhythmias	0.1036
Troponin I	50 ng/mL	Ischemic heart disease	0.2760
	<=0.03	Congestive heart failure, nonhypertensive	0.1658
		Respiratory failure; insufficiency; arrest	0.1371
		Renal failure	0.1174
		Cardiomyopathy	0.1098
		Shock	0.1080
		Pleurisy	0.1063
		Cardiomegaly	0.1007
		Abnormal serum enzyme levels	0.1004
Lipase	1200 U/L	Diseases of pancreas	0.1311
	(10-60)	Chronic liver disease and cirrhosis	0.0827
		Alcohol-related disorders	0.0766
Cholesterol	500 mg/dL	Menopausal and postmenopausal disorders	0.0898

	(115-200)	Osteoporosis, osteopenia, & pathological fractures	0.0688
		Abnormal findings on mammogram or breast exam	0.0674
Vitamin B12	50 pg/mL	Vitamin deficiency	0.0478
	(180-1000)	Known or suspected fetal abnormality	0.0456
		Other conditions of the mother complicating pregnancy	0.0417
Vitamin B12	1500 pg/mL	Chronic liver disease and cirrhosis	0.0803
	(180-1000)	Fluid, electrolyte, & acid-base balance disorders	0.0792
		Other anemias	0.0785
PCV	30%	Other anemias	0.2179
	(35-45)	Respiratory failure; insufficiency; arrest	0.1441
		Fluid, electrolyte, & acid-base balance disorders	0.1424
		Fever of unknown origin	0.1296
		Pulmonary collapse; interstitial/compensatory emphysema	0.1219
		Protein-calorie malnutrition	0.1188
		Renal failure	0.1159
		Pleurisy	0.1149
		Bacterial infection NOS	0.1136
		Septicemia	0.1116
		Pneumonia	0.1003

Table 4.2. Top correlated PheWAS codes at varying levels of mean corpuscular volume (MCV) using Spearman's correlation.

Value (fL) (normal 80-100)	PheWAS Code Description	Correlation
60 (low)	Lack of normal physiological development	0.0972
	Known or suspected fetal abnormality	0.0670
	Iron deficiency anemias	0.0663
75 (slightly low)	Lack of normal physiological development	0.0972
	Known or suspected fetal abnormality	0.0662
	Acute upper respiratory infections	0.0628
90 (normal)	Disorders of lipoid metabolism	0.0888
	Hypertension	0.0610
	Pain in joint	0.0553
105 (slightly high)	Other perinatal conditions	0.1271
	Short gestation; low birth weight; and fetal growth retardation	0.1141
	Alcohol-related disorders	0.0748
120 (high)	Other perinatal conditions	0.1430
	Short gestation; low birth weight; and fetal growth retardation	0.1342
	Alcohol-related disorders	0.0744

Figure 4.2. Spearman's ρ for different transplant procedures at seven different blood levels of tacrolimus and cyclosporine. Bars across the top of plots show the desired blood levels of each drug to achieve therapeutic benefit.



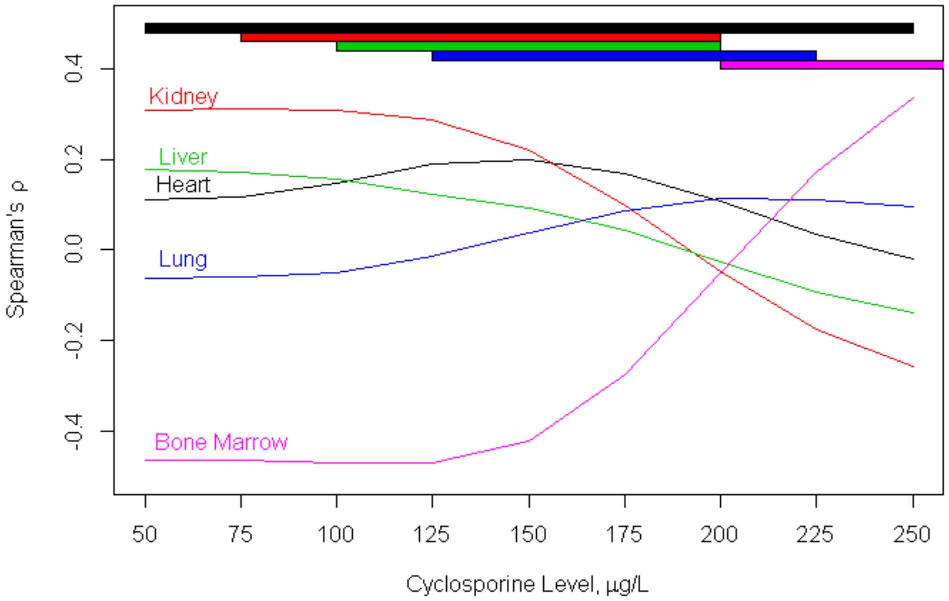


Table 4.3. Coefficients maintained in elastic net models at cross-validatedλ_{1se}, and cross validated model mean squared error (MSE).

Analyte	Target	PheWAS Code Description	β
(Model MSE)	(Normal)		
Glucose	450 mg/dL	None	-
(0.0017)	(70-100)		
Creatinine	5.9 mg/dL	Renal failure	0.4158
(0.0018)	(0.70-1.50)	Hypertension	0.0312
		Ischemic heart disease	0.0269
		Diabetes mellitus	0.0131
		Congestive heart failure, nonhypertensive	0.0066
		Respiratory failure; insufficiency; arrest	-0.0045
		Short gestation; low birth weight; and fetal growth retardation	-0.0099
		Other perinatal conditions	- 0.0209
		Cardiac & circulatory congenital anomalies	-0.0227
Troponin I	0.8 ng/mL	Ischemic heart disease	-0.0526
(0.0143)	(<=0.03)		
Troponin I	50 ng/mL	None	-
(0.0025)	(<=0.03)		
Lipase	1200 U/L	None	-
(0.0011)			

Cholesterol	500 mg/dL	None	-
(0.0006)	(115-200)		
Vitamin B12	50 pg/mL	Mood disorders	0.0355
(0.0195)	(180-1000)	Substance addiction and disorders	0.0074
		Intracranial hemorrhage	0.0057
		Anxiety, phobic & dissociative disorders	0.0037
		Cerebrovascular disease	0.0036
		For remaining coefficients, see Appendix C.	
Vitamin B12	1500 pg/mL	None	-
(0.0118)	(180-1000)		
PCV	30%	Cancer of other female genital organs	0.8287
(0.0298)	(35-45)	Chemotherapy	0.7697
		Cancer of kidney and urinary organs	0.6907
		Cancer of bone & connective tissue	0.6682
		Known or suspected fetal abnormality	0.6182
		Early or threatened labor	0.6060
		For remaining coefficients, see Appendix D.	

Exploring the Data

By including a normal measure of hemoglobin A1C along with the elevated glucose result, my method was able to identify a correlation with the diagnosis code for abnormal glucose measurements that was weaker when the only information available was an elevated glucose. Diabetes mellitus, the most highly correlated diagnosis code without information on hemoglobin A1C, no longer breaks the correlation threshold of 0.1 (Table 4.4).

Unlike the penalized regression model predicting the similarity of records to the target of just high glucose, the elastic net model of the target containing both a high glucose and normal hemoglobin A1C retained 176 correlation coefficients greater than zero at the λ_{1se}(Table 4.5, Appendix E).

Using PCV, RDW and MCV, I was able to identify some known associations between these combinations and known anemia phenotypes. However, many of the correlation coefficients were below my threshold of 0.1 (Table 4.6)

Table 4.4. Inclusion of a normal hemoglobin A1C induces a different set of observed correlations.

Analytes	Target (Normal)	PheWAS Code Description	Correlation
Glucose	450 mg/dL	Diabetes mellitus	0.3573
	(70-100)	Hypertension	0.1309
		Ischemic heart disease	0.1301
Glucose, HbA1C	450 mg/dL, 5.5%	Abnormal glucose	0.1366
	(70-100; 4.0-6.5)	Hypertension	0.1272
		Ischemic heart disease	0.1018

Table 4.5. Coefficients maintained in elastic net model of high glucose and normal hemoglobin A1C at cross-validated λ_{1se}.

Analytes (Model MSE)	Target (Normal)	PheWAS Code Description	β
Glucose, HbA1C	450 mg/dL, 5.5%	Gestational diabetes	0.2115
(0.0031)	(70-100; 4.0-6.5)	Abnormal glucose	0.2507
		Disorders of lipoid metabolism	0.1759
		Heart valve disorders	0.1532
		Sleep disorders	0.1071
		Overweight	0.1051
		For remaining coefficients, see Appendix E.	

Table 4.6. Top correlated PheWAS codes at varying levels of packed cell volume (PCV), red cell distribution width (RDW), and mean corpuscular volume (MCV). A low PCV is indicative of an anemia. RDW and MCV together are often used to classify anemias and point to particular etiologies.

Analytes	Target (Normal)	PheWAS Code Description	Correlation
PCV, RDW	30%, 13%	Known or suspected fetal abnormality	0.1033
	(35-45; 11.5-14.5)	Early or threatened labor	0.0857
		Other conditions of the mother complicating pregnancy	0.0714
PCV, RDW, MCV	30%, 13%, 60 fL	Known or suspected fetal abnormality	0.1138
	(35-45; 11.5-14.5; 80-100)	Early or threatened labor	0.0947
		Other conditions of the mother complicating pregnancy	0.0818
PCV, RDW, MCV	30%, 13%, 75 fL	Known or suspected fetal abnormality	0.1178
	(35-45; 11.5-14.5; 80-100)	Early or threatened labor	0.0987
		Other conditions of the mother complicating pregnancy	0.0896
PCV, RDW, MCV	30%, 13%, 90 fL	Known or suspected fetal abnormality	0.0757
	(35-45; 11.5-14.5; 80-100)	Early or threatened labor	0.0619
		Fracture of the vertebral column without mention of spinal cord injury	0.0469
PCV, RDW, MCV	30%, 13%, 105 fL	Known or suspected fetal abnormality	0.0673
	(35-45; 11.5-14.5; 80-100)	Fracture of the vertebral column without mention of spinal cord injury	0.0561
		Early or threatened labor	0.0556

PCV, RDW, MCV	30%, 13%, 120 fL (35-45; 11.5-14.5; 80-100)	Known or suspected fetal abnormality	0.0916
		Early or threatened labor	0.0758
		Fracture of the vertebral column without mention of spinal cord injury	0.0609
PCV, RDW	30%, 17%	Other anemias	0.2407
	(35-45; 11.5-14.5)	Respiratory failure; insufficiency; arrest	0.1822
		Fluid, electrolyte, & acid-base balance disorders	0.1797
PCV, RDW, MCV	30%, 17%, 60 fL (35-45; 11.5-14.5; 80-100)	Other anemias	0.2351
		Respiratory failure; insufficiency; arrest	0.1751
		Fluid, electrolyte, & acid-base balance disorders	0.1734
PCV, RDW, MCV	30%, 17%, 75 fL (35-45; 11.5-14.5; 80-100)	Other anemias	0.2126
		Respiratory failure; insufficiency; arrest	0.1582
		Fluid, electrolyte, & acid-base balance disorders	0.1571
PCV, RDW, MCV	30%, 17%, 90 fL (35-45; 11.5-14.5; 80-100)	Other anemias	0.1683
		Respiratory failure; insufficiency; arrest	0.1453
		Fluid, electrolyte, & acid-base balance disorders	0.1249
PCV, RDW, MCV	30%, 17%, 105 fL (35-45; 11.5-14.5; 80-100)	Other anemias	0.2332
		Respiratory failure; insufficiency; arrest	0.1808
		Fluid, electrolyte, & acid-base balance disorders	0.1792

PCV, RDW, MCV	30%, 17%, 120 fL	Other anemias	0.2411
	(35-45; 11.5-14.5; 80-100)	Respiratory failure; insufficiency; arrest	0.1836
		Fluid, electrolyte, & acid-base balance disorders	0.1817

Discussion

Using continuous data representations, I was able to recover known associations between combinations of laboratory results and phenotypes of interest. Using penalized regression, I demonstrated that it is possible to use linear combinations of PheWAS codes to accurately predict specific values of multiple laboratory tests simultaneously. I was able to abstract away some of the difficulties in modeling electronic health data that arise from irregularity and asynchrony using continuous, longitudinal transformations of the data.

Testing the approach

In the single laboratory value correlation studies, the most positively correlated PheWAS codes have face validity for known associations. Elevated glucose, for example, is a defining feature of diabetes mellitus. The other top hits, hypertension and ischemic heart disease, are common comorbidities of diabetes [116]. Hypertension and ischemic heart disease are also known to be associated with renal failure, the primary cause of elevated creatinine [117,118]. Lipase elevated to ten-times the upper limit of normal is strongly correlated with diseases of the pancreas [119], but also chronic liver disease and alcohol abuse [120].

While the correlations identified for troponin and cholesterol may not be intuitively correct, a review of the literature suggests that they may be valid findings. Although elevated troponin is most often considered in the context of acute myocardial infarction, it is also associated with renal failure, congestive heart failure, and pulmonary embolism [121]. Menopause is known to increase cholesterol levels in women [122] and studies suggest there may be links between high cholesterol and both osteoporosis and breast cancer [123,124].

It is clear from the example of mean corpuscular volume that different levels of a laboratory result are associated with different phenotypes. While it is widely known that a common cause of microcytosis is iron deficiency and a common cause of macrocytosis is alcoholism, exploring an association between these two phenotypes without the ability to target specific laboratory values would have required two models; one for the association between MCV and alcoholism, and one for the association between MCV and iron deficiency. Here, the one model is able to identify both relationships, dependent only on the specified target lab values. One caveat to these interpretations is that while MCV is largely homogenous in the adult population, it varies significantly across a lifespan, especially in neonates, children, and teenagers. It is possible that some of the signal I detected, such as problems associated with pregnancy or failure to thrive, were driven by one or more of the age groups within the population. As I did not collect the ages of the study population, it is difficult to say this definitively.

Exploring correlations between the intensities of diagnosis codes for transplant surgeries and the blood levels of anti-rejection drugs tacrolimus and cyclosporine, I was able to loosely recover clinical guidelines for the therapeutic drug levels for each surgery [110,111]. However, my method did not perfectly identify the clinical guidelines. One major reason for the discrepancy between my findings and clinical guidelines may be that the therapeutic level for each surgery changes as a function of time since the operation. Calculating the correlations using only cross-sections of the continuous functions, it was impossible for my method to be able to identify that dimension. In spite of this known limitation, my method still was nonetheless able to identify rough regions of therapeutic levels for tacrolimus and cyclosporine.

The penalized regression models allowed for exploration of the relationships between laboratory results and PheWAS codes in the context of other lab-code relationships. Surprisingly, diagnosis codes that were strongly correlated with a laboratory of interest when not accounting for other associations, such as glucose and diabetes or lipase and pancreatitis, were occasionally included in regression models of the same problem. In some other cases, like the slightly elevated creatinine and slightly decreased PCV, significantly more coefficients were included in the regression model than would be expected based on the correlation coefficients.

Perhaps one reason for this discrepancy is the extreme nature of some of the values I selected. Using PheWAS diagnosis codes for diabetes and pancreatitis to predict elevated glucose and lipase levels may not have been included in the models because such extreme cases made up a very small percentage of the populations. Conversely, models predicting the slightly elevated creatinine and slightly decreased PCV from PheWAS diagnosis codes maintained a significant number of predictors in the models. This could be because there are more than just one or two diagnosis codes that are necessary in order to predict these values. In other words, there may be more than one etiology for these lab abnormalities.

Exploring the Data

The ability of this method to handle targets with more than one laboratory value is one of its most promising features. As demonstrated by the example of combining elevated glucose and normal glycosylated hemoglobin, adding additional constraints on the laboratory target can drastically change which PheWAS codes are found to be correlated. Unconstrained by any other information, a glucose measurement of 450 mg/dL would strongly suggest a diabetic patient, potentially one in an acute exacerbation. However, including the information that their hemoglobin A1C (a measure of long-term glucose control) is normal makes the diagnosis code of chronic diabetes less likely, and increases the correlation to the PheWAS code for an abnormal glucose measurement.

As with elevated creatinine and slightly decreased PCV in the univariate sense, the elastic net model kept the coefficients for many PheWAS codes when predicting the combination of high glucose and normal hemoglobin A1C. The largest coefficient in the model is gestational diabetes, an acute metabolic syndrome that occurs during pregnancy. This result makes sense, as this disease could easily lead to an increased glucose and normal hemoglobin A1C.

Several other diagnosis codes that were kept in this model seem to share a common relationship to acutely elevated glucose; namely, they are either transplant surgeries or conditions that could reasonably lead to transplant surgeries. With these surgeries, patients would be required to take anti-rejection medications, including steroids, which are known to acutely elevate glucose levels.

Using the combined laboratory values for PCV, RDW and MCV provided a less clear result. At RDW levels of 13%, the dominant correlated phenotypes were fetal abnormalities and early labor. When RDW levels were 17%, the correlated phenotypes were anemias, respiratory failure, and acid-base disorders. This is not entirely as one would expect; because we set the value of PCV to 30% across all comparisons, every one of the instances should have returned some indication of anemia. However, the lower RDW values seem to be driving the correlation with pregnancy-related outcomes. Across the range of MCV values, it does not seem that MCV contributes meaningfully to the correlated phenotypes after RDW and PCV are considered. Again, this result could be due to uneven age distributions in this sample, but which would be difficult to determine with the data I collected.

Compared to previous methods of identifying associations between findings and diseases, using continuous data representations allows many advantages. In order to achieve the same type of analysis without a continuous representation, a researcher would have to make at least two decisions about how to use their data. First, they would have to determine how temporally close together two clinical events would have to be in order to be considered simultaneous. Second, they would also need to decide how close in value a laboratory result would need to be to the target to be considered identical. The answer to both of these questions has traditionally been binning of both time and laboratory result variables. However, as noted above, this type of approach is an approximation for the type of analysis I am able to perform using continuous data representations and my similarity measure.

There are some limitations of this approach. As with all exploratory data analysis, it is entirely possible that many of the associations discovered are simply data artifacts. The same analysis could be run in a separate set of clinical records, or even another hospital's record, to determine if the findings replicate. A review of medical literature may be able to show whether there is prior evidence for the correlations I have uncovered. Finally, associations that are identified in this way could serve as hypotheses for designing other experiments to test for replication.

In its current incarnation, it is impossible to determine whether a diagnosis preceded or followed a particular set of lab results. Unfortunately, this removes all possibility of identifying which associated findings may be used as risk factors in prediction or prognosis. It would be possible to overcome this limitation by retaining the entire estimated function for all lab results and PheWAS codes for all patients, which would allow the user to determine how two correlated events are temporally related. However, this would have led to a significant increase in computational demand, as well as required adjustments to the model formulation in order to account for intra-record correlations.

My threshold of 0.1 for flagging correlations as interesting was determined by trial and error on early experiments. It is likely, however, that there are a host of considerations that should go into determining the appropriate correlation cutoff for each query. For instance, values that were more extreme often had higher correlation with known associated diseases. This was not always the case, as extremely high vitamin B12 values did not have any correlation coefficients over my threshold and could be due to the lower rate of vitamin B12 testing among this clinical population. Further work is required to better understand the relationships between strength of association, magnitude of deviation from the population mean, and the prevalence of test orders.

It is also likely that the decision to require at least three PheWAS codes in order to generate a trace washed out some of the correlations that would have been found if I had included traces for these codes. This may explain why the slightly elevated troponin measurement was not highly associated with ischemic heart disease (the PheWAS code which subsumes myocardial infarction), even though this diagnosis is the most likely etiology of an elevated troponin. Perhaps the acute nature of a myocardial infarction, combined with the decision to ignore PheWAS codes with fewer than three entries, limited my ability to find this known association. Even so, ischemic heart disease is the most positively correlated PheWAS code for troponins that are sufficiently elevated.

The granularity of PheWAS available for this work also likely limited the kinds of associations that I was able to identify. In this set, the code for diabetes mellitus subsumes both insulin-dependent and non-insulin dependent forms, as well as the acute event of diabetic ketoacidosis. There is also no PheWAS code in this dictionary for a normal pregnancy. As a result, labs which are elevated in a fair proportion of normal pregnancies may have falsely shown up as associated with complications of pregnancy or congenital problems with the newborn, assuming these complications do not change the underlying pregnancy physiology which elevates those specific labs in the first place. Future work to identify more specific associations will require a more precise vocabulary of diagnosis codes, as well as the inclusion of other types of data, such as medication administrations, vital signs and demographic information.

CHAPTER V

DISCUSSION

The major challenges to making use of health data for identifying more precise phenotypes can be tackled by one of two approaches: 1) developing new methodologies to analyze the irregular, asynchronous nature of the data, or 2) abstracting and transforming the data to be amenable to standard analysis methodologies. In this dissertation I have explored some of the properties of various methods to address these issues, and demonstrated that each may have its place in particular circumstances.

I have shown that, in the case of classifying records by presence or absence of high-specificity procedure codes or demographics, low-complexity abstraction models to alleviate these problems are an efficient method of encoding health data. These data representations also allow for the creation of models that can utilize non-specific, diffuse information spread throughout the health record, and provide classifications with respectable discrimination, calibration and confidence.

Using simple data abstraction models to more accurately identify patients with a phenotype of interest could be a low-cost, simple way to improve the quality of populations used for phenotyping analysis. Such an approach could even be used to impute missing data, which commonly arise because of lack of interoperability between clinical record systems. Such low-cost, simple methods are appealing, and could potentially have large returns in terms of the usability of clinical data.

While I have demonstrated that simple data representations can be used to accurately identify patients with phenotypes of interest, I have not fully explored using the continuous data representations from Chapter IV in a similar manner. Preliminary results suggest that, at least in the paradigm of a random forest or similar classifier, such a longitudinal representation may not provide additional improvement in discrimination.

One may question why, when continuous representations proved so useful in targeting specific combinations of laboratory results in Chapter IV, they do not greatly outperform simpler methods in predicting high-specificity binary phenotypes. One possible explanation has already been suggested: namely, that the random forest model employed requires too much data to make a complete representation of the problem, and methods that compress the longitudinal record into small dimensional space are more efficient.

Another possible explanation may be that these two tasks are exploring two different types of phenotypes. In Chapter III, I had defined my outcomes of interest and wanted to determine if there was evidence in the record that the event had ever happened. In Chapter IV, I was less interested in whether an event had ever happened, and wanted to see which phenotype codes were associated with particular laboratory results. Because laboratory results can change over time, it made sense to look at diagnosis codes over time as well. Presuming a constant level for the phenotypes of interest throughout the patient's trajectory would have likely dispersed any signal throughout the medical record, making association mining nearly impossible.

I have also demonstrated the utility of using continuous longitudinal data abstraction models of health data, obviating the need for binning time variables when modeling health record data which is captured irregularly and asynchronously. Calculations and models can be built at any time points over the period of interest because of the specification of continuous functions over the input space; all time points have either an observed or estimated value for the entity of interest.

I have shown that these continuous representations, because of their ability to abstract away irregularity and asynchrony, can be used to query against combinations of exact laboratory values.

Unlike previous methods, this allows for the identification of correlations between unique sets of clinical findings and phenotypes of interest.

Querying against specific clinical findings has a clear potential use in clinical decision support. The inspiration for this approach came largely from the use case of a perplexed physician, unsure how to interpret uncommon, confusing combinations of laboratory values. While it would not make sense for a seasoned physician to query against well-known associations, it may prove beneficial to augment their clinical knowledge with information about the most likely reasons for their patients' difficult-to-diagnose complaints or ambiguous test results.

Another potential use of such a method may manifest as decision support for ordering laboratory tests. While my method can currently identify associations, it is imaginable that a modified version of my method could be used to 1) identify the highest probability diagnoses, and 2) identify the laboratory test that has the highest likelihood of differentiating between the most likely diseases, perhaps through estimating the information value of particular tests. The principled use of laboratory tests and medication trials could help to decrease the cost of medical care by decreasing uncertainty, a timely goal given the ever increasing cost of medicine worldwide.

Open Questions

In my work, I selected cross-sections from each of the records, where each cross-section contained the estimated function value for all laboratory results and diagnosis code intensity curves. This was done in order to remove the need to address intra-record correlation. However, given time and computing power, it would be feasible to calculate correlations on not just cross-sections of records, but on the entire records themselves. Similarity measures could be computed somewhat equivalently, the exception being that instead of a single value per record, this measure would yield a function of similarity values for a record over time. Using appropriate transformation approaches such as Fisher's z, it would be possible to combine and average these correlations, thus allowing the use of all records *and* all the data points within a specific record.

One particularly interesting opportunity is the question of whether temporal relations other than simultaneity can be explored using continuous data representations. Were it possible to calculate correlations between similarity measures and PheWAS diagnosis code intensities over an entire record, as I just discussed, then it would also be possible to calculate cross-variance between the function of similarity measures and the PheWAS code intensities. This might allow for identification of clinical laboratory entities that occur either before or after a rise in the intensity of a diagnosis code. Using this type of approach, it is possible that my task of recovering clinical guidelines for anti-rejection medications might be improved.

Gaussian process regression is a method that has been used to quantify the uncertainty around point estimates of a function. Given the time demands of modeling clinical data in this way, I elected to use simpler methods that do not include this uncertainty term and even remove information about when the observed data points occurred. However, it is likely that information about the exact location of observed data points and the estimated uncertainty throughout the function would provide additional uses for the utility of these methods.

Methods utilizing continuous representations of medical data can be applied to more than just structured elements. As mentioned in Chapter II, there are several additional types of medical data, such as images and free text forms. To learn from these types of data, one approach has been to extract features from their structure. With these features extracted, it is entirely possible to model the occurrence of these features using continuous representations, just as I did with structured laboratory results and diagnosis codes. In this way, heterogeneous data sources such as clinical concepts encoded in free text or visual features from radiology images could be seamlessly combined with structured data elements, all in a way that would be immediately computable by machine learning algorithms, allowing researchers the ability to efficiently and automatically perform analysis on large complex medical data sets.

Conclusion

This dissertation demonstrates that it is possible to overcome some of the problems of medical data sparsity, irregularity and asynchrony by modeling clinical data at different levels of abstraction and using samples from those models as substrates to machine learning algorithms. Modeling clinical data using summary measures such as counts or means is an efficient way to encode data, and these representations can be used to build highly discriminative classification models. Modeling clinical data as continuous functions from which samples can be drawn alleviates the complications that arise from the irregular and asynchronous nature of the clinical environment. Samples from these functions can be used as the substrates for standard learning algorithms. The methods I have proposed here show the advantages of modeling medical data by overcoming some of the challenges that hamper wider use of machine learning in medicine.

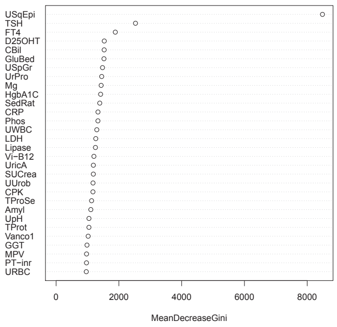
APPENDIX A.

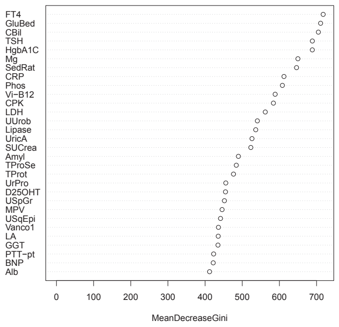
Lists of CPT and ICD-9 codes used to identify records with outcomes of interest. All codes are CPT codes unless marked with a "*".

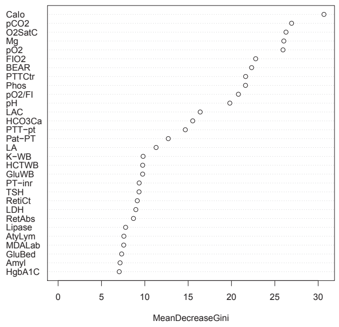
Outcome	Codes	Outcome	Codes
Appendectomy	44950	Hip Replacement	27090
	44955		27091
	44960		27125
	44970		27130
	44979		27132
			27134
Cholecystectomy	47562		27136
	47563		27137
	47564		27138
	47570	Kidney Transplant	50360
	47579		50365
	47600		
	47605	Knee Replacement	27438
	47610		27446
	47612		27447
	47620		27486
			27487
Hemorrhoid Surgery	46083		27488
	46200		
	46220	Pancreatectomy	48140
	46221		48145
	46230		48146
	46250		48148
	46255		48150
	46257		48152
	46258		48153
	46260		
	46261	Splenectomy	38100
	46262		38101
			38102
Liver Transplant	47135		38115
	47136		38120
	*50.51		
	*50.59		
	*v42.7		

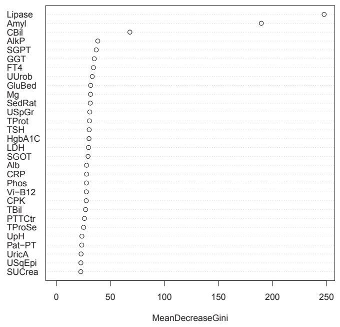
APPENDIX B.

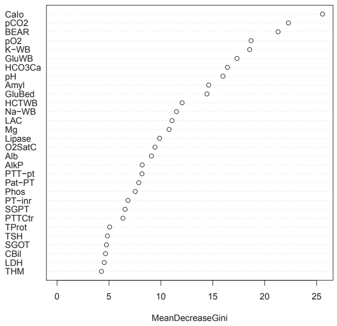
Full set of variable importance plots for seven different representations and thirteen different classification tasks.

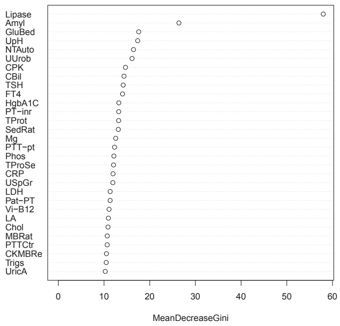

Variable Importance for Binary Abstraction Model Predicting Sex

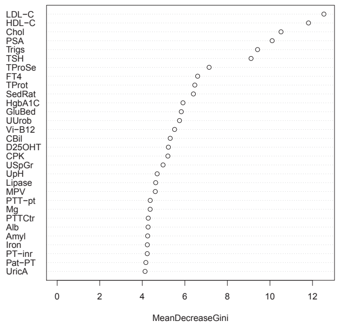

Variable Importance for Binary Abstraction Model Predicting Race

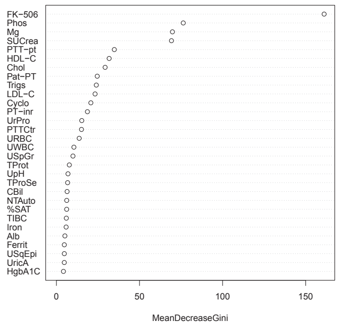

Variable Importance for Binary Abstraction Model Predicting Splenectomy

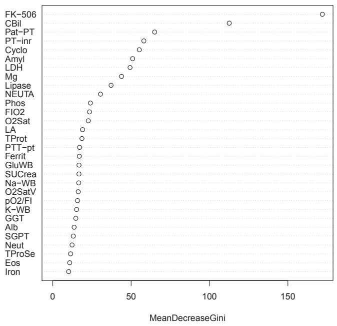

Variable Importance for Binary Abstraction Model Predicting Cholecystectomy

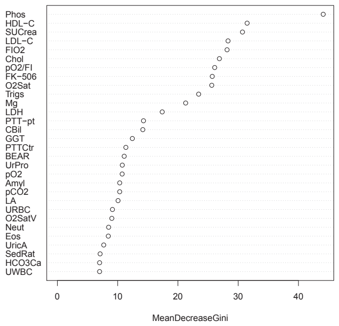

Variable Importance for Binary Abstraction Model Predicting Pancreatectomy

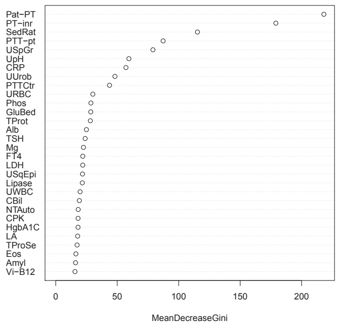

Variable Importance for Binary Abstraction Model Predicting Appendectomy

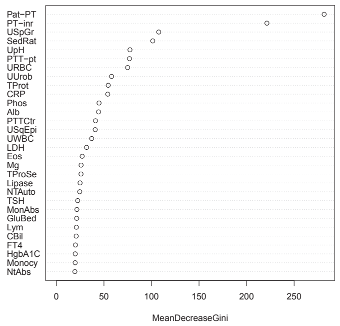

Variable Importance for Binary Abstraction Model Predicting Hemorrhoid Surgery

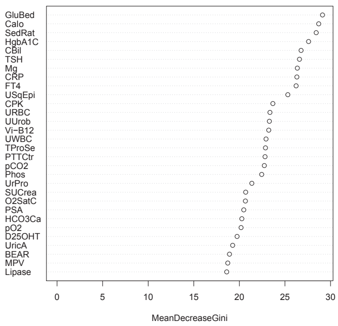

Variable Importance for Binary Abstraction Model Predicting Kidney Transplant


Variable Importance for Binary Abstraction Model Predicting Liver Transplant

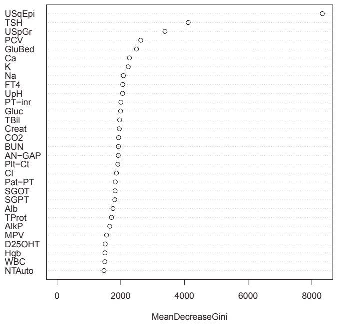

Variable Importance for Binary Abstraction Model Predicting Kidney vs Liver Transplant


Variable Importance for Binary Abstraction Model Predicting Hip Replacement

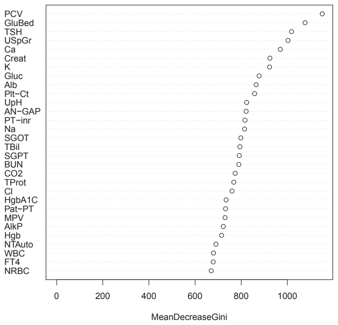

Variable Importance for Binary Abstraction Model Predicting Knee Replacement


Variable Importance for Binary Abstraction Model Predicting Hip vs Knee Replacement

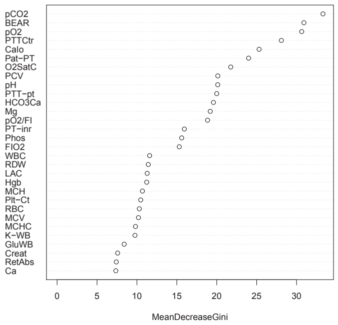
Variable Importance for Counts Abstraction Model Predicting Sex



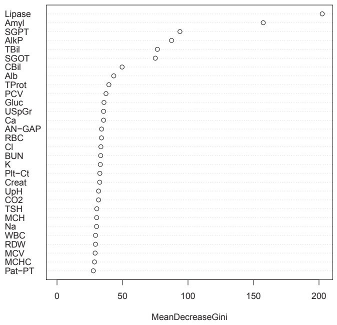
Variable Importance for Counts Abstraction Model Predicting Race



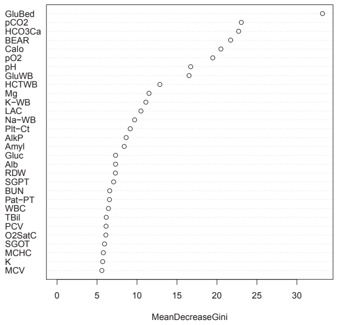
Variable Importance for Counts Abstraction Model Predicting Splenectomy



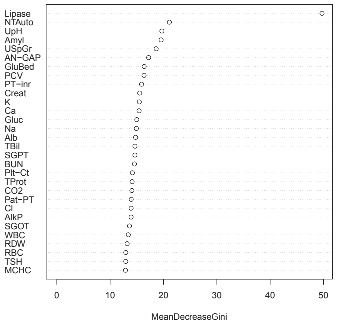
Variable Importance for Counts Abstraction Model Predicting Cholecystectomy



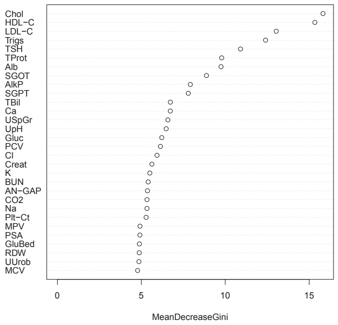
Variable Importance for Counts Abstraction Model Predicting Pancreatectomy



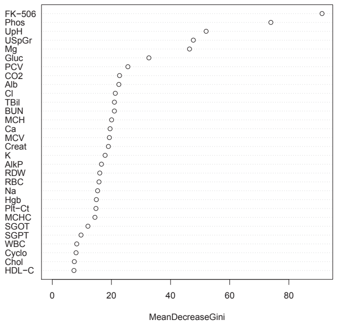
Variable Importance for Counts Abstraction Model Predicting Appendectomy



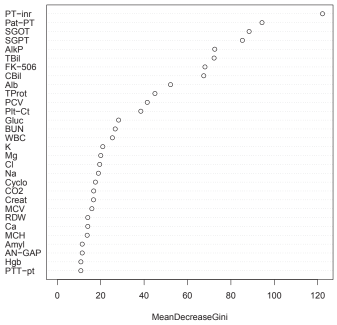
Variable Importance for Counts Abstraction Model Predicting Hemorrhoid Surgery



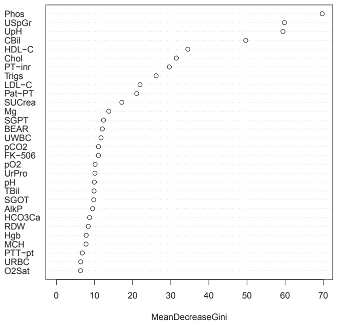
Variable Importance for Counts Abstraction Model Predicting Kidney Transplant



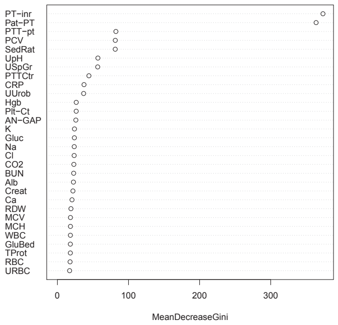
Variable Importance for Counts Abstraction Model Predicting Liver Transplant



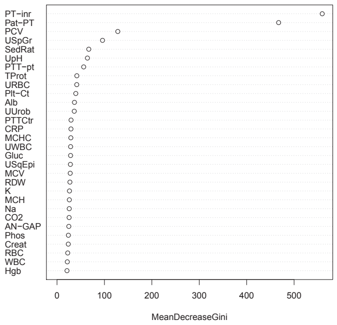
Variable Importance for Counts Abstraction Model Predicting Kidney vs Liver Transplant



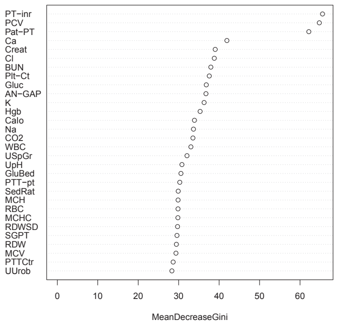
Variable Importance for Counts Abstraction Model Predicting Hip Replacement



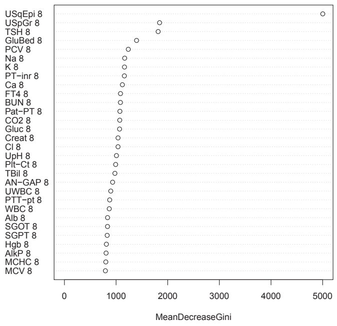
Variable Importance for Counts Abstraction Model Predicting Knee Replacement



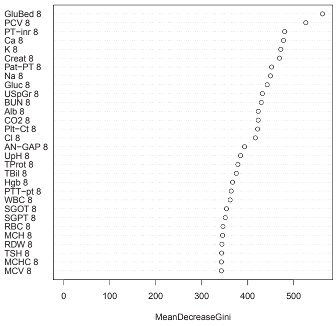
Variable Importance for Counts Abstraction Model Predicting Hip vs Knee Replacement



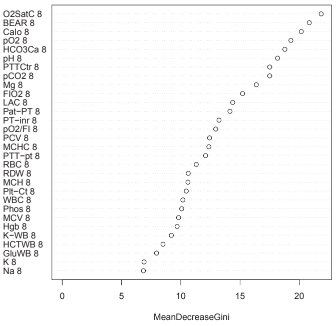
Variable Importance for Counts Per Year Abstraction Model Predicting Sex



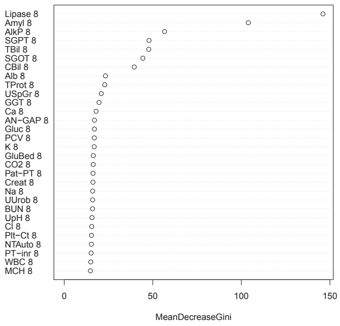
Variable Importance for Counts Per Year Abstraction Model Predicting Race



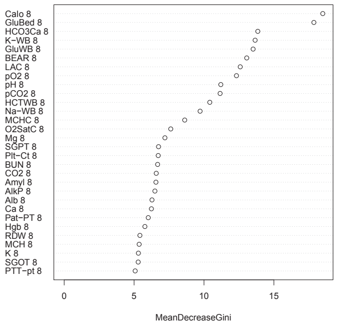
Variable Importance for Counts Per Year Abstraction Model Predicting Splenectomy



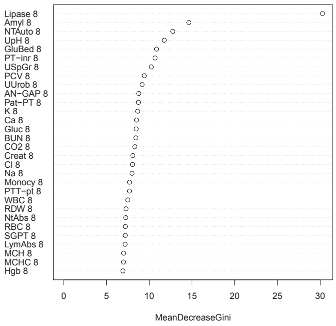
Variable Importance for Counts Per Year Abstraction Model Predicting Cholecystectomy



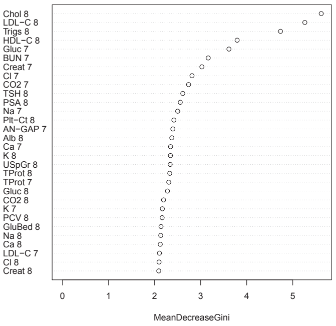
Variable Importance for Counts Per Year Abstraction Model Predicting Pancreatectomy



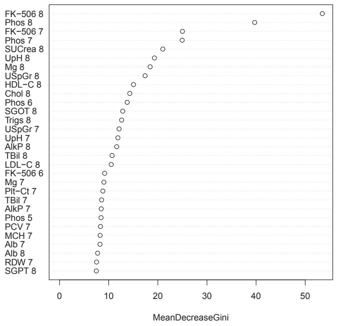
Variable Importance for Counts Per Year Abstraction Model Predicting Appendectomy



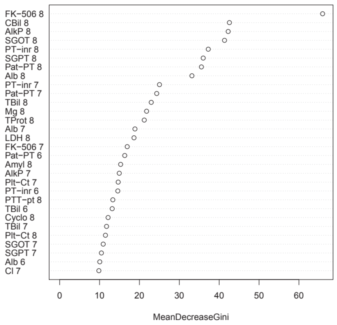
Variable Importance for Counts Per Year Abstraction Model Predicting Hemorrhoid Surgery



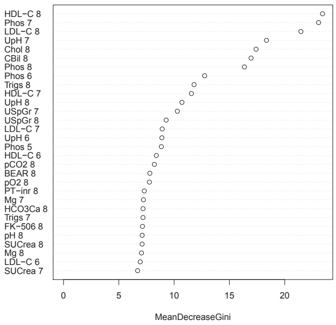
Variable Importance for Counts Per Year Abstraction Model Predicting Kidney Transplant



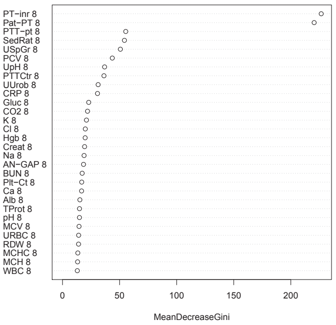
Variable Importance for Counts Per Year Abstraction Model Predicting Liver Transplant



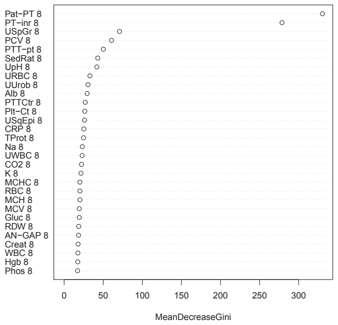
Variable Importance for Counts Per Year Abstraction Model Predicting Kidney vs Liver Transplant



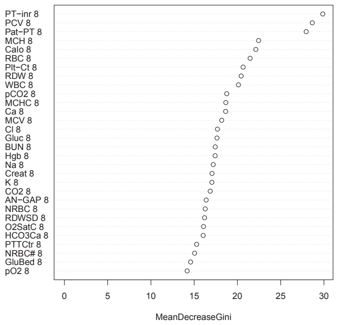
Variable Importance for Counts Per Year Abstraction Model Predicting Hip Replacement

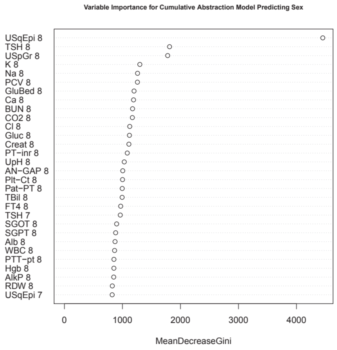


Variable Importance for Counts Per Year Abstraction Model Predicting Knee Replacement

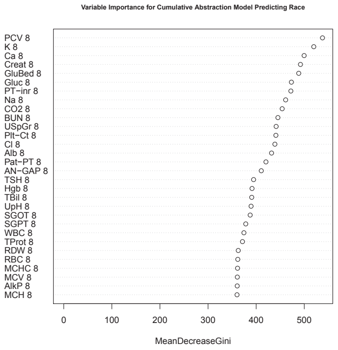


Variable Importance for Counts Per Year Abstraction Model Predicting Hip vs Knee Replacement

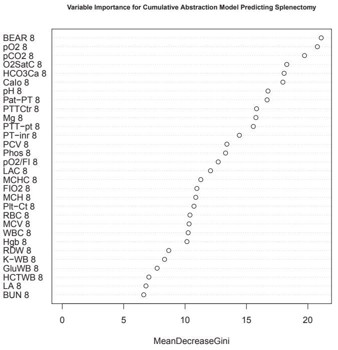




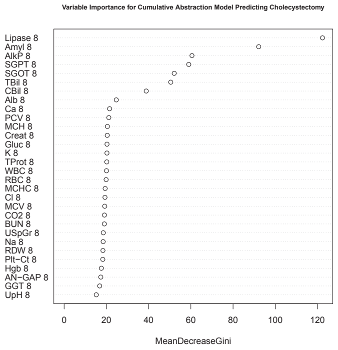
Variable Importance for Cumulative Abstraction Model Predicting Sex



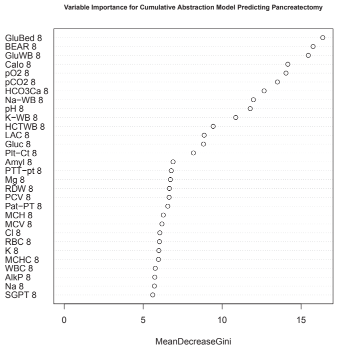
Variable Importance for Cumulative Abstraction Model Predicting Race



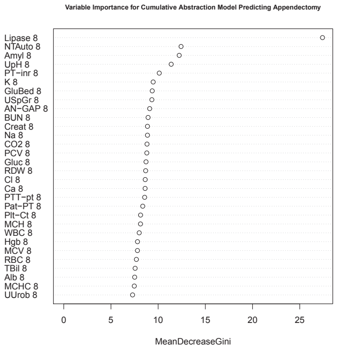
Variable Importance for Cumulative Abstraction Model Predicting Splenectomy



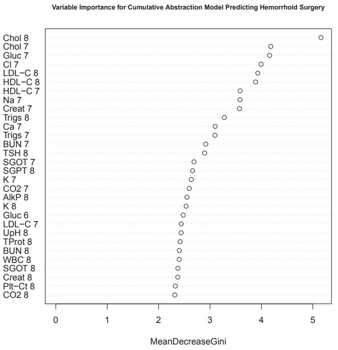
Variable Importance for Cumulative Abstraction Model Predicting Cholecystectomy



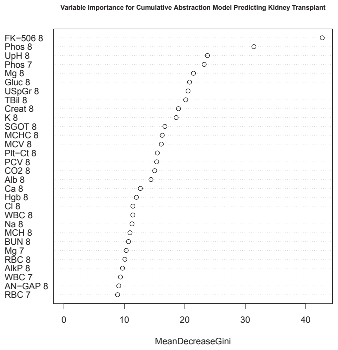
Variable Importance for Cumulative Abstraction Model Predicting Pancreatectomy



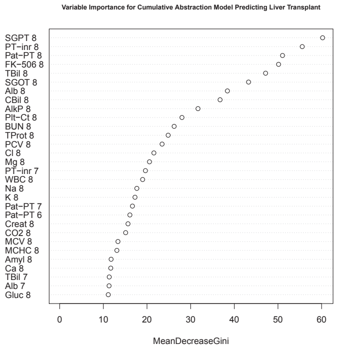
Variable Importance for Cumulative Abstraction Model Predicting Appendectomy



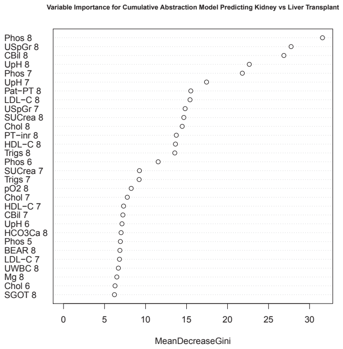
Variable Importance for Cumulative Abstraction Model Predicting Hemorrhoid Surgery



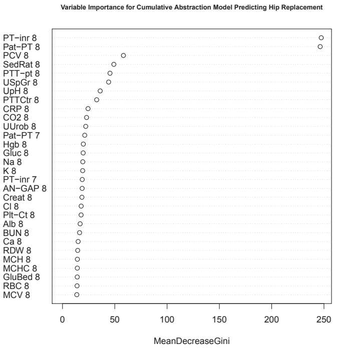
Variable Importance for Cumulative Abstraction Model Predicting Kidney Transplant



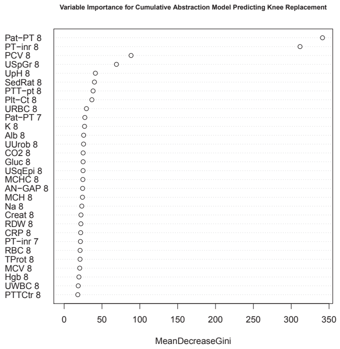
Variable Importance for Cumulative Abstraction Model Predicting Liver Transplant



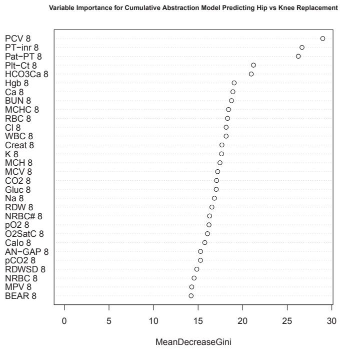
Variable Importance for Cumulative Abstraction Model Predicting Kidney vs Liver Transplant



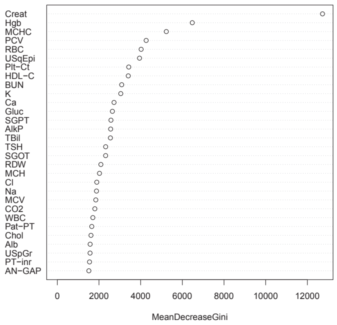
Variable Importance for Cumulative Abstraction Model Predicting Hip Replacement

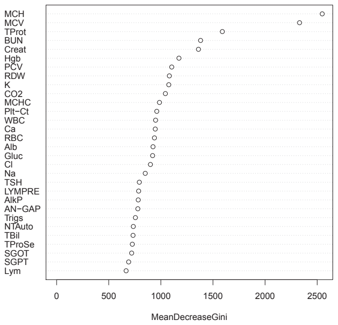


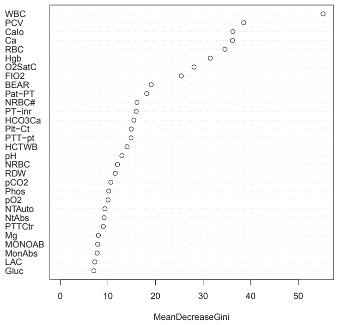
Variable Importance for Cumulative Abstraction Model Predicting Knee Replacement

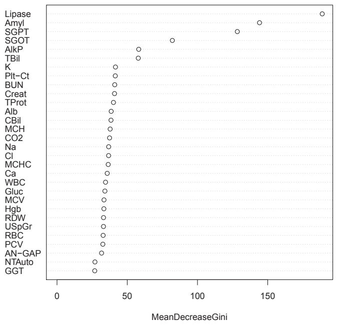


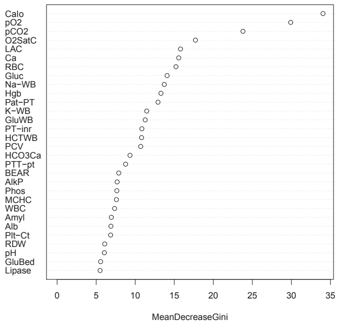
Variable Importance for Cumulative Abstraction Model Predicting Hip vs Knee Replacement

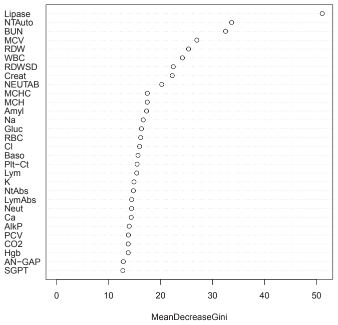

Variable Importance for Means Abstraction Model Predicting Sex

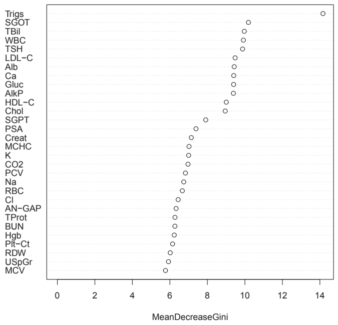

Variable Importance for Means Abstraction Model Predicting Race

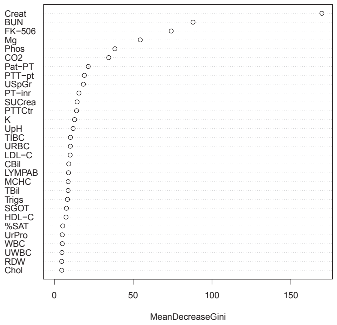

Variable Importance for Means Abstraction Model Predicting Splenectomy

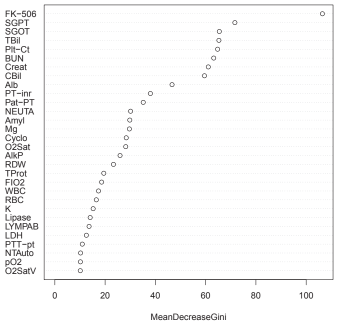

Variable Importance for Means Abstraction Model Predicting Cholecystectomy

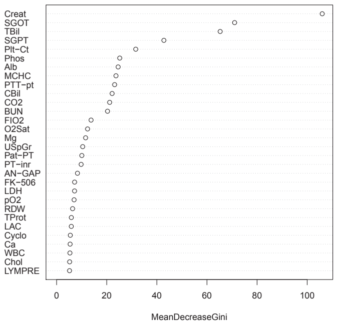

Variable Importance for Means Abstraction Model Predicting Pancreatectomy

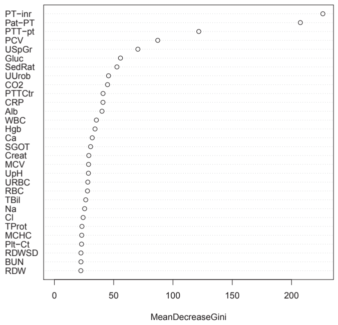

Variable Importance for Means Abstraction Model Predicting Appendectomy

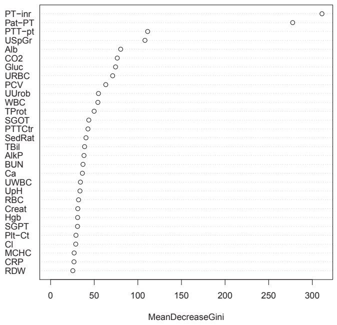

Variable Importance for Means Abstraction Model Predicting Hemorrhoid Surgery

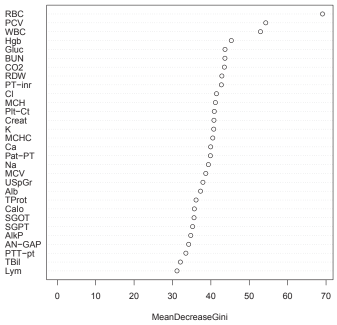

Variable Importance for Means Abstraction Model Predicting Kidney Transplant

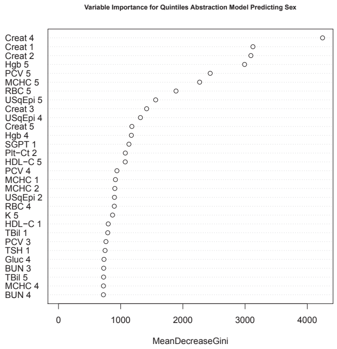

Variable Importance for Means Abstraction Model Predicting Liver Transplant

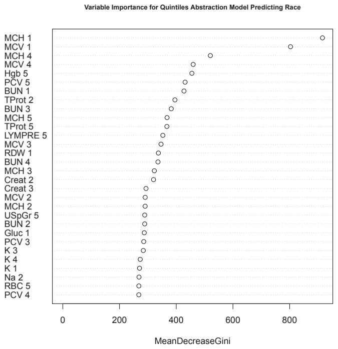

Variable Importance for Means Abstraction Model Predicting Kidney vs Liver Transplant

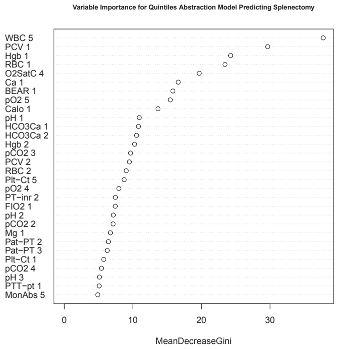

Variable Importance for Means Abstraction Model Predicting Hip Replacement

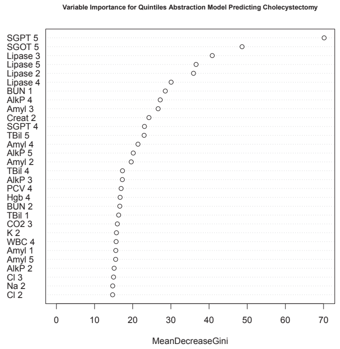

Variable Importance for Means Abstraction Model Predicting Knee Replacement

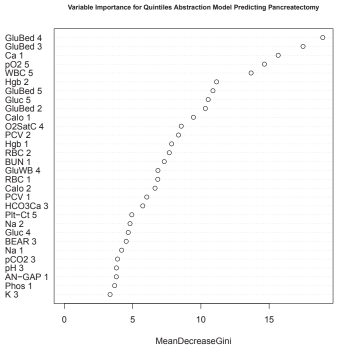

Variable Importance for Means Abstraction Model Predicting Hip vs Knee Replacement

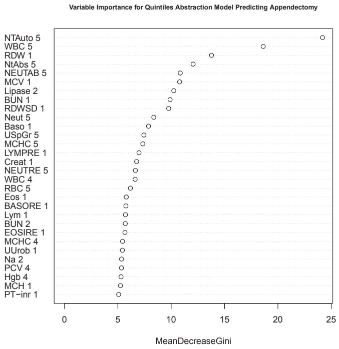

Variable Importance for Quintiles Abstraction Model Predicting Sex

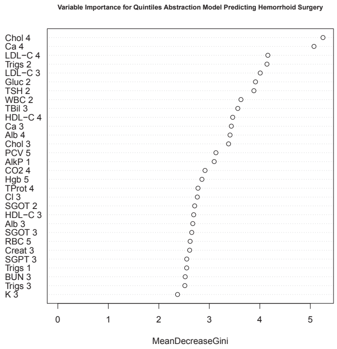

Variable Importance for Quintiles Abstraction Model Predicting Race

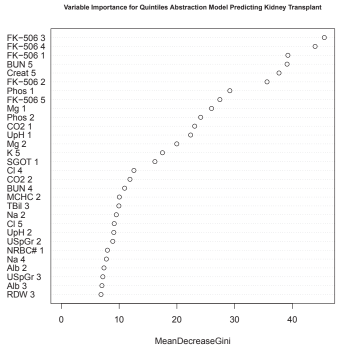

Variable Importance for Quintiles Abstraction Model Predicting Splenectomy

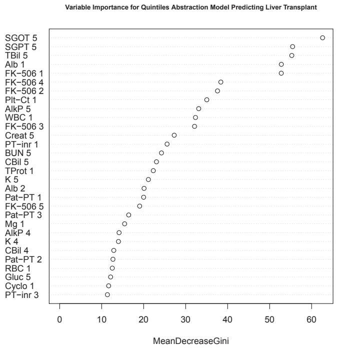

Variable Importance for Quintiles Abstraction Model Predicting Cholecystectomy

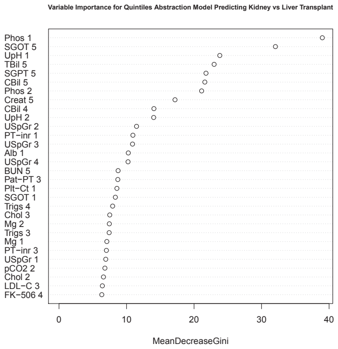

Variable Importance for Quintiles Abstraction Model Predicting Pancreatectomy

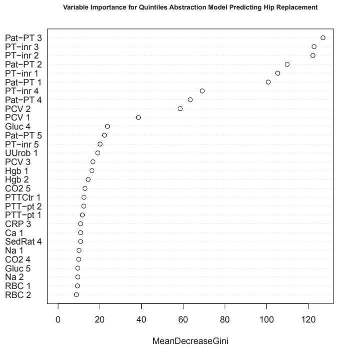

Variable Importance for Quintiles Abstraction Model Predicting Appendectomy

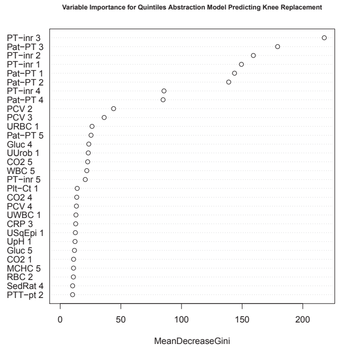

Variable Importance for Quintiles Abstraction Model Predicting Hemorrhoid Surgery

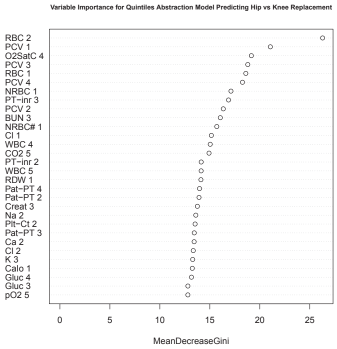

Variable Importance for Quintiles Abstraction Model Predicting Kidney Transplant

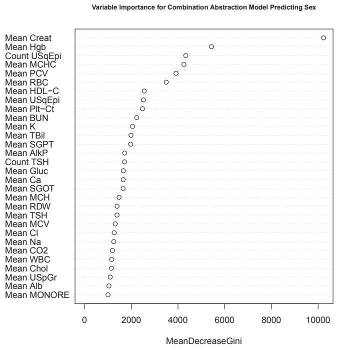

Variable Importance for Quintiles Abstraction Model Predicting Liver Transplant


Variable Importance for Quintiles Abstraction Model Predicting Kidney vs Liver Transplant

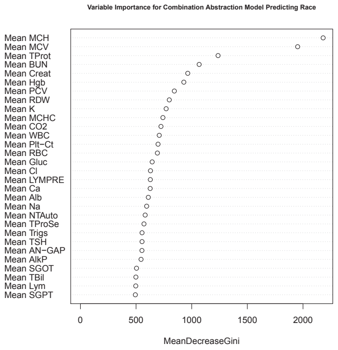

Variable Importance for Quintiles Abstraction Model Predicting Hip Replacement


Variable Importance for Quintiles Abstraction Model Predicting Knee Replacement

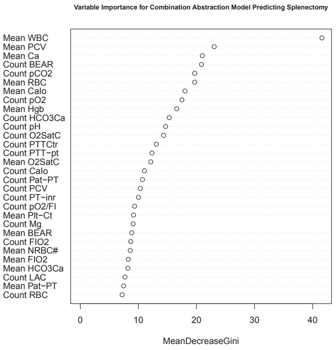

Variable Importance for Quintiles Abstraction Model Predicting Hip vs Knee Replacement



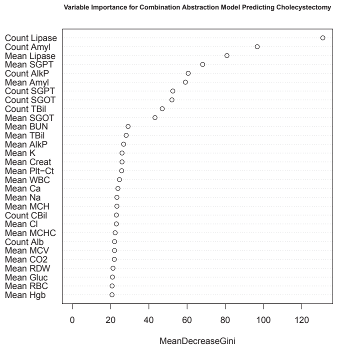
Variable Importance for Combination Abstraction Model Predicting Sex



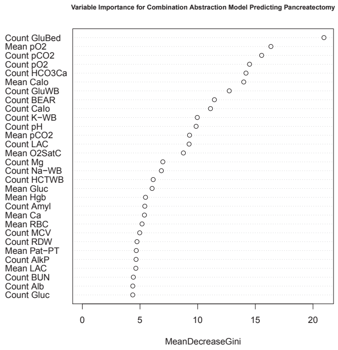
Variable Importance for Combination Abstraction Model Predicting Race



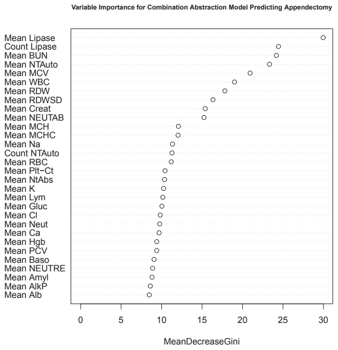
Variable Importance for Combination Abstraction Model Predicting Splenectomy



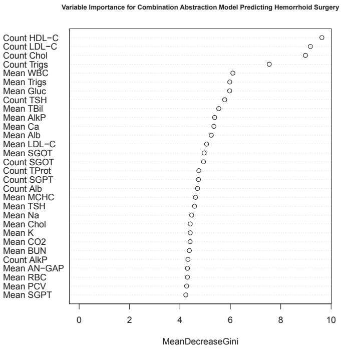
Variable Importance for Combination Abstraction Model Predicting Cholecystectomy



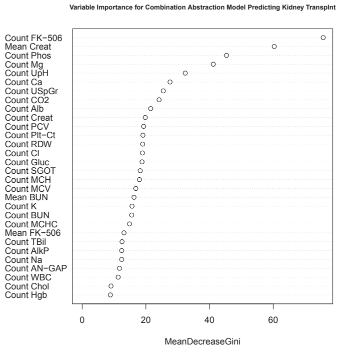
Variable Importance for Combination Abstraction Model Predicting Pancreatectomy



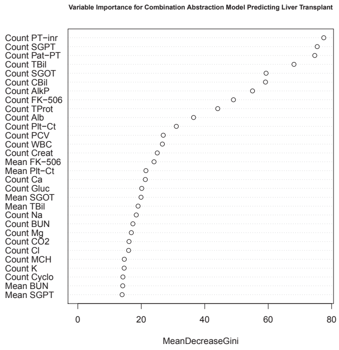
Variable Importance for Combination Abstraction Model Predicting Appendectomy



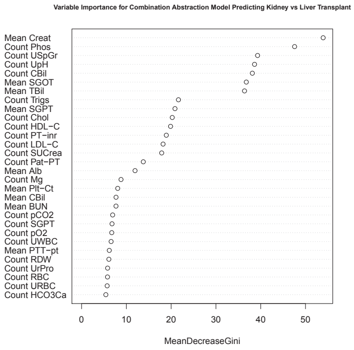
Variable Importance for Combination Abstraction Model Predicting Hemorrhoid Surgery



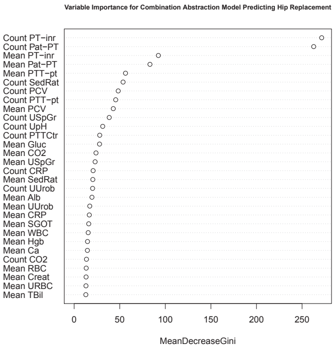
Variable Importance for Combination Abstraction Model Predicting Kidney Transplnt



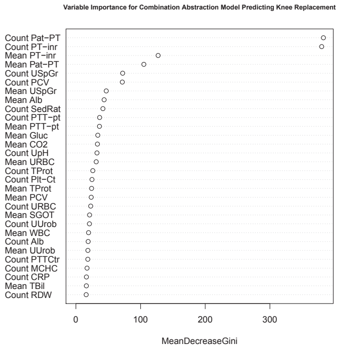
Variable Importance for Combination Abstraction Model Predicting Liver Transplant



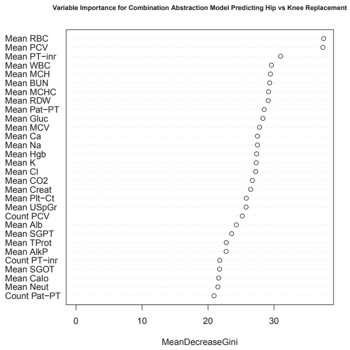
Variable Importance for Combination Abstraction Model Predicting Kidney vs Liver Transplant



Variable Importance for Combination Abstraction Model Predicting Hip Replacement



Variable Importance for Combination Abstraction Model Predicting Knee Replacement



Variable Importance for Combination Abstraction Model Predicting Hip vs Knee Replacement

APPENDIX C.

Coefficients remaining in penalized regression model predicting Vitamin B12 at 50 pg/mL.

PheWAS Code Description	β
Mood disorders	0.0355
Substance addiction and disorders	0.0074
Intracranial hemorrhage	0.0057
Anxiety, phobic & dissociative disorders	0.0037
Cerebrovascular disease	0.0036
Urinary tract infection	-0.0006
Hypothyroidism	-0.003
Bacterial infection NOS	-0.0101
Chemotherapy	-0.0133
Sepsis and SIRS	-0.0163
Pleurisy	-0.0165
Cancer of other lymphoid, histiocytic tissue	-0.0193
Secondary malignant neoplasm	-0.0198
Malaise and fatigue	-0.0213
Diabetes mellitus	-0.0237
Respiratory failure; insufficiency; arrest	-0.0265
Ascites (non-malignant)	-0.0288
Dysphagia	-0.041
Leukemia	-0.0484
Other symptoms of respiratory system	-0.0606
Viral hepatitis	-0.0651
Alcohol-related disorders	-0.0681
Fluid, electrolyte, & acid-base balance disorders	-0.0758
Other anemias	-0.077
Protein-calorie malnutrition	-0.0841
Neurological disorders	-0.0848
Pneumonia	-0.0899
Purpura and other hemorrhagic conditions	-0.0934
Congestive heart failure, nonhypertensive	-0.1082
Septicemia	-0.1447
Renal failure	-0.2396
Chronic liver disease and cirrhosis	-0.2711

APPENDIX D.

Coefficients remaining in penalized regression model predicting PCV at 30%.

PheWAS Code Description	β
Cancer of other female genital organs	0.8287
Chemotherapy	0.7697
Cancer of kidney and urinary organs	0.6907
Cancer of bone & connective tissue	0.6682
Known or suspected fetal abnormality	0.6182
Early or threatened labor	0.6060
Pancreatic cancer	0.5982
Colorectal cancer	0.5843
Fracture of lower limb	0.5519
Cancer of the upper GI tract	0.5351
Infections involving bone	0.5242
Other conditions of the mother complicating pregnancy	0.5094
Stomach cancer	0.5028
Cancer within the respiratory system	0.4985
Retinal disorders	0.4902
Fracture of ankle and foot	0.4779
Abnormality of organs & soft tissues of pelvis complicating pregnancy, childbirth, or the puerperium	0.4746
Breast cancer	0.4534
Fracture of pelvis	0.4524
Chronic ulcer of skin	0.4375
Curvature of spine	0.4320
Hereditary hemolytic anemias	0.4297
Fracture of unspecified bones	0.4245
Cervical cancer and dysplasia	0.4151
Other anemias	0.4145
Osteoarthrosis	0.4139
Acute bronchitis and bronchiolitis	0.4105
Hypertension complicating pregnancy	0.4045
Heart valve disorders	0.4044
Peripheral vascular disease	0.3953
Other aneurysm	0.3947
Iron deficiency anemias	0.3868
Cancer of the digestive organs and peritoneum	0.3692
Other biliary tract disease	0.3584
Diseases of esophagus	0.3506
Cancer of other lymphoid, histiocytic tissue	0.3434
Gestational diabetes	0.3386
Viral infection	0.3305

Secondary malignant neoplasm	0.3267
Lack of normal physiological development	0.3262
Other disorders of intestine	0.3241
Arthropathy associated with infections	0.3229
Hodgkin's disease	0.3226
Ileostomy status	0.3178
Edema	0.3175
Fever of unknown origin	0.3108
Fracture of vertebral column without mention of spinal cord injury	0.3024
Other upper respiratory disease	0.3018
Chronic liver disease and cirrhosis	0.2993
Congenital anomalies of face and neck	0.2992
Uterine cancer	0.2989
Postoperative infection	0.2969
Ischemic Heart Disease	0.2896
Muscular dystrophies and other myopathies	0.2890
Pneumonitis due to inhalation of food or vomitus	0.2870
Bone marrow or stem cell transplant	0.2868
Leukemia	0.2825
Hemorrhage during pregnancy; childbirth and postpartum	0.2824
Lymphadenitis	0.2701
Cancer of mouth	0.2671
Open wounds of extremities	0.2602
Atherosclerosis	0.2600
Nephritis; nephrosis; renal sclerosis	0.2560
Pyelonephritis	0.2552
Liver replaced by transplant	0.2539
Urinary tract infection	0.2450
Inflammatory diseases of female pelvic organs	0.2423
Pleurisy	0.2400
Nausea and vomiting	0.2381
Contusion	0.2374
Empyema and pneumothorax	0.2328
Fracture of upper limb	0.2318
Decreased white blood cell count	0.2300
Kidney replaced by transplant	0.2286
Hepatic cancer	0.2284
Open wounds of head; neck; and trunk	0.2274
Protein-calorie malnutrition	0.2243
Skull fracture and other intracranial injury	0.2239
Prostate cancer	0.2129
Disorders of the kidney & ureters	0.2125
Intracranial hemorrhage (injury)	0.2105

Disorders of liver	0.2088
Meningitis	0.2068
Fracture of ribs	0.2058
Lung disease due to external agents	0.2006
Problems associated with amniotic cavity and membranes	0.1980
Renal failure	0.1975
Retention of urine	0.1966
Diseases of pancreas	0.1962
Other symptoms of respiratory system	0.1915
Heart transplant/surgery	0.1898
Spinal stenosis	0.1880
Venous complications in pregnancy and the puerperium	0.1847
Peptic ulcer	0.1830
Esophageal cancer	0.1787
Gastrointestinal hemorrhage	0.1787
Erythematous conditions	0.1763
Ascites (non-malignant)	0.1748
Acute upper respiratory infections	0.1742
Pneumonia	0.1724
Fluid, electrolyte, & acid-base balance disorders	0.1674
Spinal cord injury without evidence of spinal bone injury	0.1674
Infective connective tissue disorders	0.1619
Otitis media & Eustachian tube disorders	0.1602
Candidiasis	0.1576
Other disorders of stomach and duodenum	0.1575
Other diseases of lung	0.1563
Septicemia	0.1515
Other disorders of bladder	0.1485
Human immunodeficiency virus	0.1475
Chronic airway obstruction	0.1467
Bacterial infection NOS	0.1449
Epilepsy, recurrent seizures, convulsions	0.1428
Infection/inflammation of internal prosthetic device, implant or graft	0.1413
Respiratory abnormalities	0.1409
Venous embolism & thrombosis	0.1395
Cancer, suspected or other	0.1383
Polyarteritis nodosa and allied conditions	0.1370
Delirium dementia and amnestic disorders	0.1359
Rash and other nonspecific skin eruption	0.1351
Cardiac conduction disorders	0.1340
Diabetes mellitus	0.1301
Other local infections of skin and subcutaneous tissue	0.1248
Abnormal movement	0.1247

Abnormal heart sounds	0.1247
Other paralytic syndromes	0.1246
Gangrene	0.1245
Other infectious diseases	0.1228
Symptoms of the muscles	0.1213
Protein plasma/amino-acid transport and metabolism disorder	0.1209
Nervous system congenital anomalies	0.1203
Other peripheral nerve disorders	0.1183
Peritonitis and retroperitoneal infections	0.1180
Encounter for long-term use of antibiotics	0.1166
Respiratory failure; insufficiency; arrest	0.1157
Excessive vomiting in pregnancy	0.1144
Intestinal obstruction without mention of hernia	0.1137
Encephalitis	0.1107
Intracranial hemorrhage	0.1074
Other and unspecified complications of birth; puerperium affecting management of mother	0.1060
Short gestation; low birth weight; and fetal growth retardation	0.1001
Pulmonary collapse; interstitial/compensatory emphysema	0.1001
Lung transplant	0.0994
Osteoporosis, osteopenia, & pathological fractures	0.0953
Influenza	0.0949
Persistent mental disorders due to other conditions	0.0894
Disorders of adrenal glands	0.0888
Infections of genitourinary tract during pregnancy	0.0887
Fracture of hand or wrist	0.0887
Early complications of trauma or procedure	0.0872
Cardiac dysrhythmias	0.0872
Inflammatory bowel disease	0.0861
Mechanical complications of cardiac/vascular device, implant, and graft	0.0857
Hypotension	0.0853
Immune disorders	0.0844
Anomalies of respiratory system, congenital	0.0827
Other complications of pregnancy NEC	0.0822
Hemangioma and lymphangioma, any site	0.0815
Abdominal pain	0.0808
Hypothyroidism	0.0792
Other nutritional deficiency	0.0789
Spondylosis and allied disorders	0.0778
Herpes simplex	0.0771
Superficial cellulitis & abscess	0.0755
Neurological disorders	0.0754
Hemorrhage or hematoma complicating a procedure	0.0746
Other symptoms involving abdomen and pelvis	0.0744

Complication of internal orthopedic device	0.0700
Noninfectious disorders of lymphatic channels	0.0700
Major puerperal infection	0.0698
Rhabdomyolysis	0.0675
Muscle weakness	0.0673
Other disorders of peritoneum	0.0673
Abnormal sputum	0.0635
Intestinal infection	0.0619
Carditis	0.0618
Dislocation	0.0615
Asthma	0.0614
Hypertension	0.0609
Disorders resulting from impaired renal function	0.0607
Other disorders of circulatory system	0.0605
Aplastic anemia	0.0583
Hepatitis NOS	0.0578
Purpura and other hemorrhagic conditions	0.0563
Other complications of the puerperium NEC	0.0555
Traumatic amputation	0.0522
Parkinson's disease	0.0507
Other cerebral degenerations	0.0501
Phlebitis and thrombophlebitis	0.0488
Alcohol-related disorders	0.0481
Miscarriage; stillbirth	0.0466
Viral hepatitis	0.0431
Other specified nonpsychotic and/or transient mental disorders	0.0431
Amyloidosis	0.0427
Congestive heart failure, nonhypertensive	0.0397
Abdominal hernia	0.0387
Neoplasm of uncertain behavior	0.0378
Congenital musculoskeletal anomalies	0.0373
Anorexia	0.0368
Infantile cerebral palsy	0.0361
Infectious & parasitic conditions complicating pregnancy	0.0302
Cancer of other endocrine glands	0.0296
Cerebral laceration and contusion	0.0293
Adverse drug events and drug allergies	0.0290
Dysphagia	0.0290
Symptoms and disorders of the joints	0.0287
Abnormal findings examination of lungs	0.0272
Arterial embolism and thrombosis	0.0270
Long-term use of anticoagulants	0.0267
Disorders of function of stomach	0.0262

Complication of amputation stump	0.0254
Abnormal serum enzyme levels	0.0243
Non-inflammatory female genital disorders	0.0238
Benign neoplasm of brain and other parts of nervous system	0.0227
Complications of labor and delivery NEC	0.0194
Symptoms involving nervous and musculoskeletal systems	0.0192
Functional digestive disorders	0.0180
Diseases of the larynx and vocal cords	0.0177
Malaise and fatigue	0.0153
Developmental delays and disorders	0.0145
Post-inflammatory pulmonary fibrosis	0.0125
Disorders of sweat glands	0.0125
Other conditions of brain	0.0121
Other disorders of the nervous system	0.0114
Swelling of limb	0.0114
Cerebrovascular disease	0.0107
CNS infection and poliomyelitis	0.0096
Coagulation defects	0.0080
Myeloproliferative disease	0.0042
Infection of the eye	0.0030
Degenerative disease of the spinal cord	-0.0006
Elevated prostate specific antigen	-0.0008
Hemiplegia	-0.0012
Sepsis and SIRS	-0.0066
Glaucoma	-0.0077
pulmonary heart disease	-0.0110
Abnormal results of function study of liver	-0.0129
Nonspecific chest pain	-0.0168
Menopausal & postmenopausal disorders	-0.0195
Digestive congenital anomalies	-0.0215
Acquired hemolytic anemias	-0.0230
Acute sinusitis	-0.0232
Substance addiction and disorders	-0.0240
Other abnormal blood chemistry	-0.0264
Disturbance of skin sensation	-0.0282
Migraine	-0.0319
Rheumatoid arthritis & related inflammatory polyarthropathies	-0.0332
Conduct disorders	-0.0340
Musculoskeletal symptoms referable to limbs	-0.0402
Shock	-0.0419
Other conditions of brain, NOS	-0.0437
Thyroid cancer	-0.0480
Disorders of synovium, tendon, and bursa	-0.0497

Acute and subacute necrosis of liver	-0.0513
Abnormal findings on mammogram or breast exam	-0.0548
Disorders of other cranial nerves	-0.0563
Multiple sclerosis	-0.0594
Schizophrenia and other psychotic disorders	-0.0636
Vitamin deficiency	-0.0664
Disorders of parathyroid gland	-0.0685
Light-headedness and vertigo	-0.0702
Mood disorders	-0.0776
Pulmonary congestion and hypostasis	-0.0871
Intestinal malabsorption	-0.0888
Anxiety, phobic & dissociative disorders	-0.0905
Pervasive developmental disorders	-0.0933
Eating disorder	-0.0977
Infections specific to the perinatal period	-0.0982
Back pain	-0.1030
Other headache syndromes	-0.1075
Tobacco use disorder	-0.1182
Nontoxic nodular goiter	-0.1229
Sleep apnea	-0.1297
Acid-base balance disorder	-0.1348
Cervicalgia	-0.1737
Diseases of sebaceous glands	-0.1803
Sleep disorders	-0.1809
Peripheral enthesopathies	-0.1861
Abnormal glucose	-0.1901
Degenerative skin conditions and other dermatoses	-0.2238
Other perinatal conditions	-0.2394
Cataract	-0.2443
Pain in joint	-0.2530
Psoriasis & related disorders	-0.2653
Testicular dysfunction	-0.2685
Elevated C-reactive protein	-0.3088
Disorders of lipoid metabolism	-0.5259
Allergic rhinitis	-0.5769

APPENDIX E.

Coefficients remaining in penalized regression model predicting glucose at 450 mg/dL and HgbA1C at 5.5%

PheWAS Code Description	β
Gestational diabetes	0.2115
Abnormal glucose	0.2057
Disorders of lipoid metabolism	0.1759
Heart valve disorders	0.1532
Sleep disorders	0.1071
Overweight	0.1051
Known or suspected fetal abnormality	0.0986
Lung transplant	0.0970
Other conditions of the mother complicating pregnancy	0.0764
Allergic rhinitis	0.0758
Other and unspecified complications of birth; puerperium affecting management of mother	0.0713
Heart transplant/surgery	0.0712
Back pain	0.0711
Tobacco use disorder	0.0666
Abnormality of organs & soft tissues of pelvis complicating pregnancy, childbirth, or the puerperium	0.0663
Pulmonary collapse; interstitial/compensatory emphysema	0.0649
Pain in joint	0.0596
Liver replaced by transplant	0.0574
Vitamin deficiency	0.0573
Ischemic Heart Disease	0.0496
Complications of labor and delivery NEC	0.0481
Hypertension	0.0420
Cardiomegaly	0.0417
Cerebrovascular disease	0.0367
Kidney replaced by transplant	0.0349
Sleep apnea	0.0342
Problems associated with amniotic cavity and membranes	0.0342
Hypothyroidism	0.0307
Long-term use of anticoagulants	0.0293
Cardiomyopathy	0.0266
Bone marrow or stem cell transplant	0.0257
Asthma	0.0257

Human immunodeficiency virus	0.0243
Neurological disorders	0.0238
Cervicalgia	0.0227
Other aneurysm	0.0219
Cardiac dysrhythmias	0.0217
Dysphagia	0.0170
Hypertension complicating pregnancy	0.0168
Myalgia and myositis NOS	0.0158
Musculoskeletal symptoms referable to limbs	0.0157
Cataract	0.0149
Renal failure	0.0137
Early or threatened labor	0.0131
Shock	0.0124
Malaise and fatigue	0.0120
Bariatric surgery	0.0098
Nonspecific chest pain	0.0098
Fluid, electrolyte, & acid-base balance disorders	0.0083
Light-headedness and vertigo	0.0078
Venous embolism & thrombosis	0.0076
Carditis	0.0066
Coma	0.0056
Degenerative skin conditions and other dermatoses	0.0045
Pleurisy	0.0043
Ovarian dysfunction	0.0031
Chronic liver disease and cirrhosis	0.0030
Symptoms/disorders of the urinary system	0.0023
Intracranial hemorrhage	0.0021
Hyperplasia of prostate	0.0020
Peripheral enthesopathies	0.0020
Other specified nonpsychotic and/or transient mental disorders	0.0015
Elevated transaminase or LDH	0.0014
Other symptoms of respiratory system	0.0013
Disorders of menstruation	0.0006
Cystic fibrosis	0.0005
Infection/inflammation of internal prosthetic device, implant or graft	0.0000
Encephalitis	-0.0002
Acute upper respiratory infections	-0.0003

Other headache syndromes	-0.0006
Cancer of mouth	-0.0008
Jaundice	-0.0009
Contusion	-0.0012
Disorders of the kidney & ureters	-0.0014
Abnormal sputum	-0.0022
Inflammatory and toxic neuropathy	-0.0022
Symptoms of the muscles	-0.0027
Developmental delays and disorders	-0.0030
Pyelonephritis	-0.0033
Fracture of ankle and foot	-0.0033
Infective connective tissue disorders	-0.0036
Cardiac conduction disorders	-0.0039
Disorders of liver	-0.0041
Mood disorders	-0.0043
Chronic airway obstruction	-0.0046
Viral infection	-0.0049
Stomach cancer	-0.0052
Peritonitis and retroperitoneal infections	-0.0058
pulmonary heart disease	-0.0060
Urinary tract infection	-0.0062
Other conditions of brain, NOS	-0.0066
Other symptoms involving abdomen and pelvis	-0.0066
Mycoses	-0.0068
Other abnormal blood chemistry	-0.0069
Adverse drug events and drug allergies	-0.0074
Sepsis and SIRS	-0.0074
Other anemias	-0.0078
Other cerebral degenerations	-0.0081
Uterine cancer	-0.0082
Respiratory failure; insufficiency; arrest	-0.0090
Cancer of other female genital organs	-0.0090
Delirium dementia and amnestic disorders	-0.0096
Genitourinary congenital anomalies	-0.0097
Purpura and other hemorrhagic conditions	-0.0101
Chemotherapy	-0.0102
Suicidal ideation or attempt	-0.0103

Leukemia	-0.0104
Pancreatic cancer	-0.0109
Bacterial infection NOS	-0.0109
Intracranial hemorrhage (injury)	-0.0111
Substance addiction and disorders	-0.0120
Arthropathy associated with infections	-0.0124
Cancer of other lymphoid, histiocytic tissue	-0.0125
Spinal cord injury without evidence of spinal bone injury	-0.0133
Other pulmonary inflammation or edema	-0.0134
Congestive heart failure, nonhypertensive	-0.0135
Diseases of respiratory system NEC	-0.0136
Poisoning by analgesics, antipyretics, and antirheumatics	-0.0147
Pneumonia	-0.0160
Abnormal heart sounds	-0.0163
Infections specific to the perinatal period	-0.0164
Gastrointestinal hemorrhage	-0.0172
Fracture of unspecified bones	-0.0177
Decreased white blood cell count	-0.0179
Diseases of white blood cells	-0.0192
Intestinal obstruction without mention of hernia	-0.0193
Pneumonitis due to inhalation of food or vomitus	-0.0195
Fracture of upper limb	-0.0198
Congenital musculoskeletal anomalies	-0.0204
Acquired hemolytic anemias	-0.0208
Cancer, suspected or other	-0.0211
Colorectal cancer	-0.0212
Skin cancer	-0.0214
Esophageal cancer	-0.0215
Open wounds of extremities	-0.0215
Epilepsy, recurrent seizures, convulsions	-0.0216
Empyema and pneumothorax	-0.0216
Fracture of vertebral column without mention of spinal cord injury	-0.0226
Other paralytic syndromes	-0.0230
Protein-calorie malnutrition	-0.0232
Abnormal movement	-0.0243
Hemiplegia	-0.0249
Eating disorder	-0.0262

Congenital anomalies of face and neck	-0.0279
Alcohol-related disorders	-0.0286
Nervous system congenital anomalies	-0.0288
Respiratory abnormalities	-0.0292
Digestive congenital anomalies	-0.0294
Rhabdomyolysis	-0.0314
Abdominal pain	-0.0317
Skull fracture and other intracranial injury	-0.0322
Short gestation; low birth weight; and fetal growth retardation	-0.0326
Acute bronchitis and bronchiolitis	-0.0327
Other disorders of circulatory system	-0.0331
Meningitis	-0.0332
Cancer within the respiratory system	-0.0334
Cancer of the upper GI tract	-0.0334
Secondary malignant neoplasm	-0.0338
Hereditary hemolytic anemias	-0.0351
Superficial cellulitis & abscess	-0.0352
Cholelithiasis and cholecystitis	-0.0364
Open wounds of head; neck; and trunk	-0.0371
Cancer of bone & connective tissue	-0.0373
Nausea and vomiting	-0.0407
Fever of unknown origin	-0.0409
Infections involving bone	-0.0411
Muscular dystrophies and other myopathies	-0.0416
Fracture of ribs	-0.0420
Malignant neoplasm of brain and nervous system	-0.0436
Other perinatal conditions	-0.0440
Fracture of pelvis	-0.0481
Inflammatory bowel disease	-0.0520
Cardiac & circulatory congenital anomalies	-0.0529
Cancer of kidney and urinary organs	-0.0535
Lack of normal physiological development	-0.0541
Fracture of lower limb	-0.0550
Diabetes mellitus	-0.1500

REFERENCES

[1] Office of the National Coordinator for Health Information Technology. What information does an electronic health record (EHR) contain? [Internet] 2013. https://www.healthit.gov/providers-professionals/faqs/what-information-does-electronic-health-record-ehr-contain (accessed January 1, 2016).

[2] Menachemi N, Collum TH. Benefits and drawbacks of electronic health record systems. Risk Manag Healthc Policy 2011;4:47–55. doi:10.2147/RMHP.S12985.

[3] Gellert GA, Hill V, Bruner K, Maciaz G, Saucedo L, Catzoela L, et al. Successful Implementation of Clinical Information Technology. Appl Clin Inform 2015;6:698–715. doi:10.4338/ACI-2015-06-SOA-0067.

[4] Niazkhani Z, Pirnejad H, van der Sijs H, Aarts J. Evaluating the medication process in the context of CPOE use: The significance of working around the system. Int J Med Inform 2011;80:490–506. doi:10.1016/j.ijmedinf.2011.03.009.

[5] Garg AX, Adhikari NKJ, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA 2005;293:1223–38. doi:10.1001/jama.293.10.1223.

[6] Hunt DL, Haynes RB, Hanna SE, Smith K. Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review. JAMA 1998;280:1339–46.

[7] Walker J, Pan E, Johnston D, Adler-Milstein J, Bates DW, Middleton B. The value of health care information exchange and interoperability. Health Aff (Millwood) 2005;Suppl Web :W5–10 – W5–18. doi:10.1377/hlthaff.w5.10.

[8] Sadeghi-Bazargani H, Tabrizi JS, Azami-Aghdash S. Barriers to evidence-based medicine: a systematic review. J Eval Clin Pract 2014;20:793–802. doi:10.1111/jep.12222.

[9] Murphy E V. Clinical decision support: effectiveness in improving quality processes and clinical outcomes and factors that may influence success. Yale J Biol Med 2014;87:187–97.

[10] Frisse ME, Johnson KB, Nian H, Davison CL, Gadd CS, Unertl KM, et al. The financial impact of health information exchange on emergency department care. J Am Med Inform Assoc 2011;19:328–33. doi:10.1136/amiajnl-2011-000394.

[11] Saef S, Melvin C, Carr C. Impact of a Health Information Exchange on Resource Use and Medicare-Allowable Reimbursements at 11 Emergency Departments in a Midsized City. West J Emerg Med 2014;15:777–85. doi:10.5811/westjem.2014.9.21311.

[12] Chacour Bahous M, Shadmi E. Health information exchange and information gaps in referrals to a pediatric emergency department. Int J Med Inform 2016;87:68–74. doi:10.1016/j.ijmedinf.2015.12.011.

[13] Richesson RL, Krischer J. Data standards in clinical research: gaps, overlaps, challenges and future directions. J Am Med Inform Assoc 2007;14:687–96. doi:10.1197/jamia.M2470.

[14] Stewart WF, Shah NR, Selna MJ, Paulus RA, Walker JM. Bridging The Inferential Gap: The Electronic Health Record And Clinical Evidence. Health Aff 2007;26:w181–91. doi:10.1377/hlthaff.26.2.w181.

[15] Yen J, Chiu W, Chu S, Hsu M-H. Secondary use of health data. J Formos Med Assoc 2016;115:137–8. doi:10.1016/j.jfma.2015.03.006.

[16] Fleming NS, Culler SD, McCorkle R, Becker ER, Ballard DJ. The Financial And Nonfinancial Costs Of Implementing Electronic Health Records In Primary Care Practices. Health Aff 2011;30:481–9. doi:10.1377/hlthaff.2010.0768.

[17] Rosenthal GE. The role of pragmatic clinical trials in the evolution of learning health systems. Trans Am Clin Climatol Assoc 2014;125:204–16; discussion 217–8.

[18] Colditz G, Winter A. Clinical trial design in the era of comparative effectiveness research. Open Access J Clin Trials 2014;Volume 6:101. doi:10.2147/OAJCT.S39758.

[19] McMurdo MET, Roberts H, Parker S, Wyatt N, May H, Goodman C, et al. Improving recruitment of older people to research through good practice. Age Ageing 2011;40:659–65. doi:10.1093/ageing/afr115.

[20] Friedman C, Rubin J, Brown J, Buntin M, Corn M, Etheredge L, et al. Toward a science of learning systems: a research agenda for the high-functioning Learning Health System. J Am Med Informatics Assoc 2014:1–6. doi:10.1136/amiajnl-2014-002977.

[21] Fox P, Hendler J. The Science of Data Science. Big Data 2014;2:68–70. doi:10.1089/big.2014.0011.

[22] Horn J-F, Habert M-O, Kas A, Malek Z, Maksud P, Lacomblez L, et al. Differential automatic diagnosis between Alzheimer's disease and frontotemporal dementia based on perfusion SPECT images. Artif Intell Med 2009;47:147–58. doi:10.1016/j.artmed.2009.05.001.

[23] Baxt WG, Shofer FS, Sites FD, Hollander JE. A neural network aid for the early diagnosis of cardiac ischemia in patients presenting to the emergency department with chest pain. Ann Emerg Med 2002;40:575–83. doi:10.1067/mem.2002.129171.

[24] Colubri A, Silver T, Fradet T, Retzepi K, Fry B, Sabeti P. Transforming Clinical Data into Actionable Prognosis Models: Machine-Learning Framework and Field-Deployable App to Predict Outcome of Ebola Patients. PLoS Negl Trop Dis 2016;10:e0004549. doi:10.1371/journal.pntd.0004549.

[25] Halpern Y, Horng S, Choi Y, Sontag D. Electronic medical record phenotyping using the anchor and learn framework. J Am Med Informatics Assoc 2016;23:731–40. doi:10.1093/jamia/ocw011.

[26] Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science (80-) 2015;349:255–60. doi:10.1126/science.aaa8415.

[27] Breiman L. Random forests. Mach Learn 2001;45:5–32. doi:10.1023/A:1010933404324.

[28] Cortes C, Vapnik V. Support-Vector networks. Mach Learn 1995;20:273–97. doi:10.1023/A:1022627411411.

[29] Hill T, Marquez L, O'Connor M, Remus W. Artificial neural network models for forecasting and decision making. Int J Forecast 1994;10:5–15. doi:10.1016/0169-2070(94)90045-0.

[30] Altman NS. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. Am Stat 1992;46:175–85. doi:10.1080/00031305.1992.10475879.

[31] Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science 2000;290:2323–6. doi:10.1126/science.290.5500.2323.

[32] Cardoso J-F. Blind signal separation: statistical principles. Proc IEEE 1998;86:2009–25. doi:10.1109/5.720250.

[33] de Bruijne M. Machine learning approaches in medical image analysis: from detection to diagnosis. Med Image Anal 2016. doi:10.1016/j.media.2016.06.032.

[34] Penatti OAB, Werneck R de O, de Almeida WR, Stein B V., Pazinato D V., Mendes Júnior PR, et al. Mid-level image representations for real-time heart view plane classification of echocardiograms. Comput Biol Med 2015;66:66–81. doi:10.1016/j.compbiomed.2015.08.004.

[35] Mohammed EA, Mohamed MMA, Far BH, Naugler C. Peripheral blood smear image analysis: A comprehensive review. J Pathol Inform 2014;5:9. doi:10.4103/2153-3539.129442.

[36] Li B, Li HK. Automated Analysis of Diabetic Retinopathy Images: Principles, Recent Developments, and Emerging Trends. Curr Diab Rep 2013;13:453–9. doi:10.1007/s11892-013-0393-9.

[37] Suzuki K. A review of computer-aided diagnosis in thoracic and colonic imaging. Quant Imaging Med Surg 2012;2:163–76. doi:10.3978/j.issn.2223-4292.2012.09.02.

[38] Cote RA. Architecture of SNOMED Its Contdbudon to Medical Language Processing ARCHITECTURE OF SNOMED SNOMED is Systematized Nomenclature of Medicine. Proc Annu Symp Comput Appl Med Care 1986:74–80.

[39] Cimino JJ, Sideli R V. Using the UMLS to bring the library to the bedside. Med Decis Making n.d.;11:S116–20.

[40] Friedlin J, McDonald CJ. A natural language processing system to extract and code concepts relating to congestive heart failure from chest radiology reports. AMIA Annu Symp Proc 2006;2005:269–73. doi:86418 [pii].

[41] Sager N, Lyman M, Bucknall C, Nhan N, Tick LJ. Natural language processing and the representation of clinical data. J Am Med Inform Assoc 1994;1:142–60.

[42] Tsuruoka Y, Tsujii J, Ananiadou S. Accelerating the annotation of sparse named entities by dynamic sentence selection. BMC Bioinformatics 2008;9 Suppl 11:S8. doi:10.1186/1471-2105-9-S11-S8.

[43] Chen ES, Hripcsak G, Xu H, Markatou M, Friedman C. Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study. J Am Med Inform Assoc n.d.;15:87–98. doi:10.1197/jamia.M2401.

[44] Murff HJ, FitzHenry F, Matheny ME, Gentry N, Kotter KL, Crimin K, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA 2011;306:848–55. doi:10.1001/jama.2011.1204.

[45] Wei W-Q, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. J Am Med Inform Assoc 2016;23:e20–7. doi:10.1093/jamia/ocv130.

[46] Pestian JP, Brew C, Matykiewicz P, Hovermale DJ, Johnson N, Cohen KB, et al. A Shared Task Involving Multi-label Classification of Clinical Free Text. Proc Work BioNLP 2007 Biol Transl Clin Lang Process (BioNLP '07) 2007;1:97–104.

[47] VanHouten JP, Starmer JM, Lorenzi NM, Maron DJ, Lasko TA. Machine learning for risk prediction of acute coronary syndrome. AMIA Annu Symp Proc 2014;2014:1940–9.

[48] Cronin RM, VanHouten JP, Siew ED, Eden SK, Fihn SD, Nielson CD, et al. National Veterans Health Administration inpatient risk stratification models for hospital-acquired acute kidney injury. J Am Med Inform Assoc 2015;22:1054–71. doi:10.1093/jamia/ocv051.

[49] National Center for Health Statistics. International classification of diseases, ninth revision (ICD-9). n.d. http://www.cdc.gov/nchs/icd/icd9.htm, 1979. (accessed July 22, 2016).

[50] Stoker MR. Common errors in clinical measurement. Anaesth Intensive Care Med 2008;9:553–8. doi:10.1016/j.mpaic.2008.09.016.

[51] Botsis T, Hartvigsen G, Chen F, Weng C. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. AMIA Jt Summits Transl Sci Proc AMIA Summit Transl Sci 2010;2010:1–5.

[52] Carroll RJ. Measurement Error in Epidemiologic Studies. Wiley StatsRef Stat. Ref. Online, Chichester, UK: John Wiley & Sons, Ltd; 2014. doi:10.1002/9781118445112.stat05178.

[53] Gaspar J, Catumbela E, Marques B, Freitas A. A SYSTEMATIC REVIEW OF OUTLIERS DETECTION TECHNIQUES IN MEDICAL DATA - Preliminary Study. Proc. Int. Conf. Heal. Informatics, SciTePress - Science and and Technology Publications; 2011, p. 575–82. doi:10.5220/0003168705750582.

[54] Marlin BM, Kale DC, Khemani RG, Wetzel RC. Unsupervised pattern discovery in electronic health care data using probabilistic clustering models. Proc. 2nd ACM SIGHIT Symp. Int. Heal. informatics - IHI '12, New York, New York, USA: ACM Press; 2012, p. 389. doi:10.1145/2110363.2110408.

[55] Lasko TA, Denny JC, Levy MA. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. PLoS One 2013;8:e66341. doi:10.1371/journal.pone.0066341.

[56] He Y. Missing data analysis using multiple imputation: getting to the heart of the matter. Circ Cardiovasc Qual Outcomes 2010;3:98–105. doi:10.1161/CIRCOUTCOMES.109.875658.

[57] Wilson BJ, President V, Bock A. The benefit of using both claims data and electronic medical record data in health care analysis n.d.:1–4.

[58] CANCER RESEARCH UK. Data collection and quality implications. Regist Eval Patient Outcomes A User's Guid 3rd Ed 2015. http://www.cancerresearchuk.org/health-professional/cancer-statistics/cancer-stats-explained/data-collection-implications#collapseZero (accessed July 21, 2016).

[59] Rubin DB. Inference and missing data. Biometrika 1976;63:581–92. doi:10.1093/biomet/63.3.581.

[60] Raghunathan TE. What Do We Do with Missing Data? Some Options for Analysis of Incomplete Data. Annu Rev Public Health 2004;25:99–117. doi:10.1146/annurev.publhealth.25.102802.124410.

[61] Steyerberg EW, van Veen M. Imputation is beneficial for handling missing data in predictive models. J Clin Epidemiol 2007;60:979. doi:10.1016/j.jclinepi.2007.03.003.

[62] Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009;338:b2393. doi:10.1136/bmj.b2393.

[63] Rusanov A, Weiskopf NG, Wang S, Weng C. Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research. BMC Med Inform Decis Mak 2014;14:51. doi:10.1186/1472-6947-14-51.

[64] Huisman M. Imputation of Missing Network Data: Some Simple Procedures. Encycl. Soc. Netw. Anal. Min., New York, NY: Springer New York; 2014, p. 707–15. doi:10.1007/978-1-4614-6170-8_394.

[65] von Hippel PT. Regression with Missing Ys: An Improved Strategy for Analyzing Multiply Imputed Data. Sociol Methodol 2007;37:83–117. doi:10.1111/j.1467-9531.2007.00180.x.

[66] Dong Y, Peng C-YJ. Principled missing data methods for researchers. Springerplus 2013;2:222. doi:10.1186/2193-1801-2-222.

[67] Saria S, Rajani AK, Gould J, Koller D, Penn AA. Integration of early physiological responses predicts later illness severity in preterm infants. Sci Transl Med 2010;2:48ra65. doi:10.1126/scitranslmed.3001304.

[68] Syed Z, Stultz CM, Scirica BM, Guttag J V. Computationally Generated Cardiac Biomarkers for Risk Stratification After Acute Coronary Syndrome. Sci Transl Med 2011;3:102ra95–102ra95. doi:10.1126/scitranslmed.3002557.

[69] Huopaniemi I, Nadkarni G, Nadukuru R, Lotay V, Ellis S, Gottesman O, et al. Disease progression subtype discovery from longitudinal EMR data with a majority of missing values and unknown initial time points. AMIA Annu Symp Proc 2014;2014:709–18.

[70] Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics 2010;26:1205–10. doi:10.1093/bioinformatics/btq126.

[71] Pathak J, Wang J, Kashyap S, Basford M, Li R, Masys DR, et al. Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE Network experience. J Am Med Inform Assoc 2011;18:376–86. doi:10.1136/amiajnl-2010-000061.

[72] Ritchie MD, Denny JC, Crawford DC, Ramirez AH, Weiner JB, Pulley JM, et al. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. Am J Hum Genet 2010;86:560–72. doi:10.1016/j.ajhg.2010.03.003.

[73] Denny J, Bastarache L, Ritchie M. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat Biotechnol 2013;31:1102–10. doi:10.1038/nbt.2749.Systematic.

[74] Johnson EK, Nelson CP. Values and Pitfalls of the Use of Administrative Databases for Outcomes Assessment. J Urol 2013;190:17–8. doi:10.1016/j.juro.2013.04.048.

[75] American Diabetes Association. 2. Classification and Diagnosis of Diabetes. Diabetes Care 2015;38:S8–16. doi:10.2337/dc15-S005.

[76] de Araújo Gonçalves P, Ferreira J, Aguiar C, Seabra-Gomes R. TIMI, PURSUIT, and GRACE risk scores: sustained prognostic value and interaction with revascularization in NSTE-ACS. Eur Heart J 2005;26:865–72. doi:10.1093/eurheartj/ehi187.

[77] Anderson AE, Kerr WT, Thames A, Li T, Xiao J, Cohen MS. Electronic health record phenotyping improves detection and screening of type 2 diabetes in the general United States population: A cross-sectional, unselected, retrospective study. J Biomed Inform 2016;60:162–8. doi:10.1016/j.jbi.2015.12.006.

[78] Yu S, Liao KP, Shaw SY, Gainer VS, Churchill SE, Szolovits P, et al. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. J Am Med Informatics Assoc 2015;22:993–1000. doi:10.1093/jamia/ocv034.

[79] Ho JC, Ghosh J, Sun J. Marble: high-throughput phenotyping from elecronic health records via sparse nonnegative tensor factorization. Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '14, New York, New York, USA: ACM Press; 2014, p. 115–24. doi:10.1145/2623330.2623658.

[80] Pivovarov R, Perotte AJ, Grave E, Angiolillo J, Wiggins CH, Elhadad N. Learning probabilistic phenotypes from heterogeneous EHR data. J Biomed Inform 2015;58:156–65. doi:10.1016/j.jbi.2015.10.001.

[81] Li L, Cheng W-Y, Glicksberg BS, Gottesman O, Tamler R, Chen R, et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. Sci Transl Med 2015;7:311ra174. doi:10.1126/scitranslmed.aaa9364.

[82] Biau G, Scornet E. A random forest guided tour. TEST 2016;25:197–227. doi:10.1007/s11749-016-0481-7.

[83] Liaw A, Wiener M. Classification and Regression by randomForest. R News 2002;2:18–22.

[84] Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. Proc. 23rd Int. Conf. Mach. Learn. - ICML '06, vol. C, New York, New York, USA: ACM Press; 2006, p. 161–8. doi:10.1145/1143844.1143865.

[85] Quinlan JR. No Title. Mach Learn 1986;1:81–106. doi:10.1023/A:1022643204877.

[86] Li R, Belford G. Instability of decision tree classification algorithms. Proc Eighth ACM SIGKDD … 2002:570–5.

[87] Liu W, Yang Y. Parametric or nonparametric? A parametricness index for model selection. Ann Stat 2011;39:2074–102. doi:10.1214/11-AOS899.

[88] Chen C, Liaw A, Breiman L. Using random forest to learn imbalanced data. 2004. doi:ley.edu/sites/default/files/tech-reports/666.pdf.

[89] Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest. BMC Med Inform Decis Mak 2011;11:51. doi:10.1186/1472-6947-11-51.

[90] Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L. The use of receiver operating characteristic curves in biomedical informatics. J Biomed Inform 2005;38:404–15. doi:10.1016/j.jbi.2005.02.008.

[91] Bickel JE. Some Comparisons among Quadratic, Spherical, and Logarithmic Scoring Rules. Decis Anal 2007;4:49–65. doi:10.1287/deca.1070.0089.

[92] Addiss DG, Shaffer N, Fowler BS, Tauxe R V. The epidemiology of appendicitis and appendectomy in the United States. Am J Epidemiol 1990;132:910–25.

[93] Roden D, Pulley J, Basford M, Bernard G, Clayton E, Balser J, et al. Development of a Large-Scale De-Identified DNA Biobank to Enable Personalized Medicine. Clin Pharmacol Ther 2008;84:362–9. doi:10.1038/clpt.2008.89.

[94] R Core Team. R: A Language and Environment for Statistical Computing 2013.

[95] Borchers HW. pracma: Practical Numerical Math Functions 2015.

[96] Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics 2005;21:3940–1. doi:10.1093/bioinformatics/bti623.

[97] Lasko TA. Nonstationary Gaussian Process Regression for Evaluating Clinical Laboratory Test Sampling Strategies. Proc. Twenty-Ninth AAAI Conf. Artif. Intell., 2015, p. 1777–83.

[98] Domingos P. A few useful things to know about machine learning. Commun ACM 2012;55:78. doi:10.1145/2347736.2347755.

[99] Moler CB. Numerical Computing with MATLAB: Revised Reprint. SIAM; 2008.

[100] Rasmussen CE, Williams CKI. Gaussian Processes for Machine Learning. 2006.

[101] Lasko TA. Efficient Inference of Gaussian Process Modulated Renewal Processes with Application to Medical Event Data. Uncertain Artif Intell 2014.

[102] Lasko TA, Bajor JM. The Diagnostic Power of Distributed Information in Clinical Billing Codes. 2016.

[103] Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 2010;33:1–22. doi:10.1359/JBMR.0301229.

[104] Spearman C. The Proof and Measurement of Association between Two Things. Am J Psychol 1904;15:72. doi:10.2307/1412159.

[105] Pearson K. Note on Regression and Inheritance in the Case of Two Parents. Proc R Soc London 2006;58:240–2. doi:10.1098/rspl.1895.0041.

[106] Hauke J, Kossowski T. Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data. Quaest Geogr 2011;30:87–93. doi:10.2478/v10117-011-0021-1.

[107] Schielzeth H. Simple means to improve the interpretability of regression coefficients. Methods Ecol Evol 2010;1:103–13. doi:10.1111/j.2041-210X.2010.00012.x.

[108] Tibshirani R. Regression Shrinkage and Selection via the Lasso. J R Stat Soc Ser B (Statistical Methodol 1996;58:267–88.

[109] Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 2010;33:1–22. doi:10.1111/j.1467-9868.2005.00503.x.

[110] Kahan BD, Shaw LM, Holt D, Grevel J, Johnston A. Consensus document: Hawk's Cay meeting on therapeutic drug monitoring of cyclosporine. Clin Chem 1990;36:1510–6.

[111] Jusko WJ, Thomson AW, Fung J, McMaster P, Wong SH, Zylber-Katz E, et al. Consensus document: therapeutic monitoring of tacrolimus (FK-506). Ther Drug Monit 1995;17:606–14.

[112] R Core Team. R: A Language and Environment for Statistical Computing 2013.

[113] Chalew SA, Hempe JM, McCarter R. Clinically significant disagreement between mean blood glucose and estimated average glucose in two populations: implications for diabetes management. J Diabetes Sci Technol 2009;3:1128–35.

[114] Colon-Otero G, Menke D, Hook CC. A practical approach to the differential diagnosis and evaluation of the adult patient with macrocytic anemia. Med Clin North Am 1992;76:581–97.

[115] Massey AC. Microcytic anemia. Differential diagnosis and management of iron deficiency anemia. Med Clin North Am 1992;76:549–66.

[116] Long AN, Dagogo-Jack S. Comorbidities of Diabetes and Hypertension: Mechanisms and Approach to Target Organ Protection. J Clin Hypertens 2011;13:244–51. doi:10.1111/j.1751-7176.2011.00434.x.

[117] Shulman NB, Ford CE, Hall WD, Blaufox MD, Simon D, Langford HG, et al. Prognostic value of serum creatinine and effect of treatment of hypertension on renal function. Results from the hypertension detection and follow-up program. The Hypertension Detection and Follow-up Program Cooperative Group. Hypertension 1989;13:I80–I80. doi:10.1161/01.HYP.13.5_Suppl.I80.

[118] Weiner DE. Chronic Kidney Disease as a Risk Factor for Cardiovascular Disease and All-Cause Mortality: A Pooled Analysis of Community-Based Studies. J Am Soc Nephrol 2004;15:1307–15. doi:10.1097/01.ASN.0000123691.46138.E2.

[119] Frank B, Gottlieb K. Amylase normal, lipase elevated: is it pancreatitis?. A case series and review of the literature. Am J Gastroenterol 1999;94:463–9. doi:10.1111/j.1572-0241.1999.878_g.x.

[120] Apte M., Wilson J. Alcohol-induced pancreatic injury. Best Pract Res Clin Gastroenterol 2003;17:593–612. doi:10.1016/S1521-6918(03)00050-7.

[121] Higgins JP, Higgins JA. Elevation of cardiac troponin I indicates more than myocardial ischemia. Clin Investig Med Médecine Clin Exp 2003;26:133–47.

[122] Matthews KA, Crawford SL, Chae CU, Everson-Rose SA, Sowers MF, Sternfeld B, et al. Are Changes in Cardiovascular Disease Risk Factors in Midlife Women Due to Chronological Aging or to the Menopausal Transition? J Am Coll Cardiol 2009;54:2366–73. doi:10.1016/j.jacc.2009.10.009.

[123] Graham LS, Parhami F, Tintut Y, Kitchen CMR, Demer LL, Effros RB. Oxidized lipids enhance RANKL production by T lymphocytes: Implications for lipid-induced bone loss. Clin Immunol 2009;133:265–75. doi:10.1016/j.clim.2009.07.011.

[124] Nelson ER, Chang C, McDonnell DP. Cholesterol and breast cancer pathophysiology. Trends Endocrinol Metab 2014;25:649–55. doi:10.1016/j.tem.2014.10.001.