Understanding and Reducing Clinical Data Biases

Daniel Fort

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

under the Executive Committee

of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2015

ABSTRACT

Understanding and Reducing Clinical Data Biases

Daniel Fort

The vast amount of clinical data made available by pervasive electronic health records presents a great opportunity for reusing these data to improve the efficiency and lower the costs of clinical and translational research. A risk to reuse is potential hidden biases in clinical data. While specific studies have demonstrated benefits in reusing clinical data for research, there are significant concerns about potential clinical data biases.

This dissertation research contributes original understanding of clinical data biases. Using research data carefully collected from a patient community served by our institution as the reference standard, we examined the measurement and sampling biases in the clinical data for selected clinical variables. Our results showed that the clinical data and research data had similar summary statistical profiles, but that there were detectable differences in definitions and measurements for variables such as height, diastolic blood pressure, and diabetes status. One implication of these results is that research data can complement clinical data for clinical phenotyping. We further supported this hypothesis using diabetes as an example clinical phenotype, showing that integrated clinical and research data improved the sensitivity and positive predictive value.

Table of Contents

List of Charts, Graphs, Illustrations

Acknowledgments

Thanks to my committee: Sue Bakken, Gil Kuperman, Rachel Richesson, Erin Holve.

Special thanks to advisors Chunhua Weng and Adam Wilcox.

Funding support: R01HS019853 (PI: Bakken), 5T15LM007079 (PI: Hripcsak), R01LM009886 (PI: Weng), UL1 TR000040 (PI: Ginsberg)

My family, especially Lisa and Sophia.

Dedication

Into the life of every PhD student a thesis must fall. This is dedicated to my wife, Lisa.

the house never sees

stones all turned, laid solid. forms

of firm foundation

1. Introduction and Significance

Overview

The vast amount of clinical data made available by pervasive electronic health records presents a great opportunity for reusing these data to improve the efficiency and lower the costs of clinical and translational research. A risk to reuse is potential hidden biases in clinical data. While specific studies have demonstrated positive value in research using clinical data, there are concerns about whether they are generally usable. This thesis is comprised of three aims which address measuring bias in and validation of a clinical dataset.

Specific Aims

Aim 1: Examining Clinical Data for Bias

Examine a clinical dataset for selection and measurement bias through comparison with a higher quality research dataset.

Aim 2: Validation of Existing Datasets

Build and evaluate a method to compare datasets through the results of randomly generated hypothesis tests.

Aim 3: Addressing Gaps and Opportunities

Explore the use of more advanced techniques to address gaps and opportunities presented by the first two aims.

This thesis examines a clinical dataset for bias first through comparison of summary statistics with a research-quality dataset from the same population served by our institution. Second, we present a method to validate existing datasets by looking at the answers they provide to simple hypothesis tests rather than their summary statistics. A third aim contains a handful of studies which address potential gaps and opportunities presented by the findings of the first two aims. A summary of all three chapters, their component studies, and conclusions is presented in Table 1-1.

Table 1-1: Study research questions and conclusions

	Study	Research Question	Conclusions
Aim 1: Examining Clinical Data for Bias			
	1A	*What are the selection and measurement biases in a electronic clinical dataset as compared to a higher quality research dataset?*	Our highly structured data and point measurements from a clinical process were not significantly different than our data from structured population survey.
	1B	*What is the performance of a diabetes phenotyping algorithm and its components be investigated using patient self-reported data?*	Complex clinical variables could not be considered accurate, but components might be used for different purposes
Aim 2: Validation of Existing Datasets			
	2A	*Can we build and evaluate a method to compare datasets through the results of randomly generated hypothesis tests?*	Method was designed and prototyped. Using this method, our research dataset is no more different than our clinical dataset than random samples from the clinical dataset are from each other.
	2B	*What is the effect of data missing at random on validity analysis?*	Data missing at random at levels found in our clinical dataset had little effect on the "accuracy" of the dataset, as defined in study 2A.
Aim 3: Addressing Gaps and Opportunities			
	3A	*Can missing data in a clinical dataset be replaced so that the "accuracy" of a dataset is improved?*	Data can be replaced using many different methods. No examined method demonstrated a significant improvement.
	3B	*Can nearest neighbor matching replace matching based on identifiable data?*	Some patients can be matched between data sources. However, nearest neighbor matching was not demonstrated to be a useful replacement for more exact methods.
	3C	*Can an individual's representativeness in a dataset be usefully represented with a point statistic?*	While the idea may merit further investigation, this score of representativeness was not a meaningful statistic to calculate in this case study.

The following section presents background on the issues of bias and validation in clinical data, and the explicit gaps which the studies of this thesis were intended to address. The rest of this chapter will briefly review relevant background, methods, results, and conclusions from each of these studies as collected aims. Each aim is covered in more detail in following, separate chapters.

Background

Electronic Clinical Data

Electronic clinical data refers to the large-scale capture of information collected as part of diagnosis, treatment, and monitoring of health related conditions of a patient population. Within this thesis the term is typically used to refer to structured information (such as simple measurements like height and weight), but should also be understood to include narrative information (such as notes) and even images.

Part of the promise of electronic clinical data is the acceleration of clinical research. Computational reuse of electronic clinical data has been frequently recommended for improving efficiency and reducing cost for comparative effectiveness research[1]. The $1.1 billion for CER provided by the American Recovery and Reinvestment Act demonstrated an investment to that change[2]. $44 million of that sum is directed toward building an infrastructure for the collection and integration of multiple sources of data, from clinical and lab data to ongoing population surveys, for long-term support of future CER[3]. The capacity to reuse data for future research is important because those kinds of supported studies, particularly retrospective observational studies, can be quicker and cost up to ten times less than randomized controlled trials[1].

Current State of Electronic Clinical Data Validation

A risk to the use of electronic clinical data for research is hidden biases in the clinical data. While specific studies have demonstrated positive value in clinical data research, there are concerns about whether the data are generally usable[4-8]. Opaque data capture processes and idiosyncratic documentation behaviors of clinicians from multiple disciplines may lead instances of measurement bias where values derived a clinical process are different from a direct research measurement. A difference in the population who seek medical care versus the general residential population may introduce a selection bias when clinical data are used to estimate population statistics. Differences in which values are measured in which patients may lead to bias encoded in patterns of missing data.

These potential problems are widely acknowledged, but they are difficult to evaluate. Comparison of data with a gold standard is by far the most frequently used method for assessing potential bias[9]. As reported by Hogan and Wagner, the gold standard for most evaluations of accuracy of electronic clinical data is the paper records of the same patients[10]. While the paper records may be of higher objective quality, they still represent an internal validation of the same measurement process and can provide no insight into potential selection biases. Individual variables or small subsets of data are sometimes validated against other portions of the same data system. Internal validation of this sort has been demonstrated in evaluations such as height and weight, race and ethnicity, or completeness of a problem list. Comparisons to alternate internal sources are sometimes referred to as 'relative gold standards'[11].

Validations of more complex variables, such as disease status, have been carried out using more external sources of data like billing data, registries, and various forms of patient self-report[8, 12-16]. In these evaluations, the clinical data are typically considered the gold standard, or at least of higher relative quality, than the other data sources. However, some recent evaluations registries and clinical data fragmentation have cast doubt on the idea that the clinical data sources are always better.

A final point is that, regardless of the gold standard used, none of the evaluations reviewed in this section can report a pure selection bias. Evaluations of variables against any gold standard for the same individuals can only report measurement bias. Evaluation of variables against a different population combine potential selection and measurement biases. Only by combining the two approaches, measurement of a different population and measurement of the same individuals within that population, could selection bias be parsed out. The possibility to make such a comparison was the opportunity for this thesis, and comprises study 1A.

Gaps in Electronic Clinical Data Validation

A fundamental gap in the current domain of electronic clinical data validation is the lack of an external gold standard. Through comparison to higher-quality research data, in study 1A we validated a basic set of electronic clinical data variables such as height, weight, and blood pressure, as well as an example of a complex disease variable, diabetes. Basic measurements of a cohort such as average height and blood pressure were largely similar between clinical and research data sources. Complex measurements or labels, such as diabetes status of an individual, was not.

However, alternate components or constructions of the electronic clinical data diabetes status, also known as a phenotyping algorithm, might be useful for constructing research cohorts for various purposes, for example high sensitivity or specificity. This conclusion corroborates the results of a paper by Richesson, et al, and further review suggested gaps in the typical evaluation of phenotyping algorithms, such as the eMERGE diabetes type 2 algorithm, which could be specifically addressed by study 1B[17]. Namely, validation of electronic phenotyping algorithms typically have small sample sizes and are performed without external gold standards. Additionally, validation of these phenotyping algorithms is performed using identified cases and controls with little visibility over which individuals may be excluded by these algorithms. Study 1B validated the eMERGE Diabetes Phenotyping Algorithm using patient self-reported diabetes status as an external standard to address both of these gaps.

Validating Datasets vs. Variables

While potential problems with electronic clinical data may be widely acknowledged and difficult to validate, research using electronic clinical data continues. Lacking gold standards, recent efforts have taken a more implicit approach to validating whole datasets in the form of study result replication. This kind of evaluation represents a shift from validating a dataset by comparing summary statistics with a reference to examining whether a dataset provides the same answers as a reference dataset. Groups such as HMORN, OMOP, and DARTNet assessed the accuracy of clinical data by comparing research results derived from clinical data with those derived from randomized controlled trials[18-20]. However, these projects reflect a focus on making a new system work rather than a lack of recognition of a potential problem.

Gaps in Validating Whole Datasets

One notable gap in dataset validation using published study results is the small size or scope of the validation. The DARTNet validation, for example, consisted of replicating a single statistical hypothesis[20]. The methods in study 2A were specifically developed to address this gap by evaluating the similarity between an electronic clinical dataset and its reference with approximately one hundred two-group hypothesis tests. Study 2A demonstrated that while there are differences in the “accuracy” of our clinical dataset as compared to our research-quality dataset, where accuracy is defined in terms of similarity between datasets of results to the body of hypothesis tests, the difference is no larger than random samples of clinical data are from each other.

Gaps in Effect and Treatment of Missing Data

The fraction of patients with no data for a given variable can be quite large in a clinical dataset. The potential effects of this missing data are widely discussed but rarely evaluated[21-23]. Study 1A revealed the level of missing data in our clinical dataset was quite high but that the distribution of missing values was primarily at random. Study 2B evaluates the effect of data removed at random in our clinical dataset using the validation methods developed for study 2A. The results in our clinical dataset suggest that "accuracy", as defined for study 2A, is not largely affected by levels of missing data typically found in clinical datasets.

Further Gaps and Opportunities

The studies of Aims 1 and 2 suggest further gaps and opportunities which might be addressed by more advanced techniques. Studies 3A, 3B, and 3C investigate three such opportunities, specifically in imputing missing data in clinical datasets, improving linkage between patients in different datasets, and computing the "representativeness" of a patient in a database.

For example, the structure of study 2B can also be used to evaluate imputation of missing data. In study 3A, missing data from all datasets used in study 2B were imputed using a variety of methods. The results of each method were then compared to "accuracy" of the non-imputed, missing at random dataset. In general, no imputation method performed better than simply leaving data points missing.

Study 1A, and any subsequent study relying on matching research survey participants to their own clinical data, relied on name and birthdate as identifiers. A way to match patients within de-identified data would be useful to researchers wishing to use a multiple data sources. Study 3B used nearest-neighbor matching to investigate whether non-personally identifiable information, such as health measurements like height, weight, and average blood pressure, could be useful in matching patients between data sources. While a small proportion of patients could be matched in this fashion, the technique cannot be recommended as a substitute for more exact methods at this time.

Study 1A also identified that neither the Research nor Clinical datasets were representative of the local population as described by census data. Study 3C investigated an adaptation of a propensity score to indicate the representativeness of a patient in a clinical dataset based on demographics and health indicators. A case study using this score suggested that inclusion and exclusion criteria to a study probably do more to influence the representativeness of a study cohort than any underlying selection bias in the data source.

Setting

The work for this thesis was performed at the Columbia University Medical Center, specifically Presbyterian Hospital and its Clinical Data Warehouse (CDW). This setting also provides a unique research opportunity in the form of the Washington Heights-Inwood Informatics Infrastructure for Community-Centered Comparative Effectiveness Research (WICER). The pair of these resources allows direct comparison of the research

(WICER) and clinical (CDW) data on the same geographic population and, for a smaller subset, the same individuals. The opportunity for this comparison was the direct motivation for studies 1A and 1B. These resources will now be reviewed.

Clinical Data Warehouse

The CDW is a large relational database compiling of much of the electronic patient data captured at this institution. Notes, treatment orders, diagnoses, lab test results, billing, demographic, and administration data from both ambulatory clinic and inpatient visits are available and matched to individual patients via a unique medical record number. Data from the CDW is used to support ongoing research, recruiting, and quality improvement activities. Data from the CDW is also used to support the institution's Meaningful Use attestation, which demonstrates completeness of the data. While most structured information is mapped to an internal Medical Entities Dictionary, mappings to widely used standards such as ICD-9 codes are also present. These mappings to more widely used standards make the implementation of something like the eMERGE diabetes phenotyping algorithm relatively easy.

Washington Heights-Inwood Informatics Infrastructure for Community-Centered Comparative Effectiveness Research

Research quality data on a population within Columbia University Medical Center's catchment area provides a unique opportunity to evaluate bias in the institution's electronic clinical data. The Washington Heights/Inwood Informatics Infrastructure for

Community-Centered Comparative Effectiveness Research (WICER) Project has been conducting community-based research and collecting patient self-reported health information[24]. The overall goal of the WICER project is to understand and improve the health of the community and its aims of the WICER Project revolve around the collection and use of data from multiple sources to integrate and make data available in a research data warehouse. One of these sources is a household survey with the goal of collecting information about social determinants of health, health seeking behaviors, as well as establishing some baseline health information collected in a community setting. While much of the survey data is self-reported, blood pressure, height, and weight are measured three times each by survey administrators. The Household Survey targets 3,500 households and approximately 6,000 individuals living in zip codes 10031, 10032, 10033, 10034, and 10040. The research team selected a random, weighted sample of households from each of the eight health districts covering these zip codes. The survey population was then extended via cluster sampling to adjacent households and via network (i.e., snowball) sampling through personal contacts of individuals within the household.

Because the WICER survey was administered to the same population served as the hospital, the selection criteria for the survey can be applied to the CDW as a basis for examining selection bias. The hospital records of survey participants who also have clinical data can be used to examine the clinical data for measurement bias. By combining these two results, the selection bias and measurement bias can be separated.

The WICER survey data is not without limitations. While the population sampling began with a random seed, the subsequent cluster and snowball sampling mean the original sample deviates in nonrandom ways from the population expected from the census distribution. Specifically, the WICER survey population contains more women and Hispanic individuals than would be expected based on census results. However, this sampling deviation is an excellent example of why the research data is of higher quality than the clinical data. Though it is not representative of the local population, because the sampling methodology was well-defined the nature and direction of the deviation can be explained.

An additional limitation is that the overlap in variables between the WICER dataset and the CDW is small. However, this overlap is also the opportunity of the study. With higher quality research data to compare to our clinical dataset, the magnitude and direction of potential bias in clinical data can be quantified.

Aims

This thesis is compromised of three aims and seven component studies. Methods, results, and conclusions will be reviewed in this chapter and expanded upon in subsequent chapters.

Aim 1: Examining Clinical Data for Bias

Aim: Examine a clinical dataset for selection and measurement bias through comparison with a higher quality research dataset.

Introduction

Aim 1 was comprised of two studies which examined a limited number of variables in a clinical dataset for selection and measurement bias versus a research-quality dataset. The goal was to calculate the differences in participant selection and measurement between using existing clinical data and prospective research data collection.

Methods

In the first study, study 1A, various clinical datasets were created to replicate the selection criteria for the WICER Community Survey and some demographic aspects of its resulting research cohort. Summary values for Age, Gender, Hispanic Ethnicity, Height, Weight, BMI, Smoking Status, Systolic and Diastolic blood pressure, and Diabetes Status from the same 18 month time period were statistically compared between samples. A second set of comparisons was also made between individuals with data from both sources, allowing us to distinguish which discrepancies in summary statistics must be the result of differences in the measurement process between the clinical and research data sources and which must be the result of sampling differences. See Table 1-2 for a summary of datasets used for these comparisons.

Table 1-2: Datasets and definitions for study 1A

Dataset Name	N	Description
Clinical Raw	**78,418**	Patients living within WICER zip codes with at least one visit recorded during WICER primary data collection
Clinical Sampled (U/H)	33,847 / 56,694	Averaged random samples from Clinical Raw to replicate demographics of Household dataset.
Filtered	60,258	Members of the Clinical Raw dataset with data for at least two measured variables
Complete Case	28,752	Members of the Clinical Raw dataset with data for all measured variables.
Household	**4,069**	WICER Community Survey participants
Matched	1,279	Population of patients with both Clinical and survey data.
Weighted		Version of any above dataset, re-weighted to age and gender distribution of the 2010 Census for the WICER zip codes

In the second study, study 1B, our example of a complex clinical variable, diabetes status, was examined in more detail. Diabetes status is computationally identified in a clinical dataset using a combination of diagnoses, medications, and lab values. These components, singly and in combination, were validated against each patient's self-reported diabetes status. See Table 1-3 for a summary of datasets used in this investigation.

Table 1-3: Datasets and definitions for study 1B

Dataset Name	N	Description
Patient Self-reported	2,249	Patients with available self-reported diabetes status
General Patient Population	786,893	Patients with one visit within the last five years. Includes patient self-reported cohort.

Results and Discussion

While the research data source was considered to be of higher quality due to the rigor of its sampling and data collection procedures, the resulting research dataset did differ from the expected local population as described by census data. Multiple alternate samples were created and tested to investigate the effect of this demographic discrepancy. There were no significant differences in results between alternate samples.

There is measurement bias present in Ethnicity, Height, Diastolic blood pressure, and Diabetes Status. There is a sampling bias in Age, Gender, and Smoking Status. There was no statistically significant difference in Weight, BMI, and Systolic blood pressure. The sampling and measurement biases in clinical data suggest three categories of clinical data variable. "Completely Accurate" variables are pieces of information such as address or birthdate which should remain the same for an individual regardless of data source. "Simple Measurement" variables are those like height, weight, and blood pressure which are the result of a single measurement or simple definition. While there may be systematic bias, the magnitude should be small. These two findings suggest datasets or analyses using highly structured data (e.g. age, gender) and point measurements (e.g. weight, blood pressure) collected from a clinical process should not have meaningfully different results than data collected as part of a structured research process.

The third category of data variable is "Inferred Information", or more complex labels which rely on multiple points of clinical data to infer a status like diabetes, and was the primary focus of study 1B. These could not be considered accurate for population summary purposes in this dataset, which is to say the summary values in the clinical dataset were very different from the research dataset, but parts of the clinical phenotype can be used to design a study toward different purposes such as maximizing sensitivity, specificity, or positive predictive value.

Aim 2: Validation of Existing Datasets

Aim: Build and evaluate a method to compare datasets through the results of randomly generated hypothesis tests.

Introduction

Aim 1 established that summary statistics for structured data and point measurements in our clinical dataset were not meaningfully different than our research dataset. However, most research using clinical data lacks the opportunity to use such direct reference data on the same individuals. Instead, a recent trend has been to validate an electronic clinical database by replicating a published finding or statistical result from another dataset. In study 2A, we expanded this concept into a method for comparing datasets through the results of multiple, randomly generated, two group hypothesis tests. This effort is considered preliminary in that the method was prototyped using a limited set of clinical data variables and without temporal considerations. In study 2B, we demonstrated the potential utility of the validation method by investigating the effect of data missing-at-random, a potential bias we could not effectively measure in Aim 1.

Methods

Datasets and Variables: This validation method was built using an electronic clinical dataset, our institution's CDW, and a population research dataset, the WICER Community Survey. Data variables include age, gender, height, weight, BMI, smoking status, diastolic and systolic blood pressure. To simplify issues of temporality, only data from the same 18-month time frame as the research survey was pulled for the clinical dataset. See Table 1-4 for a summary of datasets used to build and analyze this method.

Table 1-4: Datasets and definitions for study 2A

Dataset Name	N	Description
Research	4,069	Household dataset from Study 1A
Clinical	78,418	Clinical Raw dataset from Study 1A
Clinical Sample	4,069	Random sample of Clinical dataset with equal size to the Research dataset

Hypothesis Generation: The core of this method is the use of multiple two group hypothesis tests to create and compare a highly granular portrait of the internal significant differences in a dataset. An example two group hypothesis test is the comparison of average systolic for 70-year-old men vs. 40-year-old men. We suggest that it is a more meaningful comparison of datasets to say that the average systolic blood pressure of 70-year-olds is higher than the 40-year-olds of the same dataset, rather than comparing the average systolic blood pressure of 70-year-olds between two datasets. Approximately a hundred such two group hypothesis tests were generated to form a hypothesis library.

Classification: All hypothesis tests in the hypothesis library were calculated in a reference dataset (typically the Research dataset), then in a candidate set (Clinical or

Clinical Sample dataset). Hypotheses were classified depending on whether the result from the candidate set agreed or disagreed with that of the reference. 'Accuracy' is reported as the percent of hypotheses which agree. Classification was performed using the clinical dataset, research dataset, and random samples of the clinical dataset of the same size as the research dataset. This demonstration of classification marks the endpoint of study 2A.

Data Missing-at-Random: For study 2B copies of the complete case clinical dataset (a version of the clinical dataset where every patient has at least one value for every variable) were created with some amount of the data deleted at random. The hypothesis library was computed on each of this sets and the accuracy calculated using the original, complete clinical dataset as a reference. These accuracies were used to gauge the effect of data MAR on our clinical dataset. See Table 1-5 for a summary of datasets used in this investigation.

Table 1-5: Datasets and definitions for study 2B

Dataset Name	N	Description
Reference Dataset	28,752	Complete Case dataset from Aim 1
Candidate Dataset	28,752	Copy of Reference Dataset with data deleted at random to a particular target (i.e. 10% deletion)

Results and Discussion

In study 2A, the comparison of the population research dataset to a random clinical dataset of the same size had an accuracy of .77. A baseline comparison of random clinical samples with each other had an accuracy of .81. The distribution of classified hypotheses between both of these trials was compared with a chi-square test with a p-value of .64, meaning that the difference in accuracy between our research and clinical datasets is not significantly different than random samples of clinical data are from each other. These findings support the conclusions of Aim 1, that while differences may be detected between our clinical and research datasets those differences are not larger that those due to chance.

In study 2B, accuracy was calculated for datasets with 10% to up to 99.99% of the data deleted at random. At levels found in original clinical dataset (~60%), calculated accuracy was still 90%. This finding is interesting because it suggests that data MAR at levels typically reported for clinical data may not greatly affect data quality. Study 2B was also valuable because it demonstrated the possibility of additional uses for the prototyped validation method, particularly in the idea that the method might be used to compare data subsets back to their sources.

Aim 3: Addressing Gaps and Opportunities

Aim: Explore the use of more advanced techniques to address gaps and opportunities presented by the first two aims.

Overview

Aims 1 and 2 were designed to address identified gaps in the current practice of clinical data validation. However, the execution of these aims suggested further opportunities for improvement that fell outside of their original scope. Studies 3A, 3B, and 3C investigate three such opportunities, specifically in imputing missing data in clinical datasets, improving linkage between patients in different datasets, and computing the "representativeness" of a patient in a database.

Study 3A: Imputing Missing Data

Introduction: A variety of imputation methods are widely recommended to replace missing values in datasets. While study 2B suggested that missing data was not a significant problem in our clinical dataset, the methods and material of Aim 2 allowed us to quickly evaluate the effectiveness of these imputation methods in data MAR scenarios.

Methods: Methods and datasets from study 2B were copied and reused with one addition: in study 3A the randomly removed data in the test datasets were imputed in by a variety of methods. Imputation methods included various implementations of single value replacement, linear regression, multiple imputation, kNN, and expectation maximization. As in study 2B, the accuracy of these datasets was computed against the complete case clinical dataset. An given imputation method would be useful if the imputed dataset had a higher computed accuracy than the un-imputed, missing data dataset from which it was made. See Table 1-6 for a summary of datasets used in this investigation.

Table 1-6: Datasets and definitions for study 3A

Dataset Name	N	Description
Reference Dataset	28,752	Complete Case dataset from Aim 1
Candidate Dataset	28,752	Copy of Reference Dataset with data deleted at random to a particular target
Imputed Dataset	28,752	Copy of Candidate Dataset with deleted data imputed by some specific method

Results and Discussion: No method significantly improved the computed accuracy of an imputed dataset over the un-imputed dataset. This finding was surprising given the widespread recommendations for imputing missing data. While not every possible imputation method was tested, it may be that imputation methods are designed to mimic summary statistics and distributions of a complete dataset and not to add any of the meaningful information tested by our validation method. Additionally, it may be that these methods would show some benefit over baseline in a data missing not at random scenario. This topic needs further investigation.

Study 3B: Nearest Neighbor Matching

Introduction: A key component of Aim 1 was the comparison of clinical and structured research data on the same individuals. This comparison relied primarily on matching individuals using name and birthdate and could not have been performed using de-identified data. Nearest Neighbor matching, on the other hand, could potentially match individuals between de-identified datasets by finding the closest resembling person based on all other variables.

Methods: The spatial distance between each individual in the Clinical dataset to every individual in the Research dataset (and vice versa) was calculated. The "rank" of each match was calculated as the number of spatial matches which were nearer to the individual than their "true" match based on name and birthdate. Summary statistics on these ranks were reported. See Table 1-7 for a summary of datasets used in this investigation.

Table 1-7: Datasets and definitions for study 3B

Dataset Name	N	Description
Research Dataset	4,069	Household dataset from Aim 1
Clinical Dataset	28,752	Complete Case dataset from Aim 1

Results and Discussion: Out of nearly five thousand patients with matching records, 6% of research participants had their true matching clinical record as their closest clinical record. For 75% of survey participants, the true match was within the top 1,300 (out of nearly 29,000) clinical records. With an accurate matching rate of only 6%, nearest neighbor matching is not a useful replacement for more exact matching methods.

Study 3C: Propensity Scoring for Representativeness in a Dataset

Introduction: Aim 1 revealed that neither the Clinical nor Research datasets represented the demographics of local population as captured by census data. In addition, Aim 1 demonstrated that sicker individuals were more likely to show up in the clinical dataset, suggesting that the clinical dataset is especially misleading with regards to the health of the local population. A way to display the "representativeness" of an individual based on their health and demographics might be a useful metric to help understand the generalizability of a patient cohort.

Methods: A propensity score was calculated for each patient in our Clinical dataset based on representativeness in the Research dataset using age, sex, and health covariates of obesity and hypertension risk. The utility of this propensity score was examined by looking at a case study of blood pressure measurements in a clinic setting. See Table 1-8 for a summary of datasets used in this investigation.

Table 1-8: Datasets and definitions for study 3C

Dataset Name	N	Description
Research Dataset	4,069	Household dataset from Aim 1
Clinical Dataset	28,752	Complete Case dataset from Aim 1

Results and Discussion: The median propensity score was .85, suggesting the patient cohort was fairly representative, and there was no apparent difference in results when considering representativeness of patients. The case study also highlights the fact that the inclusion criteria of the study have a great deal more effect on the representativeness of the cohort than any underlying sampling bias in the clinical dataset. While the idea may merit further investigation, this score of representativeness was not a meaningful statistic to calculate in this case study.

Conclusion

This chapter has presented brief overview of the background, methods, results, and impact of the aims and studies of this thesis. The following chapters will go into more detail on each of the three main aims.

2. Examining Clinical Data for Bias

Aim 1: Examining Clinical Data for Bias

Examine a clinical dataset for selection and measurement bias through comparison with a higher quality research dataset.

Aim 1 was comprised of two studies which examined a limited number of variables in a clinical dataset for selection and measurement bias versus a research-quality dataset. The goal was to calculate the differences in participant selection and measurement between using existing clinical data and prospective research data collection. Study 1A used comparisons of summary statistics between data sources. Study 2B examined a particular example of a complex variable, diabetes, in greater depth. Additional comparisons involving missing data and categorical analysis are also presented in this chapter.

Study 1A

Introduction

There is a need to increase the pace of Comparative Effectiveness Research (CER). The $1.1 billion for CER provided by the American Recovery and Reinvestment Act demonstrates an investment to that change[2]. $44 million of that sum is directed toward building an infrastructure for the collection and integration of multiple sources of data, from clinical and lab data to ongoing population surveys, for long-term support of future CER[3].The capacity to reuse data for future research is important because those kinds of supported studies, particularly retrospective observational studies, can be quicker and cost up to ten times less than randomized controlled trials[1].

There may be unanticipated consequences to the secondary use of data which were not collected for research, particularly clinical data. Clinical data are collected to aid clinicians in diagnosis, treatment, and monitoring of health-related conditions. However, exactly what data values are collected and how they are measured depend on the clinical need. In contrast, data collected as part of a research effort is targeted to exactly what questions need to be answered and can be controlled for consistency and completeness.

An implication of reusing existing data in new research is the possible inclusion of whatever biases are present in the original data. These biases may impact new research conclusions even if they had no effect on the original purpose for which the data were collected[25-27]. The cause of a bias could be any of several mechanisms, such as selection bias, missing data, or measurement error, but the effect is a measurable difference between the sample in a data set and the underlying population which that sample is meant to represent. If such a bias is present, conclusions drawn from the dataset may not be true for the population as a whole. The severity of any biases determines whether the data are still useful.

Understanding the effect of bias in this way is important because it explains why clinical data, while perfectly appropriate for its originally collected purpose, may not be appropriate for research. Clinical data are collected with no concern if a patient is representative. The observed data are a product of health monitoring, diagnosis, or treatment, and so may not be consistent across the patient population or measured with the same rigor for each patient.

Research data, in contrast, are collected in such a way as to preserve genuine causal relationships. Population samples are selected in order to minimize selection bias or at least deviate from the underlying population in known ways. In a well-designed study data collection is designed to be as consistent across patients as possible.

Despite possible biases, clinical data are still very tempting to use. Even before the advent of Meaningful Use guidelines, clinical data were abundant both in number of patients covered and in the quantity of data available per patient. Given the cost of prospective data collection in healthcare, existing clinical data might also much cheaper to acquire. What is needed is an assessment of the extent of bias in clinical data and whether that bias makes clinical data inappropriate for research.

The methodology for assessing bias is derived from experimental and control group comparison in trial reporting. A presentation of baseline values for study groups is expected in a simple table. In a randomized controlled trial, for example, any difference between the intervention and control groups should be due only to chance. In this case statistical comparison between groups is recommended against[28]. In the case of our clinical and research datasets, however, we suspect differences between the two groups may be due to systematic selection and measurement biases and, therefore, statistical comparison will reveal these differences.

Concern about bias in clinical data is widely reported, but rarely assessed[4-8]. Typically, summary and baseline values are reported and straightforward statistical comparisons are performed to demonstrate significant differences between populations. The difference between these assessments of bias and those in trial reporting are the choice of groups. Census data may be useful for some demographics, but cannot be used to evaluate any clinical values and are therefore limited in assessing the extent of bias in clinical data[29]. Other evaluations focus on differences between sites or sources of clinical data to sub-populations which are already in a clinical dataset[30, 31]. The weakness of these evaluations is that they can demonstrate extent but not direction of any bias. What is therefore needed is a clinical dataset which has been selected to overlap with a research dataset to allow direct comparison of demographic and baseline values. We can meet this need with the CDW and the WICER Community Survey.

Setting

This work was performed within the Columbia University Medical Center (CUMC). The source of clinical data for this proposal was the CDW. The source of research data was the Washington Heights/Inwood Informatics Infrastructure for Community-Centered Comparative Effectiveness Research (WICER) Household Survey. CUMC is based in and serves the population of the Washington Heights / Inwood region of New York, implying that a subset of the CDW population will be drawn from the same population sampled by WICER.

CDW

The CDW is a large relational database compiling of much of the electronic patient data captured at this institution. Notes, treatment orders, diagnoses, lab test results, billing, demographic, and administration data from both ambulatory clinic and inpatient visits are available and matched to individual patients via a unique medical record number. Data from the CDW is used to support ongoing research, recruiting, and quality improvement activities. Data from the CDW is also used to support the institution's Meaningful Use attestation, which demonstrates completeness of the data.

WICER Household Survey

The aims of the WICER Project revolve around the collection and use of data from multiple sources to integrate and make available in a research data warehouse. One of these sources is a household survey with the goal of collecting information about social determinants of health, health seeking behaviors, as well as establishing some baseline health information collected in a community setting. While much of the survey data is self-reported, blood pressure, height, and weight are measured three times each by survey administrators. Survey data will be combined with matching participants' clinical information to create a longitudinal health record. Information from patients with both clinical and survey data will be made available via a "Research Data Explorer" for future research[32].

The Household Survey targets 3,500 households and approximately 6,000 individuals living in zip codes 10031, 10032, 10033, 10034, and 10040. The research team selected a random, weighted sample of households from each of the eight health districts covering these zip codes. The survey population was then extended via cluster sampling to adjacent households and via network sampling through personal contacts of individuals within the household.

Multiple versions of the survey have been administered. For the three versions administered since April 2012, the data variables for this proposed work appear to have been measured the same way. Data for the study will be queried across all three versions and combined.

Table 2-1: Datasets and definitions for Study 1A

Dataset Name	N	Description
Clinical Raw	**78,418**	Patients living within WICER zip codes with at least one visit recorded during WICER primary data collection
Clinical Sampled (U/H)	33,847 / 56,694	Averaged random samples from Clinical Raw to replicate demographics of Household dataset.
Filtered	60,258	Members of the Clinical Raw dataset with data for at least two measured variables
Complete Case	28,752	Members of the Clinical Raw dataset with data for all measured variables.
Household	**4,069**	WICER Community Survey participants
Matched	1,279	Population of patients with both Clinical and survey data.
Weighted		Version of any above dataset, re-weighted to age and gender distribution of the 2010 Census for the WICER zip codes

Datasets and Processing

This section reviews datasets used for this thesis. Specifically, the exact definitions for inclusion into any dataset, exact query definitions for clinical variables, and summarization and data processing steps will be discussed. A summary of datasets and their composition is presented in Table 2-1.

Raw Clinical

The Raw Clinical dataset represents the simplest attempt to replicate the WICER Community Survey results by replicating its selection criteria directly within the CDW. For each person >18 years old on 3/1/2012, living within the 5 WICER zip codes (10031, 10032, 10033, 10034, 10040), with at least one recorded visit between 3/1/2012 and 9/1/2013, the following variables were extracted from the CDW: Birthdate, Gender, Race, Ethnicity, Smoking Status, Height, Weight, Systolic and Diastolic blood pressure, Glucose Test Values, Diabetes ICD-9 codes, HbA1c values. In case of multiple recorded values between 3/1/2012 and 9/1/2013, all values were retrieved. The time points used are the beginning and end of primary data collection for WICER.

For primary comparison with the survey dataset, the following were calculated or carried forward: Age, Gender, Race, Ethnicity, Smoking Status, Average Height, Average Weight, BMI, Average Systolic and Diastolic blood pressure, Consensus_Diabetes (>1 Diabetes ICD-9 Code AND (>1 High Glucose OR >0 High HbA1c). The choice was made to average multiple values to best represent mimicking a population-based research study.

For smoking status, the most common answer was used. Smoking status is also a special case in that, unlike the continuous variables, the effect of missing data must be accounted for. In smoking status, a missing value may be simply missing or may denote negative smoking status. While the simple prevalence of smoking status was reported, an alternate value of “Prevalence of Smoking with Reported Status was also calculated but not reported. Exact CDW mappings for each variable are presented in Table 2-2.

Raw Survey

For each individual taking the WICER Community Survey in the Household setting, the following variables were extracted: Age, Sex, Race, Hispanic, Smoking Status, Height, Weight, Systolic and Diastolic blood pressures, Diabetes Status.

Processing

For each dataset, a number of secondary processing steps were taken to inform later investigations. Some steps were common to all datasets, some steps were specific to each dataset.

Common to all sets:

- A matching indicator variable was added and set to 1 if the individual is present in both survey and clinical datasets (as defined by a matching dictionary created by Adam Wilcox) and 0 otherwise.
- A partial (excluding cholesterol) Framingham risk score was calculated.

Table 2-2: Clinical variables, definitions, and local mappings

Variable	Definition	Mapping
AVERAGE_WEIGHT	Average weight recorded for the patient.	Values of flow sheet items descriptions starting with 'vs_weight%' from FS_WEST with a recorded time between 2012-03-01 and 2013-09-30 are stored for each MRN. These values are converted into float and averaged for each MRN.
AVERAGE_HEIGHT	Average height recorded for the patient.	Values of flow sheet items descriptions starting with 'vs_height%' from FS_WEST with a recorded time between 2012-03-01 and 2013-09-30 are stored for each MRN. These values are converted into float and averaged for each MRN.
BMI	Calculated BMI for the patient.	Basic BMI formula of weight (kg) / height (m) squared is calculated for each patient with both a height and weight.
AVERAGE_SYSTOLIC	Average systolic blood pressure recorded for the patient.	Values for rows with item names equal to 'vs_amb_intnal_med_NIBP_(s)', 'vs_bp_arterial_s', and 'vs_bp_noninvasive (s)' recorded time between 2012-03-01 and 2013-09-30 are retrieved and averaged for each patient. These specific item names were supplied by Adam Wilcox as corresponding to ambulatory blood pressures.
AVERAGE_DIASTOLIC	Average diastolic blood pressure recorded for the patient.	Values for rows with item names equal to 'vs_amb_intnal_med_NIBP_(d)', 'vs_bp_arterial_d', and 'vs_bp_noninvasive (d)' recorded time between 2012-03-01 and 2013-09-30 are retrieved and averaged for each patient. These specific item names were supplied by Adam Wilcox as corresponding to ambulatory blood pressures.
SMOKING	Consensus smoking status for the patient.	Values for rows with item names equal to 'amb_tobacco', 'amb_tobacco_use_MU', 'amb_intnalmed_tobacco', 'note_sw_initas_tobacco', 'amb_fam_plan_visit_soc_tobacco', 'md_ivcard_SocHx_tobacco', 'amb_aim_tobacco_use', 'amb_obgyn_visit_tobacco_use', 'amb_fam_plan_PT_soc_tobacco', 'amb_rheumatology_tobaccouse', 'note_EDNurAssess_smoking_hx', 'note_nsg_hx_smoking', 'note_EDAdltTemp_smoking_YN', 'note_UCC_Soc_smoking', 'note_dc_hx_smoking', 'note_cardiac_surg_SocHx_smoking', 'amb_ENT_CON_soc_hx_smoking', 'note_EDRME_smoking_YN' recorded time between 2012-03-01 and 2013-09-30 are retrieved for each patient. These values are automatically recoded to Yes, No, NA (where the value is unrelated to smoking), and Unknown (where the value couldn't be otherwise parsed). The value with the greatest tally is recorded as smoking status.

- Categorical indicator variables were added consisting of BMI (Underweight, Normal Weight, Overweight, Obese 1, Obese 2, Obese 3), Age by Decade (18-24, 25-34, 35-44, 45-54, 55-64, 75-84, 85+), and Hypertension Risk (Normal, Prehypertension, Stage 1, Stage 2). For each category, each person is assigned a 1 for the variable which includes them and a 0 for all other variables.

Survey datasets:

- If duplicates of name and date of birth for participants are detected, only the first recorded survey is kept.

Clinical dataset:

- Numerous other variables were collected for each individual. These include total number of diagnoses recorded, number of high HbA1c values recorded at any time in the patient record, number of high glucose values recorded at any time in the patient record, number of normal HbA1c values, and the number of visits, diagnoses, procedures, and labs recorded each for the last 5, 3, and 1 years. Alternate ways of classifying potential diabetes status were also calculated.

Sampling Bias

Summary

This section presents the background, methods, and results of an investigation into the sampling bias in a clinical dataset. The initial effort is to attempt various strategies for replicating the WICER research sample using the CDW, both by applying the same selection criteria as used in the Community Survey and by replicating the resulting demographics of the Community Survey. While significant differences in resulting population measurements can be detected, without a matched sample it is impossible to determine whether these differences arise from sampling or measurement bias.

Research Question: Can the population sample of a research study be replicated using a clinical population? Can you retrieve a cohort with the same demographic properties?

Background

The WICER Household Survey used snowball and network sampling on top of a stratified random population seed of individuals over the age of 18 within five zipcodes (10031, 10032, 10033, 10034, 10040). The survey participant population was found to deviate from the demographics of the known census distribution. Specifically, the survey population has a higher proportion of women (.71 vs .53 in Census) and is almost entirely Hispanic. One way of assessing potential problems with clinical data reused for research is to apply the same selection criteria as a population research study and determine whether the same kinds of people are selected.

Innovation

The selection bias of clinical patient populations has been studied. What is innovative about this study is the attempt to replicate the selection criteria of an existing research study for direct comparison, followed by alternate sampling methodologies to attempt to replicate the results of the that selection criteria in a particular population (for example, ensuring higher proportion of women and Hispanic individuals rather than anyone over the age of 18).

Methods

Raw Clinical

The selection criteria of the WICER Household survey were applied to the CDW to form the Raw Clinical dataset, described above.

Clinical Sampled (H)

To replicate the higher proportion of women and Hispanic individuals in the actual WICER Survey population, the Raw Clinical dataset was randomly sampled with exact gender and ethnicity targets. The procedure was repeated ten times, and the results averaged, to form the Clinical Sampled (H) dataset.

Clinical Sampled (U)

Later steps revealed that many Hispanic individuals were being labeled as "Unknown" ethnicity, so the sampling process was repeated to allow both labeled "Hispanic" and

"Unknown" ethnicities to meet the ethnicity target. This set is the Clinical Sampled (U) dataset.

Weighted

Finally, Raw Clinical and the WICER Community were re-weighted to the expected Census distribution for the 5 WICER zipcodes for both age and gender. For completeness, the weighting procedure was extended to both Clinical Sampled Datasets (U and H)

Comparing Datasets

Samples were primarily described by the following variables: Size of the set (N), Age, Proportion Female, Proportion Hispanic, Weight (kg), Height (cm), Prevalence of Smoking, Prevalence of Smoking among Labeled Status, Systolic and Diastolic blood pressure, Prevalence of Diabetes (self-reported status in the survey population, >1 Diabetes ICD-9 code AND (>0 abnormal HbA1c OR >1 abnormal glucose)).

For continuous variables, values for each patient for each dataset were averaged and compared via t-test. For categorical variables, proportions were compared via chi-square. For datasets with multiple samples (Clinical Sampled H/U), values were also averaged across all samples for comparison. Due to the number of statistical comparisons performed, a Bonferroni-corrected p-value of 1e-4 was used.

Statistical comparison was not performed for the re-weighted samples. It is a possible, but non-trivial task to compute standard deviations of a re-weighted sample.

Results

Summary statistics for the WICER Household, Raw Clinical, Clinical Sampled (H), and Clinical Sampled (U) populations are presented in Table 2-3. WICER Household population was known to contain a higher proportion of women than the clinical population (.71 vs .62) and Hispanic individuals (.96 vs .50). The Household population is also older (50.12 vs 47.55) than the clinical population and has high diastolic blood pressure (80.95 vs 73.07). The proportion of individuals self-identifying as diabetic in the Household population (.16) is wildly divergent from the proportion of clinical patients with diabetes according to this definition (.04). Prevalence of diabetes was later investigated in much greater detail using the eMERGE Diabetes phenotyping algorithm.

Table 2-3: Summary statistics for Household, Clinical Raw, and two Clinical Sampled cohorts

File	Household		Clinical Raw	Clinical Sampled (H)	Clinical Sampled (U)
N	4069		78418	33847	56694
Age	50.12		47.55	46.47	46.98
Proportion Female	0.71		0.62	0.71	0.71
Proportion Hispanic	0.96		0.50	1.00	0.60
Weight kg	75.42		75.69	74.69	74.75
Height cm	161.25		160.34	158.91	159.13
BMI	28.20		28.10	28.30	28.30
Prevalence of Smoking	0.06		0.09	0.08	0.08
Prevalence of Smoking Among Labeled Status	0.06		0.12	0.10	0.10
Systolic	127.68		127.23	126.75	126.48
Diastolic	80.95		73.07	72.46	72.72
Prevalence of Diabetes (Survey = self-report, Clinical = >1 Diabetes ICD-9 AND >1 abnormal test)	0.16		0.04	0.04	0.04

The clinical sampling process does change some of the summary statistics. Sampled populations are approximately 0.5-1.0 years younger, 1-2cm shorter, and weigh 1kg less.

The combined Census distribution for the 5 WICER zip codes for age and gender is presented in Table 2-4. These values were used to re-weight sampled to the expected Census distribution.

Table 2-4: Census gender and age distributions for WICER zip codes

Age Range	% Male	% Female
18-24	0.08	0.07
25-34	0.11	0.11
35-44	0.09	0.09
45-54	0.08	0.09
55-64	0.06	0.08
65-74	0.03	0.05
75-84	0.02	0.03
>85	0.01	0.01
total	0.47	0.53

Summary statistics for the re-weighted WICER Household, Raw Clinical, Clinical Sampled (H), and Clinical Sampled (U) populations are presented in Table 2-5. Weighted samples are 3-6 years younger than the original samples and are 53% female. Weighted samples tend to be slightly taller, weigh a little more, and be more likely to smoke. It is possible these characteristics are more prevalent in the younger portion of the population and were magnified by the re-weighting procedure. If so, this could represent a selection bias which was not detected by direct comparison between the Research and Clinical datasets.

Table 2-5: Summary statistics for Weighted samples

	Weighted Household		Weighted Clinical	Weighted Sampled (H)	Clinical Sampled (U)
N					
Age	44.63		44.13	44.11	44.10
Proportion Female	0.53		0.53	0.53	0.53
Proportion Hispanic	0.95		0.50	1.00	0.61
Weight kg	76.96		78.24	77.99	78.00
Height cm	163.68		162.70	162.11	162.50
BMI	27.74		28.11	28.25	28.15
Prevalence of Smoking	0.06		0.10	0.10	0.09
Prevalence of Smoking Among Labeled Status	0.07		0.14	0.12	0.13
Systolic	125.48		126.76	127.34	126.70
Diastolic	80.65		73.43	73.31	73.47
Prevalence of Diabetes (Survey = self-report, Clinical = >1 Diabetes ICD-9 AND >1 abnormal test)	0.12		0.03	0.04	0.04

Results of the statistical comparison between the Raw Clinical, Clinical Sampled (H) and Clinical Sampled (U) are presented in Table 2-6. The Bonferroni-corrected p-value is 1e-4. Age, Height, Smoking Status, and Diastolic blood pressure were all strongly significantly different. The proportion of patients with diabetes was also strongly significantly different. However, this represents a very simplistic diabetes phenotype and this portion of the investigation was later expanded using the eMERGE Diabetes phenotyping algorithm.

The proportion of female and Hispanic patients in the Raw Clinical population was significantly different than the Household population. By design, this difference is removed in the two Clinical Sampled populations. Weight, BMI, and Systolic blood pressure are insignificantly different across all three samples.

Table 2-6: P-values of comparison between Raw Clinical and Clinical Sampled datasets to the Household dataset

	p-values of Raw Clinical vs Household	p-values of Clinical Sampled (H) vs Household	p-values of Clinical Sampled (U) vs Household
Age	3.58E-12	8.52E-22	3.99E-17
Proportion Female	8.47E-08	0.755	0.749
Proportion Hispanic	9.76E-180	0.054	4.91E-86
Weight kg	0.851	0.016	0.016
Height cm	3.42E-07	1.3E-14	3.06E-13
BMI	0.207	0.924	0.167
Prevalence of Smoking	1.95E-10	2.22E-05	2.31E-05
Prevalence of Smoking Among Labeled Status	4.02E-32	4.13E-16	1.72E-18
Systolic	0.164	0.005	0.0002
Diastolic	0	0	0
Prevalence of Diabetes (Survey = self-report, Clinical = >1 Diabetes ICD-9 AND >1 abnormal test)	1.32E-241	5.23E-181	8.7E-209

Discussion and Conclusion

It is possible to apply the same selection criteria to a clinical database as used in a population research study. However, there are significant discrepancies in the demographics of the resulting dataset. Even when the demographic discrepancies are accounted for (via targeted sampling), there are significant differences in variables such as height and diastolic blood pressure. What is most surprising is that there are some variables, such as weight, BMI, and systolic blood pressure, which are not significantly different regardless of sampling method.

This first stage of the investigation only examined the effect of sampling. What is not known at this point is whether the apparent discrepancies between WICER Survey participants and the Clinical population are a result of simply sampling or whether the variables themselves are being measured differently.

Measurement Bias

Summary

By using a matched cohort, a group of individuals who participated in the Community Survey and also have clinical records in the CDW, it is possible to compare the measurement of individuals by a large clinical system to the focused, methodological measurement of a research study. However, this analysis of measurement bias comes at the expense of identifying selection bias.

Research Question: Given the same group of individuals, does the process of measuring people in a clinical environment and for a primary clinical purpose result in the same summary values as a population research study?

Background

Following the first stage of the investigation, we know that applying the same sampling criteria results in a population with different demographic properties than the research study. Likewise, after correcting for the demographic differences, there remain significant differences in variables such as height and diastolic blood pressure. By limiting the variable summary comparison to a group of the same individuals, it is possible to determine which discrepancies are a result of differences in measurement (as part of a clinical process vs part of a research study) and which are a result of the differences in sampling.

Innovation

Validation of individual variables, including the ones used in this study, have been performed. These validations are rarely performed on larger groups of variables, and never with the goal of parsing discrepancies in measurement from discrepancies in sampling.

Methods

The basic data, variables, and statistical comparison remain the same as the previous investigation. However, comparison was limited to only individuals who took a WICER Community Survey who also have at least one visit during the study period (3/1/2012 to 9/1/2013). Because this step in the investigation is about differences in measurement, rather than sampling, individuals who took the survey in the clinic (ACN) setting were also included to maximize cohort size. In the instance of a survey participant matching more than one clinical record, all clinical records were included in the clinical dataset.

A survey participant is considered ‘matched’ if there is a patient in the CDW with the same name and birthdate. Original matches were provided by Adam Wilcox. The body of resulting Survey data is the Matched Survey dataset and the clinical data for the same individuals is the Matched Clinical dataset. T-tests were performed on continuous variables, and chi-square tests on categorical variables, for comparison between datasets with a Bonferroni-corrected p-value of 1e-4.

Results

Summary and comparison statistics for Clinical and Survey measurements for the matched individuals are presented in Table 2-7. Only 12 participants matched more than one clinical record. The individuals in the dataset were matched on birthdate, so the apparent (and statistically insignificant) discrepancy in age is primarily a result of calculating age from birthday. Proportion Hispanic is significantly different and a result of many self-identifying Hispanic individuals being recorded as "Unknown" ethnicity in the clinical database. Recorded weights, BMI, Smoking, and Systolic BP are insignificantly different. Heights are approximately 3cm taller in the Survey data, and statistically significantly different. Diastolic blood pressure is approximately 5 points higher in the Survey and also statistically significantly different. The proportion of patients with diabetes was also strongly significantly different, however, this represents a very simplistic diabetes phenotype and this portion of the investigation was later expanded using the eMERGE Diabetes phenotyping algorithm.

Discussion / Conclusion

The measurements in a clinical environment and for a clinical purpose were different in some variables than measurement for research purposes of the same individuals. While it was possible that these discrepancies were in fact introduced by the investigation, the fact that the variables came in pairs which were retrieved and analyzed the same way suggests that the discrepancies were present in the data themselves. For example, height and weight were identified, retrieved, and analyzed in exactly the same fashion, yet heights showed a discrepancy and weights did not. Systolic and diastolic blood pressure were the same.

Table 2-7: Summary statistics and p-value of comparison between the Matched samples

File	Matched Clinical	Matched Survey	p-values of Matched vs Matched
N	1291	1279	
Age	52.33	51.12	0.072
Proportion Female	0.79	0.78	0.963
Proportion Hispanic	0.56	0.94	8.17E-17
Weight kg	77.16	76.99	0.851
Height cm	158.23	161.31	3.42E-07
BMI	29.70	28.90	0.207
Prevalence of Smoking	0.08	0.08	0.944
Prevalence of Smoking Among Labeled Status	0.09	0.08	0.283
Systolic	128.48	127.50	0.204
Diastolic	74.34	79.24	8.68E-25
Prevalence of Diabetes (Survey = self-report, Clinical = >1 Diabetes ICD-9 AND >1 abnormal test)	0.09	0.22	3.91E-15

The more immediate value of this stage was in ethnicity and smoking. By using the same individuals it is possible to demonstrate that people who self-identify as Hispanic are being labeled incorrectly in the clinical process. Conversely, smoking status, which was very different between the unmatched Clinical and Survey populations in the prior stage of the investigation, is not significantly different here. This finding suggests that smoking status is accurately labeled in the clinical process and the difference detected in the previous stage represents a difference in the population samples and not their measurement.

Combining Results

Summary

By combining the previous two sets of results, it becomes possible to parse the differences between datasets into selection bias and measurement bias. Summary values which are different in the Matched cohort comparison must be the result of measurement bias. Summary values which are different in the sample comparison but which were not measured differently in the Matched comparison must therefore be the result of sampling bias. That nested comparison is performed in this section. The sampling and measurement biases in this clinical dataset suggest three categories of clinical variable: completely accurate, simple measurement, and inferred information. The nature and implications of these variables are discussed.

Additionally, putting the results together in this fashion allows some sensitivity analysis into alternate data point selection and summarization steps. In general none of the alternatives had a significant effect any of the results. Alternatives and their effects are discussed in detail. This further detail includes discussion about the components of the ad hoc diabetes phenotyping algorithm used in the first part of this thesis. The investigation and discussion of the ad hoc diabetes phenotyping algorithm led directly to the following, supplemental study on the eMERGE phenotyping algorithm.

Methods

Diabetes Phenotyping Algorithm

A simple clinical phenotyping method was developed for type 2 diabetes in the CDW using ICD-9 Codes, HbA1c test values, and glucose test values. Using the strictest criteria, a patient will only be identified as having diabetes if there are at least two ICD-9 codes for diabetes, at least one HbA1c test value >6.5, or at least two high glucose test values. A glucose test value is coded as high if it is >126 for a fasting glucose test or >200 otherwise. Effectiveness of labeling of each of these components was also explored.

Data Point Selection

Each clinical variable could have many data points from multiple points of measurement across time, which necessitated careful data point selection to ensure that summary data points were both representative of all data points and comparable across data sources without introducing data sampling biases. This includes an issue of temporal bias, where some data variables, such as weight, might naturally be expected to change over time. To make a comparable cross-section to the Survey dataset and to ensure the resulting data reflects not only the same sample but also the same sample at the same time, we selected only data points recorded during the 18-month WICER study period from the CDW. In this way, assuming the survey participants are measured at random throughout an 18-month period, so too are the clinical data population.

In the matched sample we had an opportunity to more finely tune the data comparison. The most direct approach is to simply select the clinical data point closest in time to the survey measurement of any given participant. Alternatives include the closest prior or subsequent data paint as well as using a single randomly selected point rather than the average of all clinical data points. While alternate data point selection options were explored, to best keep the results comparable the reported values for the matched sample were derived in the same fashion as for the sample at large.

Data Measure Selection

With representative patient sample, meaningful variables, and representative data points, the next important step for designing an unbiased verification study was to select a meaningful data measure, which seems to be the most subjective step without standard guidance. For this step, we considered two measures: (a) population-level average summary statistics; and (b) patient-level average summary statistics.

Option (a): Population-Level Average summary statistics

Multiple data values available during the study period were averaged in order to minimize any temporal effects while also allowing the use of the most number of patients. Continuous variables within each set were averaged, with one exception, and compared via t-test. The median BMI value was used for comparison as the mean summary value for the calculation of BMI is more susceptible to outliers. Choice of other "best matching" clinical data values, such as the closest prior and subsequent values in time as well as simple random choice, were also explored.

Proportions of interest, which include % female, % smoking, and % Hispanic, for the categorical variables were reported and compared with chi-square test. For some proportions there is a possibility that negative or healthy status might not be recorded and would therefore be accurately represented by missing data. Therefore for smoking and diabetes there is a second value reported: the proportion of labeled status, which excludes any patient with missing data rather than assume missing data denotes known negative status.

For the purpose of primary analysis, only the strictest, ALL criteria for diabetes diagnosis are reported, as consistent with the eMERGE criteria. However, each component of the diabetes diagnosis was examined for sensitivity, specificity, and positive predictive against the patient's self-reported diabetes status. All summary and statistical comparisons were performed in Python, using the SciPy scientific computing package for statistical comparisons.

Option (b): Patient-level Average Summary Statistics

When there is sufficient clinical data, it is possible to create a distribution of expected values for a given patient and compare the survey value to that distribution. At its simplest, the comparison is simply whether the survey value is within one standard deviation of the mean of the available clinical values. This process was performed for patients with at least five data points for the same variable recorded during the study period.

Results

Following the population summary approach, values and statistics for each data point are presented in Table 2-8. Here, the interior two columns of summary statistics are the Matched cohorts and the exterior are the Raw Clinical and Survey datasets. Variables where values for the interior columns are significantly different (with a Bonferroni corrected p-value of 1e10-4) represent instances of measurement bias. Variables where the values for the exterior columns are significantly different but the interior columns are not are instances of sampling bias. The Survey dataset tends to be slightly older and contain more women. Survey participants were almost entirely identifying as Hispanic. Sixteen percent of the survey participants self-identified as having diabetes. Measuring the Matched dataset via clinical data and primary survey collection processes broadly records the same values. There are statistically significant measurement discrepancies in Hispanic ethnicity labeling, height measurement, diastolic blood pressure, and diabetes status determination. Where the Clinical and Survey datasets differ, in age, proportion of women, and prevalence of smoking, are evidence of statistically significant differences in sample composition.

In exploring patient-level summary statistics, the number of patients with sufficient data to construct a distribution of expected blood pressures was 866. Of these, 491(57%) and 479(55%) had a survey systolic or diastolic blood pressure, respectively, greater than one standard deviation away from their clinical mean. Table 2-9 shows an example result of alternate data point selections in Systolic BP. While values are statistically significantly different from one another in this and other examples, they would not change the conclusions drawn from Table 2-7.

The sensitivity, specificity, and positive predictive value of various strategies to identify diabetes status using clinical data are presented in Table 2-10. In this simple phenotype, ALL is the intersection of three criteria and ANY is the union. The three criteria are having at least two ICD-9 codes for diabetes, one high HbA1c value, and at least two high glucose values. The rationale for requiring two of some categories is to restrict potentially spurious results. In the case of diagnostic codes, for example, a diabetes ICD-9 code might be recorded for a negative diabetes evaluation. The removal of these restrictions was also considered. The ALL criteria have the highest positive predictive value, but the lowest sensitivity. Both the ICD-9 and HbA1c-based criteria have high specificities and the ICD-9 based criteria alone have the highest F-measure for sensitivity and specificity. Proportions of patients retrieved under each qualifying criteria are consistent with published results[17].

Table 2-8: Summary statistics and p-values of comparison for both Clinical and Survey samples and the Matched samples to parse measurement and selection bias

	Raw Clinical	Matched Clinical	Matched Survey	Survey		Matched vs. Matched	Clinical vs. Survey
Age	47.55	52.33	51.12	50.12		0.072	p << .0001
Proportion Female	0.62	0.79	0.78	0.71		0.963	p << .0001
Proportion Hispanic	0.50	0.56	0.94	0.96		p << .0001	p << .0001
Weight kg	75.69	77.16	76.99	75.42		0.851	0.851
Height cm	160.34	158.23	161.31	161.25		p << .0001	p << .0001
BMI	28.10	29.70	28.90	28.20		0.207	0.207
Prevalence of Smoking	0.09	0.08	0.08	0.06		0.944	p << .0001
Systolic	127.23	128.48	127.50	127.68		0.204	0.164
Diastolic	73.07	74.34	79.24	80.95		p << .0001	p << .0001
Prevalence of Diabetes (Survey = self-report, Clinical = >1 Diabetes ICD-9 AND >1 abnormal test)	0.04	0.09	0.22	0.16		p << .0001	p << .0001

Table 2-9: Example summary statistics for alternate data point choice. While summary statistics can be significantly different from each other, alternate choices would not have changed the conclusions of the study

Systolic Blood Pressure	Research Value	Clinical Data Point Choice			
		Closest Prior	Closest Subsequent	Random Point	Mean of Available Values
N	1290	1107	962	1185	1185
Mean	127.8	127.9	130.3	129.3	128.5

Table 2-10: Sensitivity, specificity, F-measure, and Positive Predictive Value of simple diabetes phenotype and its components

Value	ALL (ICD-9 AND HbA1C AND Glucose)	ANY (ICD-9 OR HbA1C OR Glucose)	≥1 ICD-9	≥2 ICD-9	HIGH HbA1C	HIGH Glucose
Sensitivity	0.33	0.81	0.90	0.84	0.48	0.52
Specificity	0.98	0.35	0.88	0.93	0.96	0.74
F-measure	0.49	0.49	0.89	0.88	0.64	0.61
Positive Predictive Value	0.82	0.27	0.68	0.78	0.79	0.37

Discussion

Our study shows discrepancies between clinical and research data, both in sampling and measurement. Clinical measurement of some data, such as gender and BMI, accurately reproduces the research measurement and others, such as diabetes, do not. While raw results may be interesting, because of the limits of overlapping data between sets and the comparisons which could be made, the raw results may have little value outside of this case study. If these discrepancies can be considered as representative of classes of clinical data, we can abstract some idea of generalizable accuracy of clinical data as compared to primary research data. We introduce three categories of accuracy.

The first category is "completely accurate" information, such as sex, birthdate, and therefore age. These data might be considered Personally Identifiable Information (PII), or information that on its own could be used to identify an individual. This classification suggests that address, social security number, and phone number would also be accurate between datasets. While there will be instances of coding error, misreporting, or other errors, by and large these data are consistent across datasets. It should be noted that birthdate was one of the criteria by which individuals were identified for the Matched, and therefore errors in the recording of birthdate would be excluded from this analysis. Also, while PII should be accurate across datasets, this does not suggest that all demographic information, such as ethnicity, will be accurate.

The second category is 'simple measurement' information, which is the result of a clear concept or measurement process. Height, weight, systolic and diastolic blood pressure, smoking status, and ethnicity are included in this category. Here, the simplicity of the measurement or concept leads to agreement in the value between sources, and differences in the value are the result of a difference in either the concept definition or the measurement process. For example, measured heights in the Matched group differ by approximately 2.5cm or 1in, suggesting that the concept and measurement of height in the Survey sample includes shoes. Likewise, diastolic blood pressure is consistently measured 5 points higher in the Survey sample, suggesting a difference in measurement. Ethnicity, which is self-reported in the survey, is labeled by hospital staff during admission to the hospital, resulting in approximately one third of Hispanic individuals being labeled as 'Unknown' ethnicity in the Clinical sample.

The final category of accuracy is 'inferred' information, where a complex concept, such as diabetes, is inferred from multiple variables. When compared with self-reported Survey values, no single prediction or combination of variables can be considered accurate for an entire cohort. However, some results may be useful enough for a specific purpose. For example, requiring ALL criteria has a high positive predictive value and may provide a high level of accuracy within a given cohort. Conversely, using just HbA1c measurements has a high sensitivity and may be most valuable when a larger quantity of data is required for statistical power.

At least in this case study, discrepancies in the 'simple measurement' category are stable across multiple sampling methodologies. Discrepancies are also stable when samples are broken down into categories such as age by decade, obesity classification, and hypertension risk category. This stability is what would be expected if the discrepancies were the result of simple measurement error and would suggest these discrepancies represent systematic bias in the clinical data. It is possible that reported discrepancies are the result of data retrieval and processing. However, the presence of pairs of measurements such as weight/height and systolic/diastolic blood pressure, retrieved and processed in an identical manner, where one is accurate and one not, suggests the discrepancies are truly present in at data source. Due to the limitations of this case study, it is unclear how generalizable this finding may be.

The choice of exact data points may also influence study results, so care must be taken in accurately summarizing patient data. In this study, the biggest apparent difference was between closest prior and subsequent data points. The reason may be that closest prior data point represents the end of a series of blood pressures which began with a hospitalization and is, therefore, the nearest to "normal". The closest subsequent data point, however, would represent the initial data collection of a hospitalization and would likely reflect a health crisis. Furthermore, defining allowable data points in time restricts the number of patients, who qualify for comparison. Using the average value for each patient smoothens out these temporal effects and allows the use of the maximum number of patients for comparison.

Study 1B: eMERGE Type II Diabetes Phenotyping Algorithm

Introduction

One popular use case for EHR data is to identify patients for care management or research, prospectively, or as part of retrospective cohort for study. In this context, cohort identification using EHR data is known as EHR phenotyping. The Electronic Medical Records and Genomics (eMERGE) consortium is a current multi-site research network sponsored by the National Institutes of Health of the United State. This network develops precise and portable phenotyping algorithms using heterogeneous EHR data[33]. To improve algorithm portability across different EHR systems, the design and evaluation of EHR phenotyping algorithms have relied on collaboration across institutions. For example, the eMERGE Type 2 Diabetes Mellitus (DM2) Case and Control algorithms were developed collaboratively by five institutions, resulting in the identification of over three thousand cases and controls to support a genome-wide association study (GWAS) on diabetes patients[34, 35]. The algorithm uses commonly captured EHR data elements for diagnosis, medications, and lab values to identify Type 2 diabetics. The emphasis on portability imposes a tradeoff due to the inherent data quality issues of those commonly captured EHR data elements. For example, ICD-9 billing codes are a coarse representation for nuanced narrative notes, medication orders do not necessarily reflect medication adherence, and as reported by Wei et al., EHR data fragmentation could negatively impact clinical phenotyping[36]. Moreover, the EHR data may not actually reflect patient perceptions of their own health.

The eMERGE DM2 algorithm was originally validated using chart review. The expense of chart review typically limits sample size and only 50-100 each for cases and controls were reviewed in this example[34, 35]. Moreover, the chart review process does not sample from patients excluded from the case and control groups, meaning that a true sensitivity for identification of diabetes cases may not be established. Finally, chart review is still internal validation, implying the reference standard is still limited to information captured within the EHRs of related institutions[36]. Richesson et al. compared the identified individuals from different diabetes phenotyping algorithms[17]. While different algorithms might be created for different purposes, for example maximizing sensitivity for a registry versus specificity for a genetic study, the results do suggest that any given algorithm may fail to identify all diabetics in a database.

With the increasing emphasis on patient and community engagement for clinical research, self-reported diseases status has risen as an alternative data source for clinical phenotyping. These data are usually collected directly from patients, as opposed to EHR data that reflect the perceptions of health care providers. Prior studies checked the self-reported diabetes status against EHR data and achieved sensitivities around 0.75, and specificities around 0.9[12, 14, 16, 37].

While pieces of patient self-reported data have informed specific elements of clinical data used for phenotyping, such as self-reported smoking rate[38] and date of diagnosis[39], little is known about how self-reported disease status data might be useful for clinical phenotyping. Both EHR data and patient self-reported health data have advantages and disadvantages for patient identification. We faced an unusual opportunity to address this research question. The Washington Heights/Inwood Informatics Infrastructure for Community-Centered Comparative Effectiveness Research (WICER) Project has been conducting community-based research and collecting patient self-reported health information[24]. A subset of surveyed individuals have clinical information stored at the Columbia University Medical Center, allowing direct comparison of diabetes status derived from clinical data to the self-reported diabetes status. Therefore, in this study we will validate the eMERGE DM2 Case algorithm using patient-reported diabetes status. This study is part of a larger research effort to use research data to verify clinical data accuracy.

Methods

1. Data Collected by WICER

The survey collected information about social determinants of health and health seeking behaviors as well as established some baseline health information. Survey participants were explicitly asked whether they had been told they had diabetes, high blood sugar, or sugar in the urine when not pregnant. The answer to this question was extracted as the self-reported diabetes status.

2. Data Collected by the Columbia University Clinical Data Warehouse

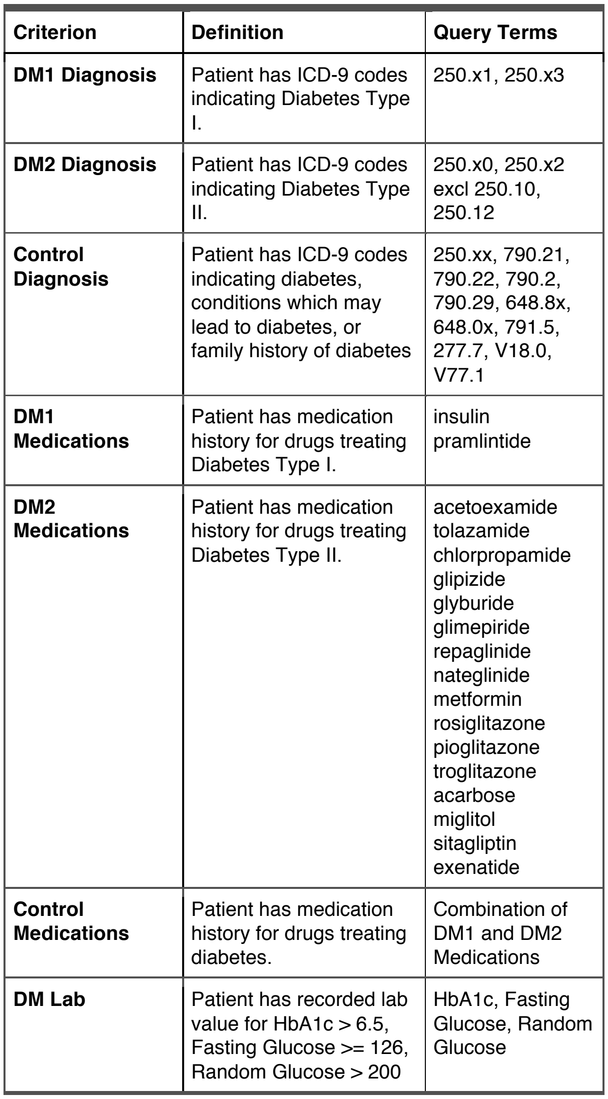
Data from the CDW were used to compute the case and control labels from the eMERGE phenotyping algorithm.

3. The eMERGE DM2 Case and Control Algorithms

As stated above, the eMERGE DM2 Case algorithm consists of three sets of criteria: diagnosis, medications, and lab values4. Diagnosis and medication criteria have components which indicate Diabetes Mellitus Type I (DM1) or Type II. Only patients with DM1 ICD-9 codes were completely excluded from the Case algorithm. For the purpose of this study, any patient reporting positive diabetes status who also had DM1 ICD-9 codes had their status reset to negative. DM1 medications only denote insulin dependence, which may also be found in DM2, and so additional logical criteria are required.

In contrast, the criteria for the eMERGE DM2 Control algorithm are very similar to the case algorithm, with the exceptions that no effort is made to distinguish between the types of diabetes (i.e., I or II), and the range of ICD-9 codes for the diagnostic criteria is expanded to include observations that co-occur with Type 2 diabetes. Criteria and their definitions are presented in Table 2-11.

Table 2-11: DM2 Phenotyping Algorithm components and definitions



Criterion	Definition	Query Terms
DM1 Diagnosis	Patient has ICD-9 codes indicating Diabetes Type I.	250.x1, 250.x3
DM2 Diagnosis	Patient has ICD-9 codes indicating Diabetes Type II.	250.x0, 250.x2 excl 250.10, 250.12
Control Diagnosis	Patient has ICD-9 codes indicating diabetes, conditions which may lead to diabetes, or family history of diabetes	250.xx, 790.21, 790.22, 790.2, 790.29, 648.8x, 648.0x, 791.5, 277.7, V18.0, V77.1
DM1 Medications	Patient has medication history for drugs treating Diabetes Type I.	insulin pramlintide
DM2 Medications	Patient has medication history for drugs treating Diabetes Type II.	acetoexamide tolazamide chlorpropamide glipizide glyburide glimepiride repaglinide nateglinide metformin rosiglitazone pioglitazone troglitazone acarbose miglitol sitagliptin exenatide
Control Medications	Patient has medication history for drugs treating diabetes.	Combination of DM1 and DM2 Medications
DM Lab	Patient has recorded lab value for HbA1c > 6.5, Fasting Glucose >= 126, Random Glucose > 200	HbA1c, Fasting Glucose, Random Glucose

4. Cohort Identification

Patient data were extracted for every patient in the CDW for 2009-13. We chose this time window to replicate the time scale used by Richesson, et al. and to accommodate the fact that the medication data in our data warehouse are not complete prior to 2009. A subset of CDW patients who also have a WICER-recorded diabetes status was identified for validation of the eMERGE DM2 Case algorithm. The remainder of the CDW population was used to investigate potential differences between the self-reported population and the general data population.

5. Data Element Extraction for Each Cohort

Table 2-12 presents the variables and definitions required for cohort identification and comparison using the eMERGE Case and Control algorithms. For each patient in a dataset, the data elements in Table 2-12 were either extracted or calculated. The self-reported diabetes status for each individual was extracted from their survey response and included in the patient level data. For the purpose of this study, any patient reporting positive diabetes status who also had DM1 ICD-9 codes had their self-reported status reset to negative.

6. Analysis Plan

Several groups of patients were collected for comparison from both the subset of patients with self-reported diabetes status and general patient population. These groups are the patients identified by the eMERGE DM2 Case algorithm (eMERGE Case), the pool of potential cases meeting any of the diagnostic, medication, or lab value criteria (Case Pool), and those patients meeting none of the criteria (Excluded). For patients with self-reported status, patients responding "Yes" and "No" were also separated for analysis. The number of patients, fraction of patients who are female, average and standard deviation for age, number of visits, and time between the first and last recorded visit for each group were reported. For groups of patients with self-reported status, the number of patients identifying as diabetic was also reported. Summary values for each group were quantitatively described and compared.

Table 2-12: Patient level data variables and definitions

Variable	Definition
Sex	Sex of the patient.
Age	Age in years on 1/1/2014.
Visits	Number of visits between 2009 and 2013.
Span	Length of time in days between first and last recorded visit.
DM1 Diagnosis	Number of ICD-9 codes meeting the Diabetes Type I diagnostic criteria.
DM2 Diagnosis	Number of ICD-9 codes meeting the Diabetes Type II diagnostic criteria.
Control Diagnosis	Number of ICD-9 codes meeting the Control algorithm diagnostic and family history exclusion criteria.
DM1 Medication	Earliest prescription date for medication meeting the Diabetes Type I medication criteria.
DM2 Medication	Earliest prescription date for medication meeting the Diabetes Type II medication criteria.
Control Medication	Number of medication orders meeting the control algorithm exclusion criteria.
Glucose Tests	Number of glucose test values recorded for the patient.
Abnormal Labs	Number of lab results high enough to indicate diabetes.
Diagnosis Criteria	1 if the patient meets the diagnostic criteria for Diabetes Type II, 0 otherwise.
Medication Criteria	1 if the patient meets the medication criteria for Diabetes Type II, 0 otherwise.
Lab Value Criteria	1 if the patient meets the labs criteria for Diabetes Type II, 0 otherwise.
Case	1 if the patient is identified by the eMERGE Case algorithm, 0 otherwise.
Control	1 if the patient is identified by the eMERGE Control algorithm, 0 otherwise.
Survey Diabetes	1 for a positive patient-reported diabetes status, 0 otherwise. Exists only in Matched Data

Sensitivity, specificity, and positive predictive value against all patient self-reported statuses were calculated for the eMERGE DM2 Case algorithm, the component criteria individually (Diagnosis, Medication, Lab), the group of patients meeting all the criteria (Diagnosis AND Medication AND Lab), and patients meeting any of the criteria (Diagnosis OR Medication OR Lab). Sensitivity, specificity, and positive predictive value were also calculated for the eMERGE DM2 Case group using just the individuals identified by the paired Control algorithm.

The eMERGE DM2 Case algorithm was expected to identify patients who do not report having diabetes, and not all patients reporting diabetes were expected to be identified by the algorithm. To investigate whether identification by the DM2 Case algorithm was a result of different subtypes of diabetes, with different patterns of comorbidities, all ICD-9 codes were pulled for each patient. ICD-9 codes were truncated at the root code level, or the whole number component of the code, and the frequencies of codes for each group were reported.

Results

We report our results in Tables 2-13 through 2-16, which includes summary statistics and demographics on specified patient groups, as well as validation statistics against all patient self-reported diabetes statuses and only those identified by the Control algorithm. See Figure 2-1 for a Venn diagram displaying the overlap between the patients identified by the eMERGE DM2 Case algorithm and those patients self-reporting positive diabetes status.

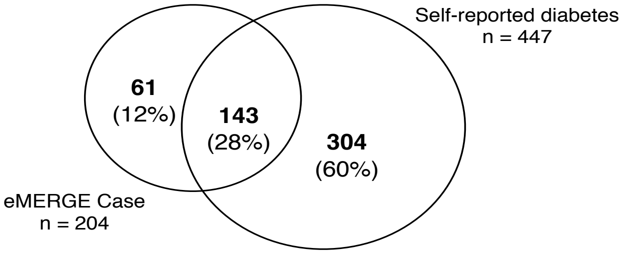


Figure 2-1: Venn diagram of overlap between patients identified by the eMERGE DM2 Case algorithm and patients self-reporting positive diabetes status

There were 2,249 WICER Survey participants with self-reported diabetes status who had at least one visit recorded at our institution within the last five years. Table 2-13 presents summary statistics and demography for patients reporting diabetes and no diabetes. The patients identified by the eMERGE DM2 Case algorithm (eMERGE Case), the pool of potential cases meeting any of the diagnostic, medication, or lab value criteria (Case Pool), and those patients meeting none of the criteria (Excluded) are presented for both the patients with reported diabetes status and the general population. In patients with self-reported status, eMERGE Cases and patients in the Case Pool are, on average, more than 15 years older than the Excluded group, and have twice as many recorded visits. The same difference is more than 24 years in the general patient population, with three times as many recorded visits. Patients with reported status are more likely to be female, as expected, but follow the same trend with regard to age and visits, albeit with 1.8-3.5x as many visits. While patients with reported status do tend to be older than the general population in general (46.1 vs. 36.4), those in the respective Case Pools are approximately the same age (61.8 vs. 61.0).

Table 2-13: Cohort demography and characteristics. Groups are patients identified by the eMERGE DM2 Case algorithm (eMERGE Case), pool of potential cases meeting any of the diagnostic, medication, or lab value criteria (Case Pool), and those patients meeting none of the criteria (Excluded). Patients answering "Yes" or "No" to diabetes status are also reported.

Cohort	Group	N	Patient-reported Diabetes Count	Fraction Female	Average Age (SD Age)	Average Visits (SD Visits)	Average Time between First and Last Visit (SD Time)
Patient-reported Diabetes Status	Yes	447	447	0.76	62.0 (12.1)	40.3 (45.4)	1223.6 (665.3)
	No	1,802	0	0.79	48.0 (16.9)	24.4 (33.6)	1052.2 (654.6)
	eMERGE Case	204	143	0.72	62.4 (12.3)	34.8 (36.7)	1293.6 (568.5)
	Case Pool	670	387	0.76	61.8 (13.0)	43.3 (45.9)	1285.1 (520.1)
	Excluded + Control	1,579	60	0.79	46.1 (16.3)	20.9 (29.7)	1159.0 (564.9)
General Patient Population	eMERGE Case	25,310	n/a	0.50	65.8 (15.2)	18.7 (29.6)	902.1 (641.0)
	Case Pool	106,569	n/a	0.50	61.0 (21.3)	19.0 (32.2)	848.4 (649.7)
	Excluded + Control	680,324	n/a	0.58	36.4 (22.8)	5.8 (11.6)	677.3 (589.2)

Table 2-14 shows the validation statistics against self-reported status. Sensitivity and specificity for the eMERGE phenotyping algorithm were .32 and .97, respectively, while positive predictive value was .70. The highest positive predictive value (.85) was achieved by requiring all criteria (Diagnosis AND Medication AND Lab). This combination also has the highest specificity (.98). While the highest sensitivity (.87) was achieved by the least restrictive combination (Diagnosis OR Medication OR Lab), the sensitivity of the combination requiring all criteria (.55) was still higher than that of the eMERGE algorithm.

Table 2-14: Positive predictive value, sensitivity, and specificity for the eMERGE DM2 Case algorithm, the component criteria individually (Diagnosis, Medication, Lab), the group of patients meeting all the criteria (Diagnosis AND Medication AND Lab), and patients meeting any of the criteria (Diagnosis OR Medication OR Lab).

Set	N	Patient-reportedDiabetes Count	Positive Predictive Value	Sensitivity	Specificity
eMERGE Case	204	143	0.70	0.32	0.97
Diagnosis	517	369	0.71	0.83	0.92
Medication	320	260	0.81	0.58	0.97
Labs	549	330	0.60	0.74	0.88
Diagnosis AND Medication AND Lab	291	246	0.85	0.55	0.98
Diagnosis OR Medication OR Lab	670	387	0.58	0.87	0.84

Validation statistics were also computed for the eMERGE DM2 Case algorithm using only the eMERGE DM2 Control patients for comparison. These results are presented in Table 2-15. As a pair the DM2 Case and Control algorithms excluded 1,449 patients, reducing the pool of analyzable patients to 800. The majority of self-identified diabetes patients fell into the excluded group, which raised the apparent sensitivity of the eMERGE DM2 Case algorithm to .93. However, the apparent specificity fell to .91.

Table 2-15: Positive predictive value, sensitivity, and specificity for the eMERGE DM2 Case algorithm using only the patients identified by the eMERGE DM2 Control algorithm.

Set	N	Patient-reportedDiabetes Count	Positive Predictive Value	Sensitivity	Specificity
eMERGE Case	204	143	0.70	0.93	0.91
eMERGE Control	596	11	n/a	n/a	n/a
Excluded	1449	293	n/a	n/a	n/a

The 15 most frequent ICD-9 codes for the intersections between the patients satisfying the eMERGE DM2 Case algorithm and the patients with positive self-identified diabetes status (+eMERGE +Self) are presented in Table 2-16. Codes for groups where the two methods disagreed (+eMERGE –Self, -eMERGE +Self) are presented in the same table as well as codes for the group of patients with no identification for diabetes (-eMERGE – Self). Note that DM1 and DM2 share the same root code (250) and no steps were taken to distinguish between types in this analysis. In general, the rank order of codes by frequency, as well as their general prevalence, is the same for the three diabetic groups regardless of how they were identified. The prevalence for diabetes ICD-9 codes is notably high in these groups. Prevalence for many of these codes is very different from patients without any indication of diabetes. Other comorbidities which are at least twice as prevalent in a diabetes group as in the non-diabetes group are hypertension, high cholesterol, diseases of the esophagus, and obesity. Patients with some identification for diabetes resemble the non-diabetic, general patient population in the prevalence of codes for follow-up examination, special investigations or examinations.

Discussion

The results of the eMERGE DM2 Case algorithm, as well as its component criteria, was validated against all patients with self-reported diabetes status, prompting several points for consideration. We will discuss issues surrounding the generalizability of the patients with self-reported diabetes status to the general patient population, discrepancies between identification from the eMERGE DM2 Case algorithm and the self-reported statuses, and the potential contributions of patient self-reported data to EHR phenotyping.

Table 2-16: Prevalence of comorbidities for group of patients identified by the eMERGE DM2 Case algorithm (+eMERGE +Self), groups where the two methods disagreed (+eMERGE –Self, -eMERGE +Self), and the group of patients with no identification for diabetes (-eMERGE –Self).

ICD9 Root Code	Root Code Description	+eMERGE +Self (n= 143)	+eMERGE -Self (n = 61)	-eMERGE +Self (n = 304)	-eMERGE -Self (n = 1,275)
250	Diabetes mellitus	0.99	0.93	0.74	0.05
401	Essential hypertension	0.86	0.85	0.79	0.34
272	Disorders of lipid metabolism	0.65	0.67	0.63	0.21
786	Symptoms involving respiratory system	0.48	0.47	0.44	0.31
V67	Follow-up examination	0.46	0.44	0.48	0.44
V76	Special screening for malignant neoplasms	0.46	0.43	0.54	0.30
724	Other and unspecified disorders of the back	0.41	0.33	0.37	0.28
V72	Special investigations and examinations	0.39	0.39	0.49	0.47
789	Abdominal pain	0.38	0.43	0.39	0.34
780	General Symptoms	0.36	0.43	0.40	0.28
719	Other and unspecified disorders of joint	0.35	0.43	0.39	0.27
530	Diseases of the esophagus	0.35	0.36	0.29	0.16
729	Disorders of the soft tissue	0.34	0.33	0.39	0.23
278	Obesity	0.33	0.43	0.40	0.21
V04	Need for prophylactic vaccination and inoculation against single disease	0.31	0.49	0.48	0.25

Patient Comparison and Generalizability

One concern with this dataset is the patients with self-reported diabetes status, those who participated in the WICER Community Survey, are known to differ from the general population in several ways. The group is older, containing more women, and is mostly Hispanic. However, the portion of these patients with positive indications for diabetes do resemble their counterparts in the general patient population in terms of age, and the relatively increased number of recorded visits, as shown in Table 2-14. These findings suggest that the characteristics of patients with diabetes do not depend on the population from which they are drawn.

In Table 2-16, ICD-9 codes for diabetes are the most frequently represented in patients with some identification, either by the eMERGE DM2 Case algorithm or self-report, for diabetes, as expected. However, there are some discrepancies. The relatively lower prevalence of diabetes ICD-9 codes in the portion of self-reporting patients not identified by the eMERGE DM2 Case algorithm may indicate self-report inaccuracies or the effect of missing data in this group. The 5% prevalence of diabetes ICD-9 codes in the the group with no identification for diabetes (-eMERGE –Self) may be a result of codes specific for DM1 which were filtered out by the DM2 case algorithm and not in that analysis.

Discrepancies in Identifying Diabetes

The eMERGE DM2 Case algorithm is known to perform well against case review and does achieve very high specificity in this evaluation. The algorithm performs less well in selecting all of the individuals who self-report having diabetes, and this may be for many reasons. First, the case algorithm is restrictive in order to limit the inclusion of DM1 patients. While steps were taken to exclude any patients who obviously had DM1, some of the patients who remain in the pool of potential cases may be rightfully excluded for this reason. Second, the non-selected patients may be incorrect about their diabetes status, though this is probably unlikely as this group of patients resembles the selected patients in patterns of visits and other demographics as well as the presence and frequency of comorbitidies. Moreover, if a large number of patients were in fact incorrect about their diabetes status, we would expect to see more discovered by the control selection algorithm. Lastly, and suggested by Wei, et al., the non-selected patients may be the product of data fragmentation, which is to say they do not have enough of their healthcare data consolidated in our system to allow identification by the eMERGE DM2 Case algorithm. For example, 83% of the self-reporting diabetic patients have at least a ICD-9 code for DM2 in our data warehouse, but at least 60% of those fail to be identified by the eMERGE DM2 Case algorithm for lack of sufficient clinical evidence for that diagnosis.

The more interesting group may be those patients selected by the eMERGE DM2 Case algorithm who do not self-identify as having diabetes. They have met the algorithm's stringent inclusion criteria, have visit patterns, other demographics, and comorbidities in common with the self-identifying diabetic patients, suggesting they do have diabetes. That these patients seem to not be aware they have diabetes may have large implications to their treatment, adherence to that treatment, and their engagement with any treatment. Pacheco reported that only approximately half of the patients identified by the eMERGE DM2 algorithm at Northwestern had diabetes as part of the patient’s problem list, further suggesting that this effect is not confined to the patient13.

Contribution of Patient Self-reported Data

There are pros and cons to both EHR data and patient self-reported data (Table 2-17) which point to how the two data sources might complement each other. EHR data is very heterogenous, with many data types, but that data may have issue such as missingness and inaccuracies that limit their secondary use for research. The more common elements have successfully been used for patient phenotyping algorithms, but that does not necessarily imply the algorithms have high sensitivity. In contrast, patient self-reported data reflects the patient's perception of their health status and may imply higher patient engagement in treatment, but may also be inaccurate and does not imply there is a useful quantity of clinical data at any one institution.

Table 2-17: Pros and cons of EHR and Patient self-report data sources

	Data Source	
	EHR	Patient self-report
Pro	Heterogenous data types support high specificity.	Reflects patient perception.
	Common, standardized elements support portability.	Might imply higher patient engagement.
Con	High rate of missingness.	Does not imply useful quantity of clinical data.
	May only reflect encounter with a single institution	Patient perception may not be clinically accurate.

The best use of patient self-reported status may be augmenting EHR-based phenotyping algorithms. Phenotyping algorithms like the eMERGE DM2 algorithm typically require multiple criteria for successful identification of a disease and in our study the majority of patients who self-reported positive diabetes status did not have enough data in our system to be selected by the DM2 Case algorithm. Yet, 87% of them did have at least one ICD-9 code, medication order, or lab result to support a diagnosis of diabetes. If patient self-reported status could be standardized and used in addition to commonly captured EHR data elements for phenoyping algorithms, our study suggests the number of patients identified by such algorithms could be greatly increased.

This recommendation comes with two caveats, however. First, the contribution of patient self-reported status to phenotyping algorithms for research will depend on the needs of that research. If clinical data are important, as in a retrospective observational study, then patients who cannot be identified from their data alone may not be useful. Approaches such as the eMERGE DM2 Case algorithm would therefore be the best way to identify meaningful cases within a data source. On the other, if the goal is to simply identify as many patients with a disease or status as possible, for a potential prospective study or a GWAS, then self-reported data would be a valuable addition.

The second caveat is the issue of standardization. The portability of phenotyping algorithms relies on the use of common and standardized EHR data elements, such as ICD-9 codes. If the source of patient self-reported disease status is not standardized down to the exact wording of the question being answered, then the results may not be comparable and the resulting algorithm may not be portable. For example, the source of patient self-reported diabetes status in our study did not distinguish between DM1 and DM2. While steps were taken to address this limitation, the exact results of this study would probably be different if the survey question had specifically addressed DM2 alone. Therefore, any potential phenotyping algorithm built using our data might not perform the same on a data source with a patient self-reported data source specific to DM2.

Limitations

This study has several limitations. First, relatively few people were surveyed compared to the size of the large volume of patients in the EHRs. While the patients with self-reported status do appear to resemble identified cases from the general patient population, the population taking the WICER Community Survey is known to be older, and contain a higher proportion of women and Hispanic individuals. Additionally, the WICER Community Survey does not distinguish between DM1 and DM2. While obvious DM1 cases were removed from the dataset, it is unknown what percentage of the remaining patients may have DM1.

Additional Comparisons: Missing Data

Research Question: How prevalent is missing data in a clinical dataset? How does removing patients with missing data change the clinical dataset?

Background

One large limitation of clinical data sources is a high rate of missing data and, put simply, sicker patients have more data. Research has been done into how the prevalence of clinical data in sicker patients may influence research cohorts. While statistics on missing data were collected at all points of previous investigation in this document, the rates of missing data have not been examined or analyzed. Additionally, one common strategy to deal with missing data in a clinical dataset is to simply remove any patient missing any data. Here, the effect of such a method on the Raw Clinical dataset is also examined.

Innovation

There is little innovation in this section. It does confirm published patterns of missing data in clinical data sources.

Methods

Proportions of patients lacking any data points for each variable were reported. Chi-square test was performed to test whether these proportions were significantly different between the Raw Clinical and Survey datasets.

A Complete Case was defined as any patient with a Gender, Weight, Height, Smoking Status, and Systolic and Diastolic blood pressure. Ethnicity was excluded due to known problem in labeling. BMI was excluded as a calculated variable depending on height and weight. Diabetes was excluded because of the simplistic and unreliable diabetes phenotype used at this stage. Age was excluded as it cannot be missing according to the database design. The Complete Case dataset was created containing only patients meeting this complete case definition.

An additional Filtered dataset was created to look at the intermediate case, where patients with a great deal of missing data were excluded but some missing data values were allowed. In the Filtered dataset, all patients have at least two of the variables present of the complete case definition. Summary and comparison statistics for the Complete Case and Filtered dataset, as compared to the Raw Clinical dataset, were reported.

Rates of missing data were also examined in terms of other variables, such as age. If the rates of missing data vary highly depending on other variables, then that variable is

MNAR. On the other hand, if rates of missing data do not depend on other variables then the data may be MAR.

Results

The proportions of missing data for each variable are presented in Table 2-18. Rates of missing data in the survey are very low, by design. Rates of missing data in the Raw Clinical dataset depend on the variable, from .16 for Systolic and Diastolic blood pressures to .59 for Height and Weight. Labels for diabetes status were missing at a much higher rate, however, diabetes was considered in more detail separately. Proportions of missing data were significantly different between Raw Clinical and Household datasets for every variable.

Table 2-18: Rates of missing data in Survey and Clinical datasets

	Survey	Clinical
missing GENDER	0.01	0.00
missing ETHNICITY	0.02	0.32
missing WEIGHT	0.01	0.53
missing HEIGHT	0.01	0.59
missing BMI	0.02	0.59
missing SMOKING	0.03	0.27
missing SYSTOLIC	0.02	0.16
missing DIASTOLIC	0.03	0.16
missing ALL DIABETES	0.02	0.87

Table 2-19 contains the summary statistics for the Filtered and Complete datasets, as well as the Raw Clinical dataset for comparison. The statistics for the Filtered dataset are virtually unchanged from the Raw Clinical dataset. Less than half of the patients (28,752 vs. 78,418) in the Raw Clinical dataset meet the definition for Complete Case. Complete cases are nearly 4 years older (51.30 vs 47.55), contain more women, are slightly heavier and shorter, but have slightly lower blood pressure.

Table 2-19: Summary statistics for Filtered and Complete Case datasets as compared to Raw Clinical

File	Clinical Raw	Filtered (>2 values)	Complete Case
N	78418	60258	28752
Age	47.55	47.33	51.30
Proportion Female	0.62	0.62	0.68
Proportion Hispanic	0.50	0.52	0.49
Weight kg	75.69	75.69	76.07
Height cm	160.34	160.18	159.97
BMI	28.10	28.20	28.30
Prevalence of Smoking	0.09	0.12	0.10
Prevalence of Smoking Among Labeled Status	0.12	0.12	0.10
Systolic	127.23	127.09	125.93
Diastolic	73.07	73.06	72.54
Prevalence of Diabetes (Survey = self-report, Clinical = >1 Diabetes ICD-9 AND >1 abnormal test)	0.04	0.05	0.09

Figure 2-2 shows the rates of missing data for one variable (systolic blood pressure) vs age by decade. This figure is representative of other graphs of missing data and was chosen to illustrate that, while the rates of missing data do fluctuate with age, those fluctuations are dwarfed by the underlying rate of missing data at all ages. For this reason, the Clinical dataset is categorized as having data missing primarily at random, which will be treated as MAR in following studies.

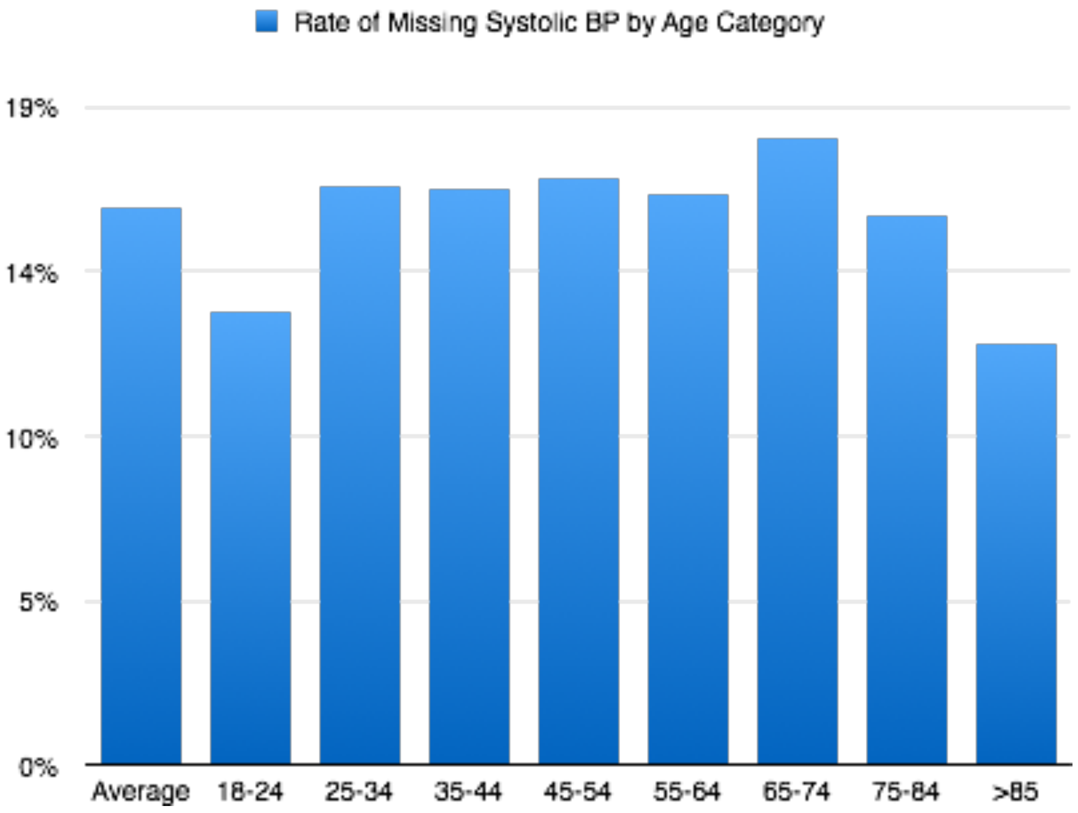


Figure 2-2: Rate of missing systolic blood pressure by age in the Clinical dataset

Conclusion/Discussion

Proportions of missing data are quite high (up to .59) for some variables and are significantly different from the research population. Again, what is surprising is that measurements at the population cohort level are in fact stable and statistically insignificantly different in some variables.

Most patients are missing some data (Filtered) and fewer than half the patients met the complete case definition. It is not surprising that the removal of patients missing most data had little effect on the summary statistics for the Filtered dataset as patients with very little data also make very little impact on the dataset. Restricting the dataset to patients meeting the complete case definition, however, did have measurable effects on the same summary statistics.

The high rates of missing data are the reason behind using population aggregate statistics for previous stages. The rate of missing data for height and weight, for example, would mean excluding approximately 35% of the Matched sample, or nearly 400 individuals. By aggregating over the whole population, we preserve all available data for analysis.

The other implication of the high rates of missing data are the significant differences in the Complete Case dataset. While data are classified as primarily missing at random, the examination of the Complete Case dataset reveals the limits of that classification. The members of the Complete Case dataset are older and likely sicker than the general patient population, and this difference results from requiring merely a single clinical data point in a handful of clinical variables. Larger data requirements for inclusion should be expected to result a much larger sampling bias. Conversely, if a researcher wants to limit sampling bias then any proposed cohort or analysis should be made as flexible to missing data as possible.

Additional Comparisons: Categorical Analysis

Research Question: So far, these investigations have looked at population aggregate statistics, which may mask more focused differences (such as within groups of 18-24 years olds). Do the trends identified above hold true smaller groups?

Background

Each row of each dataset was also assigned an indicator variable for BMI, Age by Decade, and Hypertension risk. These categories are useful because they aggregate continuous variables into clinically meaningful subgroups. While trends and discrepancies identified in previous stages of this investigation may be interesting, it is possible there are more significant differences hidden at these more highly granular levels.

Methods

Categorical indicator variables were added consisting of BMI (Underweight, Normal Weight, Overweight, Obese 1, Obese 2, Obese 3), Age by Decade (18-24, 25-34, 35-44, 45-54, 55-64, 75-84, 85+), and Hypertension Risk (Normal, Prehypertension, Stage 1, Stage 2). For each category, each clinical patient or survey participant is assigned a 1 for the variable which includes them and a 0 for all other variables.

All datasets were split along each category to be compared side by side and examined for notable differences from prior recorded conclusions.

Results

The results are too large to be displayed in this format. A sample result is included in Table 2-20. Here, the summary statistics of each dataset are presented for only those members with a recorded age between 18 and 24.

Table 2-20: Example Categorical Comparison. Differences in summary statistics remain present, but at typically smaller magnitudes than at the cohort level

Category	18-24				
Data	Raw Clinical	Clinical Sampled (H)	Clinical Sampled (U)	Household	ACN
N	10148	50156	78324	390	114
Age	21.63	21.56	21.6	20.53	20.54
Proportion Female	0.6152	0.7124	0.7067	0.6385	0.8158
Hispanic	0.5681	1	0.6473	0.9179	0.8947
Weight kg	72.4754	71.5152	71.476	74.2689	70.9576
Height cm	163.1921	161.5481	162.1624	165.0253	163.0572

Considering age, the broader trends between datasets remain true. Points which break the broader trends are areas where the datasets unexpectedly agree. For example, the proportion female is not significantly different in the categories of age 25-34, 35-44, 45-54, 75-84, and >85 even though it remains significantly different at the aggregate level.

In the BMI categories, Raw Clinical and the Household Survey population align very closely. In the Normal weighted category, smoking, gender, and even the crude diabetes label used here are not significantly different.

In the Hypertension Risk categories, the broader trends between datasets hold true. In the Pre-Hypertension Category, Systolic BP is significantly different, while gender is not.

Discussion/Conclusion

For the most part, all broad conclusions about discrepancies between sampling and measurement hold true at higher granularities. What is surprising is that, where the prior results are not true, the datasets actually become more similar at these higher levels of granularity. For example, the differences between the members of the Raw Clinical and Household datasets for patients between the ages of 25-34 may be smaller than between all the members of the entire dataset.

These exceptions suggest that maybe differences in summary statistics across entire datasets may not have as large of an effect at the level of a specific research hypothesis which considers only one of these categories. It was this conclusion which led directly into the idea for a dataset validity analysis using a highly granular portrait of the significant differences which may be present in a dataset.

3. Dataset Validity Analysis

Aim 2: Validation of Existing Datasets

Aim: Build and evaluate a method to compare datasets through the results of randomly generated hypothesis tests.

Aim 1 established that summary statistics for structured data and point measurements in our clinical dataset were not meaningfully different than our research dataset. However, most research using clinical data lacks the opportunity to use such direct reference data on the same individuals. Instead, a recent trend has been to validate an electronic clinical database by replicating a published finding or statistical result from another dataset. In study 2A, we expanded this concept into a method for comparing datasets through the results of multiple, randomly generated, two group hypothesis tests. This effort is considered preliminary in that the method was prototyped using a limited set of clinical data variables and without temporal considerations. In study 2B, we demonstrated the potential utility of the validation method by investigating the effect of data missing-at-random, a potential bias we could not effectively measure in Aim 1.

Study 2A: Introduction

Despite potential problems and biases in electronic clinical data, they are widely used for research. We have demonstrated that, with a unique set of circumstances like the overlap of the WICER Community Survey with the population of the CDW, it is possible to examine a clinical dataset for specific selection and measurement biases. However, the lack of those unique circumstances does not mean that clinical datasets used for research remain unexamined. Clinical datasets are examined, or validated, in different ways depending on the resources available. These validation steps may include comparing data from different sites within a federated research network or validating through replicating the result of an existing research trial. Several projects and their approaches to dataset validation will now be reviewed, followed by an overview of the dataset validity analysis we have implemented.

SENTINEL and HMORN were designed to integrate clinical data with claims data to make up for gaps in the analysis of claims data alone[29, 40]. Because members' claims warehouses typically comprise multiple clinical locations, this goal required the creation of a federated system where local data elements are mapped to a common data model which can be queried across multiple sites. Again, data normalization and bias issues are acknowledged and the solution is transparency to the end user[40]. These projects validated their datasets through comparison with expected census distributions, similar to our weighting steps in the previous aim. Sociodemographic characteristics for the research population in one region were compared to census results for the same region with a result of no significant differences, suggesting that a clinical sample selection may not represent a bias in their population[39].

DARTNet is another project to build a federated data system designed to compile clinically enriched data for CER. The improvement over a project like HMORN was the goal of improving the quality and reliability of data by limiting mapping to only around 150 common data elements, though local mapping may be time consuming and there is concern about bias issues because of it[20, 41]. To investigate potential data bias issues, pilot studies using the system were explicitly chosen to replicate a set of previous studies so research outcomes could be verified against a true baseline. The goal was to identify greater effectiveness of combination therapies over single drug therapy for diabetes on common diabetes disease markers, a task which required the identification of a cohort as well as the monitoring of their health status and treatment protocols[41]. Results were "highly similar" to published findings and this pilot study step represents a true acknowledgment of potential biases associated with secondary use of data as well as a compromise towards its use. This kind of study result replication represents a shift from comparing differences in the content of datasets, i.e. summary statistics, to comparing the answers a dataset can provide.

Our method for dataset validity analysis explicitly expands upon this final example of validation. However, where projects like DARTNet validate a dataset on a single research conclusion, we use a set of randomly generated two group hypothesis tests to create a highly granular portrait of the answers a dataset might provide. In the first aim, a clinical dataset was compared to a research dataset by asking, for example, whether the average systolic blood pressure of 70 year-olds was the same between datasets. In this aim we ask questions like whether the average systolic blood pressure of 70 year-olds is greater than 40 year-olds in both datasets. In this way, the actual numbers do not matter, only that the datasets answer the same or different. For an illustration of this example, see Figure 3-1.

Table 3-1 contains common concepts and their definitions for this section. Building and using a method for validity analysis such as this requires several steps, which will now be reviewed in detail.

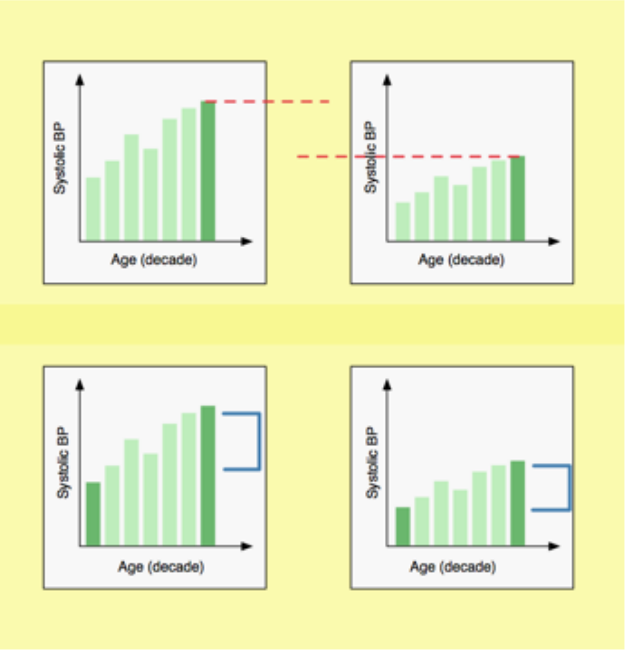


Figure 3-1: Example of testing for answers rather than comparing summary statistics. Perhaps it is more meaningful if the same difference, in this example that 70 year-olds have higher blood pressure than 40 year olds, is present in two datasets rather than whether their summary statistics are the same.

Table 3-1: Validity Analysis concepts and definitions

Concept	**Definition**
Hypothesis	Any randomly generated two group hypothesis test in the hypothesis set.
Hypothesis Set	Set of hypotheses that, when evaluated on a dataset, provide a highly granular portrait of the differences within a dataset.
Candidate Dataset	Dataset where the hypothesis set is being evaluated, answers to be classified against the reference dataset.
Reference Dataset	Dataset where the hypothesis set was generated and whose answers will serve as the basis for classifying a candidate dataset. In these studies, the reference dataset is assumed to be of equal or higher quality to the candidate dataset. For example, the research dataset vs. the clinical dataset.
Clinical Dataset	Raw Clinical Dataset from Aim 1.
Research Dataset	WICER Household Survey dataset from Aim 1.
Clinical Sample	A dataset of equal size to the research dataset, randomly drawn from the clinical dataset.

Dataset Validity Analysis Methods

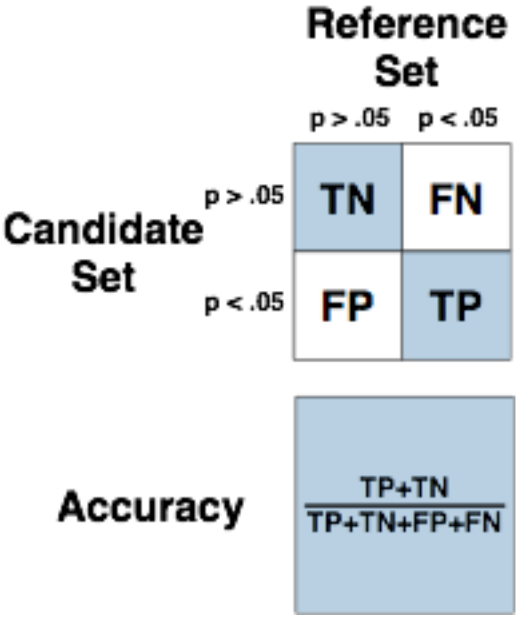
For the purpose of this investigation, a 'testing hypothesis' will be limited to a simple two-group comparison test (either chi-square or t-test depending on whether the outcome variable is categorical or continuous, respectively) between cohorts defined in terms of one or more other categorical variables. For example, is the systolic blood pressure of women older than 65 the same as men of the same age?

Potential test hypotheses will be randomly generated with the following procedure. Two variables will be chosen at random to serve as independent variables to define a cohort. If the variable chosen has a continuous range, then what is actually referenced is one of the categorical values for that variable. For example, if systolic blood pressure is chosen then what will actually define the cohort is one of the hypertension risk categories. Three examples from the randomly generated hypothesis set are shown in Table 3-2.

Table 3-2: Example randomly generate hypotheses from the hypothesis set

Is the PROPORTION of SMOKERS in MEN with STAGE 1 HYPERTENSION vs. PREHYPERTENSION different?
Is the MEAN of DIASTOLIC BP of PATIENTS >85 with SEVERE OBESITY vs. OBESITY different?
Is the PROPORTION of WOMEN in SMOKERS AGE 65-74 vs. 75-84 different?

The testing hypothesis is then classified by whether the null hypothesis is accepted or rejected (with a significance threshold of .05) when the test is performed in the candidate set and the base set. A hypothesis test which results in a p-value of <.05 in the candidate set and in the base set is considered a true positive. One which has a p-value of >.05 in both sets is considered a true negative. Each of these results is considered "accurate" in that the candidate dataset provides the same answer as the base set.



Candidate Set	Reference Set p > .05	Reference Set p < .05
p > .05	TN	FN
p < .05	FP	TP

Accuracy $\frac{TP+TN}{TP+TN+FP+FN}$

Figure 3-2: Classification matrix and Accuracy calculation for Validity Analysis

For example, the mean systolic blood pressure of women older than 65 is compared with men older than 65 with a p-value of .65 in some candidate set and .37 in the base set. This test is classified as a true negative as no significant difference was detected in either set. This hypothesis test supports the "accuracy" of the candidate set because the candidate set has provided the same answer as the base set. A high proportion of accurate hypothesis tests supports the validity of that candidate set. Hypothesis tests which result in different answers from the base set are coded as false positive or false negative depending on whether the test result in a p-value of <.05 or >.05 in the candidate set, respectively, and are considered "inaccurate" results regardless of direction. This classification matrix is presented in Figure 3-2.

The hypothesis generating procedure was performed using the Complete Case clinical dataset as the candidate set and the Research Dataset as the base set. The procedure was repeated until there were at least five unique hypotheses in each class.

Comparing Two Datasets

Introduction

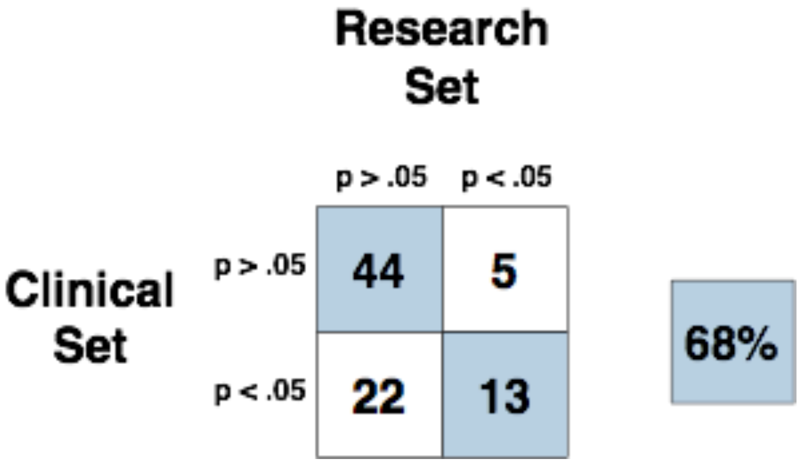
In prior work, the summary statistics of two datasets were compared to establish the magnitude and direction of selection and measurement bias in clinical data. The discovered differences, while statistically significant and with some caveats, are probably not meaningful differences. The reason the multi-hypothesis validation procedure was devised and implemented was to investigate how similar the clinical and research datasets are in their internal differences and in how much they agree in the answers they might provide.

This section discusses how multi-hypothesis validation was applied to the Clinical and Research datasets from prior work, as well as additional comparisons to explore the impact of the results. Specifically, the Research (n=4,069) and Clinical (n=78,418) datasets were compared using multi-hypothesis validation procedure described above. The datasets agree on 57/84 hypotheses, for an accuracy of 68%.

However, while outcomes were encouraging, there were concerns that the greater size and power of the clinical dataset might skew the results. Therefore, 40 rounds of comparisons were made between the Research dataset and an equal sized (n=4,069) random sample of the Clinical dataset. In addition, 40 rounds of comparisons were made between equal sized (n=4,069) random samples of the Clinical dataset to determine the effect of the random sampling. Accuracy was 77% and 81% for research vs. clinical sample and clinical sample vs. clinical sample, respectively. The distribution of results was compared with chi-square test with a p-value of .64, suggesting that the results from the Research dataset are no more different from the clinical dataset than samples of the clinical dataset are from each other.

Results

Each hypothesis in the hypothesis set was evaluated on the two datasets and the result classified as described in the previous section. Aggregate results of this multi-hypothesis validation are presented in Figure 3-3. In direct comparison of the clinical and research datasets, a significant difference was detected in both datasets for 44 hypotheses. A significant difference was not detected it both datasets in 13 hypotheses, leading to a 68% overall agreement between datasets.



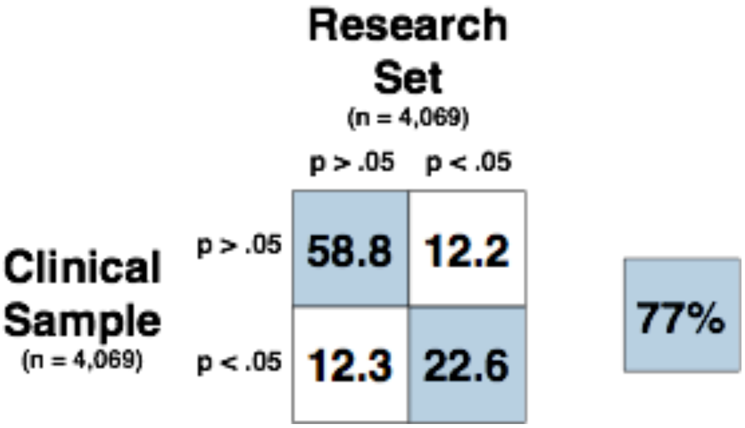
		Research Set	
		p > .05	p < .05
Clinical Set	p > .05	44	5
	p < .05	22	13

68%

Figure 3-3: Hypothesis classification results and accuracy for Clinical dataset vs. Research dataset

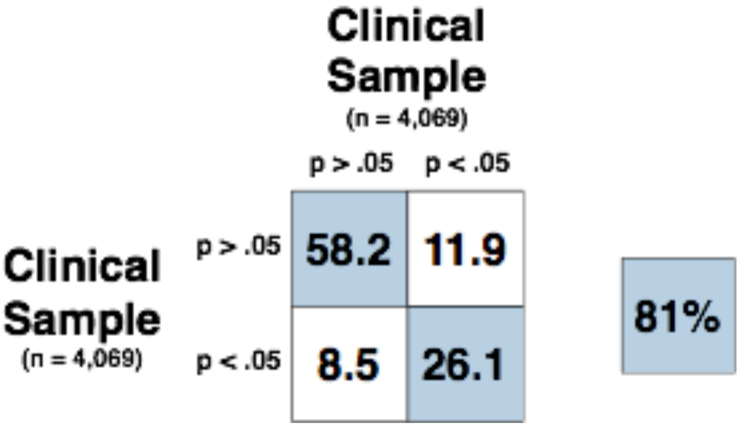
In the clinical dataset, a significant difference was detected in 22 hypotheses where there was not a difference in the research set. This may be due to the larger power in the clinical set due to its greater number of members (n=78,418 in Clinical, vs n=4,069 in Research). To investigate this possibility, and compare the datasets more fairly, forty random subsets of the clinical dataset were compared to the research dataset, the results averaged, and presented in Figure 3-4. The lower power of the clinical sample lead to fewer false positives (12.3 vs 22) and a higher rate of agreement between datasets of (77%).

However, the use of the smaller clinical samples introduced a new question: was the clinical sampling procedure introducing any new problems? To investigate this possibility, random clinical samples of the same size (n=4,069) were compared with each other and the results of forty comparisons averaged. These results are presented in Figure 3-5.



		Research Set (n = 4,069)		
		p > .05	p < .05	
Clinical Sample (n = 4,069)	p > .05	58.8	12.2	77%
	p < .05	12.3	22.6	

Figure 3-4: Hypothesis classification results and Accuracy for Clinical Sample vs. Research dataset



		Clinical Sample (n = 4,069)		
		p > .05	p < .05	
Clinical Sample (n = 4,069)	p > .05	58.2	11.9	81%
	p < .05	8.5	26.1	

Figure 3-5: Hypothesis classification results and Accuracy for Clinical Sample vs. Clinical Sample

As they were drawn from the same clinical dataset, the comparison of clinical samples should result in the highest possible rate of accuracy. What is interesting is that this rate of accuracy, and indeed the distribution of classified hypotheses, between the comparison of clinical samples (Figure 3-5) and the comparison of the Research dataset with clinical samples (Figure 3-4) are not very dissimilar. The similarity can be evaluated statistically with a chi-square test using the results of the clinical sample comparison as the expected frequencies and the comparison of the Research dataset to clinical samples as the observed, experimental frequencies.

The result of the chi-square test is a p-value of 0.64, meaning the classification of hypotheses between the Research datasets and any clinical sample of the same size is not significantly different from that between two random clinical samples.

Limitations

What has been presented is an analytic method to explore similarities and differences between datasets in terms of the answers they provide to statistical tests. Choices were made in the exploration and evaluation of this method which may have impacted the results. Because the goal was to compare the Research and Clinical sets, again only a limited set of overlapping variables could be used. Next, hypotheses were generated using the same Research and Clinical datasets, meaning all evaluations were predicated on there being detectable differences between the sets. Lastly, while the clinical sample comparison can establish an upper bound on "accuracy" between datasets, this evaluation stops short of determining what level of accuracy is needed for confident research using clinical data.

Discussion and Recommendations

This evaluation demonstrates that, regardless of statistically significant differences in summary statistics, Research data and Clinical data provide largely similar answers to random statistical tests. Moreover, the difference between Research and Clinical data is no larger than the difference between random clinical data samples drawn from the same dataset. These findings suggest that electronic clinical data might be used with equal confidence as research data on the same population.

However, there are larger implications to how this analytic method might be applied. The goal is validation of a dataset, which requires a dataset for comparison and an available hypothesis library. Given a large publicly available clinical dataset, researchers could generate their own hypothesis libraries using all overlapping data. The benefit of this option is that the dataset validation can be tailored. For example, hypotheses could be generated using only outcomes of interest for the proposed research questions. The downside of this option is that it requires the availability of a large, highly granular and inclusive dataset.

A different, related approach is the publication of a common hypothesis library and the results when performed on a particular database. The advantage here is that no data need be shared yet the same comparisons can be made between datasets. There might be problems with power issues due to different database sizes, as were encountered in this analysis, but they might be overcome by publishing exact p-value results and simply scaling the significance threshold accordingly. The disadvantage to publishing just a hypothesis library and results is that the generated hypotheses would have to be limited to common data elements to ensure generalizability between clinical datasets. This was essentially the approach taken in this evaluation between a research and clinical dataset, demonstrating the approach is possible.

A more limited, but possibly more useful, application of this analytic method is for local validation. A medical institution might be interested in validating a random subset of data for preliminary research use or ensuring that a specialized cohort has representative clinical data. In these cases it would be very useful to compare the limited datasets to other random clinical datasets of the same size. Such an internal validation might establish the extent to which the answers from a de-identified random subset might be trusted without resorting to the use of identified data, or whether the results from a more specific research cohort might be generalized to the larger patient population.

Study 2B: Data Missing at Random

Summary

Data MAR should not bias a dataset, however, the levels of missing data present in our clinical dataset are quite high. Using the dataset validity analysis method developed for this aim, the effect of data MAR can be investigated. At levels of missing data up to 60%, or the highest level present in our clinical dataset, the dataset still results in 90% of the same answers as the original version of the dataset with no missing data. The more surprising finding is that even at 99.9% data deletion, if a significant difference between two groups can be detected, then that difference was almost certainly present in the complete dataset.

Research Question: What is the effect of data MAR on validity analysis?

Introduction

This thesis has primarily focused on the differences between a clinical dataset and research dataset in terms of sampling and measurement bias. The third major issue is that of missing data. The rate of missing data in the clinical dataset from virtually zero for some variables (sex, age) to quite high for others (height, weight). A description of missing data for more complex, inferred variables such as diabetes depends on the phenotyping algorithm being used and, particularly, how many different sources of information are required to support the diabetes label.

There are three generally accepted classes of bias due to data incompleteness: Missing at Random, Missing at Random, Missing Not at Random, and Missing Completely at Random[42, 43].

Missing data bias is generally classified as Missing at Random (MAR) or Missing Not at Random (MNAR). Data MAR are, as the name implies, missing throughout the dataset at a rate which does not depend on any recorded factor. A subset of data MAR are data Missing Completely at Random (MCAR) in which data are missing with no dependence on any factor recorded or otherwise. The data MCAR is difficult to demonstrate outside of artificially limited datasets. While data MAR does decrease the sample size, and therefore "resolution", for detecting difference between groups, it adds no directed bias.

Data MNAR, however, are where the rate of missing data depends on some other recorded variable. Data MNAR can lead to biases which affect research outcomes because here the rate of missing data is unbalanced. If, for example, younger people were more likely to refuse participation in a blood pressure survey and blood pressure tended to increase with age, then that survey might report an average age and blood pressure higher than the true underlying population.

As previously reported in this document, the effect of missing data in the clinical dataset is primarily MAR. However, while data MAR does not bias a dataset as defined above, data MAR does degrade the power of a dataset in possibly unpredictable ways. The most common approach to dealing with missing clinical data is to simply exclude any patient with missing data, called the Complete Case approach. The advantage is that any issues with missing data are excluded from the dataset. The disadvantage is patients with some missing data may have useful data to contribute to an aggregate analysis, and are discarded.

The effects of restricting patient cohorts to those with more data have been previously investigated[44, 45]. While the effect of particular data requirements were not investigated in this thesis, some similar broad trends were recognized in the analysis of the Filtered and Complete Case datasets. Given the method of validating datasets described in the previous section, there is an opportunity to evaluate the effect of data MAR on a clinical dataset. Table 3-3 presents concepts and definitions for this section.

Table 3-3: Concepts and definitions for examining the effect of data MAR in a clinical dataset

Concept	Definition
Clinical Data Point	Any clinical observation for a variable within the dataset. Patients may have multiple clinical data points per variable.
Summary Data Point	The average of all clinical data points for a variable for each patient. Values of summary data points are evaluated as part of the validity analysis.
Target	Any one of a range of data deletion targets from 10% to 99.99%. The target represents the chance that each clinical data point will be deleted.
Candidate Set	For this study, candidate sets are targeted, deleted datasets where clinical data points have been deleted up to a given target.
Reference Set	For this study, the reference set is the Complete Case dataset where each patient has at least one clinical data point for each variable.

Given the method of validating datasets described in the previous section, there is an opportunity to evaluate the effect of data MAR on a clinical dataset.

Methods

Dataset

A 'clinical data point' is a single observation for a patient, for some variable, in the clinical data source. A ‘summary data point’ is the average of all clinical data points for a patient. The intent is to start with a dataset where every patient has at least one clinical data point per variable, and therefore a summary data point for each variable, known as the Complete Case set. While not the same as the original Clinical dataset, the Complete Case set does have the advantage of providing a gold standard for a subsequent evaluation of missing data.

A secondary consideration is to restrict data choice to what was used for the first aim. One reason for this restriction is that defining and acquiring clinical data is a painstaking and time consuming process. Another is that this restriction means the effects of missing data and imputation will be investigated in the same context as the first aim, where the quantity of missing data could be determined in a "real world" scenario. A third consideration is that restricting the dataset to the same variables as previously used also means the same hypothesis library can be used as was generated in the previous evaluation.

Therefore, this investigation uses a Complete Case dataset of clinical patients with at least one clinical data point for Age, Sex, Height, Weight, BMI, Systolic and Diastolic blood pressures. If multiple clinical data points are available for a patient, they were averaged into a 'summary data point'. Validation set hypothesis tests will be performed on the summary values. There are approximately thirty thousand patients in this Complete Case set.

Data Deletion

There are multiple ways to create a data MAR scenario. What is needed is a way to delete clinical data up to a target value (50%, 90%, etc.) For this investigation, treat each patient's summary data value as a collection of clinical data points. For example, each patient has one summary data value for weight which is the mean of many recorded clinical data values. For each trial, a candidate set is created where each clinical data point has a chance of deletion equal to the target. Clinical data points are then summarized and recorded in the same way as the base set. Therefore, the target for each set represents the approximate percentage of clinical data points which have been deleted. This process is repeated on ten sets for each target, and the results averaged.

Hypotheses and Testing

The hypothesis library generated in the previous step was used in the evaluation of missing data. Each hypothesis test is evaluated on the candidate set and that result compared to the result of that hypothesis test on the base set. A hypothesis test which results in a p-value of <.05 in the candidate set and in the base set is considered a true positive. One which has a p-value of >.05 in both sets is considered a true negative. Each of these results is considered "accurate" in that the candidate dataset provides the same answer as the base set. Hypothesis tests which result in different answers from the base set are coded as false positive or false negative depending on whether the test result in a p-value of <.05 or >.05 in the candidate set, respectively.

For this investigation accuracy is reported as the percentage of hypothesis tests whose results are either true positive or true negative as compared to the base set, or (TP + TN) / (TP + TN + FP + FN). Positive predictive value is the percentage of hypothesis tests which result in a p-value <.05 in the candidate set which are also true positives, or TP / (TP + FP).

Alternative methods of data deletion were considered, performed, and evaluated. The results were not significantly different. This method was chosen because it more closely maintains the underlying patterns of missingness in the clinical data points.

Results

See Figure 3-6 for accuracy and positive predictive value of validation vs. target data deletion value. Up until 30% data removal there is very little difference in validation. Up until 80% data removal, the vast majority of validation hypotheses will still have the same results. Accuracy quickly degrades past 80% data deletion.

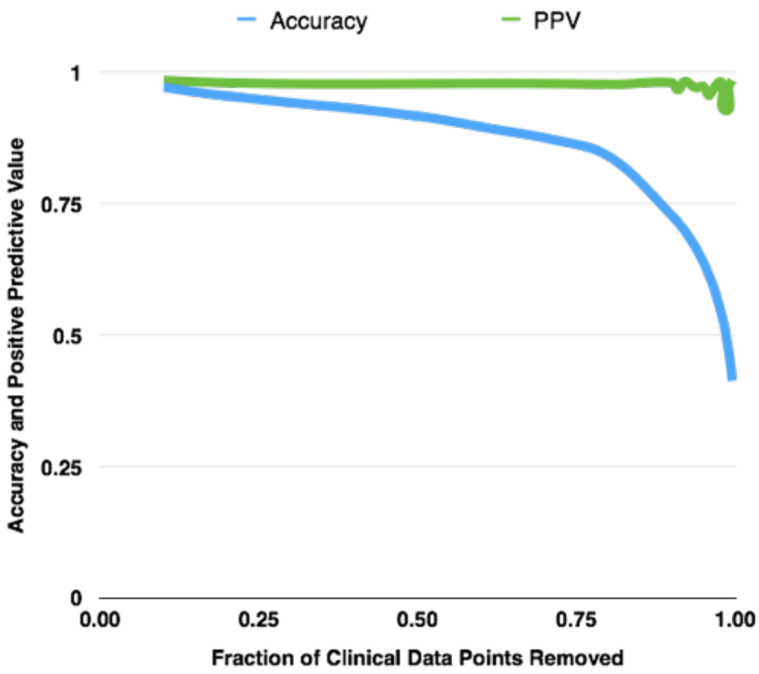


Figure 3-6: Graph of Accuracy and Positive Predictive Value for targeted, deleted datasets vs Complete Case dataset

Even at near total data deletion, there is a sizable fraction of hypotheses which still compute “correctly”, which is to say the same answer to a hypothesis test is calculated in the targeted, deleted dataset as in the base, complete dataset. However, these hypotheses are instances where no significant difference was could be detected and that was also true in the base, complete set. While these results remain true, because so many hypotheses are now incorrectly reporting no significant difference, they cannot be trusted.

The surprising finding is that the rate of false positive results is very, very low at all levels of missing data, implying that even in a very degraded dataset, any two group hypothesis test which indicates a significant difference between groups is likely correct.

Limitations

This investigation had several limitations, some in common with the rest of this thesis and some pertaining specifically to this investigation. The more general limitations include the reuse of the same dataset and variables as previous investigations. The result of those choices is a hypotheses library limited to the same variables. While it meant the same hypotheses library could be reused, the broader dynamics of this analytic method could not be explored.

Conclusions and Recommendations

The biggest takeaway from this investigation is that data MAR have very little effect on this analytic method until data are missing at very high levels. While the amount of missing data in the clinical dataset may seem objectively high (~50% for some variables), it is valuable to note that this amount is less than the level of missing data required to significantly affect the validity of the dataset. In other words, the effect of missing data should not be a concern for researchers using this particular clinical dataset. If the level of missing data, and especially the fact the data appear to be primarily MAR, are common across clinical datasets, then typical levels of missing data should generally not be a concern for researchers using electronic clinical data.

This recommendation has important implications. If missing data has no significant effect on the validity of a dataset, efforts should be taken to retain patients with missing data in clinical datasets. If care is taken to include analyses which can be performed on some level of aggregate statistics, such as those randomly chosen by the hypothesis generation process, then power can be added to a dataset even if the patients being added are missing data.

The positive predictive value result should further contribute to confidence in using datasets with missing data for research. At all levels of missing data, even up to levels approaching 100% data deletion, if a difference could be statistically significantly detected, then that difference was almost certainly statistically significant in the original dataset. The steadily decreasing accuracy demonstrates that as the level of missing data increases, the number of hypotheses where a statistically significant result is also decreasing and that negative predictive value becomes poor. These implications suggest that structured clinical datasets with >90% missing data can still be used for research. However, in work with these datasets only statistically significant differences should be reported as non-significant differences have much lower likelihood of being accurate.

4. Addressing Gaps and Opportunities

Aim 3: Addressing Gaps and Opportunities

Aim: Explore the use of more advanced techniques to address gaps and opportunities presented by the first two aims.

Aims 1 and 2 were designed to address identified gaps in the current practice of clinical data validation. However, the execution of these aims suggested further opportunities for improvement that fell outside of their original scope. Studies 3A, 3B, and 3C investigate three such opportunities, specifically in imputing missing data in clinical datasets, improving linkage between patients in different datasets, and computing the "representativeness" of a patient in a database.

Study 3A: Imputing Missing Data in a Clinical Dataset

Research Question: Can missing data in a clinical dataset be replaced so that the accuracy of a dataset is improved?

Methods Background

Multiple methods have been recommended for imputing missing data in datasets. They range from using a single summary value to substitute for all missing data to complex machine learning methods which seek to replicate not only summary statistics but the underlying distribution. This section reviews the background and recommendations concerning this range of imputation methods.

Complete case analysis is the simplest approach to dealing with missing data. With complete case analysis, only records where values for all data variables are recorded are included[42, 46-49]. Complete case analysis neatly sidesteps the problem of missing data by ignoring, which can be an effective strategy when data are MAR. However, when data are MNAR, complete case analysis precisely preserves whatever problems are present. A subset of complete case analysis is case restriction, where the criteria for inclusion are much more rigidly defined in order to limit potential biases by excluding known confounders[25, 50]. For example, a retrospective study on heart disease my exclude any patient record with any indication for heart disease or history of compatible symptoms. Much like a live randomized, controlled trial, case restriction can strengthen argument for a causal link between exposure and outcome, but the generalizability of the finding may be compromised. Complete case analysis was used as a baseline in this investigation.

The next step in complexity involves replacing all missing data of a type with a single, simple value, commonly known as single value replacement[42, 46-49]. Most of the methods in this class, such as Last Observation Carried Forward (LCOF), are designed to deal with gaps in time series data and dropouts. However, some can be adapted to single data fields. For instance, the mean or median value of a variable across the entire population can be substituted for any missing value. When comparing two groups, the mean or median of each group can be substituted for missing values within that group. A more conservative variation is to use the mean or median value of the control group alone for all missing values, thereby ensuring that any finding is biased toward the null hypothesis.

The third class are estimating methods, such as linear regression. Typically, regression coefficients are used to estimate the values of missing data[46-48, 51]. The advantage of this method is that substituted values are arguably more suited to the case to which they are supplied. However, the substitution is only as strong as the regression model.

Highly computational approaches exist which use more complex assumptions about the distribution of variables in order to make substitutions. Previously mentioned methods substitute the most likely value for any missing data point, for different definitions and calculations of likely. However, the likelihood of all substituted data values being exactly the most likely (for example, exactly the mean) is small. These more complex methods take that point into account and create substituted data points more holistically by looking at the range of values.

The first such method is multiple imputation[42, 46, 47, 49, 50]. A substitute for each missing data value is drawn randomly from all values present for a given variable. The process is repeated multiple times to create a pool of imputed data sets. Results from queries over all datasets are averaged. The advantage is that the extremities and the shape of the distribution are preserved exactly as well as they are represented by existing values. The disadvantage is that only the exact values actually present are used as replacements.

K Nearest Neighbor (kNN) is a method similar to multiple imputation where the possible replacement values are limited to those of the "nearest" rows in terms of spatial distance.

Here, spatial distance is a way of computing similarity using all available data. In this method the intuition is values from similar patients will be closer to the true value missing for the patient in question than a randomly chosen value from the entire set. The downside of this method is that it may fail when there are many data variables, or dimensions, used for the distance calculation. This limitation, which was not a factor in the current investigation, might be side-stepped by limiting the variables for the distance calculation.

Expectation Maximization, on the other hand, draws values from an artificial distribution based on the values which are present[42, 46, 49, 52]. Missing data values of any variable will be replaced with the result of an expectation maximization algorithm, taking into account the distribution of that variable and the likelihood of all other replacements made. The advantage over multiple imputation is that substituted values are smoother and do not literally rely on existing values.

Evaluation of Methods for Reducing Bias in Clinical Data

As mentioned previously, while there is a strong history of use of presented methods in other contexts, there is not a great deal of evaluation of these methods on clinical data. Two such evaluations and a meta-analysis of missing data method application are presented here. The two evaluations are characteristic of those which are performed. One reason they may be so rare is that, lacking research quality reference data, evaluation is limited to essentially a sensitivity analysis on a targeted study's conclusions. What these evaluations demonstrate is that when a simple data replacement method is sufficient to change a study's statistical conclusion, then there was probably a significant bias in the study's original data.

The first evaluation takes the form of an analysis of basic missing data mitigation strategies on longitudinal data from a weight-loss trial[22]. Ware, et al., began with a look at missing data, starting from a complete case set, which should mimic the rigorous inclusion standards of a complex clinical dataset. While they did not have access to full data, they could add cases to the analysis through the application of basic methods such as LCOF and First Observation Carried Forward (FOCF). LOCF is a method for missing data replacement in a time series where the last recorded value is used to substitute for the missing data values at the end. FOCF is a related method where the first recorded observation is used as the substitute. Each of these methods was applied to the weight loss dataset to augment the number of cases and compare results to only using complete cases. The study demonstrated that while the trend of the data was preserved the statistical significance of the conclusion was not, suggesting study dropout was a case of data MNAR.

An evaluation by Raboud, et al., of slightly more complex methods on nonrandom missing data in a study of CD4 counts following antiretroviral therapy concluded that "missing data ... can result in underestimation of treatment effects"[21]. Dropout is common in HIV trials due to poor compliance, treatment toxicity, and other causes and the result of that dropout can mean data MNAR. This evaluation again began with simple value replacements such as LOFC but stepped forward to methods such as imputation with regression-predicted values. While in this instance, using a complete case baseline would result in an underestimation of treatment effects, the authors acknowledge that because dropout was correlated with treatment group, which is to say missing not at random, the effect could have been easily inverted for a different set of treatment effects. Again, bias could not be directly demonstrated, but the change in apparent treatment effect is suggestive of data MNAR.

Rather than examining a single study, Molnar, et al., took a systematic review approach to the application of a LOFC approach to the treatment of missing data in the domain of dementia therapiesol[23]. They examined the quality of included research studies with regard to the treatment of missing data and case selection with the conclusion that "published results of some trials may be inaccurate." However, this review derives its conclusion from an analysis of which included studies used any form of missing data mitigating methodology and the degree to which that methodology, if present, was applied. The data itself was not examined. What is interesting about Molnar's approach is the acknowledgement that LOFC is only successful when data are MAR and can introduce greater bias otherwise. Studies under analysis were penalized for the application of LOCF without a discussion of whether it was an appropriate method to use.

While these three studies are applicable to proposed work, they do highlight some gaps in current research. One is that only a limited set of methods is ever directly compared using the same data, making it difficult to evaluate the relative strength of one method against another. The second is that, lacking reference data, it is impossible to directly evaluate the extent of bias. Rather, evaluation of a study is treated as a sensitivity analysis where a study's conclusion is thrown into doubt if application of some method is enough to significantly change it. There is a need for an evaluation of clinical data against research quality reference data as well as an evaluation of multiple methods for addressing bias performed on the same dataset.

Methods

Datasets

This investigation makes use of the targeted, deleted datasets created for the prior investigation of the effect of data MAR on the validity analysis. Ten datasets were created for each missing data target, where the targets range from 10% to 99.99% and the target represents the chance of being deleted for each clinical data point. Patients with multiple clinical data points per variable have their clinical data points averaged to make a summary data point. The starting set for these targeted, deleted datasets is the Complete Case dataset, a set restricted to patients with at least one clinical data point for Age, Sex, Height, Weight, BMI, Systolic and Diastolic blood pressures.

Hypothesis Set

Like the targeted, deleted datasets, the same hypothesis library generated in the previous section was used in this analysis.

Imputation Methods

The following methods were used to replace missing data in a copy of each targeted, deleted dataset.

Simple Mean, Median

All missing values are replaced with the mean or median value, respectively, for that variable across the entire dataset. Replacement was performed in Python.

Conserve Mean, Median

In the two cohorts defined by the current hypothesis test in the validity analysis, missing values are replaced with the mean or median value, respectively, of the first cohort. The intent is to err on the side of minimizing the difference between cohorts. Replacement was performed in Python.

Target Mean, Median

In the two cohorts defined by the current hypothesis test in the validity analysis, missing values are replaced with the mean or median value, respectively, of the cohort. The intent is to tailor the value replacement to the cohort being studied. Replacement was performed in Python.

Linear Imputation

A linear model is constructed out of available data and missing values are filled using this model. This method was applied with the lmImpute function in the Imputation package for R. Due to limitations with the number of data points which could be used, as well as the fact that this function was not configured for categorical variables, the results of this method are not reported.

kNN

Replacement value is the mean of the five nearest patients as calculated by spatial similarity. This method was applied with the kNNImpute function of the Imputation package for R.

GBM

GBM is a technique to impute missing values when large quantities of categorical and numerical data are present. Expectation maximization is performed with boosted trees. This method was applied with the gbmImpute function in the Imputation package for R.

Imputation Application and Reporting

There are ten targeted, deleted datasets for each target. For each of these sets, a copy is created with the missing summary data values imputed by one of the methods described above. Each dataset is then processed to assign the categorical variables such as cardiac risk and obesity. As in the previous investigation, these candidate sets are validated against the original Complete Case dataset using the existing hypothesis library. For this investigation, only accuracy is reported and compared to a baseline accuracy of the accuracy of the targeted, deleted dataset without imputation or replacement of missing data.

Results

Accuracy for Simple Mean and Median, Conserve Mean and Median, Target Mean and Median, kNN, and GBM imputation methods in terms of the level of missing data is presented in Figure 4-1. The baseline used for comparison of imputation efficacy is the accuracy of the targeted, deleted dataset without imputation or replacement of missing data.

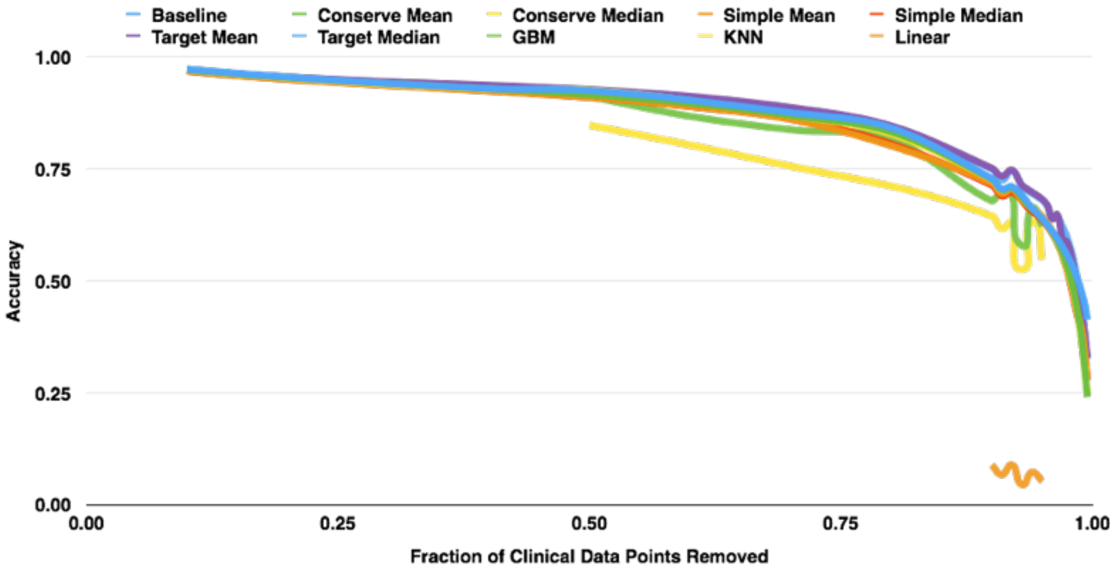


Figure 4-1: Graph of Accuracy for various imputation methods vs. the Complete Case Dataset. No method performs significantly better than Baseline, which is accuracy calculated with missing data in place.

In general, no imputation method performed better at maintaining the accuracy of hypothesis test results as compared to the original, complete dataset. At around data removal of .9, Target Mean and Median have a slight improvement in accuracy over baseline. The more complicated imputation method of kNN is actually significantly worse than baseline and all of the simpler methods.

Limitations

This investigation reuses the same datasets, hypotheses, and validation method as previous investigations, and shares the same limitations. While a good cross-section of methods, at differing levels of complexity, were evaluated, this investigation was not exhaustive. It is possible a more specialized, focused method could maintain accuracy at high levels of missing data. In addition, these methods were only applied to a data MAR scenario. While the data MAR here replicates a broadly true clinical data scenario, these methods may also perform better in a data MNAR scenario.

Discussion and Contributions

In general, no imputation method performed better than baseline. This may be because the goal of these methods, from the very simple mean replacements to the very complex machine learning approaches, is to replicate the summary statistics and distributions of the existing data. In the validation method reused in this investigation, what is being tested is whether randomly defined cohorts have a significant difference. In other words, imputation methods may be very good at replicating summary statistics but do not have any effect at preserving or replicating the differences between tiny subsets of the dataset.

The conclusion of the prior investigation was that data missingness has very little effect on validating a dataset until present at very high levels, that missing data at the patient level should not exclude a patient from a dataset. Given the conclusion of that study and the result of this study, it is recommended that no missing data replacement be performed for clinical data intended for research. The caveat to the recommendation is that this investigation was only performed using data MAR random scenarios. While clinical data was primarily MAR in the clinical dataset used here, the class of missing data in a clinical dataset should be determined before deciding whether to replace missing data by any method.

While ultimately negative, the results of this investigation are still a significant contribution. As reviewed in the background section, ranges of data replacement and imputation methods have been recommended for clinical data. However, these methods are not evaluated in terms of maintaining accuracy of a dataset in the face of increasing levels of missing data, nor are they ever evaluated against each other. Secondly, the result is surprising. It was the expectation for this investigation that missing data replacement would be beneficial and that more complex methods would perform better than simpler options. In reality, no method performs better than leaving the missing data in place and the recommendation of this investigation is to not replace missing data when the data is MAR.

Study 3B: Nearest Neighbor Matching between Clinical and Survey Records

Research Question: Can nearest neighbor matching replace matching based on identifiable data?

Background

In WICER, survey participants are matched to their own clinical records primarily by looking at name and birthdate. While additional manual efforts raise the number of matches, some survey participants do not have a clinical record in the CDW. It might be possible to assign such survey participants a "closest" record whose demographics and health measurements are literally the nearest to those recorded by the survey. A further application of this method is that, if successful, it could be used to match patients within de-identified datasets without using a research identifier.

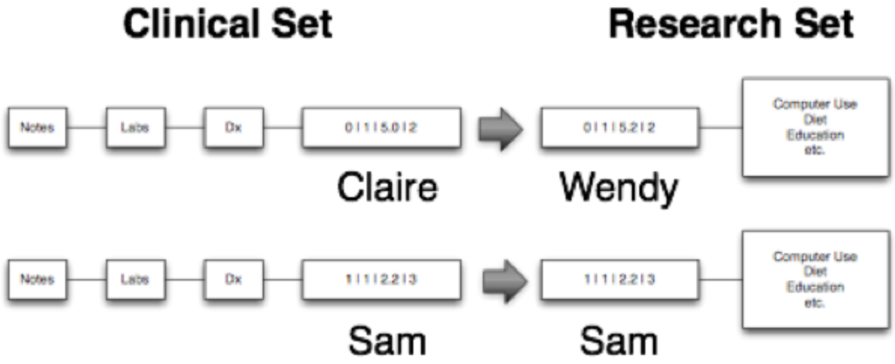


Figure 4-2: Graphical example of nearest-neighbor matching

An example of this nearest-neighbor matching is presented in Figure 4-2. Here, Sam on the bottom represents a survey participant whose clinical measurements are nearly identical to his survey measurements. The “distance” between these two measurement vectors is virtually zero, making his closest clinical match his “true” match, or himself. The record on the top is a composite patient of an individual named Wendy, who was interviewed for the Household Survey, and Claire, who has a record in the CDW and lives in Washington Heights. The data for their demographics, BMI, blood pressure, etc. are almost identical, suggesting that someone like Wendy might have a clinical record like that of Claire.

Methods

The nearest neighbor matching method relies on complete data for each record, so each data source was limited to only those records with complete data for age, sex, height, weight, BMI, systolic and diastolic blood pressure. Additionally, the clinical data source was limited to Hispanic or Unknown ethnicities.

Each record is treated as a vector. For each survey record, the spatial distance to every clinical record is calculated and sorted based on distance. Each records place in the list is known as its "rank". The record with the smallest distance is the best match, but the process is evaluated by determining the number of records between the closest match based on spatial distance and the true match.

Results

This process was evaluated by looking at survey participants who do have a matching clinical record based on name and birthdate, then calculating how many clinical records stand between the true match and the closest match. Out of nearly five thousand matches, 6% of survey participants had their true matching clinical record as their closest clinical record. For 75% of survey participants, the true match was within the top 1,300 (out of nearly 29,000) clinical records.

The reverse matching rates (clinical record to nearest survey) were also calculated. The same 6% most closely match their true survey record. However, 75% percentile matching is approximately 2500 (out of nearly 5000). Full results are presented below.

```
SURVEYS MATCHING TO CLINICAL
Average rank of 1881 of 28749.

5th percentile rank is 0.
25th percentile rank is 22.
50th percentile rank is 185.
75th percentile rank is 1283.
95th percentile rank is 11718.

6 percent had a rank of 1.
18 percent had a rank of less than 10.
41 percent had a rank of less than 100.
71 percent had a rank of less than 1000.
82 percent had a rank of less than 2000.


CLINICAL MATCHEDS TO SURVEY
Average rank of 1599 of 4872.

5th percentile rank is 0.
25th percentile rank is 56.
50th percentile rank is 2025.
75th percentile rank is 2535.
95th percentile rank is 3568.

6 percent had a rank of 1.
16 percent had a rank of less than 10.
27 percent had a rank of less than 100.
40 percent had a rank of less than 1000.
49 percent had a rank of less than 2000.
```

Discussion and Conclusion

Six percent of records match exactly themselves in both directions. Survey records are much more likely to be spatially close to their true matches in the clinical data source than the opposite. This is true in terms of the absolute rank, but, considering how many fewer survey records are available as potential matches, much worse in relative terms. This finding suggests that the members of the pool of survey records more closely resemble each other than the members of the pool of clinical records. While it may mean that in some sense it does not matter which survey record is matched to, it is certainly more difficult to find true matches.

In contrast, the much larger pool of clinical records contains a greater variability and distance between individuals, meaning the matching process does a much better job of finding true matches. While this reverse matching is still better than chance, it is not as close in relative terms as the performance of the original direction, which was surveys matching against the clinical records. This finding suggests that the survey records are more homogeneous than the clinical records, which is backed up by the demographics of the survey population. Furthermore, it may be that the clinical matches are not as useful as they appear on the surface. If the 1,300 clinical records to which one survey participants closely matches are the same clinical records for every survey participant, then the matching is effectively meaningless.

Additional limitations on this method are the known, albeit small, biases in measurement between the clinical and research measurements at the cohort level, and the especially high variability at the patient level between these measurements. While the presence of these limitations may make the 6% true match rate surprisingly high, it is not high enough to consider nearest-neighbor matching a viable substitute for more exact matching methods.

Study 3C: Propensity Score to Indicate Representativeness in a Dataset

Research Question: Can an individual's representativeness in a dataset be usefully represented with a point statistic?

Background

A propensity score is the marginal probability of treatment, effect, or inclusion in a group given any number of possible covariates. For example, in a hypertension study it may be that older people are simultaneously more likely to suffer from a heart attack and take a certain drug. A simple investigation into the drug's effects might conclude the drug was a risk factor for heart attacks. A propensity score based approach would group study participants by likelihood of taking the drug rather than simply exposure to the drug. Older people who took the drug would be pooled with any other group of people who were also very likely to take the drug. These people would only be compared to individuals from the same group who did not take the drug. Similarly, individuals from groups which were very unlikely to take the drug would only be compared to individuals from those groups which did take the drug. In this way, the effect of exposure to the drug is separated from whatever factors confound that exposure.

The intuition behind this study was that a similar score might be useful to account for sampling bias in a dataset. If a clinical dataset contains significant sampling bias, presumably favoring sicker individuals than the general population, then research conclusions from that dataset might not be generalizable to a healthier population. A propensity score for representativeness in a dataset could be used to stratify the dataset prior to analysis. If a research conclusion is true for the over-represented segment of the dataset as well as the under-represented segment, then it is likely to be generalizable to the general population.

Given selection bias in a clinical dataset, a 'bias propensity score' would account for the marginal probability of inclusion in a dataset given any number of potential health covariates. With access to the source for clinical data and the Household Survey population, it is possible to estimate the representativeness of any person in the clinical dataset based on the number of people similar to them in the research dataset. In this way, the representativeness of any arbitrary cohort drawn from clinical database could be determined.

Methods

The representativeness of a person in a dataset is calculated as the fraction of the dataset composed of individuals of the same gender, age by decade, hypertension risk category, and BMI category. The bias propensity score for a person in the clinical dataset is calculated as their representativeness in the clinical dataset divided by the representativeness in the research dataset. In the event that a particular intersection of categories was not represented in the research dataset, its representativeness was set to 0.5. A propensity score >1.0 implies that an individual is over-represented, or more common, in the clinical dataset than the research dataset. A propensity score <1.0 implies that an individual is actually less represented. The bias propensity score is different than a simple case weight, as was used to re-weight the cohorts to a census distribution in the first aim, because it also includes covariates like hypertension risk and BMI categories which have direct health implications.

The performance of the bias propensity score was investigated via case study. A cohort was requested to study the difference between measured blood pressures between the ACN survey setting and the clinical record. The bias propensity score was calculated for every member of the cohort and the relationship between the scores and the results examined in that context.

Results and Discussion

The case study research cohort had 511 members and a median bias propensity score of .85. Bias propensity score had a range of .04 to 5.66. The top and bottom quintile of the cohort as grouped by propensity score did not have significantly different results.

This case study demonstrates that a bias propensity score is fairly easy to calculate and could meaningfully indicate the representativeness of any individual in a dataset. The median propensity score indicates that this cohort is fairly representative of a research cohort, containing individuals only slightly less represented in the clinical dataset than the research dataset.

However, this case study also demonstrates the limitations of this approach. While a bias propensity score might be interesting, in this case study there was no noticeable effect on study outcomes. Also, it demonstrates how the selection criteria for a research study have a great deal more effect on the makeup of a research cohort than any underlying sampling bias in the dataset. While the utility of a bias propensity score should be investigated in other cohorts, it was not a useful statistic to calculate in this case study.

5. Conclusions and Future Work

Conclusions

This thesis examines issues and opportunities surrounding electronic clinical data for research. In the first aim we examined a clinical dataset for sampling and measurement bias primarily through comparison to a research quality dataset drawn from the same population. While there is only a direct overlap between the datasets in a few variables, there was enough of an overlap to detect some significant, but probably not meaningful, differences between the datasets. We reported some considerations for replicating work of this kind, most important being the inclusion of a matched set, or a set of individuals with records in both data sources, to parse the difference between measurement and sampling bias.

The more interesting outcome of the first aim is the idea of three categories of clinical variable. We identified completely accurate, simple measurement, and inferred information as the three categories of clinical variable. The completely accurate category encompasses information like addresses, phone numbers, and other personally identifying information, which might be expected to remain the same for an individual for a given time. The second category, simple measurement, includes height, weight, and blood pressures, or variables with a simple definition that derive from a single measurement. While there may be systematic differences, they were typically small in our datasets. These two findings suggest datasets or analyses using highly structured data (e.g. age, gender) and point measurements (e.g. weight, blood pressure) collected from a clinical process should not have meaningfully different results than data collected as part of a structured research process.

The third category is inferred information, or clinical variables which draw from multiple sources to infer a complex status such as diabetes. These could not be considered accurate for population summary purposes in this dataset, which is to say the summary values in the clinical dataset were very different from the research dataset, but parts of the clinical phenotype can be used to design a study toward different purposes such as maximizing sensitivity, specificity, or positive predictive value.

To further investigate the dynamics of an inferred variable, we used the eMERGE Type 2 Diabetes Phenotyping Algorithm. eMERGE is a consortium which aims to build precise and portable phenotyping algorithms for electronic clinical data. In this instance, the diabetes phenotyping algorithm uses a combination of diagnoses, medications, and lab results to infer a diabetes for a genetic study, appropriately trading off sensitivity to support a research goal of high specificity. However, this phenotype should not be used, for example, to populate a diabetes registry where high sensitivity should be the primary concern. By validating the components of the eMERGE diabetes phenotyping algorithm, singly and in simple combinations, we can demonstrate how the same criteria might be repurposed to other research goals.

Aim 2 revolved around building and demonstrating a method to validate existing datasets for research. Rather than comparing the summary statistics for variables in a dataset, it may be more interesting and meaningful to compare the answers a dataset might provide. For example, it might be more meaningful that the mean blood pressure 70 year-olds is higher than 40 year-olds in a candidate dataset and that is also true in a reference dataset. By randomly generating and evaluating many such hypotheses, we can create a highly granular portrait of a datasets and the similarity of the answers they might provide.

The conclusion of this validity analysis method on our clinical and research datasets is that the clinical data is no more different than the research dataset in terms of the answers it provides than two random clinical data samples are from each other. The validity analysis method was also used to investigate the effect of data MAR on a clinical dataset. At levels common in our clinical dataset, or up to 60% of the data missing, the dataset still scores 90%. Even at 99.9% of data removed, if a significant difference can be detected in the dataset, that difference is almost certainly present in the complete dataset. Again, the results of the second aim bolster the result of the first aim that electronic clinical datasets comprised of structured, simple measurements might be effectively used for research regardless of concerns about measurement and sampling biases, and bias due to missing data.

A third aim collects three studies which investigate the use of more advanced techniques on gaps and opportunities for using electronic clinical data for research. These studies involved evaluating missing data imputation methods using the existing deleted data datasets, whether nearest-neighbor vector matching could be used to substitute for more exact matching methods based on name and birthdate, and whether propensity scores can be used to meaningfully indicate a patient's representativeness in a dataset. While none of these studies ultimately had a positive result, they demonstrate a willingness to look at the edges of the utility of electronic clinical data and suggest directions for future research.

Future Work

While the results and conclusions of this thesis support the idea that electronic clinical data might be used for research, the limited scope may negatively impact the generalizability of the conclusions. Instead, the impact of this thesis might be as a set of analytic methods and an indication for future work.

From the first aim there are a set of considerations which describe how to use a research dataset to validate a clinical dataset. The opportunities to perform this task may be limited. However, the methods described could be used to validate various research cohorts or their sampling criteria against the database from which they were drawn. These tasks could be performed as a sensitivity analysis on a dataset and resulting measurement and sampling differences from the source dataset can be rigorously described. Like the difference in heights between the Survey and Clinical datasets, it is not as if any one answer is more correct than the other, but that the capacity to describe the source of the difference adds to the quality of a dataset.

Similarly the second aim was limited by its original purpose to the same set of overlapping variables between the Clinical and Survey datasets. However, the method suggests an immediate application in the same vein as described for the first aim. Here the validity analysis method would be used to demonstrate the accuracy of a research dataset against its data source or to demonstrate that a de-identified subset can provide the same answers as a whole dataset. In both cases the immediate next steps involve generalizing the methods used in this thesis beyond the limitations of the datasets used. Methods must be generalized to deal with more types of variables and with issues of temporality.

Contributions

Concerns about bias in electronic clinical data may be widely reported but are difficult to quantify. This thesis includes an explicit investigation of selection and measurement biases in electronic clinical data through comparison to a higher quality data source on the same individuals. Such an investigation simply would not be possible without the setting and resources available in our institution. Secondary investigations resulted in a novel method to compare datasets by the answers they provide to a library of randomly generated hypothesis tests. Use of this method supports the conclusions of the first aim, that any biases in our electronic clinical dataset were largely insignificant, but also shows promise for future work in validating arbitrary datasets for research.

6. Bibliography

1. *A First Look at the Volume and Cost of Comparative Effectiveness Research in the United States*. 2009, Academy Health.
2. *Report to the President and The Congress*. 2009, Federal Coordinating Council for Comparative Effectiveness Research.
3. *PROSPECT Studies: Building New Clinical Infrastructure for Comparative Effectiveness Research*.
4. Remontet, L., et al., *Is it possible to estimate the incidence of breast cancer from medico-administrative databases?* Eur J Epidemiol, 2008. **23**(10): p. 681-8.
5. Couris, C.M., et al., *Breast cancer incidence using administrative data: correction with sensitivity and specificity*. J Clin Epidemiol, 2009. **62**(6): p. 660-6.
6. Manuel, D.G., L.C. Rosella, and T.A. Stukel, *Importance of accurately identifying disease in studies using electronic health records*. BMJ, 2010. **341**: p. c4226.
7. Hripcsak, G., et al., *Bias associated with mining electronic health records*. J Biomed Discov Collab, 2011. **6**: p. 48-52.
8. Thygesen, L.C. and A.K. Ersboll, *When the entire population is the sample: strengths and limitations in register-based epidemiology*. Eur J Epidemiol, 2014. **29**(8): p. 551-8.
9. Weiskopf, N.G. and C. Weng, *Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research*. J Am Med Inform Assoc, 2013. **20**(1): p. 144-51.
10. Hogan, W.R. and M.M. Wagner, *Accuracy of data in computer-based patient records*. J Am Med Inform Assoc, 1997. **4**(5): p. 342-55.
11. Kahn, M.G., B.B. Eliason, and J. Bathurst, *Quantifying clinical data quality using relative gold standards*. AMIA Annu Symp Proc, 2010. **2010**: p. 356-60.
12. Kriegsman, D.M., et al., *Self-reports and general practitioner information on the presence of chronic diseases in community dwelling elderly. A study on the accuracy of patients' self-reports and on determinants of inaccuracy*. J Clin Epidemiol, 1996. **49**(12): p. 1407-17.
13. Martin, L.M., et al., *Validation of self-reported chronic conditions and health services in a managed care population*. Am J Prev Med, 2000. **18**(3): p. 215-8.
14. Okura, Y., et al., *Agreement between self-report questionnaires and medical record data was substantial for diabetes, hypertension, myocardial infarction and stroke but not for heart failure*. J Clin Epidemiol, 2004. **57**(10): p. 1096-103.
15. Kahn, M.G. and D. Ranade, *The impact of electronic medical records data sources on an adverse drug event quality measure*. J Am Med Inform Assoc, 2010. **17**(2): p. 185-91.
16. Palepu PR, e.a., *Assessment of accuracy of data obtained from patient-reported questionnaire (PRQ) compared to electronic patient records (EPR) in patients with lung cancer*. J Clin Oncol, 2013. **31**(31): p. 40.
17. Richesson, R.L., et al., *A comparison of phenotype definitions for diabetes mellitus*. J Am Med Inform Assoc, 2013. **20**(e2): p. e319-26.
18. *OMOP Design and Validation*. Available from: http://omop.fnih.org.
19. Tannen, R.L., M.G. Weiner, and D. Xie, *Use of primary care electronic medical record database in drug efficacy research on cardiovascular outcomes:*

comparison of database and randomised controlled trial findings. BMJ, 2009. **338**: p. b81.

20. Libby, A.M., et al., *Comparative effectiveness research in DARTNet primary care practices: point of care data collection on hypoglycemia and over-the-counter and herbal use among patients diagnosed with diabetes*. Med Care, 2010. **48**(6 Suppl): p. S39-44.
21. Raboud, J.M., et al., *Impact of missing data due to dropouts on estimates of the treatment effect in a randomized trial of antiretroviral therapy for HIV-infected individuals. Canadian HIV Trials Network A002 Study Group*. J Acquir Immune Defic Syndr Hum Retrovirol, 1996. **12**(1): p. 46-55.
22. Ware, J.H., *Interpreting incomplete data in studies of diet and weight loss*. N Engl J Med, 2003. **348**(21): p. 2136-7.
23. Molnar, F.J., et al., *Have last-observation-carried-forward analyses caused us to favour more toxic dementia therapies over less toxic alternatives? A systematic review*. Open Med, 2009. **3**(2): p. e31-50.
24. *WICER: Washington Heights/Inwood Informatics Infrastructure for Community-Centerede Comparative Effectiveness Research*.
25. Cox, E., et al., *Good research practices for comparative effectiveness research: approaches to mitigate bias and confounding in the design of nonrandomized studies of treatment effects using secondary data sources: the International Society for Pharmacoeconomics and Outcomes Research Good Research Practices for Retrospective Database Analysis Task Force Report--Part II*. Value Health, 2009. **12**(8): p. 1053-61.
26. Danaei, G., M. Tavakkoli, and M.A. Hernan, *Bias in observational studies of prevalent users: lessons for comparative effectiveness research from a meta-analysis of statins*. Am J Epidemiol, 2012. **175**(4): p. 250-62.
27. Ionescu-Ittu, R., M. Abrahamowicz, and L. Pilote, *Treatment effect estimates varied depending on the definition of the provider prescribing preference-based instrumental variables*. J Clin Epidemiol, 2012. **65**(2): p. 155-62.
28. *Baseline Data: CONSORT: Transparent Reporting of Trials*. . 2015.
29. Koebnick, C., et al., *Sociodemographic characteristics of members of a large, integrated health care system: comparison with US Census Bureau data*. Perm J, 2012. **16**(3): p. 37-41.
30. *OMOP OSCAR*.
31. *Annual Progress Report Reporting period: 1/1/2012 - 12/31/2012. Office of the National Coordinator for Health Information Technology: AREA 4: Secondary Use of EHR Data (SHARPn) Program*.
32. Bakken, S., *Washington Heights/Inwood Informatics Infrastructure for Community-Centered Comparative Effectiveness Research (WICER)*. 2010-2013 AHRQ ARRA Infrastructure Projects, 2013.
33. McCarty, C.A., et al., *The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies*. BMC Med Genomics, 2011. **4**: p. 13.
34. Kho, A.N., et al., *Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study*. J Am Med Inform Assoc, 2012. **19**(2): p. 212-8.

35. JT, P., *Type 2 Diabetes Mellitus*. 2012.

36. Wei, W.Q., et al., *Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus*. J Am Med Inform Assoc, 2012. **19**(2): p. 219-24.

37. Martin, G.W., D.A. Wilkinson, and B.M. Kapur, *Validation of self-reported cannabis use by urine analysis*. Addict Behav, 1988. **13**(2): p. 147-50.

38. AK, P.J.A.P.T.J.L.M.Q.J.G., *A highly specific algorithm for identifying asthma cases and controls for genome-wide association studies*. AMIA Annu Symp Proc, 2009: p. 497-501.

39. E, S.S.W.R.D.L.D.S.C.K.J.J., *CREX: Utility of a Computerized Methodology to Identify Health Conditions Using EMR for GWAS, in the Kaiser Permanente Research Program on Genes, Environment, and Health*. Clinical Medicine and Research, 2013. **11**(3): p. 149.

40. Kahn, M.G., et al., *A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research*. Med Care, 2012. **50 Suppl**: p. S21-9.

41. Pace, W.D., et al., *An electronic practice-based network for observational comparative effectiveness research*. Ann Intern Med, 2009. **151**(5): p. 338-40.

42. Schafer, J.L. and J.W. Graham, *Missing data: our view of the state of the art*. Psychol Methods, 2002. **7**(2): p. 147-77.

43. Haukoos, J.S. and C.D. Newgard, *Advanced statistics: missing data in clinical research--part 1: an introduction and conceptual framework*. Acad Emerg Med, 2007. **14**(7): p. 662-8.

44. Weiskopf, N.G., A. Rusanov, and C. Weng, *Sick patients have more data: the non-random completeness of electronic health records*. AMIA Annu Symp Proc, 2013. **2013**: p. 1472-7.

45. Rusanov, A., et al., *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**: p. 51.

46. Roth, P., *Missing Data: A Conceptual Review for Applied Psychologists*. Personnel Psychology, 1994. **47**(3).

47. Myers, W., *Handling Missing Data in Clinical Trials: An Overview*. Drug Information Journal, 2000. **34**: p. 525-533.

48. Gorelick, M.H., *Bias arising from missing data in predictive models*. J Clin Epidemiol, 2006. **59**(10): p. 1115-23.

49. Little, R.J., et al., *The prevention and treatment of missing data in clinical trials*. N Engl J Med, 2012. **367**(14): p. 1355-60.

50. Ghosh S, P.P. *Assessing Bias Associated With Missing Data from Joint Canada/U.S. Survey of Health: An Application*. in *JSM*. 2008. Denver, CO.

51. Schafer, J.A., N.K. Kjesbo, and P.P. Gleason, *Dronedarone: current evidence and future questions*. Cardiovasc Ther, 2010. **28**(1): p. 38-47.

52. Lohr, K.N., *Comparative effectiveness research methods: symposium overview and summary*. Med Care, 2010. **48**(6 Suppl): p. S3-6.