Leveraging patient-provided data to improve understanding of disease risk

Fernanda Caroline da Graça Polubriaginof

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

2018

ABSTRACT

Leveraging patient-provided data to improve understanding of disease risk

Fernanda Caroline da Graça Polubriaginof

Patient-provided data are crucial to achieving the goal of precision medicine. These data, which include family medical history, race and ethnicity, and social and behavioral determinants of health, are essential for disease risk assessment. Despite the well-established importance of patient-provided data, there are many data quality challenges that affect how this information can be used for biomedical research.

To determine how to best use patient-provided data to assess disease risk, the research reflected in this dissertation was divided into three overarching aims. In Aim 1, I focused on determining the quality of race and ethnicity, family history and smoking status in clinical databases. In Aim 2, I assessed the impact of various interventions on the quality of these data, including policy changes such as the implementation of the requirements imposed by the Meaningful Use program, and patient-facing tools for collecting and sharing information with patients. In addition to these evaluations, I also developed and evaluated a method "Relationship Inference from the Electronic Health Record" (RIFTEHR), that infers familial relationships from clinical datasets. In Aim 3, I used patient-provided data to assess disease risk both at a population level, by estimating disease heritability, and at an individual level, by identifying high-risk individuals eligible for additional screening for a common disease (diabetes mellitus) and a rare disease (celiac disease).

My research uncovered several data quality concerns for patient-provided data in clinical databases. When assessing the impact of interventions on the quality of these data,

I found that policy interventions led to more data collection, but not necessarily to better data quality. In contrast, patient-facing tools did increase the quality of the patient-provided data. In the absence of high-quality patient-provided data for family medical history, I developed and evaluated a method for inferring this information from large clinical databases. I demonstrated that electronic health record data can be used to infer familial relationships accurately. Moreover, I showed how the use of clinical data in conjunction with the inferred familial relationships could determine disease risk in two studies. In the first study, I successfully computed disease heritability estimates for 500 conditions, some of which had not been previously studied. In the second study, I identified that screening rates among family members that are considered to be at high-risk for disease development were low for both diabetes mellitus and celiac disease.

In summary, the work represented in this dissertation contributes to the understanding of the quality of patient-provided data, how interventions affect the quality of these data, and how novel methods can be applied to troves of existing clinical data to generate new knowledge to support research and clinical care.

Contents

This page intentionally left blank.

List of Figures

List of Tables

This page intentionally left blank.

Acknowledgements

This dissertation would not have been possible without the help and support of many people. First, I would like to thank my advisor, Dr. David K. Vawdrey. He is the most amazing advisor one could ever ask for. He was always helpful, supportive, and encouraging. He provided me with guidance throughout the graduate program, making my experience at DBMI extremely pleasant. Working with him has always been wonderful, and I feel very lucky to be his student.

I would also like to thank my committee members. Dr. Nicholas P. Tatonetti, who taught me so much and always treated me as a colleague and member of his team. The work in this dissertation would not have been the same without his guidance, thoughts, and hard work. I am very thankful for having had the opportunity to work with him. I would also like to extend my thanks to Dr. George Hripcsak, who has always provided insightful comments, helped me define the scope of my dissertation, and has always been available to provide guidance. I also would like to thank my external committee members. Dr. Peter Stetson was always available to meet and discuss, providing me with insightful thoughts, and Dr. Genevieve Melton-Meaux, who has always been open to conversations about numerous topics. Thank you both for agreeing to be a part of my committee. Without your participation, this work would not have been possible.

In addition to my advisor and my committee members, there are many more who contributed during the past four years. I would like to thank all my co-authors, who have been instrumental in the success of all studies here presented. I would also like to thank my academic advisor, Dr. Carol Friedman, who led the training program at DBMI and provided me with guidance to successfully complete the graduate program. In addition, I would like to thank Marina Bonanno, who was always available to help with any challenges faced by the students and was always so knowledgeable, facilitating every step towards the completion of my Ph.D. degree. I would also like to thank AHRQ (R01HS021816, PI: Vawdrey), which supported the work in this dissertation.

I would also like to thank my lab partners, from both Dr. Vawdrey's and Dr. Tatonetti's lab, and other DBMI students. Special thanks to Dr. Jennifer E. Prey, who was a lab mate and became my chosen "big sister." Also special thanks to Michelle Chau, who started the program in the same year as I did, and became a wonderful friend. Additionally, I would like to thank many of the students and post-docs who were supportive, provided invaluable feedback in practice presentations, including Yun Hao, Andrew Chiang, Jonathan Chang, Amelia Averitt, Mollie McKillop, Alexandre Yahi, Dr. Rami Vanguri, and Joseph Romano. I would like to especially thank those of you who also became invaluable co-authors in the work described here.

Thank you to all my friends, who were instrumental in supporting me throughout my career.

I would like to extend a very special thank you to my family. I would like to thank my parents, Regina da Graça Polubriaginof and Paulo Domingos Polubriaginof, and my sister, Luiza Cristine da Graça Polubriaginof, for your unconditional love and support. Without

you, I would have never made so far. You have always been there for me, during highs and lows, even when we were miles apart. While being far away is difficult, distance has never been a problem. You have always been so present and supportive and always helped me overcome every challenge. You taught me that everything is possible; you just have to believe it. I am so grateful for having the most wonderful family. I would also like to thank my fiancé, Dr. Silis Y. Jiang, who was always the most caring and supportive. He was there for me at every step, helping me in every way, and making me happy on a daily basis. I am very lucky to have had you by my side. There are not enough words to express my gratitude to the four of you. I am very fortunate to have you in my life.

Lastly, I would like to thank the patients from NewYork-Presbyterian Hospital (Columbia University Medical Center and Weill Cornell Medical Center) and from Mount Sinai Health System. Without you, none of this work would have been possible.

This page intentionally left blank.

To my parents, Regina da Graça Polubriaginof and Paulo Domingos Polubriaginof, who were always present, and without their support none of this would have been possible. I am so grateful for your endless love.

This page intentionally left blank.

Chapter 1

Introduction

Patient-provided information, such as family medical history, is crucial to achieving the goal of precision medicine. Precision medicine focuses on disease prevention and treatment while accounting for a patient's variability (Collins and Varmus 2015). In precision medicine, there is a need for understanding not only genetic causes of disease but also the impact of environmental and behavioral factors on the disease (Aronson and Rehm 2015; Cutting 2010; Maher 2008). Patient-provided information, such as family medical history, self-reported race and ethnicity, social and behavioral determinants of health, and past medical and surgical history, are important pieces of information that directly impact the risk of disease. For example, women with a family history of breast cancer can be at higher risk of developing the disease, and therefore, they are candidates for additional and/or early screening (Ozanne et al. 2009). Similar approaches are well established for a variety of conditions, including prevalent disorders, such as osteoporosis (U.S. Preventive Services Task Force 2011) and lipid disorders in adults (Helfand and Carson 2008). With a better understanding of disease risk, clinicians can personalize the care they deliver by risk-adjusted disease screening, prevention, and early diagnosis. While there has been increased interest in precision medicine and the patient-provided information that drives it (Adams and Petersen 2016; Aronson and Rehm 2015; Collins and Varmus 2015), there is a lack of research

on methods to most effectively capture patient-provided information in clinical databases.

Clinical databases, including those derived from electronic health records (EHRs), are an important resource for biomedical research and have previously been utilized to shed light on disease processes (Boland et al. 2015; Coopey et al. 2012; Hripcsak et al. 2016; Li et al. 2015; Ritchie, Andrade, and Kuivaniemi 2015; Wei and Denny 2015), including genetics (Kohane 2011; Polubriaginof et al. 2017; Wang et al. 2017), and on drug effectiveness and interactions (Dudley, Deshpande, and Butte 2011; Lorberbaum et al. 2016b; Tatonetti et al. 2012). Notwithstanding the utility of EHR data for research activities, there are concerns regarding the quality of these data (Ahmad et al. 2017; Aronsky and Haug 2000; Arts et al. 2002; Brennan and Stead 2000; Brown, Kahn, and Toh 2013; Hasan and Padman 2006; Hersh et al. 2013; Hogan and Wagner 1997; Hripcsak et al. 2011b; Kahn, Eliason, and Bathurst 2010; Lei 1991; Rusanov et al. 2014; Thiru, Hassey, and Sullivan 2003). Limited research has been conducted to assess the quality of patient-provided information in the EHR (Arsoniadis et al. 2015; Booth, Prevost, and Gulliford 2013; Chen et al. 2014; Lee, Grobe, and Tiro 2015; Melton et al. 2010), especially related to race and ethnicity, family history, and smoking status. Because patient-provided information is frequently stored in EHRs, it is critical to assess the quality of this information specifically.

Previous research has demonstrated that patients are an important and underutilized source of information (Ball and Lillis 2001; Staroselsky et al. 2006), and that patient-provided information can be used to overcome data quality issues (Arsoniadis et al. 2015; Staroselsky et al. 2006, 2008; Wu et al. 2014). Some efforts have been made to enhance the collection and use of patient-provided data. Those include broad policy initiatives such as the Meaningful Use financial incentive program in the United States, which included, for

example, standardization of the collection of family history (Centers for Medicare & Medicaid Services 2014a), smoking status (Centers for Medicare & Medicaid Services 2010), and race and ethnicity (Centers for Medicare & Medicaid Services 2014b). In addition, patient-facing tools, such as online patient portals, have been implemented in diverse clinical settings. These portals allow patients to review their clinical information as available in the EHR, and in some cases to record information that is fed back to the EHR (Cimino, Patel, and Kushniruk 2001; Collins et al. 2011; Delbanco et al. 2010; Greenhalgh et al. 2008; Grossman et al. 2017; Halamka, Mandl, and Tang 2008; Hassol et al. 2004; Kaelber et al. 2008; Leveille et al. 2012; Nazi et al. 2010; Nazi et al. 2015; Pyper et al. 2004; Ralston et al. 2007; Reti et al. 2010; Tang and Lee 2009; Walker et al. 2011). These patient-facing tools have been employed to help maintain up-to-date records (Staroselsky et al. 2006, 2008), to promote disease screening and prevention through preventive health services (Murabito et al. 2001; Reid et al. 2009; Staroselsky et al. 2006), to facilitate the assessment of disease risk and healthcare disparities (Chin 2015; Douglas et al. 2015; Kressin 2015; Woods, Evans, and Frisbee 2016), to manage disease symptoms (Basch et al. 2009, 2017; Pakhomov et al. 2008; Weingart et al. 2005), and to improve the medication reconciliation process (Dullabh et al. 2014; Finkelstein 2006; Kogut et al. 2014; Weingart et al. 2008).

While a substantial body of knowledge related to patient-provided data and patient-facing tools exists, many questions remain. Three critical knowledge gaps are 1) a lack of awareness of the quality of patient-provided information, particularly for race and ethnicity, family history, and smoking status, 2) a limited understanding of the impact that various interventions have on data quality, including policy decisions such as the Meaningful Use program, and 3) the question of whether EHR data can be used in combination with patient-

provided information to assess disease risk.

The purpose of this thesis is to establish methods and tools 1) for assessing the quality of patient-provided data in clinical databases, 2) for studying the impact that distinct interventions have on influencing the quality of patient-provided data, and 3) for using patient-provided information to better understand disease risk and better inform clinical decisions. I explored the quality of patient-provided information, specifically race and ethnicity, family history, and smoking status. I evaluated the impact of various interventions designed to improve the quality of this information. Finally, I developed a novel method that uses patient-provided information to extract familial relationships from existing clinical databases, and I demonstrated how these relationships, in combination with EHR data, may be used to better understand disease risk and support clinical research.

1.1 Specific aims

Aim 1: Assess the quality of patient-provided data in clinical databases.

Observational databases, including those containing EHR data, are a valuable resource for biomedical research (Boland et al. 2015; Coopey et al. 2012; Dudley, Deshpande, and Butte 2011; Hripcsak et al. 2016; Kohane 2011; Li et al. 2015; Ritchie, Andrade, and Kuivaniemi 2015; Wang et al. 2017; Wei and Denny 2015). However, there are concerns regarding the quality of EHR data, since these data are primarily collected as part of clinical care, and not for research purposes (Weiner and Embi 2009; Weiskopf and Weng 2013). Broadly, in this aim, I studied various data quality dimensions of patient-provided data such as race and ethnicity, family history, and smoking status. The purpose of the analysis was to assess the completeness, correctness, concordance, and plausibility of patient-provided data stored in clinical databases and to identify opportunities for improvement.

To achieve this goal, I performed three studies. In the first, I investigated the quality of race and ethnicity information in multiple clinical databases, both at a local level and national level, based on completeness, correctness, and concordance using a combination of data sources. In the second study, I focused on determining completeness of family history information in the EHR by analyzing a sample of observations recorded using structured and free-text note templates from a large academic medical center. And finally, in the third study, I investigated the quality of smoking status information at the same medical center, examining completeness, correctness, and plausibility of the data recorded. The results of

the three studies included in this Aim provided generalizable knowledge about the quality of patient-provided data in EHRs.

Aim 2: Evaluate methods for improving quality of patient-provided data.

Results from the studies in Aim 1 demonstrated generally poor data quality in race and ethnicity, family history, and smoking status based on the dimensions of quality analyzed. These results showed a need to consider whether various types of interventions could improve the quality of patient-provided data. Previous research on patient-provided information sought to improve data collection and/or data quality using patient-facing tools or by extracting important concepts from narrative text using natural language processing techniques (Arsoniadis et al. 2015; Baumgart, Postula, and Knaus 2015; Bill et al. 2014; Chen et al. 2012, 2014; Cohn et al. 2010; Feero 2013; Giovanni and Murray 2010; Hoyt et al. 2013; Hulse et al. 2010; Masterson Creber et al. 2016; Melton et al. 2010; Murray et al. 2013; Orlando et al. 2013; Ozanne et al. 2009; Peace, Valdez, and Lutz 2012; Pyper et al. 2004; Staroselsky et al. 2006, 2008; Wang et al. 2016; Wilson et al. 2012a; Wu et al. 2014, 2015; Yoon, Scheuner, and Khoury 2003; Yoon et al. 2009). Some of the endeavors to improve the quality of patient-provided information occurred at the national level through policymaking efforts, such as the Meaningful Use program, while others occurred in local institutions with the development of new applications, such as patient-facing tools. While these interventions have been implemented to varying degrees, there has been little research assessing whether data quality has improved. In this Aim, I studied the impact of

policymaking efforts, patient-facing tools and informatics interventions on the quality of patient-provided information.

To evaluate methods for improving quality of patient-provided data, I performed three studies in Aim 2. In the first study, I analyzed the impact of the federal Meaningful Use program and the Hospital Consumer Assessment of Healthcare Providers and Services (HCAHPS) survey in the quality of race and ethnicity information in the EHR. Additionally, I analyzed the effect of a local informatics intervention that allowed patients to review and update their race and ethnicity information obtained from the EHR. In the second study, I developed and evaluated a method that extracts familial relationships from clinical databases using emergency contact information, a type of patient-provided data. Currently, clinical databases do not record familial relationships, an important piece of information when studying disease risk. In the third study, I analyzed the impact of the Meaningful Use program on the completeness, correctness, and plausibility of smoking status in the EHR. The results of these studies provided insight into the impact of various types of interventions on the quality of patient-provided data.

Aim 3: Use patient-provided data to assess disease risk.

Accurately collecting patients' family health history is important for the implementation of precision medicine in the clinical setting (Aronson and Rehm 2015; Guttmacher, Collins, and Carmona 2004). The predictive value of family history for any given trait is directly related to the fraction of phenotypic variance attributable to genetic factors, called heritability (Tenesa and Haley 2013; Visscher, Hill, and Wray 2008), as well as to shared

environmental factors. Knowledge of disease heritability combined with family history information is clinically useful for identifying risk factors, estimating disease risk, customizing treatment, and tailoring patient care (Chatterjee, Shi, and García-Closas 2016). Unfortunately, heritability studies are typically time-consuming and costly. They are often designed as prospective family-based studies and therefore have limited sample sizes, limiting their power to detect associations. Further, identification of high-risk individuals in the clinical setting can be complex. The proper identification of high-risk patients is a crucial factor for the adherence to guidelines that target early screening and modified treatment for patients considered at risk for disease development. However, the study of screening rates is challenging due to the necessity of identifying patients that would have been deemed to be high-risk but were not recognized as such during a clinical encounter.

Fortuitously, EHRs provide a valuable resource of clinical information that is currently underutilized in genetic and clinical studies. EHRs are now widely adopted, capturing clinical data for millions of individuals as part of clinical care. The use of EHR data in heritability studies can overcome many of the challenges previously described, by increasing sample sizes, thus enabling researchers to study diseases and relationships between diseases that are currently poorly understood, and generate research hypothesis that can then be tested using traditional genetic studies. EHR data, combined with health information technology, can also support clinical research and clinical care. Identification of high-risk individuals in existing clinical databases can facilitate better and easier measurement of adherence to clinical guidelines.

In this Aim, I studied the use of EHR data to conduct genetic and clinical studies. Specifically, I used the familial relationships identified as part of Aim 2 to perform two studies

which estimated disease heritability and measured adherence to clinical guidelines. In the first study, I assessed disease heritability using clinical traits, both dichotomous and quantitative, recorded in the EHR. This study was conducted in three hospital systems from New York. In the second study, I measure the adherence to guidelines rate for two distinct conditions, diabetes mellitus and celiac disease. The results of these studies provided insight into the heritability of multiple clinical traits, demonstrating that EHRs are a valuable resource of clinical phenotypes. Further, it affirmed the ability to use EHR data to support clinical care by identifying patients that are at high-risk for disease development.

1.2 Significance and Contributions

This dissertation includes the following contributions: 1) an investigation of the pervasiveness of data quality issues in patient-provided information, 2) measurement of the effectiveness of different interventions to improve data quality of patient-provided information, 3) development and evaluation of a novel method to identify familial relationships from existing clinical databases using patient-provided data, and 4) generation of new knowledge based on patient-provided information.

In the first Aim, I extensively explored the data quality of patient-provided information, which is crucial to operationalizing precision medicine. Previous research has primarily focused on the quality of clinical data; whereas, in this Aim, I focused on studying the quality of patient-provided data in clinical databases. The results of this Aim showed significant data quality issues for race and ethnicity both at local institutions but also among national databases. For family history, I identified data completeness issues among family history information that was recorded in free-text format, which I found to be the strongly preferred method for recording and storing patient information. While assessing the quality of smoking status, I found many implausible changes to patients' smoking status. Together, these data quality issues point to challenges in the way patient-information is collected and stored. Furthermore, interventions to improve data quality certainly appear to have merit in light of the findings of this Aim.

In the second Aim, I assessed the impact of different initiatives on the quality of patient-provided data. I explored the effects of different types of interventions, such as policymaking efforts, patient-facing tools, and informatics algorithms, on improving the quality of

the information collected. In relation to race and ethnicity, I found that high-level policy changes, such as moving to a two-question format to capture race and ethnicity, increased collection of race and ethnicity information, but did not necessarily improve the quality of that data. However, giving patients the opportunity to review and correct the information was, in comparison, more successful. Similarly, assessing the impact of policy change on smoking status showed more data collection, but also more data quality issues. Results from Aim 1 showed that family history was poorly recorded in the EHR, and therefore, I assessed the feasibility of inferring this information from other sources as a way to overcome the completeness issues. In this Aim, I demonstrated for the first time the feasibility of using patients' emergency contact information, in combination with clinical data to accurately infer family history. Overall, in this aim, I demonstrated various ways interventions can be implemented to overcome data quality issues.

In the third Aim, I used patient-provided information along with the method to infer familial relationships developed in Aim 2 to generate new understanding of disease characteristics and disease management, demonstrating that the method developed opened up new avenues for biomedical research. This Aim focused on two aspects, the measure of disease heritability, and assessment of screening rates among high-risk individuals. Both of these areas of research are traditionally resource-intensive and lie in two distinct fields: genetics, and medicine. The results of this Aim showed that EHR data can be used to support genetic studies by providing an opportunity to study conditions previously not investigated due to the availability of larger sample sizes in EHRs. Further, this Aim also demonstrated the potential of EHR data to be used to assess screening rates among high-risk patients. The results of both of these studies demonstrate that EHR data has utility far beyond supporting

clinical care.

The contributions of this thesis include 1) an overview of the quality of patient-provided information in clinical databases, 2) an assessment of the impact of different interventions types on the quality of patient-provided data, 3) the development and evaluation a novel method that uses patient-provided information to generate an unique database that can support biomedical research, and 4) the use of readily available data to understand disease risk and assess disease screening rates among high-risk individuals. The results of the presented studies contributed to the understanding of data quality issues concerns regarding patient-provided information in clinical databases. Further, the studies assessed how interventions currently affect the quality of patient-provided data, ultimately building the foundational work of how to better collect patient-provided information as part of healthcare encounters and how EHRs could be designed to overcome these challenges. These studies also developed and validated a novel method that relies on data readily available in EHRs to extract familial relationships from existing clinical databases. And lastly, these studies demonstrated that the use of inferred family history information can support the execution of large clinical and genetic studies, having a significant impact on biomedical research.

1.3 Limitations

This thesis has several limitations. All studies presented in this thesis heavily relied on data available in the EHR, and therefore, are subject to limitations of EHR data to conduct research. Fragmentation of care, for example, is an important limitation when using EHR data for research. Patients often receive treatment at multiple healthcare systems, and therefore, the information available in a single institution may be incomplete. Studies from Aim 1 (studies 2 and 3), Aim 2 (studies 1 and 3), and Aim 3 (study 2) were performed at a single, large, urban academic medical center, and as such, the findings may not be generalizable to other institutions. Aim 2 (study 1) had a small sample size due to patient recruitment constraints, limiting the generalizability of the findings. Additionally, recruitment was only conducted with English-speaking participants, limiting the patient population in the study. Studies conducted in Aim 3 extracted phenotypes from billing codes and did not necessarily developed a careful EHR-phenotype for each analyzed trait. Biases in billing documentation may have affected the presented results.

This page intentionally left blank.

Chapter 2

Review of the Literature

Some consider that disease risk prediction began in 1948 with the Framingham Heart Study. This study successfully developed a risk prediction model for cardiovascular disease using clinical data (Dawber, Meadors, and Moore 1951). Since this study, risk assessment models have been developed for a variety of diseases, creating the basis of what is now called "precision medicine." Precision medicine focuses on disease prevention and treatment while considering individuals variability in genes, environment, and lifestyle (Collins and Varmus 2015; *U S National Library of Medicine, Precision medicine*). Despite the growing importance of genomics and genetic sequencing, a crucial part of precision medicine is patient-provided information. This information is a key element for individualizing disease screening, diagnosis, and treatment (Guttmacher, Collins, and Carmona 2004). Some patient-provided information, such as a patient's medical history, family history, allergies, and medication use, can guide clinicians to determine the best course of action based on the risk of disease development for the patient.

Patient-provided data are data collected directly from patients or caregivers (Basch 2010; Hirsch and Abernethy 2013), and include many important elements in clinical care. Previous work demonstrated patient-provided data to be critical in multiple tasks pertaining to clinical care, including disease risk assessment (Berry et al. 1997; Claus, Risch,

and Thompson 1994; Ozanne et al. 2013; Tyrer, Duffy, and Cuzick 2004; Wu and Orlando 2015), and medication reconciliation (Dullabh et al. 2014; Staroselsky et al. 2008; Weingart et al. 2008). Further, others have demonstrated that patient-provided information can support activities related to symptom management (Basch et al. 2009; Pakhomov et al. 2008; Weingart et al. 2005) and to disease screening and prevention (Murabito et al. 2001; Reid et al. 2009; Staroselsky et al. 2006). Patient-provided data have also been shown to be valuable for broader tasks, such as assessment of healthcare disparities in healthcare systems (Chin 2015; Douglas et al. 2015; Kressin 2015; Woods, Evans, and Frisbee 2016). Family history, allergies, adherence to treatment plans and preventive services are common examples. Less common examples may not be perceived as patient-provided information, such as chief complaint, history of present illness, and demographic information, such as race and ethnicity and emergency contact information. Patient-provided data are stored in patients' records in EHRs, often recorded by physicians or other care providers during or after a clinical encounter (Weiner and Embi 2009; Weiskopf and Weng 2013). Many previous studies have raised issues regarding the data quality of this information in the EHR. Incompleteness and incorrectness are two data quality issues that have frequently been reported in studies focusing on the quality of patient-provided data (Ball and Lillis 2001; Douglas et al. 2015; Kaplan 2014; Klinger et al. 2015; Kressin 2015; Lee, Grobe, and Tiro 2015; Polubriaginof, Tatonetti, and Vawdrey 2015; Qureshi et al. 2009; Staroselsky et al. 2006; Welch, Dere, and Schiffman 2015).

Figure 2.1 shows the traditional process of collecting patient-provided data. Given that this information is now commonly stored in the EHR, reuse of this data is increasingly common, leading to discoveries and an improved understanding of the patient (Boland et

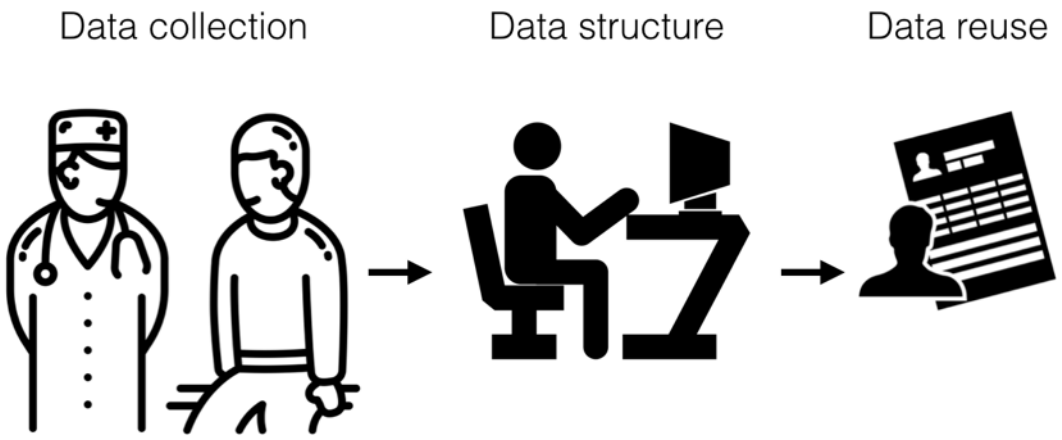

Figure 2.1: The traditional process of collecting patient-provided information.

al. 2015; Coopey et al. 2012; Dudley, Deshpande, and Butte 2011; Hripcsak et al. 2016; Kohane 2011; Li et al. 2015; Lorberbaum et al. 2016b; Polubriaginof et al. 2017; Ritchie, Andrade, and Kuivaniemi 2015; Tatonetti et al. 2012; Wang et al. 2017; Wei and Denny 2015). Informatics interventions to improve quality and therefore data reuse can target each one of these steps. Changes can be implemented to modify how data are collected and structured, leading to improved data reuse and the generation of new knowledge.

Data collection

The first step in the process is the collection of patient-provided information. Traditionally, practitioners have used the patient's appointment to collect a variety of patient-provided information. As time for patient encounters has become compressed and filled with other tasks, opportunities to collect and record patient-provided information have decreased. Advancements in technology allow patients to provide this data directly to providers through patient-facing tools (Bardes 2012; Warner 2010). Such tools include

registration forms, which can be implemented in paper or electronic format, and patient portals, which often also provide patients with access to some of their clinical information in addition to collecting information from them (Cimino, Patel, and Kushniruk 2001; Collins et al. 2011; Delbanco et al. 2010; Greenhalgh et al. 2008; Grossman et al. 2017; Halamka, Mandl, and Tang 2008; Hassol et al. 2004; Kaelber et al. 2008; Leveille et al. 2012; Nazi et al. 2010; Nazi et al. 2015; Pyper et al. 2004; Ralston et al. 2007; Reti et al. 2010; Tang and Lee 2009; Walker et al. 2011).

Patient-facing tools can be used both in inpatient and outpatient care settings, or even at home. These tools can be an avenue for patients to review and provide health information, reducing practitioner burden and saving time during the patient visit, while simultaneously engaging patients in their care (Arar et al. 2011; Epstein et al. 2010; Otte-Trojel et al. 2014). Previous research studies have used patient portals to improve data quality of past medical, surgical and social history (Arsoniadis et al. 2015), quality of EHR medication lists (Staroselsky et al. 2008), quality of medication information (Weingart et al. 2008), collection of smoking status (Baumgart, Postula, and Knaus 2015), and quality of family history data (Baumgart, Postula, and Knaus 2015; Volk et al. 2007). Others have used other patient-facing tools such as tablet applications, websites, and surveys to collect and improve quality of patient-provided information (Facio et al. 2010; Hamilton et al. 2009; Klinger et al. 2015; Pakhomov et al. 2008; Staroselsky et al. 2006; Sweet et al. 2014; Wu et al. 2015; Yoon et al. 2009). The use of these tools can also be beneficial for patients. There is also considerable evidence showing that the use of patient portals have improved patient engagement (Greenfield et al. 1988; Hack, Degner, and Dyck 1994; Kaplan et al. 1995) and patient outcomes (Arar et al. 2011; Davis Giardina et al. 2014; Dwamena and

Rovner 2012; Epstein et al. 2010; Otte-Trojel et al. 2014).

Regardless of how the data is collected, patients play an important role in providing their information, and they are often an underutilized resource (Ball and Lillis 2001; Staroselsky et al. 2006). Others have emphasized the importance of patient participation by stating that "patient and physician must meet as equals, bringing different knowledge, needs, and concerns" (Bardes 2012). Previous research studies have shown that patients are not only willing to provide their information, but they find it to be a useful experience (Arar et al. 2011; Wu et al. 2013). Furthermore, multiple studies have shown that the quality of the data collected by patients is superior to data collected in the routine clinical practice (Cohn et al. 2010; Frezzo et al. 2003; Jones, McGhee, and McGhee 1992; Porter et al. 2000; Reid et al. 2009; Selvachandran et al. 2002; Sweet, Bradley, and Westman 2002; Wuerdeman et al. 2005), demonstrating the importance of giving patients the opportunity to directly provide their information.

Data structure

It is important to understand how patient-provided data are stored in clinical documentation. Traditionally, patients' medical information was kept on paper charts. The purpose of these charts was to facilitate physicians' clinical reasoning and management of the course of treatment while communicating important decisions and relevant information to other members of the clinical team (Cusack et al. 2013; Engle 1991; Siegler 2010). Over time, documentation had to incorporate not only clinically meaningful information, but also a variety of elements pertaining to billing and reimbursement, regulations and accreditations,

and legal requirements (Cusack et al. 2013; Hagland 2011; Wasserman 2011). In the U.S., the Health Information Technology for Economic and Clinical Health (HITECH) Act, authorized $27 billion in incentives for EHR adoption, resulting in the EHR adoption rates to increase rapidly, resulting in 95% of hospitals reportedly using EHRs by 2016 (Conn 2016; Health Information Technology 2017). The transition from manual clinical documentation to EHR facilitated the reuse of clinical data for biomedical research, quality initiatives, and other purposes (Hammond et al. 1980; Schriger et al. 1997, 2000).

Along with the implementation of EHR systems, there is a trend toward collecting information using structured documentation, which facilitates reporting and data reuse but often results in more cumbersome information entry processes, compared to narrative text. As part of the HITECH Act, the federal government developed the Meaningful Use program, a financial incentive program supporting the adoption of EHRs (*CMS Electronic Health Records Incentive Programs*). The Meaningful Use program requires a set of data elements to be collected in a structured format. Many of these elements are patient-provided information, such as race and ethnicity (Centers for Medicare & Medicaid Services 2014b), family history (Centers for Medicare & Medicaid Services 2014a), and smoking status (Centers for Medicare & Medicaid Services 2010). The program also determined the use of standards and terminologies to store this information in an effort to standardize the data collected across organizations.

Rosenbloom and colleagues have described a tension between structure documentation and expressiveness (Rosenbloom et al. 2011). The use of free-text in clinical documentation versus structured data necessitates a trade-off between expressivity, flexibility, efficiency, and ease of data reuse. Many practitioners prefer to use free-text documenta-

tion to express subtleties and uncertainty; elements that structure documentation often does not support. While policy change can require the documentation of patient-provided data and storage in the EHR, concerns regarding the data quality of the information collected remains, and the impact of implementation of new policy and tools is often poorly understood.

Data reuse

With the broad availability of clinical data in electronic format, there has been a growing interest in reusing clinical data collected during patient encounters in research. Traditional clinical trials often rely on patient recruitment, a process that is labor-intensive and costly (Drennan 2002; Institute of Medicine (US) Forum on Drug Discovery, Development, and Translation 2010). EHRs store clinical data, including patient-provided data, that is collected on a daily basis as part of clinical care. Use of EHR data can overcome many of the challenges faced by traditional studies by enabling larger sample sizes with sharply decreased costs. Data reuse can also support reproducibility of findings.

While EHR data are required to be kept in secure databases with limited access to safeguard protected health information, initiatives such as OHDSI (Hripcsak et al. 2015) and i2b2 (Murphy et al. 2010) have developed strategies to support research reproducibility while maintaining privacy (Yuan et al. 2017). These initiatives created open-source frameworks that allow disparate teams of researchers to run the same analyses on separate private databases, and combine the results with confidence in order to effectively achieve the benefits of data aggregation on closed data. Using the OHDSI platform, Hripcsak and colleagues

conducted the largest observational research study focusing on the characterization of treatment pathways for disease, leveraging the medical records and administrative claims data of 250 million patients (Hripcsak et al. 2016).

In addition to EHR data, other initiatives have taken place. In 2009, President Obama created the Open Government Directive, an initiative created to make the government more transparent. As part of this initiative, federal, state and local authorities began to release de-identified health data. The data released was publicly accessible, available in multiple formats, free of charge, and has unlimited use and distribution rights (*Open Government Directive*). As a result of this initiative, there are currently more than 3,500 open datasets available at HealthData.gov (*HealthData.gov*). In New York State, the use of open data has already positively impacted emergency response during hurricanes, changes in policies and medical education curriculum, patient safety, and accountability of healthcare costs (Martin, Helbig, and Shah 2014).

2.1 Patient-provided data

Even though clinical databases contain valuable information that is increasingly available, the use of this information in areas such as genetic research has been limited. In this dissertation, I focus on the study of three types of patient-provided information: race and ethnicity, family history, and smoking status.

Race and Ethnicity

Race and ethnicity are collected for many reasons, including for clinical, administrative, and research purposes. Clinically, race and ethnicity are commonly used for estimating disease risk (Gail et al. 1989; Levey et al. 2009; Stevens et al. 2006) and for assessing racial and ethnic health disparities (Dorsey et al. 2014; Douglas et al. 2015; Kressin 2015; LaVeist, Gaskin, and Richard 2011). From an administrative standpoint, the Centers for Medicare and Medicaid Services, through the Meaningful Use incentive program, requires standardized collection of patients' race and ethnicity (Centers for Medicare & Medicaid Services 2014b). This is because race and ethnicity are a part of a patient's socioeconomic status, which has been discussed as a method for risk adjusting in payment reform (Buntin and Ayanian 2017; Committee on Accounting for Socioeconomic Status in Medicare Payment Programs et al. 2016). From a research perspective, studies frequently report patients' demographic information, including race and ethnicity.

Race and ethnicity can be collected from patients in a variety of formats and by a variety of personnel. This information is often collected either verbally or through patient-facing tools, such as intake forms completed during a clinical encounter. Race and ethnicity are ei-

ther directly entered or transcribed from intake forms into the EHR (Adler and Stead 2015). However, there are many challenges to the collection of race and ethnicity information that may degrade the quality of this data in the EHR (Blustein 1994; Chakkalakal et al. 2015; Gomez and Glaser 2006; Hamilton et al. 2009; Lee, Grobe, and Tiro 2015; Moscou et al. 2003). Cultural insensitivity and lack of understanding of the importance of race and ethnicity information are major challenges to collecting race and ethnicity information in the hospital setting. Verbally asking patients their race and ethnicity can be an uncomfortable situation for both healthcare workers and patients (Baker et al. 2007). A previous study conducted in 2014 reported that registration personnel felt inadequately trained to ask patients' race and ethnicity (Berry et al. 2014). Secondly, there is a lack of understanding of why this information is collected and how it will be used. A study conducted in 2005 showed that information that is not known to be used by others is not accurately collected (Nelson et al. 2005). Registration personnel are often unaware of the importance of race and ethnicity and also do not know who uses the information. This lack of awareness presents a barrier to registration personnel asking patients their race and ethnicity. From a patient's perspective, the question is often unexpected and may not come with an explanation of how the information will be used and why it is important.

In the U.S., some effort has gone into standardizing the data structures for storing race and ethnicity information. The Meaningful Use program specifies that race and ethnicity data collection should follow the standard developed by the Office of Management and Budget (OMB) (OMB 1997). According to this standard, race and ethnicity information can be collected in either a single-question or in a two-question format. Despite this option, there has been limited research on the impact of using a single-question or the two-question

format on data quality. Furthermore, the OMB also established that patient-provided information be considered the gold standard for the collection of race and ethnicity data. In this respect, there has been little effort to compare race and ethnicity information directly provided by patients with the corresponding data stored in the EHR.

Family History

Family history has always been considered "a core element of clinical care" (Berg et al. 2009) and has been described as being a free genetic tool that almost every patient has access to (Guttmacher, Collins, and Carmona 2004). Since the Human Genome Project, new genomic tools have been described (Guttmacher and Collins 2003); however, family history remains critical for identifying patients that may be at higher risk to develop disease. Family history provides information that enables individualized disease diagnosis, treatment, and prevention.

Several studies have shown that family history is an important element in deciding clinical care. Knowing that a patient is at increased risk of developing a disease based on family history enables disease prevention that can vary from intensive screening to prophylactic surgery. It can also facilitate earlier diagnosis and more tailored treatment. For example, current guidelines from the American Cancer Society define criteria for MRI eligibility in addition to mammography for breast cancer screening (Saslow et al. 2007; Smith, Cokkinides, and Brawley 2012). The guidelines recommend that patients that have a lifetime risk of breast cancer greater than 20% by BRCAPRO (Berry et al. 1997), Tyrer-Cuzick (Tyrer, Duffy, and Cuzick 2004), or Claus (Claus, Risch, and Thompson 1994) models

should have screening MRI in addition to mammography. Each of these models was developed using different methods, different populations and different risk factors, and each of them was developed to predict different outcomes, but all of them heavily rely on family history and presence of risk factors (Ozanne et al. 2013). This is just one of many examples of the importance of family history information for clinical care.

Given the variety of guidelines and models available, clinical decision support systems have been developed to help clinicians deliver precision medicine. One of the earliest research studies about the use of clinical decision support (CDS) systems to support precision medicine was conducted by Emery and colleagues in 1999. The study identified that the CDS systems available were not appropriate for use in a primary care setting (Emery 1999). To address this problem, a system was developed to record and interpret family history data in the primary care clinic. The system included family history relevant to breast, ovarian and colorectal cancer (Emery et al. 1999). Over time, other clinical decision support systems were developed to manage other types of cancers, such as colorectal cancer and Lynch syndrome, and all of these systems used family history information to provide disease risk assessment (Welch and Kawamoto 2013).

Currently, the U.S. Preventive Services Task Force (USPSTF) recommends risk assessment based on family history for some conditions such as screening for BRCA mutation and BRCA-related cancers (Moyer and U.S. Preventive Services Task Force 2014), osteoporosis (U.S. Preventive Services Task Force 2011), and lipid disorders in adults (Helfand and Carson 2008). A 2007 report commissioned by the Agency for Healthcare Research and Quality (AHRQ) recommended that collection of family history information should include diseases in first-degree relatives and second-degree relatives from both the maternal

and paternal side, the relatives' age at the time of disease diagnosis, and each relatives' race and ethnicity (Qureshi et al. 2007).

Family history can be collected through several modalities; however, two common methods for capturing family history information include directly from the patient using either free-text or semi-structured patient-facing tools. Some examples of tools that have been deployed to help collect family history include *Family Healthware* from the Centers for Disease Control and Prevention (Yoon et al. 2009), *Family HealthLink* from The Ohio State University Medical Center (Sweet et al. 2014), *Health Heritage* from University of Virginia Health System (Baumgart, Postula, and Knaus 2015; Cohn et al. 2010), *Hughes RiskApps* from the Massachusetts General Hospital (Ozanne et al. 2009), *OurFamilyHealth* from Intermountain Healthcare (Hulse et al. 2011), and *MeTree* from Duke University (Orlando et al. 2011, 2013; Wu et al. 2015). Some of these tools not only capture family history but also perform disease risk assessment based on the family history data. For example, Family Healthware, a web-based tool developed by the Centers for Disease Control and Prevention focused on the collection of family history to assess familial risk for six conditions: heart disease, stroke, diabetes, colorectal cancer, breast cancer, and ovarian cancer (Yoon et al. 2009). Previous studies describing tools for collecting family health history highlighted data collection issues, such as time limitations. A study conducted in 2011 described that the average time required for patient to input family history information using a web-based tool was 15 minutes, in a range from 3 to 45 minutes (Owens et al. 2011).

Given the importance and utility of family history data, several initiatives have been put implemented to increase the structured collection of family history. The goal of these initiatives is to enable precision medicine, which requires accurate and detailed family his-

tory data. For example, Stage 2 of the Meaningful Use program included a requirement of clinicians to use structured data entry for family history. Under the program, eligible hospitals must have recorded at least one structured family history data element for at least one first-degree relative for 20% of their patients (Centers for Medicare & Medicaid Services 2014a).

In addition to the Meaningful Use program, other federal initiatives have focused on collecting family history from the patient, such as the U.S. Surgeon General's My Family Health Portrait (*My Family Health Portrait*). This initiative promotes individuals to share their information with family members and healthcare providers through a web-based application. While most interventions have focused on collecting structured family history, previous research has identified a tradeoff between strictly structured data, which promotes reuse and standardization, and free-text documentation, which promotes expressiveness (Rosenbloom et al. 2011). In fact, there has been little prior research comparing the impact of data structures on the data quality of family history.

Smoking Status

Smoking is an important risk factor for multiple diseases, including cardiovascular diseases and numerous types of cancer. It remains the number one cause of preventable death in the United States, being responsible for more than 480,000 deaths annually (National Center for Chronic Disease Prevention and Health Promotion Office on Smoking and Health 2014). To provide patients with the resources to quit smoking, the collection of patients' smoking status during clinical encounters is critical (Boyle, Solberg, and Fiore

2014). Smoking cessation can be difficult, and clinical visits are opportunities to intervene and recommend smoking cessation programs and therapies. Obtaining a patient's smoking status is a crucial step in beginning smoking cessation interventions and monitoring progress (Caplan, Stout, and Blumenthal 2011).

Similar to other types of patient-provided data, smoking status can be collected in a variety of formats. Usually, this information is collected during a clinical encounter, by clinicians verbally asking patients about their smoking status. This information is then entered into the EHR, usually as part of clinical notes. It may seem that recording updates to smoking status in a timely and accurate manner would be straightforward using modern EHRs; however, previous research has shown that social and behavioral determinants of health are often overlooked during clinical encounters (Adler and Stead 2015). The challenges related to the appropriate collection and documentation of smoking status include lack of standard terminology and granularity for data collection, shifting cultural attitudes regarding tobacco use, and potentially frequent changes in individuals' smoking behavior (Committee on the Recommended Social and Behavioral Domains and Measures for Electronic Health Records, Board on Population Health and Public Health Practice, and Institute of, Medicine 2015; Winden et al. 2015).

Due to the positive impact of smoking cessation in health, tobacco control policies such as smoke-free legislation, tobacco taxation, and smoking cessation services have been implemented and have been shown to have substantial benefits in children's health (Faber et al. 2017).

Given the clinical importance of recording smoking status, the Meaningful Use financial incentive program for EHR adoption in the U.S. included a requirement for healthcare

providers to capture patients' smoking status electronically in a structured fashion (Centers for Medicare & Medicaid Services 2010). Meaningful Use has helped to standardize data collection of smoking status and other information. However, previous efforts to improve patient-provided data types, such as race and ethnicity (Klinger et al. 2015; Lee, Grobe, and Tiro 2015) and family history (Polubriaginof, Tatonetti, and Vawdrey 2015; Powell et al. 2013) did not necessarily improve data quality through the adoption of standards for representing information, but it often resulted in increased documentation of these data types. There is little knowledge whether the data quality of smoking status improved after Meaningful Use requirements were implemented.

2.2 Knowledge gaps

This dissertation focused on several knowledge gaps related to patient-provided data. First, there have been limited studies focusing on whether some patient-provided data types are accurately captured in the EHR. Patient-provided data is traditionally collected by clinicians as part of clinical encounters and due to time constraints during clinical visits, the appropriate collection and documentation of important information may be overlooked.

Second, due to the importance of patient-provided data, initiatives such as the Meaningful Use program had the objective of ensuring that patient-provided data would be collected and recorded in the EHR in a structured format. However, limited work has been performed to measure the impact of the Meaningful Use program in the quality of patient-provided data. Concerns regarding data quality of structured data in EHRs have been previously raised. Previous studies have discussed the trade-off between the flexibility of free-text documentation versus the easy reuse of structured data. The mandatory collection of these data in a structured format may decrease data quality as clinicians have to rush more with structured documentation, losing documentation flexibility which may lead to reduced expressiveness, and therefore poorer data quality.

Third, while patients are considered the reference standard for patient-provided data, limited work has been done in assessing the impact of patients directly providing data as compared to patient-provided data being recorded by clinical or administrative staff during hospital encounters. Despite numerous patient-facing tools being developed, allowing patients to contribute information to their care, there have been few attempts to assess the quality of the information provided by patients in comparison to the data available in the

EHR.

Fourth, there has been little work focusing on the use of informatics methods to infer patient-provided data from clinical databases. Large amounts of clinical data are now available due to the broad adoption of EHRs, and these data could be used to infer additional information and ultimately support research.

Fifth, there has been little work focusing on strategies to overcome the challenges of data quality and biases in patient-provided information stored in the EHR for knowledge generation. Clinical databases are a valuable resource of information, and the use of these data should account for biases and data quality concerns when generating new knowledge and supporting clinical research.

Chapter 3

Aim 1 - Assess the quality of patient-provided data in clinical databases

The purpose of this aim was to evaluate the quality of patient-provided data in clinical databases. With the growing adoption of EHRs, research studies conducted on observational data can complement prospective studies. The use of EHR data not only provides increased sample sizes and access to diverse populations but also allows for hypothesis testing prior to the implementation of prospective studies (Benson and Hartz 2000; Berger et al. 2009; Concato, Shah, and Horwitz 2000; Madigan et al. 2014; Ryan et al. 2012). EHR data have been used to shed light on disease processes (Boland et al. 2015; Coopey et al. 2012; Hripcsak et al. 2016; Li et al. 2015; Ritchie, Andrade, and Kuivaniemi 2015; Wei and Denny 2015), including genetics (Kohane 2011; Polubriaginof et al. 2017; Wang et al. 2017), and on drug effectiveness and interactions (Dudley, Deshpande, and Butte 2011; Lorberbaum et al. 2016b; Tatonetti et al. 2012). However, there are concerns regarding the quality of EHR data. Multiple research studies have demonstrated quality concerns with EHR data (Ahmad et al. 2017; Aronsky and Haug 2000; Arts et al. 2002; Brennan and Stead 2000; Brown, Kahn, and Toh 2013; Hasan and Padman 2006; Hersh et al. 2013; Hogan and Wagner 1997; Hripcsak et al. 2011b; Kahn, Eliason, and Bathurst 2010; Lei 1991; Rusanov et al. 2014; Thiru, Hassey, and Sullivan 2003). For example, a study conducted in 2017

Dimension	Description
Completeness	Information being available
Correctness	Information being truthful
Concordance	Information from different data elements being in agreement
Plausibility	Information being feasible
Currency	Information being recorded in a timely manner

Table 3.1: Description of the five dimensions of data quality.

analyzed the validity of cardiovascular data by comparing EHR data to data collected by standardized research approaches in a cohort study. Overall, this study demonstrated that some clinical features were better documented in the EHR than others, and therefore, data quality concerns should be considered when using existing clinical databases for research (Ahmad et al. 2017).

In this Aim, I analyzed dimensions of data quality for three distinct types of patient-provided data: race and ethnicity, family history and smoking status. The purpose of the analysis was to understand how reliable this information was in clinical databases, and uncover opportunities for improvement. Data quality is defined along 5 dimensions: 1) completeness, 2) correctness, 3) concordance, 4) plausibility and 5) currency (Weiskopf and Weng 2013). Table 3.1 shows the definitions of each dimension of data quality. The studies from this Aim describe data quality in terms of these dimensions.

3.1 Aim 1.1 - Assessing the quality of race and ethnicity data collected in clinical databases

Background

Race and ethnicity information has long been collected by U.S. hospitals (Adler and Stead 2015; Hasnain-Wynia, Pierce, and Pittman 2004), and this information is frequently reported in observational studies that use electronic health record (EHR) data. Race and ethnicity are commonly used for estimating disease risk (Gail et al. 1989; Levey et al. 2009; Stevens et al. 2006) and for assessing racial and ethnic health disparities (Dorsey et al. 2014; Douglas et al. 2015; Kressin 2015; LaVeist, Gaskin, and Richard 2011). The use of race and ethnicity as a proxy for socioeconomic or as a marker for disparities in health care is being increasingly discussed (Bach et al. 2004; Buntin and Ayanian 2017; Committee on Accounting for Socioeconomic Status in Medicare Payment Programs et al. 2016).

The United States Centers for Medicare and Medicaid Services (CMS) "Meaningful Use" financial incentive program for EHR adoption includes the collection of patients' race and ethnicity as one of its requirements (Centers for Medicare & Medicaid Services 2014b). The Meaningful Use program adopted as a model for race and ethnicity data collection the standard developed by the Office of Management and Budget (OMB) (OMB 1997). According to this standard, race and ethnicity information can be collected in either a single question or in a two-question format. It also established that patient-provided information be considered the gold standard for the collection of race and ethnicity data.

Race and ethnicity are typically collected in healthcare settings at the time of registration. Although EHRs now include structured fields for collection of patient-provided data such as race and ethnicity information, previous studies report this information is often not accurate in the EHR (Blustein 1994; Chakkalakal et al. 2015; Gomez and Glaser 2006; Hamilton et al. 2009; Lee, Grobe, and Tiro 2015; Moscou et al. 2003; Polubriaginof, Tatonetti, and Vawdrey 2015). A possible way to improve the quality of race and ethnicity information in clinical databases is to have patients report this information themselves via a paper or electronic form.

Objectives

The purpose of this study was to evaluate data quality of race and ethnicity data nationally as well as in a large healthcare system in New York. Specifically, I evaluated the completeness, correctness and concordance of race and ethnicity information in clinical databases.

Research Questions

- *What proportion of race and ethnicity data is clinically informative in observational clinical databases?*
- *How do updates in race and ethnicity fields over time change data quality of this information in EHR?*

Methods

Data

United States National Databases

I analyzed data from two large observational health databases: HCUP and the Optum Labs Data Warehouse. The HCUP (Healthcare Cost and Utilization Project) database is a hospital transactional database created by AHRQ that includes over 90 million inpatient, emergency visits, and ambulatory surgery encounters from multiple hospitals in the United States (*Healthcare Cost and Utilization Project (HCUP)*). The Optum Labs Data Warehouse is an administrative claims database of more than 70 million commercially insured and Medicare Advantage enrollees, with greatest representation in the Midwest and South US census regions (*Optum Data Assets*; Wallace et al. 2014).

I examined HCUP data from 2000 to 2011 and Optum data from 2000 to 2016. The two databases stored race and ethnicity information using slightly different categories. Both included "White," "Black or African American," "Hispanic or Latino" and "Unknown" as categories, so I only included only these four options and reported the remaining groups collectively as "Other." A detailed description of these datasets, including the categories used to collect race and ethnicity information, sample size and timeframes, is shown in Table 3.2.

Academic Healthcare System in New York City

I conducted a retrospective analysis of race and ethnicity data recorded for patients that had at least one inpatient, outpatient, or emergency department visit from January 2014

Dataset	Description	Timeframe	Number of patients	Black or African American	White	Hispanic or Latino	Other Race/ Ethnicity	Unknown Race
Observational health databases								
HCUP	Healthcare Cost and Utilization Project (HCUP) is a hospital transactional database that includes inpatient and emergency visits	2000 — 2011	91,983,358	10.75%	52.13%	9.40%	2.42%	25.31%
OPTUM	Health claims database for members of United Healthcare which includes patients enrolled in commercial plans, Medicaid and Legacy Medicare Choice	2000 — 2016	73,992,364	7.45%	53.21%	9.69%	3.64%	26.00%
EHR*	Data from the electronic health record (EHR) of a academic healthcare system in New York City including inpatient, outpatient and emergency department visits	2014 — 2015	2,338,421	6.09%	23.55%	9.51%	7.10%	57.88%

Table 3.2: Description of the data sources, including timeframes and race and ethnicity categories. *Numbers do not sum to 100% because race and ethnicity were collected separetely.

through December 2015 at an academic health system that serves a racially and ethnically diverse population in 10 hospital campuses in and around New York City, including a quaternary care hospital. The Ambulatory Care Network (ACN) consists of 14 primary care practice sites and more than 50 specialty care clinics. The academic health system provides millions of visits annually, including 2.2 million outpatient visits, 286,000 emergency department visits, and 126,000 inpatient discharges.

Race and ethnicity data were collected by the health system in one of two ways: 1) patients completed paper forms as part of the registration process, or 2) registration clerks verbally asked patients about their race and ethnicity. To collect race and ethnicity, the health system used a two-question format, the first field capturing the patient's race ("American Indian or Alaska Native," "Asian," "Black or African American," "Native Hawaiian or Other Pacific Islander," "White," "Unknown," "Other," or "Declined to Answer"), and the second field capturing the patient's ethnicity ("Hispanic or Latino," "Not Hispanic or Latino," "Declined to Answer," or "Unknown"). Race and ethnicity information were collected at every encounter and stored in a centralized location in the EHR.

From the same academic health system, I examined data from the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) Survey administered to patients who had a hospital stay from January 2014 through December 2015. The HCAHPS Survey was sent via U.S. Mail after hospital discharge. To collect race and ethnicity, the survey used a two-question format, with one field capturing race ("White," "Black or African American," "Asian," "Native Hawaiian or other Pacific Islander," "American Indian or Alaska Native"), and the second field capturing ethnicity (Hispanic or Latino origin or Not Hispanic).

Analysis

Since each data source collected race and ethnicity using different categories, I described groups that were available in all data sources and reported the remaining groups collectively as "Other." I performed descriptive statistics for "White," "Black or African American," "Hispanic or Latino," and "Other racial and ethnic groups." Patients classified as "Unknown," "Other" or "Declined to Answer" were considered to have clinically uninformative data; I combined these categories into a larger group designated as "Uninformative" for further analysis. Completeness was assessed based on the percentage of "Uninformative" race and ethnicity in the database. Using EHR data, I also calculated descriptive statistics on the frequency of race/ethnicity pairs, since these two fields are highly correlated.

Changes in race and ethnicity information in the EHR

I analyzed changes to race and ethnicity recorded for the same patient over multiple visits, using system logs from the EHR. Patients with two or more visits during the study period were included in this analysis. I reported descriptive statistics on changes of race and ethnicity pairs over time.

A race/ethnicity pair was recorded during each clinical encounter. To quantify the frequency of changes recorded for a patient's race and ethnicity, each race and ethnicity pair was scored based on the amount of information it contained. Race and ethnicity pairs were broken down into concept pairs, with one concept for race and another for ethnicity, and each concept was scored individually. Each informative concept received a score of 1,

and each uninformative concept received a score of 0. For example, “White”, “Hispanic” would receive a score of 2 since both the race and ethnicity concepts are informative. Likewise, “White” with “Unknown” ethnicity would receive a score of 1, and “Unknown” race, “Unknown” ethnicity received a score of 0. The scores were compared for each pair chronologically.

The changes in the content of the race and ethnicity pairs were classified as: information loss, neutral, or information gain. If the patient had the same score in the previous and current visit (i.e., the difference between the previous and current race/ethnicity score was 0), it was considered to be neutral. If the score from the second visit was greater than the previous visit, it was considered information gain. Finally, if the score from the second visit was less than the previous visit, it was considered information loss. I reported descriptive statistics of the aggregated scores.

Comparison to patient-provided data

I assumed patient-reported data to be the reference standard for race and ethnicity data collection. To assess differences between patient-reported race and ethnicity information and data from observational databases, I evaluated race and ethnicity reported in the HCAHPS survey. Because I had patient-level data from the New York academic healthcare system from both the EHR and the HCAHPS survey I reported the concordance between the patient’s race and ethnicity information in the EHR and the self-reported from HCAHPS.

Comparison to Census data

To assess how well the data from the EHR and the HCAHPS survey represented the population of the community in which the academic health system was located, I compared the EHR and HCAHPS race and ethnicity distribution by ZIP Code of the patient's home address to the race and ethnicity distribution for that ZIP Code as reported by the U.S. Census from the American Community Survey 5-Year Demographic and Housing Estimates. For each ZIP Code, I calculated the percent difference between the Census data and the EHR data for each race and ethnicity category. For this analysis, I included ZIP Codes that had at least 50 patients in the EHR and HCAHPS data.

Results

United States National Databases

There were 165,975,722 combined patient records in the HCUP and Optum databases. Of these, 25.3% and 26.0%, respectively, had uninformative race and ethnicity (Table 3.2).

Academic Healthcare System in New York City

In the New York academic health system, 2,338,421 patients had at least one visit during the two-year study period. As shown in Table Table 3.2, 57.9% of patients did not have race or ethnicity identified in the EHR. The distribution of all race-ethnicity pairs is described in Table 3.3.

Race	Ethnicity		
	Hispanic or Latino	Not Hispanic or Latino	Uninformative
Asian	0.05%	1.28%	1.78%
Black or African American	0.70%	3.00%	2.39%
White	3.27%	9.51%	10.77%
Native Hawaiian or Pacific Islander	0.10%	0.07%	0.06%
American Indian or Alaska Native	< 0.01%	0.04%	0.08%
Uninformative	5.38%	3.64%	57.88%

Table 3.3: Frequency of race and ethnicity pairs in academic health system electronic health record. *Patients designated as Hispanic in single question format surveys were assumed to be White-Hispanic.

Changes in race and ethnicity information in the EHR

I identified 1,205,796 patients who had more than one visit to the academic health system. There were 161,114 modifications made to race or ethnicity fields in the EHR for 147,061 distinct patients (12% of total population). There were 0.13 changes to race and ethnicity fields made per patient, on average, over the two-year study period (max=18).

Modifications to race or ethnicity often improved completeness (i.e., a change was made from an 'uninformative' concept to a specific race or ethnicity category), but this was not always the case. Overall, I observed that 60% of the changes made in race and ethnicity improved completeness (information gain), 31% of the changes resulted in information loss, and 9% of the changes were information neutral.

The most frequent change resulting in information gain was an update of previously documented race "Unknown" and ethnicity "Unknown" to race "White" and ethnicity "Not Hispanic;" the most frequent change resulting in information loss was a modification from

race “White” and ethnicity “Hispanic” to race “Unknown” and ethnicity “Unknown;” and lastly the most frequent change that did not affect the amount of race and ethnicity information collected was an update from race “Unknown” and ethnicity “Unknown” to race “Declined to answer” and ethnicity “Declined to answer.”

Comparison to patient-provided data

During the study period, 25,664 unique patients responded to the HCAHPS survey. Of those, 1,255 patients completed the survey more than once, and 356 had conflicting self-reported race and ethnicity information.

After excluding cases with conflicting self-reported race and ethnicity information, 86.3% provided meaningful race or ethnicity data from a total of 25,308 patients. Among these patients, race and ethnicity information from the EHR was available for 25,014 patients.

Among patients with both self-reported and EHR race and ethnicity information, 16,625 (66.5%) patients provided race or ethnicity information that was discordant with data recorded in the EHR. Table 3.4 provides a list of the most common discrepancies between EHR and self-reported race and ethnicity data. While 6,540 had both race and ethnicity as “Uninformative,” self-reported data provided meaningful race or ethnicity information for 5,533 of these patients, 84.6% of patients that did not otherwise have meaningful information recorded.

Self-reported Race	EHR Race	Self-reported Ethnicity	EHR Ethnicity	Frequency
White	Uninformative	Not Hispanic	Uninformative	20.54%
White	White	Not Hispanic	Uninformative	19.92%
White	Uninformative	Not Hispanic	Not Hispanic	6.88%
Uninformative	Uninformative	Uninformative	Hispanic	4.00%
Uninformative	White	Uninformative	Not Hispanic	3.06%
White	Uninformative	Uninformative	Uninformative	2.98%
Asian	Asian	Not Hispanic	Uninformative	2.53%
Asian	Uninformative	Not Hispanic	Uninformative	2.45%
Black	Uninformative	Not Hispanic	Uninformative	2.33%
Uninformative	White	Uninformative	Hispanic	2.26%

Table 3.4: Ten most common discrepancies between EHR and self-reported data.

Comparison to Census data

There were 44 ZIP Codes with more than 100 patients in the EHR and HCAHPS datasets. When comparing the distribution of race and ethnicity categories between the EHR, HCAHPS and the Census datasets, I observed that, on average, the EHR data contained a higher proportion of uninformative race than the Census (63% vs. 14%, Figure 3.1). However, when performing the same comparison using patient-reported information, I observed that the rate of uninformative race in the HCAHPS dataset was similar to the Census dataset (18.1% vs. 14%, Figure 3.1). Table 3.5 contains the distribution of race and ethnicity categories for each ZIP Code included in the analysis.

ZIP Code	Census					EHR					HCAHPS				
	N	White	Black	Uninformative race	Hispanic	N	White	Black	Uninformative race	Hispanic	N	White	Black	Uninformative race	Hispanic
10032	62685	30.3%	20.7%	37.1%	68.6%	53942	19.6%	9.5%	68.9%	31.3%	679	31.5%	20.9%	39.6%	7.1%
10021	43573	83.6%	1.7%	1.9%	6.7%	40563	33.4%	1.5%	61.3%	1.5%	586	84.5%	0.5%	9.0%	1.2%
10033	59844	40.6%	9.3%	40.6%	68.8%	38306	26.7%	6.0%	65.9%	37.5%	611	44.4%	11.1%	37.6%	8.7%
10025	96068	64.9%	14.1%	6.8%	22.2%	34970	31.1%	4.3%	61.3%	5.0%	352	69.9%	6.5%	12.8%	5.1%
10463	71981	44.0%	14.7%	31.5%	48.5%	32571	31.2%	8.2%	58.9%	29.5%	708	56.9%	11.2%	25.4%	11.9%
10023	60586	82.2%	3.1%	2.3%	9.7%	31172	35.7%	1.7%	58.7%	1.7%	374	81.0%	2.4%	10.2%	2.4%
10024	58391	81.4%	5.9%	2.1%	12.9%	30462	35.0%	1.8%	60.6%	2.1%	335	81.2%	1.8%	9.9%	2.4%
10040	44378	36.7%	9.4%	45.1%	70.5%	28545	27.5%	6.1%	64.9%	40.3%	531	40.7%	11.5%	38.8%	8.5%
10031	59244	28.2%	33.4%	19.4%	52.9%	27849	17.6%	13.8%	67.3%	29.0%	307	22.2%	30.3%	38.8%	5.2%
10034	43405	37.1%	8.5%	45.2%	74.5%	27259	27.0%	7.0%	64.7%	42.1%	519	41.4%	12.0%	38.3%	10.8%
10128	61927	81.3%	5.1%	1.1%	11.3%	25696	33.4%	2.1%	60.7%	2.3%	342	81.0%	2.3%	10.2%	3.5%
10065	30237	86.5%	3.0%	0.5%	5.3%	25163	32.3%	2.1%	61.0%	2.9%	332	78.3%	0.3%	11.5%	1.5%
10028	46883	89.2%	1.0%	0.4%	5.3%	25039	35.5%	1.4%	60.1%	1.3%	319	85.9%	0.6%	7.5%	0.6%
10022	30607	85.6%	1.9%	1.0%	5.9%	22520	30.9%	1.4%	64.5%	1.6%	349	82.2%	0.9%	11.5%	1.7%
10002	80736	29.6%	8.6%	14.8%	26.4%	19446	13.6%	5.3%	59.0%	14.2%	298	27.9%	1.7%	19.8%	10.4%
10027	64413	28.8%	43.2%	14.5%	23.1%	19119	17.2%	14.3%	65.8%	12.1%	154	39.6%	31.2%	19.5%	5.2%
10453	80081	9.8%	31.9%	52.8%	68.8%	16566	15.8%	14.5%	68.7%	39.8%	184	21.7%	26.1%	42.9%	12.0%
10452	75559	12.9%	33.5%	48.2%	66.7%	16522	15.1%	13.8%	70.0%	38.7%	182	19.2%	24.2%	50.0%	9.9%
11201	58350	64.6%	13.5%	6.6%	13.2%	16302	31.7%	4.2%	59.1%	2.4%	199	69.9%	4.5%	9.1%	3.0%
10468	73637	13.5%	21.8%	54.7%	72.1%	15875	16.8%	10.8%	70.9%	43.0%	227	35.2%	19.4%	35.7%	11.0%
10011	52349	79.2%	3.7%	2.0%	10.2%	15759	29.9%	1.6%	65.3%	1.9%	155	76.8%	0.7%	7.7%	1.3%
10003	57112	77.6%	4.4%	1.4%	7.2%	14274	28.9%	1.7%	65.9%	1.5%	137	75.2%	2.2%	6.6%	0.0%
10019	38830	71.4%	5.6%	4.0%	16.2%	14125	26.9%	2.4%	65.1%	3.9%	149	72.5%	2.7%	12.1%	6.0%
10016	50641	71.9%	3.0%	4.5%	9.5%	14013	25.6%	2.1%	67.0%	1.8%	132	72.7%	0.8%	7.6%	3.8%
11215	70818	75.3%	4.7%	6.4%	17.7%	13530	35.0%	2.9%	58.5%	3.2%	122	74.6%	2.5%	6.6%	1.6%
10583	39996	81.4%	1.7%	0.8%	4.3%	13408	36.2%	1.3%	59.4%	7.1%	106	87.7%	0.0%	6.6%	0.9%
10029	79251	30.9%	28.9%	28.0%	49.0%	12985	15.8%	13.3%	67.5%	18.2%	114	48.3%	16.7%	29.0%	25.4%
11211	97772	77.5%	6.1%	8.7%	24.8%	12421	30.9%	2.0%	64.4%	5.8%	124	75.0%	1.6%	10.5%	8.9%
10038	21464	50.8%	6.9%	6.8%	14.8%	12388	25.5%	8.8%	53.6%	13.6%	146	51.4%	6.9%	13.0%	17.8%
10075	25756	89.6%	1.3%	2.0%	7.6%	12371	40.7%	1.0%	54.6%	1.4%	190	81.1%	1.1%	9.5%	0.5%
10456	91868	8.5%	40.1%	46.8%	59.7%	12185	12.9%	17.2%	69.0%	33.0%	115	21.7%	29.6%	43.5%	7.8%
11375	70723	62.8%	3.4%	3.8%	13.1%	11597	26.8%	1.6%	61.6%	3.2%	167	61.7%	0.6%	10.8%	3.0%
10013	27415	55.0%	4.7%	1.8%	7.2%	10731	30.9%	2.3%	54.7%	2.0%	139	59.7%	3.6%	7.9%	1.4%
10039	26697	12.9%	66.0%	14.3%	29.6%	10283	9.8%	27.9%	61.1%	16.9%	127	10.2%	60.6%	22.8%	4.7%
11385	102209	82.8%	2.1%	6.3%	43.9%	8979	29.1%	1.9%	65.4%	14.0%	113	64.6%	1.8%	18.6%	13.3%
11230	88589	69.6%	7.4%	7.0%	9.6%	8696	30.1%	3.1%	64.0%	2.2%	101	69.3%	5.9%	13.9%	4.0%
10471	21727	71.8%	8.0%	10.3%	23.0%	8659	39.8%	5.5%	52.4%	12.4%	212	76.9%	4.3%	16.5%	6.6%
10010	31447	72.8%	6.2%	2.0%	9.6%	8621	28.4%	2.2%	65.0%	1.7%	104	78.9%	1.9%	7.7%	1.0%
11377	90615	47.9%	2.7%	9.2%	38.8%	8305	20.7%	2.4%	66.6%	12.6%	112	58.0%	0.9%	11.6%	5.4%
11235	76668	79.5%	3.0%	2.4%	8.4%	7874	28.2%	1.5%	66.8%	1.6%	125	77.6%	1.6%	10.4%	1.6%
11209	70803	76.2%	2.0%	6.5%	16.3%	6986	30.2%	2.2%	61.6%	3.6%	116	77.6%	0.0%	10.3%	4.3%
11219	97670	65.6%	1.3%	7.1%	12.8%	6618	27.6%	1.2%	59.0%	2.5%	103	65.1%	1.0%	6.8%	1.0%
11234	95912	42.2%	47.7%	3.1%	9.1%	6454	22.6%	14.7%	60.6%	2.7%	115	57.4%	15.7%	17.4%	4.4%
10314	87600	77.3%	4.7%	2.4%	13.7%	5918	23.5%	1.6%	70.8%	3.0%	123	78.1%	0.8%	12.2%	3.3%

Table 3.5: Comparison of Census, EHR and HCAHPS racial and ethnic distribution among 44 ZIP Codes that contained at least 100 patients in the EHR and HCAHPS datasets.

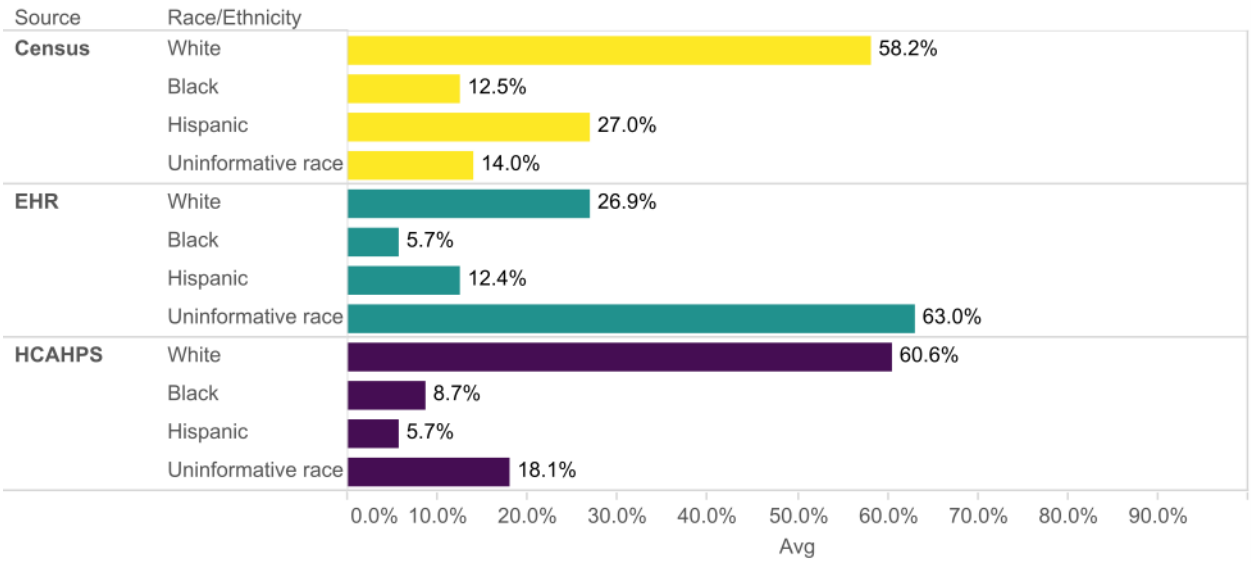

Figure 3.1: Comparison of the average Census, EHR and HCAHPS racial and ethnic distribution among 44 ZIP Codes that contained at least 100 patients in the EHR and HCAHPS datasets.

Discussion

Accurate collection of race and ethnicity information is key to recognizing disparities that affect racial and ethnic minority populations (Dorsey et al. 2014; Douglas et al. 2015; Kressin 2015; LaVeist, Gaskin, and Richard 2011). Furthermore, this information can be used to perform disease risk assessment both for individuals and populations (Gail et al. 1989; Levey et al. 2009; Stevens et al. 2006). Despite its importance, previous studies have reported challenges in collecting race and ethnicity data (Blustein 1994; Chakkalakal et al. 2015; Gomez and Glaser 2006; Hamilton et al. 2009; Lee, Grobe, and Tiro 2015; Moscou et al. 2003). For example, a study conducted in 2015 reported data quality issues by comparing patients' race and ethnicity information from different data sources within the same institution (Lee, Grobe, and Tiro 2015).

In this study, a large proportion of patients did not have informative documentation

regarding their race and ethnicity, either in the national observational databases or in the urban academic health system. These findings suggest that it is challenging to capture this information despite the inclusion of race and ethnicity data collection as part of the U.S. Meaningful Use program. When analyzing race and ethnicity data in the EHR from a single institution, changes over time did not always improve the data quality of race and ethnicity. Indeed, information loss occurred in 31% of updates.

Previous studies have illuminated some of the challenges of obtaining race and ethnicity from patients in the healthcare delivery setting. First, verbally asking patients their race and ethnicity may be perceived as a sensitive topic by both hospital personnel and patients (Baker et al. 2007; Berry et al. 2014). Second, there is a general lack of understanding of why this information is collected and how it will be used (Nelson et al. 2005). This lack of understanding poses a barrier to registration personnel asking patients their race and ethnicity. From the patients' perspective, the question is often unexpected and may not be framed with an explanation of how the information will be used and why it is important.

Some have argued that collecting race and ethnicity in the healthcare setting is increasingly unnecessary in the context of inexpensive genetic testing (Ng et al. 2008). Race and ethnicity have been used in medicine as a proxy to genetics. However, it is well established that traits occur in gradients rather than in pre-determined race categories. Currently, with the increased number of mixed populations, heritage can be more informative than the racial category itself. Additionally, with the improvements in genetics and the decreased cost of genetic testing, in the foreseeable future, we could rely on genetic testing instead genetic proxies for determination of disease risk. However, genetic testing availability will not facilitate the elimination of health disparities that have social determinants. Therefore,

until health equity is achieved, collection of race and ethnicity data is necessary to measure health disparities.

Until we transition to an era of ubiquitous genetic testing and health equity, one way to improve the quality and completeness of patient demographics in electronic health records is to allow patients to review and request updates to their information. In this study, the HCAHPS survey, a sources of patient-provided information had high rates of completeness for race and ethnicity, with only 7.1% of the records documented as "Unknown." This finding suggests that patients are willing to provide their race and ethnicity information when they have the opportunity to do so. A study conducted at one Veterans Affairs (VA) Medical Center compared patient-reported race and ethnicity information to the data available in the EHR. Investigators mailed 300 surveys to select patients that received care primarily at the VA clinic. Of the completed surveys, 15.7% contained race and ethnicity information discordant from the EHR (Hamilton et al. 2009). I compared race and ethnicity information available in the EHR to data from HCAHPS survey. Among patients with survey data, 86.3% provided informative race and ethnicity information and 66.5% of the answers were discordant with the EHR data. More than 84% of patients with uninformative race and ethnicity in the EHR provided meaningful information in the survey.

Patient-facing tools give patients the opportunity to fill out or review their information directly, removing some of the cultural sensitivity of having someone verbally asking for this information. A previous study demonstrated improvement in race and ethnicity data quality after using a custom patient portal application on a tablet computer to allow patients to review their demographic information (Polubriaginof et al. 2016). Interestingly, when self-reporting, many Hispanic patients did not seem to consider themselves to belong to

any of the OMB-defined race categories, as the majority identified their race as 'Other' and their ethnicity as "Hispanic or Latino" when self-reporting. Such phenomena have been previously described (Berry et al. 2014; Bhalla, Yongue, and Currie 2012; Markus 2008; Robbin 1999) and this behavior raises questions about the efficacy of the two-question format (i.e., collecting race and ethnicity as separate fields) that is now widely used, as well as the meaning of the constructs of "race" and "ethnicity" for patients. These findings suggest that patient-facing tools that allow patients to provide race and ethnicity information before, during, or after their healthcare encounters could markedly improve data quality. This could be accomplished in many ways, but one useful method is to use patient portals.

In summary, race and ethnicity provide valuable information for precision medicine and critical information for efforts to eliminate socially determined health disparities. However, the quality of these data is concerning. While the use of genetics is not feasible at a population level, the use of patient-facing tools have the potential of dramatically improving its quality and ultimately facilitate disease risk assessment and identification and monitoring of health disparities.

Conclusion

This study demonstrates that collection of race and ethnicity, particularly among diverse populations, can be problematic. Poor data quality for race and ethnicity can negatively impact clinical care decisions that are based on disease risk adjustment models incorporating race and ethnicity. Moreover, incomplete or inaccurate race and ethnicity data prevents public health professionals and policy-makers from measuring and reducing racial and eth-

nic healthcare disparities. To address these challenges, we recommend patient-reported data be used to improve quality and completeness of race and ethnicity.

3.2 Aim 1.2 - Assessing the quality of family history data collected in clinical databases

Background

Family history has always been considered critical component in care delivery (Berg et al. 2009) and it is described as a free genetic tool that almost every patient has access to (Guttmacher, Collins, and Carmona 2004). Since the Human Genome Project, new genomic tools have been described (Guttmacher and Collins 2003); however, family history remains critical for identifying patients that may be at higher risk to develop disease. Family history provides information that enables individualized disease diagnosis, treatment, and prevention.

Several studies have shown that family history is an important element in determining the appropriate clinical care. Knowing that a patient is at increased risk of developing a disease based on family history enables disease prevention that can vary from intensive screening to prophylactic surgery, early diagnosis and/or early and tailored treatment (Berry et al. 1997; Claus, Risch, and Thompson 1994; Saslow et al. 2007; Smith, Cokkinides, and Brawley 2012; Tyrer, Duffy, and Cuzick 2004). Currently, the U.S. Preventive Services Task Force (USPSTF) recommends risk assessment based on family history for some conditions such as screening for BRCA mutation and BRCA-related cancers (Moyer and U.S. Preventive Services Task Force 2014), osteoporosis (U.S. Preventive Services Task Force 2011), and lipid disorders in adults. (Helfand and Carson 2008) A 2007 report commissioned by the Agency for Healthcare Research and Quality (AHRQ) recommended that

collection of family history information should include diseases in first-degree relatives and second-degree relatives from both the maternal and paternal side, the relatives' age at the time of disease diagnosis, and each relatives' race and ethnicity (Qureshi et al. 2007).

In addition to guidelines and recommendations encouraging the family history data being incorporated into clinical practice, several initiatives have been put in place to increase the structured collection of family history. These goals of these initiatives are focused on collecting data to enable precision medicine, where there is a need for accurate and detailed family history data. For example, Stage 2 of the Meaningful Use program included a requirement of clinicians to use structured data entry for family history. Eligible hospitals had to have for 20% of their patients at least one structured family history data element, for at least one first-degree relative in the electronic health record (Centers for Medicare & Medicaid Services 2014a). A key hallmark in this initiative is the strict data structuring involved, and yet, previous studies on data quality have shown that clinicians describe a need for free-text documentation for expressiveness of documentation (Rosenbloom et al. 2011). The contrast between clinician desires and goals of federal initiatives present challenges to finding optimal ways to collect and store family history.

Objectives

The purpose of this study was to assess the quality of family history data captured in an established commercial EHR system at a large academic medical center. This study focused on the differences between family history data collected using structured fields vs. free-text. Data quality was measured in terms of completeness.

Research Questions

- *Does the method (free-text vs structured template) of capturing family history information impact the quality of family health history data in the EHR?*

Methods

With Institutional Review Board approval, I conducted a retrospective analysis of data from the Allscripts EHR (Allscripts Corp., Chicago IL) used at NewYork-Presbyterian Hospital/Columbia University Medical Center from 2007 to 2014.

This study focused on the differences between family history data collected using structured fields vs. free-text. Each note template in the EHR contained one or more "observation" data elements. An observation could be a text box, a Boolean (e.g., a checkbox or radio button), or numeric value. Text boxes could be fully free-text, or they could be constrained to enumerated data types, allowing only options from a predefined list, such as "low," "medium," or "high." The EHR system contained 1,560 active templates for documentation. Each of these templates contained one to several hundred discrete observations. Observations had an internal code and description specified using a configuration tool in the EHR. While the EHR vendor provided some predefined observations, the vast majority were locally defined and do not comport with any existing standard terminology. There were 140,038 observations defined in the EHR; of those, 653 had an internal code containing the words "fam hx" or "family hist."

I identified the note templates that contained these observations and queried the EHR database to identify the number of times each note template was used, as well as the number

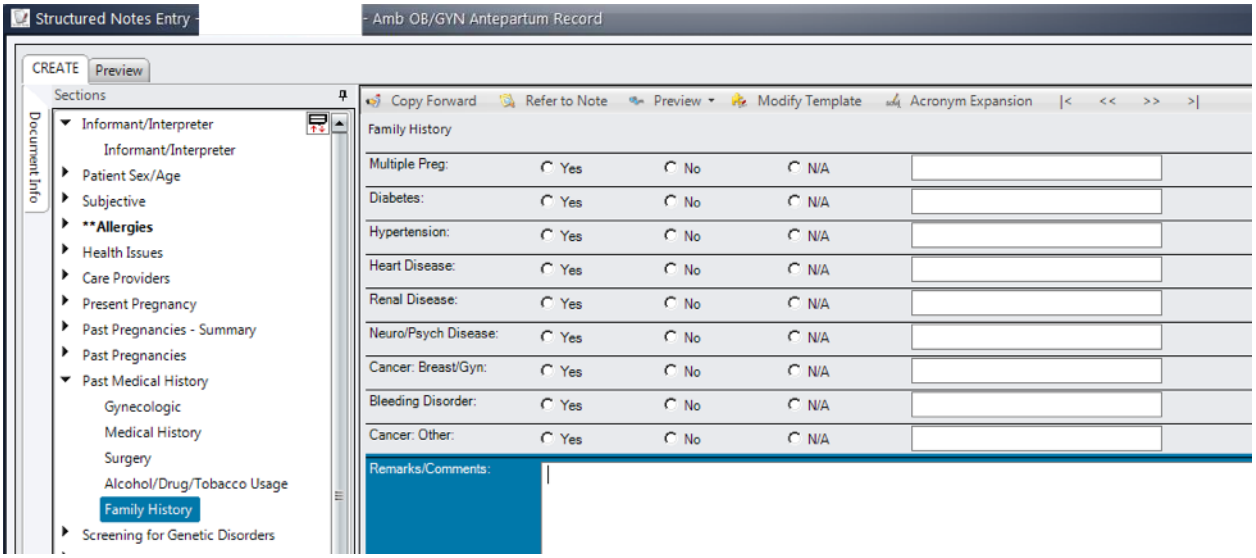

Figure 3.2: Ambulatory OB/GYN Antepartum Record: the most-used template note that contained structured family history observations.

of unique patients who had at least one of these observations recorded. The number of times each note template was used varied from 1 to 79,505, and number of unique patients varied from 1 to 67,276 for each note template. The note templates that contained the most commonly used free-text and structured text observations were selected for further analysis.

The most-used note template that contained structured family history observations was the Ambulatory OB/GYN Antepartum Record (Figure 3.2). This note template was used in our institution for obstetric patients in the institution's ambulatory care network. Overall, this note template was used 79,505 times for 67,276 unique patients. The most-used free-text family history observation was the Neurology Admission Note (Figure 3.3). This note was used for every patient admitted to the neurology service. The Neurology Admission Note was used 49,656 times for 22,642 unique patients.

For both the Ambulatory OB/GYN Antepartum Record and the Neurology Admission Note templates, 10,000 family history observation entries were randomly selected from

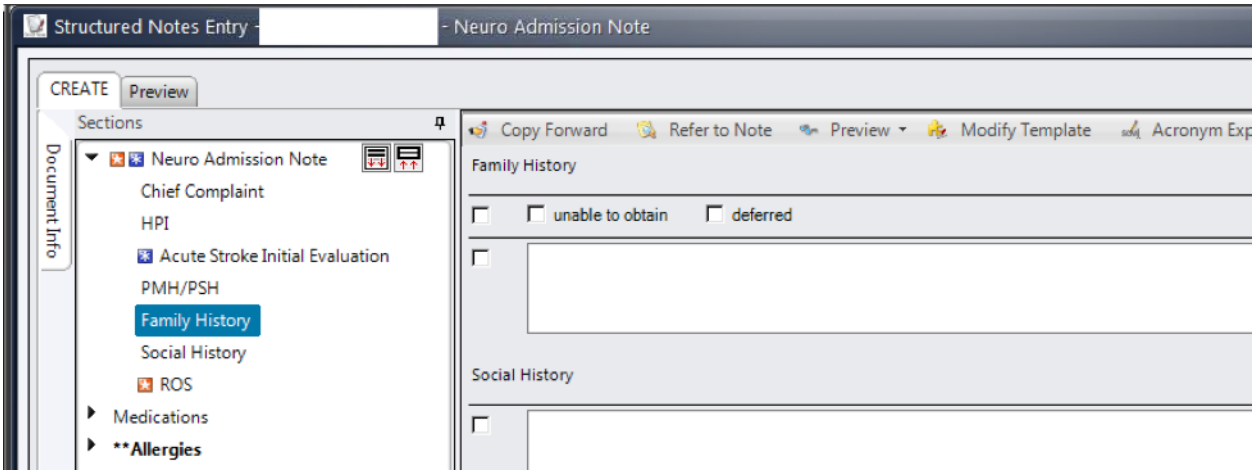

Figure 3.3: Neurology Admission Note: the most-used template note that contained free-text family history observation.

notes between 2007 and 2014. Manual annotation by a clinical expert (FP) based on pre-determined categories was performed in all 10,000 free-text observations, as well as in all structured observations that occurred more than once (9,121 observations). The categories were defined based on the content of information in the observations and the standards endorsed by AHRQ. (Qureshi et al. 2007) The categories that were used are 1) presence of disease in specified relative(s), 2) presence of disease in unspecified relative(s), 3) absence of disease and 4) other (Table 3.6). The annotation results were compared between the datasets. I performed descriptive statistics for each group, reporting the frequency of each category per group. Data quality was assessed based on completeness, where the presence of more detailed information was considered to be more complete. For example, records that captured presence of disease in a specified relative were considered more complete than records where the affected family member was not recorded.

Family History Categories	Definition	Examples
Presence of disease in specified relative(s)	When family history of disease is reported paired with a relative	"Mother: hepatic cancer; Brother: colon cancer"
		"Hypertension Mother"
Presence of disease in unspecified relative(s)	When a family history of disease is reported by itself without affected relative information	"History of diabetes, hypertension, MI, strokes."
		"Diabetes"
Absence of disease	When family history of disease is negated	"No dementia; No strokes."
		"Diabetes Denies"
Other	Miscellaneous responses	"Non-contributory"
		"None"
		"no Arabic translator"
		"No family at bedside and pt nonverbal"
		"Pt adopted, unknown family history"

Table 3.6: The traditional process of collecting patient-provided information.

Results

The results of the manual annotation of family history observations is summarized in Figure 3.4. Overall, the majority of observations (58.7%) captured by the Neurology Admission Note (free-text) included information regarding family history of disease along with the family member affected. However, when analyzing data from the observations from the Ambulatory OB/GYN Antepartum Record (structured), only 5.2% contained information specifying the patient's relative. In contrast, 7.3% of the observations from the Neurology Admission Note (free-text) contained information categorized as "Presence of disease in unspecified relative(s)," and 50.1% of the observations from the Ambulatory OB/GYN Antepartum Record (structured) captured this type of information.

Furthermore, 27.5% of the observations from the Neurology Admission Note (free-text) captured information about the absence of family history of a certain disease, while only 0.9% of the observations from the Ambulatory OB/GYN Antepartum Record (structured) captured information with this level of detail.

A large proportion (39.2%) of the observations from the Ambulatory OB/GYN Antepartum Record (structured) were classified as "Other." The vast majority of these cases referred to family history described as "N/A." Such description provides no information of the patient's family history. In contrast, 7.2% of the observations from the Neurology Admission Note (free-text) were classified as "Other." These observations often described that patients were not verbal and therefore family health history could not be collected or simply described as "None."

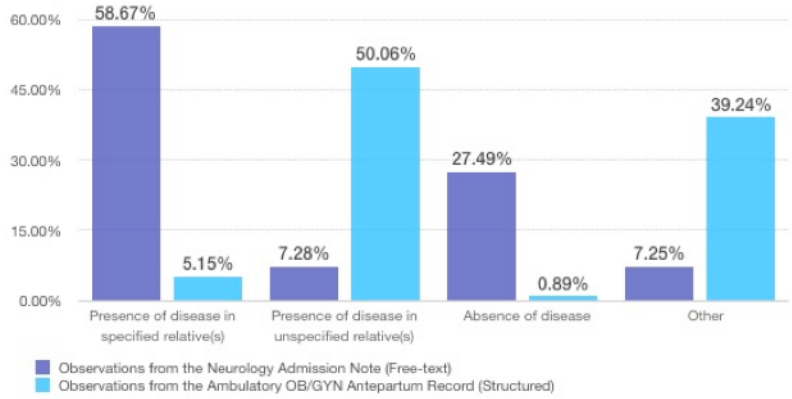

Figure 3.4: Comparison of categories from free-text (purple) and structured (blue) family history observations.

Discussion

When analyzing family history data collected using structured vs. free-text data elements, the annotations revealed that there was a considerable difference between the content of the family history information collected.

Overall, notes that used the free-text template were more comprehensive and often contained more useful information compared with structured templates. The free-text note template captured information regarding the family history of disease along with the family member affected more frequently than the structured note template, 58.7% vs. 5.2%, respectively. In the structured template shown in Figure 3.2, text could be optionally entered in the free-text box on the right, which could be used to capture additional information such as relatives, deny presence conditions, or even to record other types of information such as age, type of cancer, etc. While both note templates provided the opportunity to collect additional information in free-text, the free-text note template captured affected relatives more frequently; however, neither consistently captured other relevant details, such as the age of

onset and vital status of the relatives.

Similarly, the free-text note template captured the absence of family history of disease more consistently than the structured template. Even though the structured template included the option "No" for the set of diseases of interest, information on the absence of family history of disease was rarely reported.

Additionally, both note templates were used to record information not pertinent to family history. Those observations classified as "Other" included indicators such as "intubated," and "no family at bedside and pt nonverbal." These are important pieces of information about the patient but should not be reported as part of the family history section. The majority of the cases classified as "Other" in the structured template referred to family history described as "N/A." It was unclear what was meant by "N/A." Possible interpretations were that this could indicate that a patient would not or could not inform, had no knowledge or even that such questions were not asked.

While notes that used the free-text note template captured more comprehensive family history information, neither template captured complete family history as recommended by AHRQ (Qureshi et al. 2007). Despite the well-known and well-described importance of family history, several barriers exist in its collection and analysis, as well as in its use for personalized management based on patients' risk assessment. Barriers to collect family history can be classified in two major categories: clinician-related and patient-related.

Clinician-related barriers include lack of time to obtain, organize and analyze family history information; lack of resources and lack of reimbursement for such activity; underestimation of the value of family history data by the clinician; lack of expertise in obtaining and analyzing family history information; lack of standards for family history collection;

and lack of clear guidelines to assess patient risk based on family history. The first, and perhaps the most critical barrier for family history collection is lack of time to obtain, organize and analyze family history information (Green 2007; Guttmacher, Collins, and Carmona 2004; Rich et al. 2004; Scheuner et al. 2009; Sussner, Jandorf, and Valdimarsdottir 2011; Wilson et al. 2012a). Obtaining complete and accurate family history information, organizing it in a pedigree and analyzing family history data is extremely time-consuming. Furthermore, it is not sufficient to collect family history from patients only once. It is important to regularly update family history information, analyze it, and reconcile conflicting information. A 1989 study surveying four genetic clinics reported that the time patients spent in the first consultation varied from 3–5.5 hours, with over half of this time spent before or after the patient's appointment (Bernhardt and Pyeritz 1989). A 2011 study demonstrated that while the majority of clinicians (77.5%) reported collecting cancer family history on their patients, only 26.0% included minimum adequate cancer family history. Furthermore, 57.4% of clinicians updated family history information just once a year, and 22.2% of clinicians never updated family history information for their patients at all. When questioned about the barriers to collecting cancer family history, clinicians reported lack of time as the primary issue (Sussner, Jandorf, and Valdimarsdottir 2011). The study focused on cancer family history, but it demonstrated how challenging family history is to collect and maintain, in general. Lack of resources and reimbursement for family history collection is another important barrier (Green 2007; Rich et al. 2004; Scheuner et al. 2009; Wilson et al. 2012a). Clinicians are not reimbursed for the time spent on family history collection and risk assessment. In fact, in 2009, lack of incentives from the government was being described as one of the challenges prohibiting adequate collection of family history (Sussner,

Jandorf, and Valdimarsdottir 2011). In addition to misaligned incentives, lack of standards has also been reported as a challenge in this area (Ginsburg and Willard 2009; Green 2007; Sussner, Jandorf, and Valdimarsdottir 2011). The lack of standards for data elements, terminology, structure, interoperability, and clinical decision support rules for family history data is a huge obstacle to implement it in the clinical workflow. This point is underscored by the existence of multiple EHR templates available to assist physicians in capturing family history data. Furthermore, limited knowledge and lack of expertise in obtaining and analyzing family history by clinicians is another barrier that has been described in several studies (Ginsburg and Willard 2009; Green 2007; Guttmacher, Collins, and Carmona 2004; Scheuner, Sieverding, and Shekelle 2008; Scheuner et al. 2009; Sussner, Jandorf, and Valdimarsdottir 2011).

There are also barriers to collecting accurate family history data on the patient side. These include uncertainty about biological family composition; uncertainty about the health history of family members; inaccuracies in patient recall, language-related and cultural factors. Clinicians often cite uncertainty about biological family composition as a challenge when collecting family histories, especially in cases where the patient is part of a large biological family (Green 2007; Peace, Valdez, and Lutz 2012; Sussner, Jandorf, and Valdimarsdottir 2011). Language-related and cultural factors can also be a factor that negatively affects collection (Sussner, Jandorf, and Valdimarsdottir 2011).

There are several initiatives to facilitate and encourage the collection and use of family history data. These initiatives are focused on the use of these data for precision medicine, where the need for accurate and detailed family history data is great. Three such initiatives are: Stage 2 of the federal "Meaningful Use" EHR financial incentive program, the U.S.

Surgeon General's "My Family Health Portrait," and the HL7 Clinical Genomics Family History/Pedigree Model.

Meaningful Use Stage 2

Stage 2 of the Meaningful Use program included a requirement of clinician's to use structured data entry for family history. Eligible hospitals had to have for 20% of their patients at least one structured family history data element, for at least one first-degree relative in the electronic health record (Centers for Medicare & Medicaid Services 2014a). As discussed above, lack of incentives to collect family history is described as an important barrier. The Meaningful Use program is a strong incentive for U.S. hospitals to collect family history information. Although the determined measure of at least one structure data entry for at least one first-degree relative is far from what is considered complete family history, it is a start.

U.S. Surgeon General's "My Family Health Portrait"

Since family history data collection is extremely time-consuming, innovative tools have been created to facilitate this process. Some are leveraging patient-facing tools to collect family history data (Cohn et al. 2010; Giovanni and Murray 2010; Hulse et al. 2011; Murray et al. 2013; Orlando et al. 2013; Ozanne et al. 2009; Wu et al. 2015; Yoon et al. 2009). The U.S. Surgeon General's My Family Health Portrait, (*My Family Health Portrait*) a federal initiative to collect family history, is a website that allows patients to collect family history information and share their information with family members and healthcare providers. A study conducted in 2011 described that the average time taken to input family history

information was 15 minutes, in a range from 3 to 45 minutes (Owens et al. 2011). Instead of having a healthcare provider questioning patients about their family history, patients can enter their own data, saving clinician time–the major barrier for family history data collection. This practice also engage patients in their care and gives them time to review their family information and contact relatives and question them about information that they may not know. Engaging patients in this fashion encourages more accurate family history information. Of note, one advantage of using electronic questionnaires is that certain questions can be made mandatory, and branching logic can be employed. In contrast, in a clinical encounter, the doctor may forget to ask certain questions or may skip questions due to lack of time.

HL7 Clinical Genomics Family History/Pedigree Model

It is important to emphasize that to fully represent family history information, data representation must be multidimensional since it is necessary to not only capture the disease but also the relative affected, age of onset, and cancer type if applicable. Moreover, development of standards to support interoperability is essential for sharing data for clinical care and clinical research purposes. In the domain of family history, HL7 has a workgroup that specifically works on development of models for representing family history. The workgroup has developed the HL7 Clinical Genomics Family History (Pedigree) Model (*HL7 Version 3 Implementation Guide: Family History/Pedigree Interoperability, Release 1 - US Realm*). It is a standard for capturing data within a system as well as to transmit family history data between systems. It includes patient's family and familial relationships, diseases, genetic data and risk analysis. This HL7 standard is already used by the U.S. Surgeon Gen-

eral's My Family Health Portrait application, and there is reason to believe that it will be important for EHR vendors and other stakeholders to adopt this standard moving forward.

Limitations

The study analyzed only observations contained in two note templates (out of a total of 1,560 available) in our EHR. One note template was used in the ambulatory care setting and the other template was used for hospital admissions. Although the templates were selected based on the fact that they were the most frequently used templates at our institution, it is unclear if analysis of other EHR templates would yield different conclusions. Additionally, manual annotations were conducted by a single clinical expert, limiting the evaluation of the annotations.

Conclusion

In summary, this study focused on the differences between family history data collected using structured fields vs. free-text. While observations from the free-text note template were more comprehensive than structured observations, neither was ideal for capturing patients' complete family history. 58.7% of observations from the free-text note template captured information regarding the family history of disease along with the family member affected vs. only 5.2% from the structured note template. However, neither consistently captured other relevant details, such as the age of onset and vital status of the relatives. Numerous efforts have been made to collect family history data in the electronic format and to facilitate its use in the clinical setting, but several barriers remain unsolved. Patient-facing

tools for collecting family history data may improve data quality of family history in EHRs.

3.3 Aim 1.3 - Assessing the quality of smoking status collected in clinical databases

Background

Smoking is an important risk factor for multiple diseases, including cardiovascular diseases and numerous types of cancer. It remains the number one cause of preventable death in the United States (National Center for Chronic Disease Prevention and Health Promotion Office on Smoking and Health 2014). The collection of patients' smoking status during clinical encounters is critical to providing patients with resources to quit smoking. Smoking cessation can be difficult, and clinical visits are opportunities to intervene and recommend smoking cessation programs and therapies. Obtaining a patient's smoking status is a crucial step in beginning smoking cessation interventions and monitoring progress (Caplan, Stout, and Blumenthal 2011). It may seem that recording updates to smoking status in a timely and accurate manner would be straightforward using modern electronic health records. This may not be the case for several reasons, including lack of standard terminology and granularity for data collection, shifting cultural attitudes regarding tobacco use, and potentially frequent changes in individuals' smoking behavior (Committee on the Recommended Social and Behavioral Domains and Measures for Electronic Health Records, Board on Population Health and Public Health Practice, and Institute of, Medicine 2015; Winden et al. 2015). As the American author, Mark Twain, famously quipped, "Giving up smoking is the easiest thing in the world. I know because I've done it thousands of times."

Given the clinical importance of smoking status, the "Meaningful Use" financial in-

centive program for electronic health record (EHR) adoption in the U.S. included a requirement for healthcare providers to capture patients' smoking status electronically in structured fashion (Centers for Medicare & Medicaid Services 2010). Meaningful Use has helped to standardize data collection of smoking status and other information. However, even with improved standards for representing information, data quality issues have persisted in many patient-provided data types, such as race and ethnicity (Klinger et al. 2015; Lee, Grobe, and Tiro 2015) and family history (Polubriaginof, Tatonetti, and Vawdrey 2015; Powell et al. 2013). Previous studies on data quality have shown that clinicians describe a need for free-text documentation for expressiveness of documentation; however, these affordances challenge data reuse (Rosenbloom et al. 2011).

Appreciating the challenges associated with data quality and the balance between the expressiveness of free-text and the benefits of structured data, I set out to answer a very simple question: how many of our hospital's patients are known to be active smokers? I undertook a study to analyze how smoking status is currently being collected in a large academic medical center and to evaluate the quality of this data in EHRs.

Objectives

The purpose of this study was to assess how smoking status was collected in a large academic medical center, and to evaluate the quality of smoking status data in EHRs. Specifically, this study assessed the completeness, concordance and plausibility of smoking status recorded during hospitalizations.

Research Questions

- *How is smoking status information captured in the EHR?*
- *What is the quality of smoking status information captured in the EHR?*

Methods

I conducted a retrospective analysis of smoking status data from the Allscripts Sunrise EHR (Allscripts Corp., Chicago IL) used at NewYork-Presbyterian Hospital/Columbia University Medical Center. All patients who had at least one hospital discharge during 2016 were included in the study.

The EHR system contained hundreds of active templates for documentation. Each of these templates contained one-to-several hundred discrete observations. An observation could be a text box, a Boolean (e.g., a checkbox or radio button), or numeric value. I identified observations in which the description contains the stemmed words "smok," "cigar" or "tobacco" and queried the EHR database to identify the number of times each one of these observations was recorded during the study period. Smoking status information was recorded either as free-text (i.e., when the parameter is "Cigarettes (packs per day)") or structured (i.e., selected from a picklist) observations. The picklist was often shown as a set of radio buttons, such as with a parameter labeled "Tobacco Use / Smoking Status," and one possible choice in the picklist was "Never smoker."

Prior exploratory analysis showed that 94% of patients had at least one structured smoking status observation recorded in a structured field. Based on this finding, I only used the structured data for the remainder of the analyses. Some of the structured observations

Clinically Actionable Smoking Status Categories	EHR Documented Categories	Number of Observations
Never Smoker	**Never Smoker**	**67,052**
	Smoker (No)	12,979
	Patient Denies	560
Current Smoker	**Current every day smoker**	**5,188**
	Current some day smoker	**1,418**
	Light smoker	**714**
	Heavy Smoker	**267**
	Smoker, current status unknown	**676**
	Smoker (Yes)	1,968
Former Smoker	**Former smoker**	**16,307**
	Ex-smoker	5
	Quit / Stopped	8,275
Unknown Smoking Status	**Unknown if ever smoked**	**16,514**
	Unknown	58
	Unable to assess	63
	N/A / None	569

Table 3.7: Description of the mapping from smoking status categories as recorded in the EHR to the four clinically actionable categories. Smoking status categories documented in the EHR that utilize the standard criteria defined by the Meaningful Use program are highlighted in bold.

captured smoking status following the standard criteria set by the Meaningful Use program. The program specifies eight distinct categories for collecting smoking status: "Current every day smoker," "Current some day smoker," "Former smoker," "Never smoker," "Smoker, current status unknown," "Unknown if ever smoked," "Heavy tobacco smoker," and "Light tobacco smoker." CentersforMedicareandMedicaidServicesCMS:2014wn I classified smoking status from the EHR into one of four clinically actionable categories: "Current smoker," "Former smoker," "Never smoker," and "Unknown smoking status" as described in Table 3.7.

Overall, data quality of smoking status was assessed based on the percentage of patients with consistent and informative smoking status available (i.e., not classified as "Unknown" in the database, or not conflicting if recorded multiple times). Patients with discrepancies in recording smoking status, such as in the previous example, were classified as: plausible

Descriptor	Sample size
Patients	48,909
Hospital Encounters	64,451
Encounters with smoking status recorded	62,988 (98%)
Encounters with smoking status recorded discrepantly	19,176 (30%)

Table 3.8: Description of the mapping from smoking status categories as recorded in the EHR to the four clinically actionable categories. Smoking status categories documented in the EHR that utilize the standard criteria defined by the Meaningful Use program are highlighted in bold.

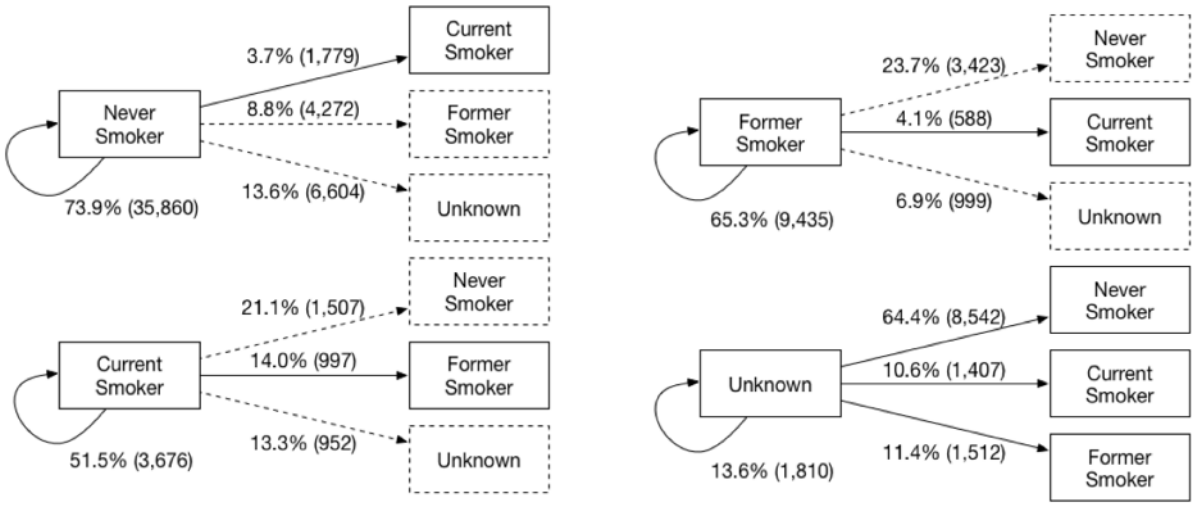

Figure 3.5: Changes of smoking status overtime. Dashed changes demonstrate implausible discrepancies and continuous lines represent plausible changes in longitudinal data. The number of changes recorded in our sample is reported in parentheses and the percentage it represents for each category is included in the Figure.

and implausible. Plausible cases occurred when the change was feasible to happen such as a change from "Never smoker" to "Current smoker"), and implausible occurred when the conflict was not possible to happen or in cases where there was a loss of information. For example, a change from "Former smoker" to "Never smoker"). More examples of this classification are illustrated in Figure 3.5.

Additionally, I analyzed the discrepancies in smoking status generated when inconsistent information was recorded for a patient during the hospital visit in different clinical notes in the EHR. I also investigated the discrepancies of smoking status recorded by different provider types (e.g., nurses, medical doctors, care coordinators, social workers). To assess whether provider type had an impact on the number of discrepancies observed, I calculated the number of distinct providers' roles recording smoking status for each admission. I then compared the number of distinct provider roles for patients with and without discrepancies in the recorded smoking status.

I also calculated the time interval (in days) between smoking status documentation events to better understand the distribution of the data during the one-year study period. For example, a time interval of zero means that two observations were recorded on the same day, and a time difference of one indicates that a second observation was recorded one day after the first observation.

Results

Overall, I reviewed 48,909 patients having 64,451 hospital encounters in the one-year study period. I identified 203,048 observations of smoking status for 47,849 unique patients across 62,988 distinct hospital encounters. No smoking status documentation was identified for 1,463 visits from 1,060 distinct patients, representing 2% of the number of hospital encounters and 2% of the overall number of patients. In other words, 98% of patients and 98% of hospital visits had documentation regarding the patient's smoking status. Of those records with smoking status, 59,663 visits (93%) from 45,822 patients (94%) had this

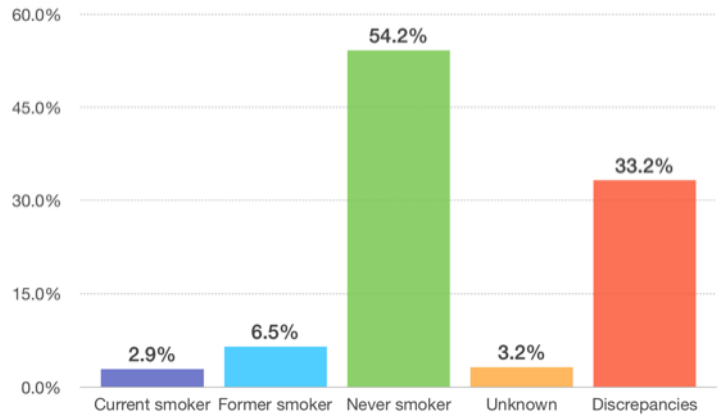

Figure 3.6: Smoking status of patients seen in 2016.

information recorded in structured format. After pre-processing, 45,771 patients (94%), including 59,593 visits (92%) and 129,134 observations were classified into four distinct smoking status categories. The number of observations and the mapping to the simplified smoking categories are described in Table 3.7. The description of smoking status data during the one-year study period is described in Table 3.8. Patients had an average of 1.3 visits/patient during the study period, with the maximum number of visits a single patient being 23 visits.

How many patients are smokers?

Overall, 54.2% of the patients in our sample were classified as non-smokers, 6.5% as former smokers, 2.9% as current smokers and 3.2% as having unknown smoking status. The remaining 33.2% of patients had at least one discrepant assessment of smoking status documented. I determined that only 63.6% of our study population had a consistent, unchanging smoking status during the one-year study period (Figure 3.6).

Longitudinal One-year Review

Overall, 15,048 patients (32.9%) had smoking status recorded in a single note, and 30,723 patients (67.1%) had more than one note with documentation regarding smoking status. Among the patients with more than one note with smoking status documented, I identified 83,363 changes in documented smoking status collected longitudinally during the one-year study period.

Among the changes in smoking status documentation, 32,582 (39.1%) had a conflicting smoking status. These discrepancies were observed in records from 15,207 distinct patients, representing 33.2% of our study population. However, because the study used longitudinal data and smoking status is not a static concept (i.e., it can change over time), some of these discrepancies are feasible. For example, someone that never smoked can become a smoker. Others, however, are implausible. For example, logically, a "never smoker" cannot become a "former smoker", nor can a "current smoker" become a "never smoker," unless some of the data were recorded incorrectly. Other changes are plausible but not good from a data quality standpoint. Having a patient with documentation regarding smoking status and later not having smoking status (smoking status as "unknown") demonstrates loss of information. Implausible changes as well as changes from a well-defined smoking status to an uninformative category were considered discrepancies due to data quality issues. I identified 17,757 discrepancies (implausible changes and loss of information changes), which constituted 54.5% of changes, in 10,836 distinct patients. These discrepancies are represented in Figure 3.5 as dashed lines, while the other changes are represented in continuous lines.

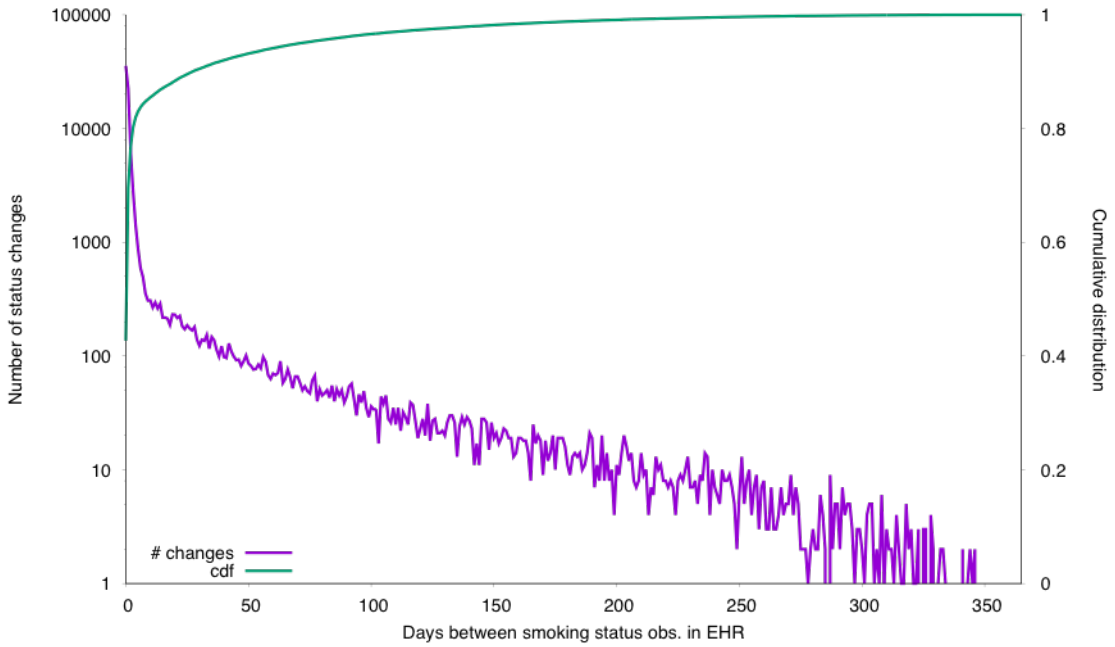

Figure 3.7: Number of smoking status changes by time interval documentation. Time interval is measured in days. Number of status changes is represented in logarithmic scale. CDF = cumulative distribution function.

On average, the time interval between different smoking status documentation was 11 days, with a minimum of 0 days (i.e, same-day documentation), and a maximum of 362 days. Most patients (80.6%) had a time interval between documentation events of less than or equal to 10 days, with 61.1% of patients having a subsequent documentation event within one day of the previous event. (Figure 3.7)

Duplicate Assessments During the Same Hospital Encounter

While it is plausible to observe changes in smoking status over the course of one year, smoking status should not change during the same hospital encounter. Given this rationale, I considered all changes during a hospital encounter to reflect a data quality issue, since smoking status should be consistent throughout a single admission. During the study pe-

Number of distinct provider roles	Encounters without discrepancies	Encounters with discrepancies
1	45.97% (8,791)	26.15% (6,650)
2	48.08% (9,195)	52.49% (13,345)
3	5.91% (1,130)	20.94% (5,323)
> 3	0.05% (9)	0% (1)

Table 3.9: Description of smoking status data during the one-year study period.

riod, I identified 59,663 distinct encounters from 45,822 patients. Of those, 32.2% of the hospital encounters (19,176 visits) had at least one conflicting smoking status recorded, which includes 14,798 patients (32.3% of our cohort of patients).

Discrepancies Among Various Provider Roles

For patients with a smoking status recorded in a structured field, 70.8% were documented as part of nursing notes, 12.9% came from social work notes, 11.6% from physician notes, and the remaining (4.7%) from notes entered by other health care professionals.

Among hospital encounters with more than one assessment of smoking status, encounters with documentation from a single role of provider (e.g., nurse) had fewer discrepancies compared with encounters containing smoking status assessments from providers with disparate roles (Table 3.9). For example, if multiple nurses documented smoking status during an admission, the number of distinct provider roles would be equal to one. However, if multiple nurses and multiple physicians documented this information, then the number of distinct provider roles would be two.

Discussion

The Centers for Medicare and Medicaid Services (CMS) Meaningful Use program requires participating healthcare providers to record patients' smoking status in a structured fashion (Centers for Medicare & Medicaid Services 2010). The program specifies eight distinct categories for collecting smoking status: "Current every day smoker," "Current some day smoker," "Former smoker," "Never smoker," "Smoker, current status unknown," "Unknown if ever smoked," "Heavy tobacco smoker," and "Light tobacco smoker."

I identified smoking status assessments (either represented in free-text or structured fields) for 98% of patients and 98% of visits. When focusing on structured documentation, I observed that 94% of patients and 92% of visits had at least one structured smoking status observation recorded. When analyzing smoking status data in the EHR, I transformed the Meaningful Use categories and other smoking status assessments into four clinically actionable categories: "Current smoker," "Former smoker," "Never smoker," and "Unknown smoking status." I observed that a 33.2% of the patients had inconsistencies in the documented smoking status during the one-year study period and 32.3% of the patients had at least one discrepancy during a single visit. These discrepancies suggest that reliable information on smoking status may not be available for a large number of patients.

Going back to the initial question of "how many patients are current smokers?" – the answer is, I do not know. Based on the analysis conducted in this study, more than half of the patients during the one-year study period were recorded consistently as non-smokers and just 2.9% were recorded consistently as current smokers. In contrast, other population-based studies estimate that 15.1% of adult Americans smoke (Jamal et al. 2016). One-third

of the studied population had inconsistencies in their smoking status, making the determination of tobacco use for these patients difficult. While smoking status was documented in 98% of hospital encounters (and therefore the criteria of Meaningful Use were satisfied), our one-year sample of hospital encounters did not contain consistent smoking status information for 36.4% of patients.

Despite the well-known and well-described importance of collecting smoking status, our institution's EHR did not have a centralized location to store smoking status information. Smoking status was collected as part of clinical notes, in either structured or free-text format. The fact that disparate healthcare providers recorded this information in several different notes resulted in many inconsistencies across notes. Further complicating the matter, different note templates allowed for different granularities of smoking status data collection. Some templates included a free-text box that allowed clinicians to enter details such as intensity of smoking, number of cigarettes per day, or when the patient stopped smoking. Other templates had only the Meaningful Use-required structured fields embedded.

Since I used longitudinal data, and smoking status is not a static concept (i.e., it can change over time), I classified smoking status changes into two distinct categories: plausible and implausible. In this study, implausible changes constituted 21.3% of all changes. Previous research has also identified consistency issues regarding tobacco use recorded in different notes in EHR systems. For example, in 2016 a research study used natural language processing to parse clinical notes and extract smoking status from various clinical notes. The authors identified several inconsistencies when comparing smoking status recorded in clinical notes (Wang et al. 2016).

Inconsistencies can be attributed to challenges in the data collection process, includ-

ing clinician-related and patient-related factors. Clinicians may not inquire at all about a patient's smoking status, or they might ask the question in a manner that leads to bias in the patient's answer. Depending on how clinicians phrase the question, patients may not feel comfortable answering. On the other hand, patients may have their own motivations to be less than truthful when providing smoking status information to clinicians, or they may inexplicably provide different smoking status responses depending on the person asking.

I conducted an analysis to identify whether hospital encounters with more than one clinical note without discrepancies were more likely to have documentation from a single provider role than encounters with discrepancies. Interestingly, I identified that encounters with multiple notes documented by the same type of provider had less discrepancies than patients with documentation from multiple types of providers. The difference I observed in discrepancies may be explained by the fact that clinicians usually do not read notes from other clinicians' roles. Previous studies have shown that most clinical notes are not read by the entire clinical team (Hripcsak et al. 2011a). Instead, clinicians may be more inclined to read clinical notes from their peers (i.e., within the same provider role). While it is important for multiple providers to assess patients' smoking status, barriers to accessing previously documented information regarding tobacco use by healthcare providers may increase vulnerabilities that allow discrepancies to propagate.

One limitation of our analysis was the use of data from only a single year and from only a single healthcare system. During a one-year period within our EHR system, I found that 33.2% of patients had discrepancies in documentation of smoking status. Furthermore, 54.5% of those inconsistencies were deemed implausible (Figure 3.5). Most patients had changes recorded within 10 days of the previous smoking status assessment. Given the

short time difference between documentation events, even plausible changes (e.g., converting from "current smoker" to "former smoker") seem unlikely. These data quality issues demonstrate just some of the considerable challenges healthcare providers and secondary users of EHR data. If I have difficulty in identifying a single meaningful and consistent smoking status using only one-year worth of data, the use and sharing of multiple years of data present even bigger challenges. For example, for encounters with conflicting smoking statuses, which one should be used in a clinical decision support system related to smoking cessation? Or which one should be reported to external organizations? Efforts using smoking status information from EHRs, including future smoking cessation initiatives, should further investigate patients identified as "Unknown smoking status" as well as patients with discrepancies in smoking status.

Recommendations

In this study, I observed that smoking status is currently being collected as part of clinical notes by multiple healthcare providers, and for almost all patients. The categories used are not consistent across clinical notes, recording smoking status in different granularities. I propose the use of four distinct clinically actionable categories: "Never smoker," "Current smoker," "Former smoker," and "Unknown smoking status." More detailed information for each one of these could also be collected in a standardized fashion, such as "packs/day" and start and quit date. Currently, this additional information is being captured in free-text format and inconsistent across notes (e.g. some use packs/day while other record this information as cigarettes/day).

In our institution, smoking status is not stored in a centralized location, but is rather

being collected as part of disparate clinical notes. The current system of data collection of smoking status presents challenges consistently collecting this information. While it is important for multiple providers to collect patients' tobacco use information, the fact that this information is collected and stored in various notes without standardization makes it challenging for clinicians to know if the patient already provided their smoking status to other clinicians, and whether this information is longitudinally consistent. In an attempt to solve these challenges, I propose to store patients' smoking status in a centralized fashion and having clinicians verifying this information in every encounter by asking patients about tobacco use.

One way to improve the consistency and correctness of patient-reported information, such as smoking status, is to allow patients to review and update their own information. Previous studies have shown that self-reported smoking status is accurate (Patrick et al. 1994; Wagenknecht et al. 2011). This task can be facilitated by health information technology in many ways, including the use of patient portals and tablet computers for this task. Patient-facing tools have been used for collection of multiple patient-provided data types such as race and ethnicity, family history, symptoms, medication reconciliation and adherence. These studies have shown that patients are willing to provide and review their information (Dullabh et al. 2014; Pyper et al. 2004; Weingart et al. 2008), that EHR data is often incomplete or inaccurate (Ball and Lillis 2001; Douglas et al. 2015; Kaplan 2014; Klinger et al. 2015; Kressin 2015; Lee, Grobe, and Tiro 2015; Polubriaginof, Tatonetti, and Vawdrey 2015; Qureshi et al. 2009; Staroselsky et al. 2006; Welch, Dere, and Schiffman 2015), and that patients can identify discrepancies, provide useful information and help keeping records up-to-date (Arsoniadis et al. 2015; Staroselsky et al. 2006, 2008; Wu et

al. 2014). Studies have also shown that there are many benefits of involving patients in their care, including improving patient engagement, patient satisfaction, health behaviors and health status as well as helping to attract and retain patients (Arar et al. 2011; Davis Giardina et al. 2014; Dwamena and Rovner 2012; Epstein et al. 2010; Otte-Trojel et al. 2014). With patient-facing tools, patients could provide their smoking status based of the four clinically actionable categories, as described above. Patients providing this information to a computer could also mitigate the potential biases introduced by clinicians asking the question.

Conclusion

In summary, while 98% of hospital encounters at our institution during 2016 contained information regarding the patients' smoking status, 32% of the encounters had discrepancies in smoking status information. For encounters with more than one clinical note documenting smoking status information, 54% of the subsequent documentation events had implausible changes. While other sources suggest that approximately 15% of adult Americans smoke, only 2.9% of our patients were consistently documented as current smokers. This finding demonstrates that while Meaningful Use has improved data collection of smoking status in terms of completeness, we may not be appropriately identifying patients that smoke. Centralized documentation with clinically actionable smoking status categories available for data collection, and implementation of patient-facing tools that allow patients to directly record their information, may help improve data quality of smoking status in EHRs.

Chapter 4

Aim 2 - Evaluate methods for improving quality of patient-provided data

Health information technology (HIT) is often touted as a means to improve the quality and efficiency of healthcare (Chaudhry et al. 2006). A number of HIT-related interventions have focused on collecting patient-provided information, which can be categorized in three different groups: 1) broad policy initiatives, 2) patient-facing tools, and 3) algorithms and informatics methods for collecting and using patient-provided data. While these interventions have generally been successful in improving the quantity of patient-provided information, there is a limited understanding of the impact these interventions have on data quality.

Policy initiatives

HIT-related policy changes can significantly impact healthcare. A major component of the 2009 Health Information Technology for Economic and Clinical Health (HITECH) Act focused on increasing adoption of EHRs (Blumenthal and Tavenner 2010; Blumenthal 2009). In order to achieve this goal, financial incentives through the Meaningful Use program were made available to institutions in the United States, resulting in over 95% of hospitals reportedly using EHRs by 2016 (Conn 2016; Health Information Technology

2017). The Meaningful Use program required providers to collect patient-provided data such as race and ethnicity (Centers for Medicare & Medicaid Services 2014b), family history (Centers for Medicare & Medicaid Services 2014a), and smoking status (Centers for Medicare & Medicaid Services 2010) in a structured fashion in the EHR. To date, there has been little research on the impact of Meaningful Use for improving the data quality of race and ethnicity, family history, and smoking status in EHRs (Chakkalakal et al. 2015; Douglas et al. 2015; Klinger et al. 2015).

Another example of a federal policy initiative that has impacted healthcare delivery in the United States is the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) survey. Currently, the Centers for Medicare & Medicaid Services (CMS) require hospitals to administer HCAHPS surveys after patient discharge to measure patient satisfaction and experience. These measures are tied to reimbursement, which can lead to gain or loss of a percentage of the Medicare payments, transforming patient experience into a financial priority for hospitals. (*Hospital Consumer Assessment of Healthcare Providers and System*) As part of the HCAHPS survey, patients are requested to provide demographic information, along with feedback regarding their medical care and hospital stay.

While broad policy initiatives may have increased the prevalence of certain patient-provided information in EHRs, such policies have not been evaluated for their impact on the data quality of the information collected.

Patient-facing tools

HIT interventions have also been developed to supply patients with access to some of their clinical information through patient portals, and also to collect information using similar types of interventions. Online patient portals are now available in many healthcare systems, allowing patients to view laboratory test results, medications, problem lists, and health summaries, refill prescriptions, and schedule appointments (Caligtan et al. 2012; Cimino, Patel, and Kushniruk 2001; Collins et al. 2011; Greenhalgh et al. 2008; Halamka, Mandl, and Tang 2008; Hassol et al. 2004; Kaelber et al. 2008; Maher et al. 2015, 2016; Masterson Creber et al. 2016; Nazi et al. 2010; O'Leary et al. 2015; Prey, Restaino, and Vawdrey 2014; Pyper et al. 2004; Ralston et al. 2007; Reti et al. 2010; Tang and Lee 2009; Wilcox et al. 2016). More recently, with the OpenNotes initiative, some healthcare provider organizations are also enabling patients to review their own medical notes via online portals (Delbanco et al. 2010; Grossman et al. 2017; Leveille et al. 2012; Nazi et al. 2015; Walker et al. 2011). Computer applications have also been deployed by some hospitals to collect patient-provided information. Patient-facing tools exist for collecting information such as medical history, family history, preventive services information such as screening tests and vaccines, and medication reconciliation (Arsoniadis et al. 2015; Feero 2013; Giovanni and Murray 2010; Hoyt et al. 2013; Hulse et al. 2010; Murray et al. 2013; Peace, Bisanar, and Licht 2012; Pyper et al. 2004; Staroselsky et al. 2006, 2008; Wilson et al. 2012a; Wu et al. 2014; Yoon, Scheuner, and Khoury 2003).

While numerous patient-facing tools exist, questions regarding the quality of the data collected remain. Efforts have focused on providing information to patients and collecting

information from them. However, there have been limited attempts to demonstrate whether patient-provided information collected through patient-facing tools have equal or higher quality than data collected by providers.

Algorithms and informatics methods focusing on using patient-provided data

Algorithms and informatics methods have been developed to use the data already available in clinical databases to support biomedical research. Natural language processing (Friedman, Hripcsak, and Shagina 1999) and EHR phenotyping (Hripcsak and Albers 2013) are important methods that support EHR data reuse for research studies. More recently, initiatives such as Informatics for Integrating Biology & the Bedside (i2b2) (Murphy et al. 2010) and Observational Health Data Science and Informatics (OHDSI) (Hripcsak et al. 2015) sought to produce open-source frameworks that allow different teams of researchers to run the same analyses on separate clinical databases, and combine the results with confidence, an important step towards reproducibility in biomedical research. Through the OHDSI platform, Hripcsak and colleagues conducted a 250-million patient observational research study focusing on the characterization of treatment pathways for disease, leveraging the medical records and administrative claims data from multiple countries (Hripcsak et al. 2016).

While many methods and frameworks have been developed to make discoveries from EHR data, efforts assessing the quality of data available and improving these data sets using informatics methods have been scarce. The use of patient-provided information in general

has been limited, likely due to the missingness and incompleteness of these data in these datasets. There has been little research on the potential impact of informatics methods to improve availability and quality of patient-provided data in clinical databases. For instance, on one hand, family history is difficult and time-consuming to collect directly (Green 2007; Guttmacher, Collins, and Carmona 2004; Rich et al. 2004; Scheuner et al. 2009; Sussner, Jandorf, and Valdimarsdottir 2011; Wilson et al. 2012b). On the other hand, emergency contact information, a patient-provided data element that often contains family relationship information, is requested from patients at nearly all hospitals. There have been few efforts to deduce family relationship from emergency contact information to discover family history, even in research. Methods that glean useful information from data that are already commonly collected can enhance the utility of large databases, further supporting discoveries and clinical research.

4.1 Aim 2.1 - Analyze the impact of various interventions on the quality of race and ethnicity information

Background

Race and ethnicity are commonly used for estimating disease risk (Gail et al. 1989; Levey et al. 2009; Stevens et al. 2006) and for assessing health disparities (Dorsey et al. 2014; Douglas et al. 2015; Kressin 2015; LaVeist, Gaskin, and Richard 2011), and these characteristics are frequently reported in observational studies that rely on EHR data. The goal of health information technology (HIT) is to improve the quality and efficiency of healthcare. In the United States, the Meaningful Use financial incentive program required that EHRs collect patients' race and ethnicity in a structured fashion (Centers for Medicare & Medicaid Services 2014b; Rao et al. 2011). The Meaningful Use program adopted the model for race and ethnicity data collection developed by the Office of Management and Budget (OMB) (OMB 1997). According to this standard, race and ethnicity information can be collected either in a single question or in a two-question format. It also established that patient-provided information should be considered the gold standard for the collection of race and ethnicity data.

Previous research on the quality of race and ethnicity data recorded in EHRs was conducted with small groups of selected patients, and did not include analysis on the impact of the Meaningful Use program on the quality of these data (Hamilton et al. 2009; Klinger et al. 2015). Because of the critical importance of race and ethnicity information for addressing health disparities and for assessing disease risk, I undertook a study to assess how

different interventions impacted the data quality of race and ethnicity. Specifically in this study, I focused on whether data quality improved over time as a result of implementing Meaningful Use requirements, and the impact of allowing patients to directly provide or curate their race and ethnicity information.

Objectives

The purpose of this study was to analyze the impact, separately, of a policy and an informatics intervention on the quality of race and ethnicity information. Specifically, this study analyzed the impact of the Meaningful Use program and patient-facing tools, such as patient surveys and a pacing-facing tablet application, on the quality of race and ethnicity data.

Research Questions

- *What is the impact of Meaningful Use on the data quality of race and ethnicity information?*
- *Does allowing patients to provide their race and ethnicity information change the data quality of this data?*

Methods

To analyze the impact of policy change, I conducted a pre-/post- comparison of the percentage of race and ethnicity data that was informative as recorded in the EHR to analyze the impact of the Meaningful Use program. Additionally, for the patients that had both

EHR data and a completed HCAHPS survey, I compared the usefulness of the information provided by patients to what was stored in the EHR. Lastly, I analyzed the impact of a local informatics intervention in the quality of race and ethnicity information using data from a block-randomized controlled trial.

Policy change - Meaningful Use program

I conducted a retrospective analysis of race and ethnicity data recorded for unique patients who visited NewYork-Presbyterian Hospital/Columbia University Medical Center (NYPH/CUMC) from 2004 to 2014. Before Meaningful Use Stage 1, race and ethnicity data were collected using a single field. The following categories could be entered in that field: "American Indian or Alaska Native," "Asian," "Black or African American," "Hispanic or Latino," "Native Hawaiian or Other Pacific Islander," "White," "Unknown," "Other," or "Declined to Answer." In response to Meaningful Use requirements, our institution implemented a two-question format for race and ethnicity data collection, where one field captured a patient's race ("American Indian or Alaska Native," "Asian," "Black or African American," "Native Hawaiian or Other Pacific Islander," "White," "Unknown," "Other," or "Declined to Answer"), and a second field captured the patient's ethnicity ("Hispanic or Latino," "Not Hispanic or Latino," "Declined to Answer," "Unknown"). I obtained race and ethnicity data collected during the study years and from before and after Meaningful Use changes were implemented. I performed descriptive statistics on the frequency and quality of data captured using the EHR. Our institution attested compliance to Meaningful Use Stage 1 requirements at the end of 2012; therefore my analysis segmented records from 2004–2012 as pre-Meaningful Use and from 2013–2014 as post-Meaningful Use. When pa-

tients had multiple visits to the institution, I only used data from the most recent visit. Patients classified as "Unknown," "Other" or "Declined to Answer" were considered to have clinically uninformative data. I combined these categories into a larger category designated as "Uninformative." Quality was assessed based on the percentage of "Uninformative" race and/or ethnicity in the database.

Patient-facing tools – United States National Databases

As previously described in Aim 1.1, I conducted a retrospective study to assess the quality of race and ethnicity information using data from HCUP and Optum Labs Data Warehouse, two large observational databases. To assess the difference in race and ethnicity data quality between national databases and patient-provided datasets, in addition to the HCUP and Optum databases, I examined the dataset generated from the National Health and Nutrition Examination Survey (NHANES). NHANES collects data from 5,000 U.S. adults and children per year (Disease Control and Prevention 2007). Among other information, it collects race and ethnicity in a single-question format, with the response coded as "White," "Black or African American," "Hispanic or Latino," "Not Hispanic or Latino," or "Unknown." I used NHANES data from 1999 to 2011 as a source of patient-provided race and ethnicity data, and reported the percentage of patients with uninformative race and ethnicity data.

Patient-facing tools - Academic Healthcare System in New York City

As reported in Aim 1.1, I previously assessed the quality of race and ethnicity data from the EHR at an academic health system that serves a racially and ethnically diverse popula-

tion in 10 hospital campuses in and around New York City. To assess differences between patient-reported race and ethnicity information and data from the EHR, I conducted a retrospective analysis using data from the EHR of the academic medical center and from the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) Survey. I identified patients who had at least one hospital encounter from January 2014 through December 2015. Among these patients, I identified patients who had answered the HCAHPS survey. The HCAHPS survey collected demographic information from hospital patients in the form of a document sent via U.S. Mail after discharge. To collect race and ethnicity data, the survey used a two-question format to collect race ("White," "Black or African American," "Asian," "Native Hawaiian or other Pacific Islander," "American Indian or Alaska Native"), and ethnicity ("Not Spanish/Hispanic/Latino," "Puerto Rican," "Mexican, Mexican American, Chicano," "Cuban," and "Other Spanish/Hispanic/Latino"). I assumed patient-reported data to be the reference standard for race and ethnicity data collection. I reported the concordance rate between the patient's race and ethnicity information in the EHR and the self-reported information from the HCAHPS survey.

To assess the impact of allowing patients to review their race and ethnicity information directly, I utilized a custom patient portal application on a tablet computer where patients reviewed and corrected their race and ethnicity information obtained from the institution's EHR. As part of a randomized controlled trial, I recruited 65 patients who were admitted through the emergency department and provided them with the tablet computer to review their demographics information. I analyzed and reported descriptive statistics on the number of patients that make modifications to their records as well as the most common changes made.

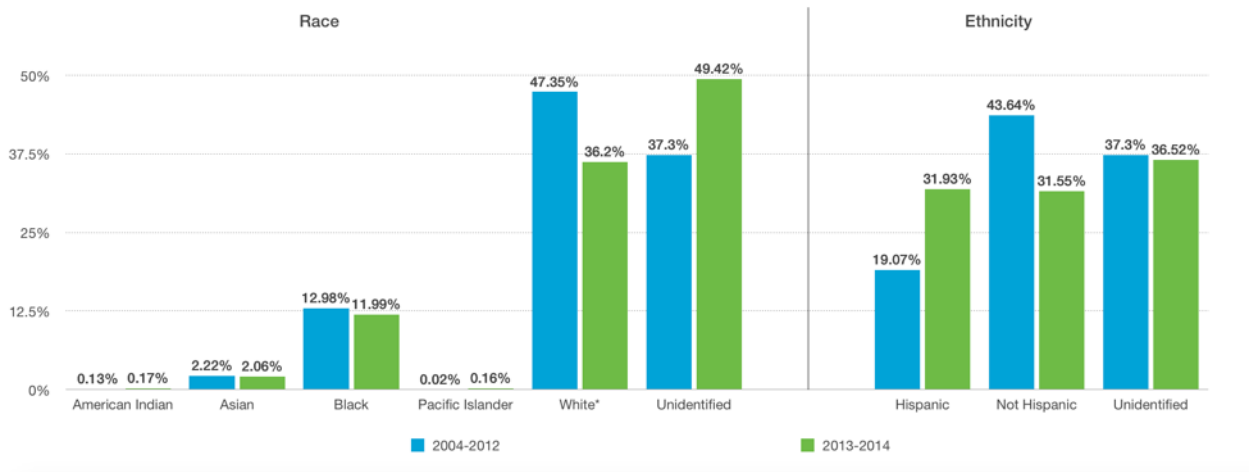

Figure 4.1: Frequency of race and ethnicity categories before and after Meaningful Use attestation. *Patients designated as Hispanic in single question format surveys were assumed to be White-Hispanic.

Results

Policy change - Meaningful Use program

As shown in Figure 4.1, before Meaningful Use was implemented (from 2004–2012), 37.3% of patients did not have race or ethnicity identified in the EHR. After Meaningful Use was implemented, 49.4% of patients did not have an identified race, and 36.5% did not have an identified ethnicity. However, we observed a significant increase in the percentage of patients classified as Hispanic or Latino after Meaningful Use implementation (19.1% to 31.9%, p=2.62e-97).

Patient-facing tools – United States National Databases

There were 165,975,722 combined patient records in the HCUP and Optum databases. Of these, 25.3% and 26.0%, respectively, had uninformative race and ethnicity. There were 71,916 records in the NHANES survey, and only 6.4% contained uninformative race and

ethnicity. Table 4.1 includes a description of the data sources along with the race and ethnicity categories.

Patient-facing tools - Academic Healthcare System in New York City

As reported in Aim 1.1, 2,338,421 patients had at least one visit during the two-year study period and 57.9% of patients did not have race or ethnicity identified in the EHR. During the study period, 25,664 unique patients responded to the HCAHPS survey. Of those, 1,255 patients completed the survey more than once, and 356 (28.4%) had conflicting self-reported race and ethnicity information. After excluding cases with conflicting self-reported race and ethnicity information, 86.3% provided meaningful race or ethnicity data from a total of 25,308 patients. Among these patients, race and ethnicity information from the EHR was available for 25,014 patients.

Among patients with both self-reported and EHR race and ethnicity information, 16,625 (66.5%) patients provided race or ethnicity information that was discordant with data recorded in the EHR. While 6,540 had both race and ethnicity as "Uninformative" in the EHR, the self-reported data provided meaningful race or ethnicity information for 5,533 (84.6%) of these patients, that did not otherwise have meaningful information recorded.

In the randomized trial that assessed patient-reported demographic data entered at the time of hospital admission, 65 patients were recruited. Of those, 35 (53.85%) made changes to their race and/or ethnicity (30 patients edited both, four patients edited ethnicity only, and one patient edited race only). Analysis of all of the "Uninformative" categories for race and ethnicity demonstrated that the majority of patients were willing to provide their information. Among the randomized trial study patients, 32 had "Uninformative" race information

Dataset	Description	Timeframe	Number of patients	Black or African American	White	Hispanic or Latino	Other Race/ Ethnicity	Unknown Race
Observational health databases								
HCUP	Healthcare Cost and Utilization Project (HCUP) is a hospital transactional database that includes inpatient and emergency visits	2000 — 2011	91,983,358	10.75%	52.13%	9.40%	2.42%	25.31%
OPTUM	Health claims database for members of United Healthcare which includes patients enrolled in commercial plans, Medicaid and Legacy Medicare Choice	2000 — 2016	73,992,364	7.45%	53.21%	9.69%	3.64%	26.00%
EHR*	Data from the electronic health record (EHR) of a academic healthcare system in New York City including inpatient, outpatient and emergency department visits	2014 — 2015	2,338,421	6.09%	23.55%	9.51%	7.10%	57.88%
Patient-provided databases								
NHANES	National Health and Nutrition Examination Survey (NHANES) is a database from a national survey that includes both adult and children information	1999 — 2011 (biennially)	71,916	23.82%	38.30%	24.24%	7.24%	6.40%
HCAHPS*	Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) is a database from a survey regarding patient satisfaction sent to patients via U.S. mail after a hospitalization	2014 — 2015	25,308	9.02%	63.85%	5.16%	12.00%	13.71%

* Numbers do not sum to 100% because race and ethnicity were collected separately

Table 4.1: Description of the data sources, including timeframes and race and ethnicity categories.

and 42 had "Uninformative" ethnicity documented in the EHR. After patients reviewed and optionally edited their information, the race of only six (18.75%) patients, and the ethnicity of 13 (30.95%) patients remained "Uninformative." Of the 20 patients who verified or modified their ethnicity to be "Hispanic," the majority did not consider their race to be available in the list of options; 10 patients recorded or retained their race as "Other," and two as "Unknown."

Discussion

Categories defined by the Meaningful Use program for collecting race and ethnicity are based on current standards published by the Office of Management and Budget (OMB) in 1997 (Rao et al. 2011). Based on these standards, self-reporting using two separate questions is the preferred method for collecting data on race and ethnicity (OMB 1997).

When comparing ethnicity data before and after Meaningful Use implementation, we observed a significant increase in the percentage of patients identified as Hispanic. Implementing the two-question format allowed us to better identify our Hispanic population; however, the ethnicity of 36.5% of our patients was still unidentified after Meaningful Use compared to 37.3% before Meaningful Use. In terms of race, 37.3% of patients were labeled "Unidentified" before Meaningful Use versus 49.4% after Meaningful Use. This increase in unidentified patients may be due to the difference in how data was collected in the two time periods. Based on the previous single-question format, collecting only ethnicity data would have identified a patient's race/ethnicity, whereas in the two-question format, both race and ethnicity had to be collected separately.

The appropriate collection of race and ethnicity information is key to recognizing disparities that affect minority populations (Kressin 2015). Furthermore, this information can be used to perform risk assessment both for individuals and populations. Our findings suggest that EHR changes implemented because of Meaningful Use improved the collection of race/ethnicity data for our Hispanic population; however, we still have a considerable number of patients without meaningful information for both race and ethnicity.

The comparison between survey data and data from large observational databases demonstrated that patients are willing to provide their racial information leading to higher data quality of race and ethnicity data in these databases. However, 28.4% of patients that completed the HCAHPS survey more than once reported conflicting data. Most of these cases were patients recording more than one race in one survey and reporting a single race in another survey response or patients that chose to report their ethnicity in one survey but not in another.

When comparing ethnicity data before and after patient review through an inpatient portal, we observed that patients were willing to review their information and make changes when needed. The majority of the patients with uninformative race and ethnicity in the EHR changed these values to more meaningful concepts. Interestingly, our Hispanic patients did not seem to consider themselves to belong to any of the OMB-defined race categories as the majority identified their race as "Other" and often entered "Hispanic," "Latino" or their country of origin in a free-text field.

Taking the findings from HCAHPS survey responses to demographic questions as well as the patient review of race and ethnicity data in their health record, these findings raise questions about the efficacy of the two-question format (i.e., collecting race and ethnicity

data collection as separate fields) that is now widely used, as well as the clarity of the difference between "race" and "ethnicity" for patients. While patient-facing tools do appear to capture race and ethnicity data more effectively than other methods, the current categories might be confusing or insufficient for patients to self-report.

Conclusion

This study found many challenges in the collection of race and ethnicity. Policy change efforts such as the Meaningful Use program resulted in better collection of ethnicity. However, a large proportion of patients remain without race and ethnicity information in the EHR. The use of patient-facing tools can dramatically improve the data quality of this information, potentially improving identification of health care disparities and supporting disease risk assessment. Future work could explore how to determine better race and ethnicity categories that would allow patients to consistently report their racial and ethnic background.

4.2 Aim 2.2 - Development and evaluation of a novel method to extract familial relationships from existing clinical databases using patient-provided emergency contact information

Background

Family history is one of the most important disease risk factors necessary for the implementation of precision medicine in the clinical setting (Aronson and Rehm 2015; Guttmacher, Collins, and Carmona 2004). It is critical for disease risk assessment (Berry et al. 1997; Claus, Risch, and Thompson 1994; Ozanne et al. 2013; Tyrer, Duffy, and Cuzick 2004; Wu and Orlando 2015), and appropriate disease screening and prevention (Murabito et al. 2001; Reid et al. 2009; Staroselsky et al. 2006). While several studies have shown that family history is an important element in deciding clinical care (Berry et al. 1997; Claus, Risch, and Thompson 1994; Ozanne et al. 2013; Saslow et al. 2007; Smith, Cokkinides, and Brawley 2012; Tyrer, Duffy, and Cuzick 2004), several barriers exist in its collection and analysis, as well as in its use for personalized management based on patients' risk assessment. Lack of time to obtain, organize and analyze family history information is perhaps the most critical barrier for the use of family history in clinical encounters (Green 2007; Guttmacher, Collins, and Carmona 2004; Rich et al. 2004; Scheuner et al. 2009; Sussner, Jandorf, and Valdimarsdottir 2011; Wilson et al. 2012a). These challenges ultimately result in data quality issues, particularly incompleteness and incorrectness of family history

data in the EHR. In an attempt to recover this valuable information from existing clinical databases, in this study I used emergency contact information—a type of patient-provided data collected at nearly every hospital as part of the routine admission process—to infer familial relationships. I present a novel algorithm for extracting relationships called Relationship Inference From The Electronic Health Record (RIFTEHR) and use it to infer familial relationships among patients.

Objectives

This study developed and evaluated a method that uses patient-provided data to infer familial relationships. The method used emergency contact information, a type of patient-provided data collected at nearly every hospital as part of hospital registration, to infer familial relationships. The familial relationships were evaluated using both clinical and genetic data.

Research Questions

- *Can routinely collected patient-provided administrative data from the EHR be used to identify familial relationships?*

Methods

The data for this study were obtained from the inpatient EHR used at the hospitals affiliated with three large academic medical centers in New York City: Columbia University Medical Center, Weill Cornell Medical Center, and Mount Sinai Health System. Columbia

University Medical Center and Weill Cornell Medical Center operate together as NewYork-Presbyterian Hospital and herein, I will refer to the hospitals and the data associated with them as Columbia and Weill Cornell, respectively. Similarly, I will refer to Mount Sinai Health System and its data as Mount Sinai. The study was approved by Institutional Review Boards independently at each site.

The standard operating procedures require patients who receive care at one of the three academic medical centers to provide information about an emergency contact. This information included the person's name, address, phone number, and their relationship to the patient (e.g., parent, sibling, friend). Using a method I call "Relationship Inference From The Electronic Health Record" (RIFTEHR), I used the emergency contact information to identify familial relationships in the EHR in cases where the emergency contact person had his or her record generated by an encounter with the healthcare system. Algorithmically, I then inferred additional relationships from the connectedness of the identified individuals. This information was validated against genetic data and a separate module of the EHR which documented the linkage between mothers' and their newborns' medical records.

Deriving familial relationships from emergency contact data

To match the emergency contact to the medical records, the algorithm created for each patient a list of all reported emergency contacts. Then, for each emergency contact, it attempted to identify a medical record by matching first name, last name, primary phone number, and ZIP code. First, I considered all cases with first name and filtered the table that contains all patients' information to identify records that contain the same first name. I then returned the identified records and performed the same comparison with last name, primary

phone number, and ZIP code. Subsequently, I compared the combination of two variables at a time (i.e., first name and last name, first name and primary phone number, first name and ZIP code, etc.). I then performed combinations of three variables and then of all four variables. I only considered it successful when I identified a single patient that matches to the emergency contact information given. I also captured which variables were used in the matching process for each one of the emergency contacts (i.e., first name and last name; first name, last name and phone number, etc.). The output of this algorithm contained a patient's identifier, the relationship between the patient and the matched emergency contact, the emergency contact's identifier, and a list of the variables used to perform the matching process. I used as patient identifiers the Enterprise Master Patient Index (EMPI), when available or the medical record number (MRN). EMPIs are a unique identifier created to refer to multiple MRNs across the healthcare organization. Using EMPIs allowed us to perform better in the matching process since duplicates from patients having more than one MRN were excluded.

Once the matches were identified, as a quality control step, I excluded patients with non-biological relationships (i.e., spouse, friend). Specific relationships were mapped to relationship groups (e.g., the relationship "mother" is mapped to "parent"). I then calculated the age difference between two related patients and excluded parents that were less than 10 years older than their children, children that were less than 10 years younger than their parents, grandparents that were less than 20 years older than their grandchildren, grandchildren that were less than 20 years younger than their grandparent. Since parents and grandparents must be older than their children and grandchildren, I also flipped relationships when the age difference between parent or grandparent and its child or grandchild was negative.

Specifically, the relationship "parent" became "child," and the relationship "grandparent" became "grandchild." The same process was done when the age difference between children and grandchildren was positive. I also excluded every patient that matches to 20 or more distinct emergency contacts since it is unlikely that patients have such a high number of family members as a direct emergency contact. Finally, I generated the opposite relationship for every relationship pair. For example, if I determined that A is a parent of B, the opposite relationship is that B is a child of A.

Using the matches identified, I also inferred additional relationships. The inference process was made based on familial relationship rules. For example, if patient A is the mother of patient B and patient B is the mother of patient C, then by inference I know that A is the grandmother of C and C is the grandchild of A. The rules used to perform these inferences are described in Table 4.2.

Once additional relationships are inferred, I removed ambiguous relationships such as "Parent/Aunt/Uncle" if the same pair contained a unique specific relationship, in this case, either "Parent" or "Aunt/Uncle." The same was done for "Child/Nephew/Niece," "Sibling/Cousin," "Parent/Parent-in-law," "Child/Child-in-law," "Grandaunt/Granduncle/Grandaunt-in-law/Granduncle-in-law," "Grandchild/Grandchild-in-law," "Grandnephew/Grandniece/Grandnephew-in-law/Grandniece-in-law," "Grandparent/Grandparent-in-law," "Great-grandchild/Great-grandchild-in-law," "Great-grandparent/Great-grandparent-in-law," "Nephew/Niece/Nephew-in-law/Niece-in-law," and "Sibling/Sibling-in-law."

To identify families in the datasets, I excluded all non-biological relationships such as spouses and in-laws, as well as ambiguous relationships such as "Parent/Parent-in-law."

Person 1-2	Person 2-3	Person 1-3
Parent	Aunt/Uncle	Grandaunt/Granduncle
Parent	Child	Sibling
Parent	Grandchild	Child/Nephew/Niece
Parent	Grandparent	Great-grandparent
Parent	Nephew/Niece	Cousin
Parent	Parent	Grandparent
Parent	Sibling	Aunt/Uncle
Child	Aunt/Uncle	Sibling/Sibling-in-law
Child	Child	Grandchild
Child	Grandchild	Great-grandchild
Child	Grandparent	Parent/Parent-in-law
Child	Nephew/Niece	Grandchild/Grandchild-in-law
Child	Parent	Spouse
Child	Sibling	Child
Sibling	Aunt/Uncle	Aunt/Uncle
Sibling	Child	Nephew/Niece
Sibling	Grandchild	Grandnephew/Grandniece
Sibling	Grandparent	Grandparent
Sibling	Nephew/Niece	Child/Nephew/Niece
Sibling	Parent	Parent
Sibling	Sibling	Sibling
Aunt/Uncle	Aunt/Uncle	Grandaunt/Granduncle/Grandaunt-in-law/Granduncle-in-law
Aunt/Uncle	Child	Cousin
Aunt/Uncle	Grandchild	First cousin once removed
Aunt/Uncle	Grandparent	Great-grandparent
Aunt/Uncle	Nephew/Niece	Sibling/Cousin
Aunt/Uncle	Parent	Great-grandparent/Great-grandparent-in-law
Aunt/Uncle	Sibling	Parent/Aunt/Uncle
Grandchild	Aunt/Uncle	Child/Child-in-law
Grandchild	Child	Great-grandchild
Grandchild	Grandchild	Great-great-grandchild
Grandchild	Grandparent	Spouse
Grandchild	Nephew/Niece	Great-grandchild/Great-grandchild-in-law
Grandchild	Parent	Child/Child-in-law
Grandchild	Sibling	Grandchild
Grandparent	Aunt/Uncle	Great-grandaunt/Great-granduncle
Grandparent	Child	Parent/Aunt/Uncle
Grandparent	Grandchild	Sibling/Cousin
Grandparent	Grandparent	Great-great-grandparent
Grandparent	Nephew/Niece	First cousin once removed
Grandparent	Parent	Great-grandparent
Grandparent	Sibling	Grandaunt/Granduncle
Nephew/Niece	Aunt/Uncle	Sibling/Sibling-in-law
Nephew/Niece	Child	Grandnephew/Grandniece
Nephew/Niece	Grandchild	Great-grandnephew/Great-grandniece
Nephew/Niece	Grandparent	Parent/Parent-in-law
Nephew/Niece	Nephew/Niece	Grandnephew/Grandniece/Grandnephew-in-law/Grandniece-in-law
Nephew/Niece	Parent	Sibling/Sibling-in-law
Nephew/Niece	Sibling	Nephew/Niece/Nephew-in-law/Niece-in-law

Table 4.2: Relationship inference rules.

Using both provided and inferred relationships, I created a network where each node corresponds to a patient and edges represent familial relationships. To identify different families, I decomposed the network into individual connected components.

To identify twins, I matched siblings that shared the same last name and the same date of birth. I did not have enough information to distinguish between monozygotic and dizygotic twins.

Evaluation of automatically inferred relationships

I used the EHR's mother-baby linkage as the reference standard to evaluate identified maternal relationships. Cases were considered true-positives when maternal relationships identified by RIFTEHR were also present in the EHR's mother-baby linkage table. Cases were considered false-positives cases when maternal relationships identified by our algorithm were discordant with the EHR's mother-baby linkage table. And lastly, false-negative cases occurred when a maternal relationship was captured by the EHR's mother-baby linkage but not by our method. Overall performance was evaluated by calculating sensitivity and positive predictive value (PPV). To assess if matches identified by different variables performed differently, I also computed sensitivity and PPV. I stratified the identified relationships by the number of variables used to match the emergency contact to a patient in a healthcare system (Table 4.3), as well as by the combination of variables (e.g., last name only, first name and last name, etc.) used to perform the match (Table 4.4). Additionally, I used the EHR mother-baby linkage information to infer siblings. I then used these relationships to evaluate siblings identified by RIFTEHR. Similarly to the maternal relationships evaluation, overall sibling performance was evaluated by calculating sensitivity and PPV.

	Columbia			Weill Cornell		
N of Paths	**True Positive**	**False Positive**	**PPV**	**True Positive**	**False Positive**	**PPV**
1	4340	1021	0.8096	2979	391	0.884
2	3911	355	0.9168	4114	95	0.9774
3	2438	55	0.9779	4753	53	0.989
4	2696	89	0.968	2089	63	0.9707
5	3075	16	0.9948	4219	29	0.9932
6	5840	30	0.9949	10170	19	0.9981
7	3892	10	0.9974	4100	12	0.9971
8	3105	13	0.9958	1739	19	0.9892
9	2575	6	0.9977	1451	3	0.9979
10	2460	8	0.9968	1217	5	0.9959
11	857	1	0.9988	532	3	0.9944
12	308	0	1	156	0	1
13	34	0	1	29	0	1
14	12	0	1	6	0	1

Table 4.3: Performance by number of paths.

To further evaluate the familial relationships, I used genetic data to perform analysis for kinship. Genotype data were collected from existing sources for 1,524 individuals. At Columbia, genetic data were available for 302 individuals. Data were collected from three separate sources: the Institute for Genomic Medicine, The Columbia University Medical Center Pathology Department, and the Washington Heights/Inwood Informatics Infrastructure for Comparative Effectiveness Research (WICER) project, using whole exome sequencing, Affymetrix CytoScan HD array, and the Illumina Multi-Ethnic Genotyping Array, respectively. To select single-nucleotide polymorphisms (SNPs) for kinship, minor allele frequency was filtered to greater than 5%, and genotyping rate to 99% using PLINK Purcell:2007dg. Independent SNPs were selected using the sliding window (100 SNPs) linkage disequilibrium approach. This resulted in a total of 24,752 variants from the

	Columbia					Weill Cornell				
Matched Path	**True Positive**	**False Positive**	**False Negative**	**Sensitivity**	**PPV**	**True Positive**	**False Positive**	**False Negative**	**Sensitivity**	**PPV**
first	2772	56	37922	0.0681	0.9802	1585	81	38411	0.0396	0.9514
first,last	23531	438	15289	0.6062	0.9817	27745	303	9579	0.7434	0.9892
first,last,phone	25137	191	14638	0.632	0.9925	29824	302	9071	0.7668	0.99
first,last,phone,zip	22793	169	17164	0.5704	0.9926	24222	318	14640	0.6233	0.987
first,last,zip	27345	687	11349	0.7067	0.9755	28179	588	9180	0.7543	0.9796
first,phone	25718	281	13898	0.6492	0.9892	29919	358	8856	0.7716	0.9882
first,phone,zip	23311	262	16482	0.5858	0.9889	24422	447	14252	0.6315	0.982
first,zip	8073	554	31337	0.2048	0.9358	7677	835	29978	0.2039	0.9019
last	2237	104	38683	0.0547	0.9556	1156	82	39140	0.0287	0.9338
last,phone	12968	920	26167	0.3314	0.9338	6061	551	31221	0.1626	0.9167
last,phone,zip	12062	838	27342	0.3061	0.935	5582	542	32464	0.1467	0.9115
last,zip	5013	440	35327	0.1243	0.9193	3097	690	35540	0.0802	0.8178
phone	1393	936	37659	0.0357	0.5981	988	796	35771	0.0269	0.5538
phone,zip	1914	986	37278	0.0488	0.66	1506	738	36217	0.0399	0.6711

Table 4.4: Performance by matched path.

Institute for Genomic Medicine data, 8,544 SNPS from the WICER data, and 32,938 SNPs from the Pathology Department data. PLINK was then used to calculate identity by descent (IBD) by determining $\hat{\pi}$ results $(P(IBD = 2) + 0.5 * P(IBD = 1)(proportionIBD))$ for each pair of individuals. I considered that the predicted relationship was correct if the blood relationship fraction between the two people was the same as the one expected for the predicted relationship with a margin of error of 20% of the expected blood relationships. For example, for inferred mother-child pairs, two individuals in a pair share 50% (±10%) of their genetic information, then that provides evidence that the predicted relationship is correct. Likewise, for inferred aunt-niece pairs, the two individuals are expected to share 25% (±5%). The performance was evaluated by calculating PPV.

Using the Mount Sinai data, I leveraged genome array data for 24,441 participants recruited to the Bio*Me* Biobank Program of The Charles Bronfman Institute for Personalized Medicine. Genotyped participants had a mean age 55.8 years, and approximately 61% are female. Participants self-identified as: Hispanic/Latino (45%), African American (31%), White/Caucasian (8%), Asian (6%), Mixed ancestry (6%), or Other (11%). To calculate genetic relatedness, I first merged Bio*Me* participants (N) genotyped either on the Illumina OmniExpress HumanCore (N=11,212) or Multi-Ethnic Genotype Array v1.0 (N=10,467) platforms, retaining only the intersection of sites (n) between the two arrays (n=385,531). I subsequently removed palindromic sites (n=7,215 SNPs) and sites with a missingness rate > 1% (n=517) and a MAF < 5% (n=112,537) leaving a total of 112,537 SNPs. Of 21,679 Bio*Me* participants with genotype data, emergency contact information was available for 16,341, and in 1,222 cases both family members with relationship inferred by RIFTEHR were in Bio*Me*. Pairwise genetic relationships were estimated by Identity-by-State anal-

ysis with PLINK1.9 using the *–genome* flag. Inferred relationships from RIFTEHR were compared to pairwise genetic relationships to assess performance metrics using the "caret" package with R version 3.0.3. Pairs of patients with conflicting familial relationships were analyzed based on the closest relationship available. For example, if the same pair has two distinct relationships inferred based on their emergency contact information (e.g., parent and aunt/uncle), I consider the first-degree relationship to be correct (in this case, parent) for evaluation of the relationship against genetic data. Parent-offspring and sibling relationships groups were both expected to share 50% genetic relatedness IBS ($\hat{\pi}$ mean 0.5, s.d. ± 0.1). I could distinguish between these two groups by examining the IBS measures at heterozygous (IBS1) and homozygous (IBS2) sites. Parent-offspring were defined as IBS1 > 0.75 and IBS2 < 0.25 (n=1087 pairs), full-siblings were defined as pairs that shared between 0.35 and 0.65 IBS1, and IBS2 > 0.15 and < 0.5 (n=502), monozygotic twins were defined as individuals sharing > 0.8 IBS2 (n=2). In each RIFTEHR group, I calculated positive predictive values (PPV) based on how many predicted parent-offspring and siblings met these genetic criteria. Grandparental, avuncular and half-siblings are all expected to share 25% genetic relatedness IBS ($\hat{\pi}$ mean 0.25, s.d. ± 0.05). Avuncular relationships involved one sibling and the offspring of the other sibling regardless of sex; therefore, the term avuncular refers to both aunts and uncles.

I could not distinguish these groups any further, so I calculated positive predictive values for each group based on how many total pairwise relationships met these criteria (n=976). I did not calculate PPV for cousins, grand-avuncular, great-grandparental, great-grand-avuncular, first cousin-once-removed relationships as the numbers of predicted relationships per group were low (n≤10). Finally, as negative control, I compared predicted

spousal relationships with low or no evidence of IBS sharing ($\hat{\pi} < 0.05$, < 0.1 IBS1 and < 0.1 IBS2). The Bio*Me* Biobank Program (Institutional Review Board 07–0529) operates under a Mount Sinai Institutional Review Board-approved research protocol. All study participants provided written informed consent.

As a subjective validation of all relationship types, including distant relationships such as great-grandparent, I calculated age difference between all pairs of family relatives and stratified it by relationship type. I compared the identified age differences to what would be expected in a real family structure. For example, great-grandparents should be much older than their great-grandchildren.

Results

In total, 3,550,598 patients provided 6,587,594 emergency contacts at the three medical centers. Of these, I identified the emergency contact as a patient in 2,191,695 cases (825,880 at Columbia, 573,804 at Weill Cornell and 792,011 at Mount Sinai). Of those, 1,902,827 provided 1,588,134 family members as emergency contact (488,932 at Columbia, 297,011 at Weill Cornell, and 802,191, at Mount Sinai; Table 4.5). Using these next-of-kin data, I inferred an additional 2,755,448 relationships at Columbia, 1,237,749 at Weill Cornell and 1,819,581 at Mount Sinai (Figure 4.2). Including inferences, I identified a total of 3,244,380 unique relationships at Columbia, 1,534,760 at Weill Cornell, and 2,621,772 at Mount Sinai. Inferred relationships included first to fourth-degree relatives as well as spouses and in-laws (Tables 4.5 and 4.6). I grouped individuals into families by identifying disconnected subgraphs. I found 223,307 families at Columbia containing 2 to 134 mem-

Variable	Columbia	Weill Cornell	Mount Sinai
N	682,267	437,375	783,185
Relationships	3,244,380	1,534,760	2,621,772
N provided relationships	488,932	297,011	802,191
N inferred relationships	2,755,448	1,237,749	1,819,581
Families	223,307	155,811	187,473
Gender, Female	418,657 (61.36%)	261,482 (59.78%)	449,878 (57.45%)
Age	40.15 (24.81)	39.85 (25.02)	51.44 (23.20)
Race/Ethnicity			
Black or African American	69,506 (10.19%)	30,975 (7.08%)	79,854 (10.20%)
White	123,800 (18.15%)	110,485 (25.26%)	285,559 (36.46%)
Hispanic or Latino	373,552 (54.75%)	52,087 (11.91%)	151,785 (19.38%)
Other	11,438 (1.68%)	26,687 (6.10%)	25,864 (3.30%)
Unknown/Declined to answer	103,971 (15.24%)	217,141 (49.65%)	240,123 (30.66%)
Degree of relationship			
First (i.e. child, parent)	1,388,858	814,650	798,440
Second (e.g. grandchild)	605,922	225,796	243,434
Third (e.g. great-grandparent)	432,262	137,712	136,936
Fourth (e.g. great-great-grandchild)	215,300	61,986	58,500
Other			
None (e.g. spouse, in-laws)	172,158	127,748	571,250
Unknown (e.g. parent/parent-in-law)	429,880	166,868	813,212

Table 4.5: Demographic data of the electronic health records at Columbia University Medical Center, Weill Cornell Medical Center, and Mount Sinai Health System.

bers per family. Similarly, I found 155,883 families at Weill Cornell, with up to 129 members per family and 187,473 families at Mount Sinai, with up to 57 family members. These include 4,271 families with fourth-degree relatives (i.e., families that contain first cousin once removed, great-grandaunt/great-granduncle or great-grandnephew/great-grandniece) at Columbia, 1,045 families at Weill Cornell, and 992 families at Mount Sinai.

Degree of relationship	Relationship	N Columbia	N Weill Cornell	N Mount Sinai
First	Child	482,308	298,136	252,584
	Parent	482,308	298,136	252,584
	Sibling	424,242	218,378	293,272
Second	Aunt/Uncle	185,822	65,410	75,404
	Nephew/Niece	185,822	65,410	75,404
	Grandparent	117,139	47,488	46,313
	Grandchild	117,139	47,488	46,313
Third	Cousin	148,806	37,370	27,994
	Grandaunt/Granduncle	96,675	31,764	36,069
	Grandnephew/Grandniece	96,675	31,764	36,069
	Great-grandchild	45,053	18,407	18,402
	Great-grandparent	45,053	18,407	18,402
Fourth	First cousin once removed	94,404	19,596	19,914
	Great-grandaunt/Great-granduncle	42,594	13,664	12,945
	Great-grandnephew/Great-grandniece	42,594	13,664	12,945
	Great-great-grandchild	17,854	7,531	6,348
	Great-great-grandparent	17,854	7,531	6,348
Other				
None	Child-in-law	0	278	0
	Parent-in-law	0	278	0
	Spouse	172,158	127,192	571,250
Unknown	Aunt/Uncle/Aunt-in-law/Uncle-in-law	13,220	5,234	45,950
	Child/Child-in-law	52,186	24,733	62,804
	Child/Nephew/Niece	31,818	8,078	96,925
	Grandaunt/Granduncle/Grandaunt-in-law/Granduncle-in-law	12,035	4,278	36,242
	Grandchild/Grandchild-in-law	12,876	4,578	32,781
	Grandnephew/Grandniece/Grandnephew-in-law/Grandniece-in-law	12,035	4,278	36,242
	Grandparent/Grandparent-in-law	12,876	4,578	32,781
	Great-grandchild/Great-grandchild-in-law	5,799	2,346	18,343
	Great-grandparent/Great-grandparent-in-law	5,799	2,346	18,343
	Nephew/Niece/Nephew-in-law/Niece-in-law	13,220	5,234	45,950
	Parent/Aunt/Uncle	31,818	8,078	96,925
	Parent/Parent-in-law	52,186	24,733	62,804
	Sibling/Cousin	41,270	9,142	88,956
	Sibling/Sibling-in-law	132,742	59,232	138,166

Table 4.6: Relationships by degree.

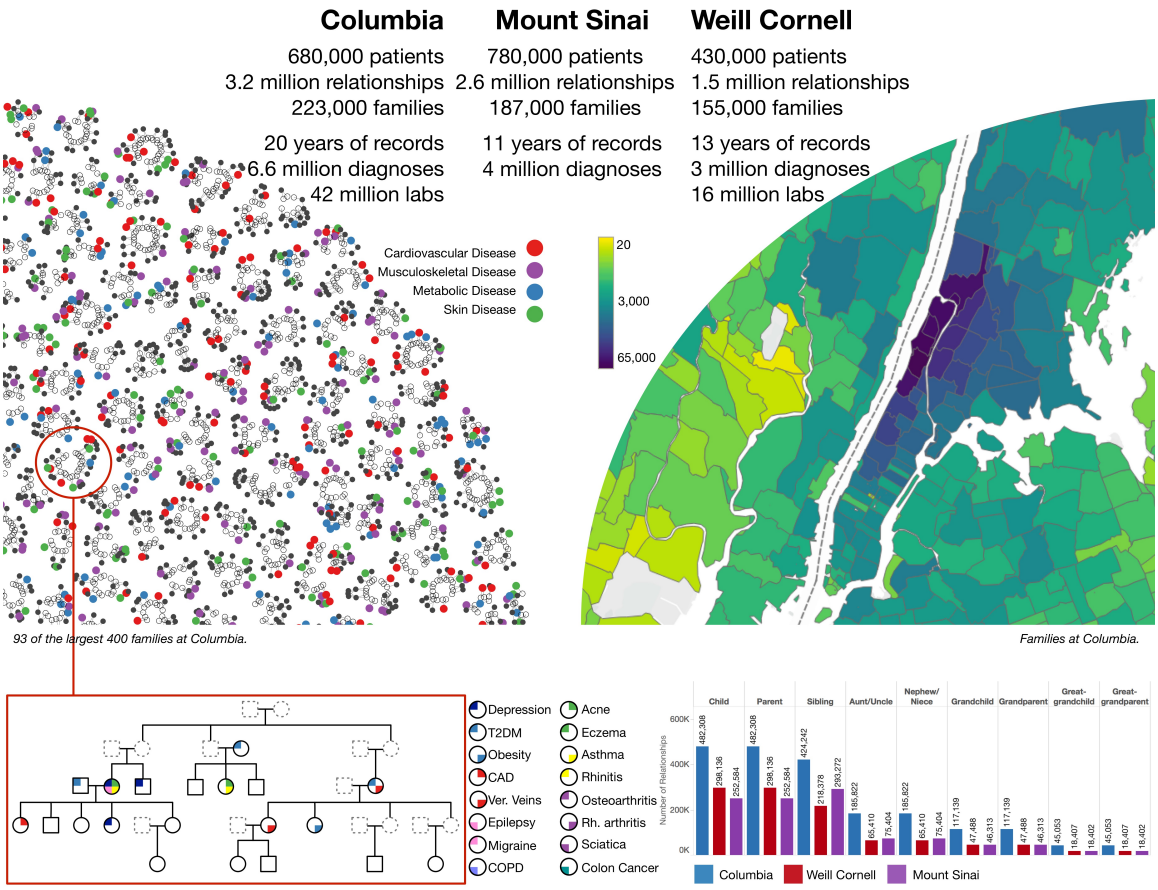

Figure 4.2: 680,000 at Columbia, 430,000 at Weill Cornell, and 780,000 at Mount Sinai reported next-of-kin data that could be identified in the institutional EHR. From these initial relationships, I was able to infer additional relationships resulting in 3.2 million patient relationships at Columbia, 1.5 million relationships at Weill Cornell, and 2.6 million relationships at Mount Sinai. A family was identified as a group of patients with no relationships outside of the group. In total, we identified 223,000 families at Columbia, 155,000 families at Weill Cornell, and 187,000 at Mount Sinai. The largest 400 families from Columbia were visualized as a graph using a force layout (Methods). Each disconnected subgraph is a family. Each node is an individual. Solid nodes represent patients in our respective EHRs. Colored nodes indicate the presence of a disease diagnosis in one of four classes: cardiovascular disease (red), musculoskeletal disease (purple), metabolic disease (blue), and skin disease (green). The top left shows 93 of the top families at Columbia. The largest family shown contains 23 individuals and the smallest, 12. I constructed detailed pedigrees for one family from Columbia (bottom left). The pedigree shown was modified for de-identification purposes. Each node is an individual. Individuals indicated by dashed lines are inferred to exist but did not exist in the EHR. The *top right* shows a map of the number of individuals from Columbia for whom relationships were identified. The colors represent the number of individuals that live in each ZIP code. The *bottom right* bar graph shows the number of individuals by relationship type for each institution. We used all disease diagnosis data and clinical pathology report data (laboratory tests) available for patients in our cohort to study genetic heritability. At Columbia, 6.6 million disease diagnoses were used to estimate heritability of dichotomous traits and 42 million laboratory tests were used to estimate heritability of quantitative traits. At Weill Cornell, 3 million disease diagnoses were used and 16 million laboratory tests and at Mount Sinai, 4 million disease diagnosis.

The relationship between mother and child was explicitly documented in the EHR for newborns delivered at Columbia and Cornell. This 'EHR mother-baby linkage' provided a reference standard for maternal relationships, allowing us to compute sensitivity and positive predictive value (PPV) of the relationship inference method. For maternal relationships, I obtained 92.9% sensitivity with 95.7% PPV at Columbia and 96.8% sensitivity with 98.3% PPV at Weill Cornell. Similarly, for siblings, I obtained 92.2% sensitivity with 98.3% PPV at Columbia and 96.5% sensitivity with 99.6% PPV at Weill Cornell (Figure 4.3A). Tables 4.3 and 4.4 present the stratified performance of the identified relationships by the number of variables used to match the emergency contact to a patient in a healthcare system, and by the combination of variables (e.g., last name only, first name and last name, etc.) used to perform the match, respectively.

I validated the identified relationships by comparison to genetically-derived relatedness (Figure 4.3). I collected data for 1,222 patients from Mount Sinai and 302 patients from Columbia for whom EHR-inferred relationships and available genetic data were consented for reuse. I included spousal relationships as a negative control using a heuristic definition of being genetically unrelated (IBS < 0.1). I estimated relatedness using PLINK (Purcell et al. 2007). At Columbia, almost all 134-predicted parent-offspring relationships had the expected genetic relatedness of 50%, and the three grandparental relationships had the expected relatedness of 25%. All 26 sibling relationships were genetically related, but four were identical twins, and three were half-siblings (Figure 4.3B). At Mount Sinai, the positive predictive value (PPV) to predict spousal relationships was 91%, 80% for parent-offspring, 66% for sibling, and 47% for grandparental and 32% for avuncular relationships (Figure 4.3D). Overall, relationships extracted from the EHR significantly correlate with

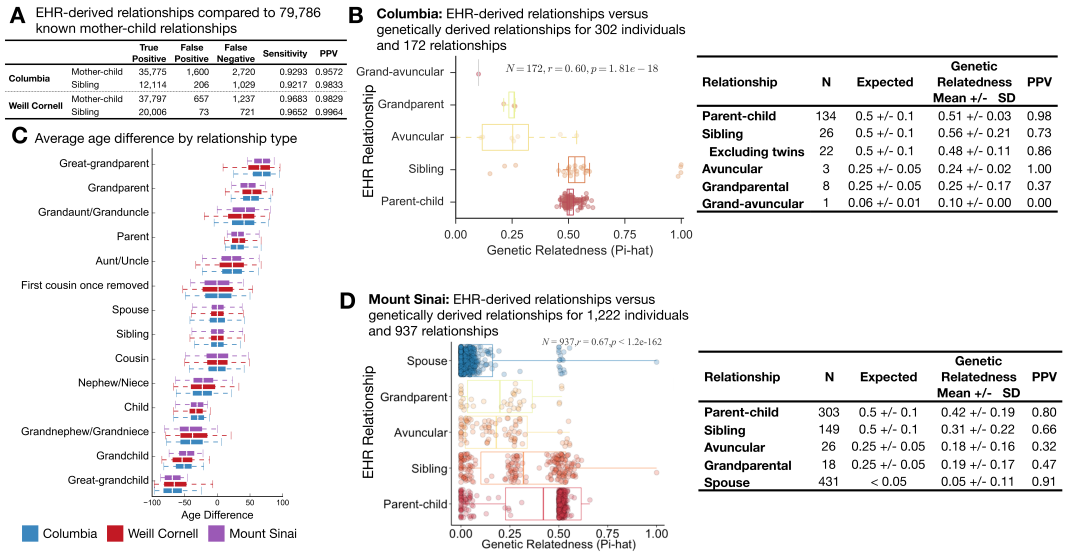

Figure 4.3: Validation of familial relationships inferred from the EHR. (A) The medical centers at both Columbia and Weill Cornell have implemented a link between the electronic health records of mother and baby at the time of birth. I used these links as a gold standard to evaluate RIFTEHR, my algorithm for automatically inferring relationships from the EHR. I also inferred siblings using the mother-baby link data. (B) Through biobanks at Columbia, 302 of the patients with identified relationships from RIFTEHR also had genetic data available and appropriately consented for use in our study. For these, RIFTEHR predicted a total of 172 relationships. Genetic relatedness was determined for each pair of individuals. Almost all 134 parent/child relationships had the expected genetic relatedness of 50% (51%±3%). Of the siblings predicted by RIFTEHR 19 were full siblings, 3 were half siblings (genetic relatedness of 25%), and 4 were identical twins. The high rate of twins in our small sample is a result of the secondary use of existing data – which was originally collected for genetic studies. Excluding these twins yields a more accurate estimate of RIFTEHR's performance (PPV=86.4%). Overall the RIFTEHR relationship and the genetic relationship were significantly correlated (r=0.60, p=1.81e-18). (C) Average age differences for each relationship type. I computed the age differences for each pair of individuals at Columbia (blue), Weill Cornell (red) and Mount Sinai (purple). The age differences are consistent across sites. (D) At Mount Sinai, I identified 1,222 patients that had familial relationships from RIFTEHR and also had genetic data available with appropriate consent for use in our study. Among these, RIFTEHR inferred 937 relationships. Genetic relatedness was determined for each individual pair and compared to the relationships inferred by RIFTEHR. RIFTEHR's performance varied from 32% to 91% PPV, being more accurate in identifying members of the nuclear family. Overall the RIFTEHR relationship and the genetic relationship were significantly correlated (r=0.67, p<1.2e-162).

the expected genetic relatedness (r=0.60, p=1.81e-18 at Columbia and r=0.67, p<1.2e-162at Mount Sinai).

Discussion

Analysis of EHR data has yielded insight into drug effectiveness and allowed precise definition of phenotypes to investigate disease processes (Birkhead, Klompas, and Shah 2015; Boland et al. 2015; Lorberbaum et al. 2016a; Ritchie, Andrade, and Kuivaniemi 2015; Tatonetti et al. 2012; Wei and Denny 2015). For the first time on a large scale, I used EHR data to infer pedigrees from patient-provided emergency contact information. I presented a novel algorithm for performing this relationship extraction, RIFTEHR, validated its performance, and applied it to the medical records of three independent institutions.

Previous research studies have used existing databases to identify twins. In 1987, a Vietnam Era (1964-1975) Twin Registry of American male-male veterans born between 1939 and 1955 was developed to provide a study sample for research evaluating the impact of Vietnam service on the medical and psychosocial aspects of health. Twins were identified using an algorithm which involved matching entries on the database for same last name, different first name, same date of birth, and similar social security number (Eisen, True, and Goldberg 1987). In 2014, researchers used a similar method to identify twins from an EHR database (Mayer et al. 2014). Unlike the methods employed in these studies, RIFTEHR identified familial relationships including distant relatives up to four generations apart, in addition to twins.

The availability of family structures in addition to clinical data has significant implica-

tions for the use of EHR data in clinical and genetic studies. EHRs are in broad use and offer an alternative to traditional phenotyping. Every day, the EHR records information for thousands of patients from drug prescriptions and disease diagnosis to clinical pathology results and physician notes. Use of EHR data presents a novel opportunity to conduct rapid and expansive genetic studies such as of disease and phenotype heritability. In particular, EHR data enables access to traits that otherwise might not be explored. Similarly, the use of EHR data with familial structures allows for large-scale clinical studies, including disease risk assessment and screening. In addition, data captured by these systems represent the diversity of the patient populations they serve, and, in ethnically diverse regions like New York City, make previously unattainable cohorts available for study (Hripcsak et al. 2016). The caveat is that EHR data are known to contain issues regarding missingness and accuracy which limits their use (Hripcsak and Albers 2013; Weiskopf and Weng 2013). Future studies should use robust methods that account for these data quality concerns.

Conclusion

We have described and validated a novel method for identifying familial relationships in medical records and used 7.4 million relationships inferred from the EHRs at three academic medical centers. The availability of family structures in addition to clinical data has significant implications for the use of EHR data in clinical and genetic studies, enables access to clinical information that otherwise might not be explored, and ultimately advancing clinical and genetics research.

4.3 Aim 2.3 - Impact of a federal initiative (Meaningful Use) on collecting patients' smoking status

Background

Smoking remains the number one cause of preventable death in the United States, responsible for more than 480,000 deaths annually (National Center for Chronic Disease Prevention and Health Promotion Office on Smoking and Health 2014). Policy change, such as tobacco control policies, smoke-free legislation, tobacco taxation, and smoking cessation services have been shown to have substantial benefits in children's health (Faber et al. 2017). In addition to these policies, obtaining a patient's smoking status during clinical encounters is a crucial step in beginning smoking cessation interventions and monitoring progress (Caplan, Stout, and Blumenthal 2011). Accurately recording smoking status during a clinical encounter may appear to be a straightforward task; however, this important behavioral determinant of health is often overlooked (Adler and Stead 2015). Given the clinical importance of recording smoking status, the Meaningful Use (MU) financial incentive program for electronic health record (EHR) adoption in the U.S. included a requirement for healthcare providers to capture patients' smoking status electronically in a structured format (Centers for Medicare & Medicaid Services 2010).

Objectives

The purpose of this study was to assess the impact of the Meaning Use program in the data quality of smoking status in a pre-/post- design with data collected over a 10-year

period in an established commercial EHR system at a large academic medical center.

Research Questions

- *How did Meaningful Use impact the quality of smoking status collected in the EHR?*

Methods

I conducted a retrospective study to analyze smoking status data *before* and *after* Meaningful Use criteria were implemented at NewYork-Presbyterian Hospital/Columbia University Medical Center. In our institution, smoking status was collected in clinical notes by several types of providers (e.g., physicians, nurses, social workers). The EHR contained thousands of note templates containing a variable number of observations. An observation could be a free-text box, a Boolean, or a numeric value. As described previously in Aim 1.3, I extracted data from observations, including structured and free-text, whose description contained the stemmed words "smok," "cigar" or "tobacco," and identified the number of times each observation was used. This analysis showed that approximately 94% of patients had at least one smoking status recorded in a structured observation. Given this finding, in this study, I limited this analysis to structured observations.

While our institution was accredited as being compliant with Meaningful Use Stage One criteria in the end of 2012, changes to the note templates were implemented throughout the preceding years. Therefore, patients that had at least one hospital admission between November 2007 and August 2017 were included in the study. I analyzed changes in the documentation pattern of smoking status during the 10-year study period. Prior to

the Meaningful Use program, smoking status was collected as part of clinical notes using locally defined templates, without standardized categories. Categories for smoking status were defined by each group responsible for developing note templates. With the implementation of the Meaningful Use program, eight distinct categories for collecting smoking status were specified: "Current every day smoker," "Current some day smoker," "Former smoker," "Never smoker," "Smoker, current status unknown," "Unknown if ever smoked," "Heavy tobacco smoker," and "Light tobacco smoker" (Centers for Medicare & Medicaid Services 2010). All observations were stored independently from each other, and not transferred to other sections of the EHR.

All smoking status observations, pre- and post-Meaningful Use, were mapped to one of four clinically meaningful categories: 1) "Current smoker," 2) "Former smoker," 3) "Never smoker," and 4) "Unknown smoking status," as described in Table 4.7. Once the categories were mapped, I examined smoking status collected over time for each patient and analyzed whether subsequent updates to smoking status were plausible or implausible. Plausible cases occurred when the change was feasible to happen such as a change from "Never smoker" to "Current smoker"), and implausible occurred when the conflict was not logically possible or in cases where there was a loss of information; for example, a change from "Former smoker" to "Never smoker." Figure 4.4 demonstrates all possible changes in smoking status along with the plausibility of each change.

Additionally, I analyzed the number of discrepancies in smoking status between clinical notes recorded during the same hospital admission for each patient. It is unlikely that patients will have changes to their smoking status during a single hospitalization; therefore, discrepancies in patients' smoking status recorded during a single hospitalization were con-

Clinically Actionable Smoking Status Categories	EHR Documented Categories
Never Smoker	**Never Smoker** Smoker (No) Patient Denies
Current Smoker	**Current every day smoker** **Current some day smoker** **Light smoker** **Heavy Smoker** **Smoker, current status unknown** Smoker (Yes)
Former Smoker	**Former smoker** Ex-smoker Quit / Stopped
Unknown Smoking Status	**Unknown if ever smoked** Unknown Unable to assess N/A / None

Table 4.7: Description of the mapping from smoking status categories as recorded in the EHR to the four clinically actionable categories. Smoking status categories documented in the EHR that utilize the standard criteria defined by the Meaningful Use program are highlighted in bold.

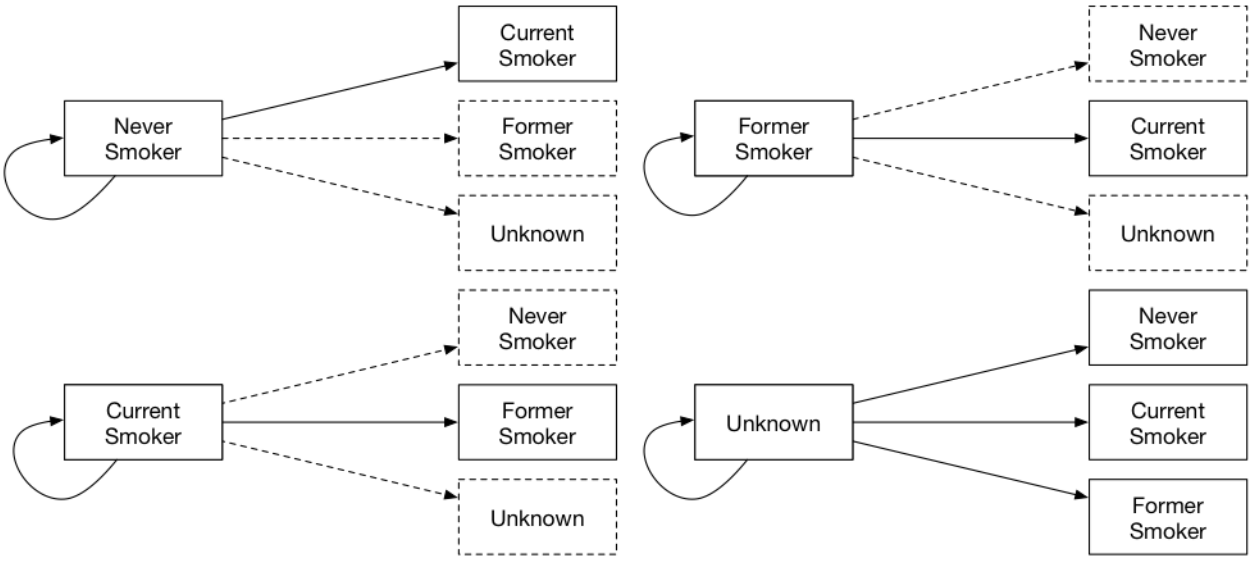

Figure 4.4: Changes of smoking status overtime. Dashed changes demonstrate implausible discrepancies and continuous lines represent plausible changes in longitudinal data..

sidered plausibility issues. I assessed quality of smoking status based on the percentage of patients with consistent and informative smoking status recorded in the EHR (i.e., not classified as "Unknown" in the database, or not conflicting if recorded multiple times). I reported the number of patients with and without smoking status, the number of times smoking status was recorded per visit, the number of different provider types (e.g., nurses, medical doctors, care coordinators, social workers) recording smoking status, the percentage of visits with discrepancies, and the number of plausible and implausible changes per year. To assess the impact of Meaningful Use on the data quality of smoking status, I compared the descriptive statistics described above during the years before and after Meaningful Use criteria were adopted.

Results

I reviewed data from 304,926 patients, who together had 529,236 hospital admissions during the 10-year study period, wherein 858,512 observations of smoking status were recorded. The accompanying Table 4.8 presents the number of patients and visits with more than a single smoking status collected, as well as the average number of times smoking status was collected, the number of provider types that collected smoking status, and the rate of discrepancies and implausible changes. As shown in Figure 4.5, over the 10-year study period, smoking status was documented increasingly frequently and by more provider types (e.g., nurses, medical doctors, care coordinators, social workers). However, the rate of discrepancies increased both at the patient and visit levels from 5% to 40% and 5% to 41%, respectively. Similarly, the rate of implausible changes increased from nearly 2% to

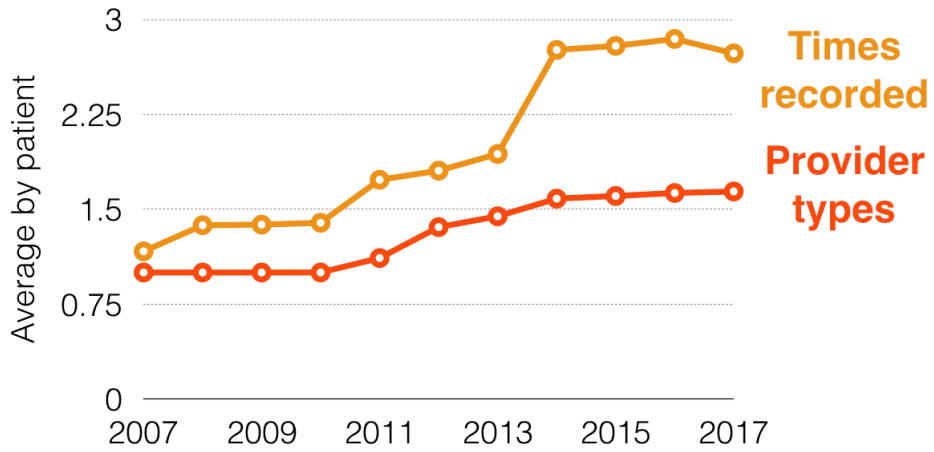

Figure 4.5: Number of times of provider types that collecting smoking status per patient.

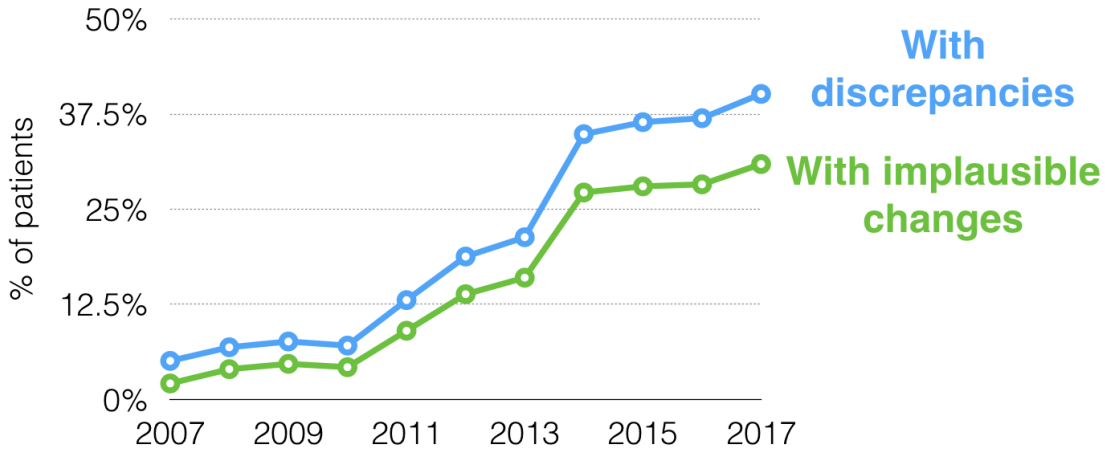

Figure 4.6: Percentage of patients with discrepancies and implausible changes in smoking status documentation.

31% (Figure 4.6).

Discussion

The Meaningful Use program specifies eight distinct categories for collecting smoking status: "Current every day smoker," "Current some day smoker," "Former smoker," "Never smoker," "Smoker, current status unknown," "Unknown if ever smoked," "Heavy tobacco

Year	Patients						Visits					
	With smoking status	With more than 1 smoking	Times recorded (avg)	Provider types (avg)	With discrepancies	With implausible changes	With smoking status	With more than 1 smoking	Times recorded (avg)	Provider types (avg)	With discrepancies	With implausible changes
2007	6,337	0.14	1.17	1.00	0.05	0.02	7,042	0.05	1.05	1.00	0.05	0.02
2008	32,788	0.22	1.37	1.00	0.07	0.04	43,022	0.05	1.05	1.00	0.07	0.04
2009	34,296	0.22	1.38	1.00	0.08	0.05	45,082	0.05	1.05	1.00	0.08	0.04
2010	38,771	0.24	1.39	1.00	0.07	0.04	50,131	0.08	1.08	1.00	0.07	0.04
2011	42,202	0.42	1.73	1.11	0.13	0.09	54,502	0.32	1.34	1.05	0.13	0.08
2012	42,658	0.46	1.81	1.36	0.19	0.14	55,064	0.37	1.40	1.33	0.20	0.13
2013	42,864	0.50	1.94	1.44	0.21	0.16	55,586	0.42	1.49	1.40	0.22	0.16
2014	45,912	0.64	2.76	1.58	0.35	0.27	60,065	0.63	2.11	1.56	0.36	0.28
2015	45,425	0.64	2.79	1.60	0.36	0.28	58,946	0.64	2.15	1.57	0.38	0.29
2016	46,193	0.65	2.85	1.63	0.37	0.28	60,282	0.65	2.18	1.59	0.38	0.29
2017	31,693	0.66	2.73	1.64	0.40	0.31	39,514	0.66	2.19	1.60	0.41	0.32

Table 4.8: Annual number of patients and visits with smoking status recorded, number of times recorded, number of different provider types recording smoking status, and rate of discrepancies and implausible changes.

smoker," and "Light tobacco smoker" (Centers for Medicare & Medicaid Services 2010). While the Meaningful Use program helped to standardize data collection of smoking status, it did not necessarily improve data quality. We observed that the number of times smoking status was collected increased over the years both at the patient level and the visit level. Because the EHR did not provide a central location to store smoking status information, different healthcare providers recorded this information in several different notes, resulting in many inconsistencies across notes.

To improve the data quality of smoking status in EHRs, I recommend that patients' smoking status be stored in a centralized fashion using clinically actionable categories. At our institution, data regarding smoking status was only available as part of clinical notes and therefore, not available in other sections of the EHR, making it challenging to identify this information in the patients' records. If smoking status were available in a centralized location, clinicians could then more easily verify this information in every encounter by asking patients about tobacco use. Future work should focus on identifying ways to overcome discrepant smoking status. To maintain and improve data quality, implausible changes and updates resulting in information loss should require explanation by the user.

Another method to improve data quality of smoking status is to involve patients directly to provide this information. Previous studies on improving patient-reported data demonstrated efficacy in improving data quality by utilizing patients to directly review and update their information using kiosks, portals, or printed forms (Caligtan et al. 2012; Cimino, Patel, and Kushniruk 2001; Collins et al. 2011; Greenhalgh et al. 2008; Halamka, Mandl, and Tang 2008; Hassol et al. 2004; Kaelber et al. 2008; Maher et al. 2015, 2016; Masterson Creber et al. 2016; Nazi et al. 2010; O'Leary et al. 2015; Prey, Restaino, and Vawdrey 2014;

Pyper et al. 2004; Ralston et al. 2007; Reti et al. 2010; Tang and Lee 2009; Wilcox et al. 2016). Eliciting this information via a computer may also mitigate the potential biases introduced by clinicians asking questions regarding smoking behavior. Given that smoking status may have negative connotations for certain patients (Gorber et al. 2009), electronically collected smoking status without direct elicitation from care providers may alleviate some hesitation from patients to provide the truth.

Conclusion

The Meaningful Use program increased data collection of smoking status; however, the quality of the information collected did not improve over time. The rate of inconsistencies and implausible changes in smoking status has risen over the years, challenging the appropriate identification of smokers. Centralized documentation with clinically actionable categories and patient-facing tools might improve the quality of smoking status in EHRs.

Chapter 5

Aim 3 - Use patient-provided data to assess disease risk

Family history is one of the most important disease risk factors necessary to implement precision medicine in the clinical setting (Aronson and Rehm 2015; Guttmacher, Collins, and Carmona 2004). It is frequently collected as part of clinical encounters, and provides information regarding the heritability of disease, along with environmental factors (Tenesa and Haley 2013; Visscher, Hill, and Wray 2008). Yet despite its importance and ubiquity in free-text form, structured family history has rarely been utilized to better understand disease risk or improve care delivery (Chatterjee, Shi, and García-Closas 2016).

Disease heritability has traditionally been determined through in-depth family studies for many reasons. For one, EHR data generally only capture positive disease cases, whereas traditional in-depth studies capture both verified disease positive cases and verified disease negative instances. Furthermore, EHR data may not be considered sufficiently accurate for research studies, and previous studies in this dissertation demonstrated data quality issues in EHR data related to family history. At the same time, EHR data holds many promises that can greatly improve upon in-depth family studies to estimate heritability. By their nature, these studies require substantial resources to carry out, and they are often limited in sample size and, subsequently, their power. A notable exception, and perhaps the largest single study of it's type, used 80,309 monozygotic and 123,382 same-sex dizygotic twins

to conclude that there is significant familial risk for prostate, melanoma, breast, ovary, and uterine cancers (Mucci et al. 2016). Another study conducted a meta-analysis of 2,748 twin studies conducted since 1955 covering 14.5 million subjects (Polderman et al. 2015). Outside of these two studies, family-based studies on disease heritability have involved much smaller sample sizes, often in the tens or hundreds.

Thus the use of EHR data for disease heritability studies presents great potential given the frequently large quantities of data available in the EHR for each patient. EHRs already in broad use offer an alternative to traditional disease phenotyping. Every day, EHRs document information for thousands of patients, from drug prescriptions and disease diagnosis to clinical pathology results and physician notes. Use of EHR data presents a novel opportunity to conduct rapid and expansive studies of disease and phenotype heritability. In particular, it enables access to traits that otherwise might not be explored. In addition, data captured by these systems represent the diversity of the patient populations they serve, and, in ethnically diverse geographies like New York City, make previously unattainable cohorts available for study (Hripcsak et al. 2016). The caveat is that these data are known to contain issues regarding missingness and accuracy which limits their use (Hripcsak and Albers 2013; Weiskopf and Weng 2013). The most critical limitation for genetic studies may be the uncontrolled ascertainment bias (Kaplan, Chambers, and Glasgow 2014). The probability that a particular trait is recorded in the EHR is not uniform across disease conditions or patients. For example, a patient that lives far from a hospital and only visits that hospital to see a specialist will most likely have incomplete records. However, a patient that lives near to a hospital may receive much of her care at that hospital, and thus, the EHR will contain relatively more complete records. Another factor that determines the presence of a trait in

patients' records is the presentation of symptoms. For example, a patient seen for a routine checkup with no symptoms is unlikely to undergo an MRI, regardless of whether he has an unruptured brain aneurysm (Bederson et al. 2000). A recent study used the first release of the UK Biobank data to estimate hundreds of heritabilities from 130,000 patients' genotype and EHR data; however, they did not account for the issues of ascertainment biases (Ge et al. 2017).

In the era of precision medicine, there has been increased focus not just genomics and gene-disease relationships but also on disease prevention and early diagnosis at cohort levels. Early diagnosis and disease prevention are often accomplished by assessing the individual risk for development of certain diseases. Family history is one of the key risk factors that enables disease risk assessment. Current clinical guidelines suggest additional or early disease screening for patients considered at high risk for the development of a variety of conditions, including cancer, cardiovascular, and gastrointestinal conditions. Given the importance of such efforts, the U.S. Preventive Services Task Force (USPSTF) recommends early or additional screening for numerous diseases. However, given the rarity of high-quality structured family history information, there has been limited effort in assessing adherence to clinical guidelines and demonstrating the potential of using EHR data to improve adherence to guidelines. To date, there has been little research on clinician adherence to the recommendation of early screening among high-risk patients (An et al. 2018; Jemal and Fedewa 2017; Solbak et al. 2018). Given that EHRs hold troves of information about diagnostic tests ordered, there is an opportunity to measure clinical guideline adherence in an automated and large-scale way.

In this Aim, I demonstrated the utility of the EHR as a resource for genetics research,

even in the absence of genetic patient data, by using extracted familial data as described in Aim 2 to estimate the heritability of clinical phenotypes, both quantitative and dichotomous. Additionally, I used these familial relationships to assess screening rates among patients considered at high risk due to family history of disease.

5.1 Aim 3.1 - Estimating disease heritability of 500 traits using electronic health records data

Background

While heritability studies have been conducted for numerous diseases (Almgren et al. 2011; Hemani et al. 2013; Lichtenstein et al. 2009; Locatelli et al. 2007; Mucci et al. 2016; Ronald and Hoekstra 2011; Sandin et al. 2014; Sullivan, Daly, and O'Donovan 2012; Sullivan, Kendler, and Neale 2003; Visscher et al. 2007), traditional genetic studies have a number of limitations, including focusing on a single racial and ethnic group. Further, these are prospective studies that take decades to recruit and observe large cohorts, at the cost of hundreds of millions of dollars. I hypothesized that EHR data could be used to overcome some of these limitations. EHRs provide a unique opportunity to increase sample sizes and conduct heritability studies for a much larger array of clinical traits. In the EHR, clinical traits such as diagnosis, procedures, and laboratory tests are collected on a daily basis as part of clinical care. Genetic research based on EHR data can be used to study multiple conditions in a short period, generate new research hypothesis that can later be tested by traditional genetic studies.

Objectives

The purpose of this study was to estimate disease heritability using data available from EHRs. To do so, I used EHR data and familial relationships extracted from EHRs, as described in Aim 2.2, to estimate disease heritability. This study was conducted at three

academic medical centers. Heritability estimates were compared across study sites. I evaluated the findings of this study by comparing heritability estimates computed using EHR data to those published by traditional genetic studies.

Research Questions

- *Can EHR data be used to identify highly heritable diseases in a highly diverse population?*
- *How can we overcome the biases and challenges of EHR data to estimate highly heritable diseases for a diverse population?*

Methods

Based on the familial relationships I identified in Aim 2.2, I computed disease heritability for all traits available in the EHR. The data for this study were obtained from the inpatient EHR used at the hospitals affiliated with three large academic medical centers in New York City: Columbia University Medical Center, Weill Cornell Medical Center, and Mount Sinai Health System. Columbia University Medical Center and Weill Cornell Medical Center operate together as NewYork-Presbyterian Hospital and herein, I will refer to the hospitals and the data associated with them as Columbia and Weill Cornell, respectively. Similarly, I will refer to Mount Sinai Health System and its data as Mount Sinai.

Phenotyping in the EHR

I used diagnostic test results, such as hemoglobin A1c (which is primarily used to measure the three-month average glucose concentration in plasma), as quantitative traits and

diagnosis billing codes (ICD codes) as dichotomous traits. I extracted the most commonly performed laboratory tests and mapped them to LOINC codes so that they could be easily matched between institutions. Each patient may have multiple laboratory reports over time. To extract a single value for each test, I collapsed all reports for each patient into a single value using the mean. This mean reflected the average value for the laboratory result for the patient. For example, I used a patient's mean blood glucose value over their lifetime instead of individual values of blood glucose.

For dichotomous traits, I used any diagnosis billing code that was used for at least 1,000 distinct patients. Any patient with evidence of that billing code in their medical record history was considered a "case." For ICD-9 codes, controls were chosen as any patient that did not have that diagnosis nor any diagnosis that shared an ancestor according to the Clinical Classifications Software (CCS).

CCS was developed by the Agency for Healthcare Research and Quality (AHRQ) and is composed of diagnoses and procedures organized in two related classification systems. In this study, I used the diagnoses classifications. The single-level system consists of 285 mutually-exclusive diagnosis categories. It enables researchers to map any of the 3,824 ICD-9-CM diagnosis codes into one of the 285 CCS categories.

CCS also has a multi-level system composed of 4 levels representing a hierarchy of the 285 categories. The first level is broken into 18 categories. To define a control group, I linked the ICD-9 codes associated with a phenotype of interest to their corresponding CCS categories using the top-level hierarchical categories. I also generated a table associating each patient to CCS categories from their diagnosis. Once this mapping was done, each phenotype was associated with one or more distinct CCS categories. I matched the CCS

categories in the multi-level system to identify the first-level parent category. I considered these top-level categories as our exclusion criteria since the control cohort for this phenotype should have no mention of any CCS under these categories in its medical records. For example, the controls for atrial fibrillation would exclude patients with cardiovascular diseases.

For conditions recorded using ICD-10 codes, I used the hierarchy from ICD-10 to identify patients for the control group. Patients that did not have the same ICD-10 code as diagnosis nor any diagnosis that shared an ancestor code were considered controls.

I curated a set of 85 phenotypes to use for training and testing the heritability algorithm. For these 85 phenotypes, I grouped closely related diagnoses codes together to increase the total number of patients (Table 5.1).

Table 5.1: Eighty-five curated phenotypes.

Phenotype	ICD9 Codes	Modifier
Acne	706.0, 706.1	
Alcoholism	303	
Alzheimer's disease	331	
Androgenic alopecia (females)	704.00, 704.01, 704.02, 704.09	
Anorexia nervosa	307.1	
Asthma	493	
Attention deficit hyperactivity disorder	314	
Autism	299	
Bipolar disorder	296.0, 296.4, 296.5, 296.6, 296.7, 296.80, 296.89	
Bladder cancer	188	
Breast cancer	174	
Bulimia nervosa	307.51	
Cancer endocrine glands	194	
Cancer Nervous system	192, 200.50	
Cancer Nervous system age >15	192, 200.50	Age=>15
Celiac disease	579	
Cervical cancer	180	
Cervix in situ cancer	180	
Chronic obstructive pulmonary disease	496	
Colon cancer	153	
Colorectum cancer	153, 154	
Coronary artery disease	414.0, 414.2	
Coronary calcification	414.4	
Corpus uteri cancer	182	
Crohn's disease	555.0, 555.1, 555.2, 555.9	
Depression	311, 296.2, 296.3	
Discoid lupus erythematosus	695.4	
Ectatic coronary lesions	447.8	
Eczema (adults)	691, 692	
Endometrial cancer	182	
Epilepsy	345	
Gallstone disease	574	
Glaucoma	365	
Graves' disease	242	
Hangover (men)	305	Sex=M
Hangover (women)	305	Sex=F
Head and neck cancer	195	
Heart disease	410-414, 420-429	
Hypertension	401-405	
Insomnia (current)	307.41	
Insomnia (lifetime)	307.42	
Irritable bowel syndrome (females)	555.0, 555.1, 555.2, 555.9, 556	Gender=F
Leukemia	208	
Leukemia age >15	208	Age=>15
Lung cancer	162	
Melanoma	172	
Migraine	346	
Nicotine dependence	305.1	
Non-Hodgkin lymphoma	202	
Obesity	278	
Osteoarthritis (Distal interphalangeal joint - DIP)	715.9	
Osteoarthritis (hip)	715.15	

(continued)

Phenotype	ICD9 Codes	Modifier
Osteoarthritis (knee and hip)	715.15, 715.16	
Osteoarthritis (knee)	715.16	
Ovarian cancer	183	
Pain	338	
Pancreas cancer	157	
Parkinson's disease	332	
Periodontitis	523	
Polycystic ovary syndrome	256.4	
Prostate cancer	185	
Psoriasis	696	
Rectal and anal cancer	154	
Rectum Cancer	154	
Renal cancer	189	
Rheumatoid arthritis	714	
Rhinitis (children)	477	
Rosacea	695.3	
Schizophrenia	295	
Sciatica	724.3	
Skin cancer nonmelanoma	173	
Stomach cancer	151	
Stroke	430, 431, 434, 436	
Systemic lupus erythematosus	710	
Systemic lupus erythematosus (first-degree relative)	710	Degree=1
Systemic lupus erythematosus (second-degree relative)	710	Degree=2
Systemic lupus erythematosus (third-degree relative)	710	Degree=3
Testicular cancer	186	
Thyroid cancer	193	
Tooth loss	525.1	
Type-1 diabetes	250.X1, 250.X3	
Type-2 diabetes	250.X0, 250.X2	
Ulcerative colitis	556	
Uterine cancer	182	
Varicose veins	454, 456	

Estimation of heritability from the Electronic Health Records

The most significant challenge when using traits defined from an observational resource, like the EHR, is the lack of ascertainment. In a heritability study, the phenotype of each study participant is, ideally, carefully evaluated and quantified. This is infeasible, however, when the cohort contains millions of patients with thousands of phenotypes. The differential probability that a given individual will be phenotyped for a study trait is the *ascertainment bias*. The bias may depend on many latent factors, including the trait being studied, the trait status of relatives, the degree to which an individual's healthcare data is contained in the EHR (which is influenced, among other factors, by geographic proximity to the hospital), and an individual's ethnicity and cultural identification. The consequence of this uncontrolled ascertainment bias is that heritability estimates will be highly dependent on the particular individuals in the study cohort. I hypothesized that repeated subsampling would be robust to biases introduced by extremely different ascertainment between families. I define the observational heritability, or h_o^2, as the average of the statistically significant sample estimates (using median). For a given trait, the procedure, which I call SOLAR*Strap*, involves sampling families, running SOLAR (Almasy and Blangero 1998) to estimate sample heritability, and rejecting or accepting the estimate based on a set of quality control criteria. Each step is detailed below.

SOLARStrap Protocol

To compute disease heritability using EHR data, I built pedigree files using the data from each one of the study sites. When building pedigree files, of the 223,307 families at

Columbia, there were 6,894 that contained conflicting relationships – where two individuals were inferred to have two different relationships. At Weill Cornell, 3,258 families out of 155,811 contained conflicts, and at Mount Sinai 25,438 families out of 187,473. These families were excluded from the heritability studies. In some cases, more than one mother or father is annotated for an individual. This could be because of duplicate patient records or errors in the EHR relationship extraction. I resolved these issues by choosing the mother or father that has more relationships in the family. The other relationship is discarded. I then constructed a master pedigree file for each site. To construct this pedigree file, I iterated through each member of each family. For each individual, I either know the mother and father from the EHR-derived relationships or not. If not known, then a new identifier was created to represent the parent. At this point, I iterated through all other family members and recorded the relationships between the new individual and each family member. I repeated this process until the entire pedigree file was filled, thus creating a master file. The master pedigree files contained 1,404,671 individuals at Columbia, 949,440 at Weill Cornell, and 863,340 at Mount Sinai.

To computer heritability estimates for each trait, I sampled an empirically-defined proportion of the available families. The number of families that are sampled combined with the prevalence of the trait defines the power of the heritability analysis. A smaller heritability can be detected with larger sample sizes. As the sample size increases towards the total number of available families, the variance in heritability will decrease, but the estimate will be less robust to bias (Figure 5.1). This is because I sampled without replacement. Based on my simulation studies, I used sample sizes of 15% and 20% of the total number of families with at least one case. I then assessed the quality of the computed estimates.

SOLAR does not converge on a solution for heritability for all samples. Errors in the pedigree or in the ascertainment of phenotypes are the most likely causes for these failures. As part of the quality control measures, I rejected any runs of SOLAR that result in no solution for the heritability. I then considered two additional criteria that must be met for a solution to be considered legitimate: edge epsilon and noise epsilon. Edge epsilon (ϵ_e) is a threshold that determines if the estimate is sufficiently close to 1 or 0. Any estimate within ϵ_e of 1 or 0 was rejected. Noise epsilon (ϵ_n) is a threshold that determines if an estimate has implausibly low error. Any estimate with implausibly low error was rejected (h^2 error is less than ϵ_n of the h^2 estimate). These hyperparameters were set using simulated heritability data. After filtering the SOLAR solutions for these criteria, I defined an additional quality control metric called the Proportion Of Significant Attempts, or POSA. POSA is defined as the number of solutions with a p value less than (α_{POSA}) divided by the total number of converged solutions (or attempts). The POSA is important because it is closely related to the power of the analysis. A fully powered analysis will have a POSA of 1, meaning that all converged estimates are statistically significant. A POSA of 0.5 means that only half of the converged estimates are statistically significant. When the families were sampled, the observed heritability was large enough to be detected with p < α_{POSA} half of the time. Or, in other words, the study was powered to detect a heritability in 50% of samplings. I demonstrated that the higher the POSA, the more accurate the heritability estimates are (Figure 5.1I). I chose a minimum POSA score, $POSA_{lower}$ and the α_{POSA} using simulations.

For those estimates that did not pass the defined quality control criteria when sampling 15% and 20% of the total number of families, I increased the number of families sampled

to 45%. The maximum sample size was defined by the limitations of SOLAR, which could only handle a maximum of 32,000 individuals per pedigree file. For each sample size, I performed 200 samplings. For each of these, I built a custom pedigree and phenotype files and ran SOLAR to estimate the heritability. I then aggregated the results and reported the median heritability with the 95% confidence interval.

For each sampling, a set of *N* families was selected. To construct the sample pedigree file, I identified all rows from the master pedigree files that corresponded to these families and created a new file from this subset.

Once the pedigree file was created, I iterated over every individual in the pedigree and used the reference trait data and demographic data to enter the phenotype status and age of the patient. If no phenotype data were available for the individual, I enter it as missing. For dichotomous traits, the trait values were either 0 (absence), 1 (presence), or *missing* and a "proband" was randomly assigned by selecting a single individual from each family that has the trait. For quantitative traits, I entered the quantitative value or missing.

I used SOLAR to estimate both quantitative and dichotomous trait heritability using a pre-defined mixed linear model. In both cases, sex and age were modeled as covariates. After the pedigree and phenotype files were loaded, the heritability of each trait was estimated with the 'polygenic –screen' command. I used the 'tdist' command in SOLAR to adjust quantitative traits that were not normally distributed. For dichotomous traits, one "proband" was chosen at random for each family. SOLAR automatically detected the presence of a dichotomous trait and converted the estimate from the observed scale to the liability scale. The heritability estimate, error on the heritability estimate, and the p-value were saved from each run for later analysis and aggregation. To investigate the relative contri-

bution of the environment to the studied phenotype, I used SOLAR to compute household effects. For this analysis, I assigned the mother ID as the household ID.

For each sampling that passed the quality control criteria previously described and met the minimum POSA score, I computed the h_o^2 as the median. The median h_o^2 corresponds to a single run of SOLAR that has passed all quality control filters. I used the 95% confidence interval as the error of the h_o^2. I found that this error is closely related to the standard error reported by SOLAR (Figure 5.1).

Preparation of data for analysis on external computing clusters

Due to the high number of heritability estimates that need to be computed, external computing resources from The Open Science Grid (OSG) and Amazon Web Services (AWS) were used. The Open Science Grid (OSG) is a massive computing resource funded by the Department of Energy and the National Science Foundation. The OSG is comprised of over 100 individual sites throughout the United States, primarily located at universities and national laboratories. The sites contain anywhere from hundreds to tens of thousands of CPU cores available for scientific research Pordes:2007ho, Sfiligoi:2009gp. AWS is used to supplement this resource, which makes available on-demand compute instances with high-performance capacity. Per institutional requirements, no protected health information or personally identifying information can be transferred to systems outside of our institutional networks. To leverage these resources for our computing task, I prepared a data subset according to the Safe Harbor guidance provided by the U.S. Department of Health and Human Services (https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html).

The following is a description of how I processed the data for Safe Harbor for each of the 18 identifiers here enumerated from (A) through (R): (A) I removed first, middle, and last names for all patients, (B) all patient address information was removed, (C) all dates were removed and all ages over 89 were coded as "90", (D) telephone numbers and (E) fax numbers were removed, (F) there were no email addresses in the subset of the clinical data, (G) there were no social security numbers in the subset of the clinical data, (H) medical record numbers were mapped to a 10 digit random number and the mapping was stored on a limited access PHI-certified server within the institutional firewall and will never be made available, (I) there were no health plan beneficiary numbers in the data subset, (J) there were no account numbers in the data subset, (K) there were no certificate or license numbers, (L) there were no vehicle numbers or serial numbers in the data subset, (M) there were no device identifiers or serial numbers, (N) there were no URLs in the data subset, (O) there were no IP addresses in the data subset, (P) there were no biometric identifiers in the data subset, (Q) there were no full-face or comparable images in our data subset, (R) there were no other uniquely identifying characteristics or numbers. All data were transferred using secure file transfer protocols using encryption and were destroyed immediately after retrieval of the results.

Validation of accuracy and robustness of SOLARStrap using Simulated Traits

To validate the accuracy and robustness of SOLAR*Strap*, I constructed a set of 4,195 families containing 14,690 individuals chosen from the families extracted from the EHR using RIFTEHR. Relationships and pedigree structures are heterogeneous across these families. I used the 'simqtl' command from SOLAR to simulate quantitative traits with heri-

tability values of 5% to 95% at 5% intervals for this pedigree. Traits were simulated for 19 different heritability values in total. To generate binary traits, a threshold for the quantitative value was chosen for each of the 19 simulations so that the prevalence of the dichotomous, or binary trait, was 15%. I used the prevalence of 15% for dichotomous traits because overall, the average prevalence of disease among patients with familial relationships was 15.9% (min-max: 8% – 37%). The result of each simulation was a phenotype file containing the family id, the individual id, and the quantitative or binary trait value.

I evaluated the quantitative and dichotomous simulated traits by running SOLAR using the simulated phenotype files for each of the 19 different values for heritability. I summarized performance using the r^2 and ran a test of significance.

I then created trait files for SOLAR*Strap*. SOLAR*Strap* is designed to use trait files that are similar to the phenotype files used by SOLAR but can contain more than one type of trait per file and more than 32,000 individuals (SOLAR's limit). I used a python script to combine the 19 heritability estimates into a single trait file.

To evaluate the accuracy of SOLAR*Strap* on quantitative traits, I ran SOLAR*Strap* on each of the 19 simulated datasets. I repeated these runs using a different sampling size (argument nfam in SOLAR*Strap*) between 100 and 700 increasing by 100. I selected the largest sample size (nfam=700) and evaluated the accuracy of SOLAR*Strap* using r-squared and tested significance using regression analysis.

When working with dichotomous traits, there are two scenarios that had to be considered to evaluate the accuracy of SOLAR*Strap*. Either 1) the cases and controls are equally known, meaning that each individual in the pedigree can be assigned to either being a case or control, or 2) the cases are higher confidence than the controls. This latter case more

closely resembles the scenario present in the electronic health records. Documentation of a disease in the EHR can be very indicative of the patient having the disease, but the absence of this documentation does not mean the patient does not have the disease. I evaluated the accuracy of SOLAR*Strap* in both cases. For the former, I included all individuals in the pedigree, and for the latter, I excluded any families where there were no cases. In the pedigrees where the cases are higher confidence than controls, I assigned a proband so that the estimate of heritability is not biased. This was accomplished by randomly selected a single individual in each family as the "proband."

To evaluate the robustness of SOLAR and SOLAR*Strap* to missing data, I chose a single simulated trait (h^2 = 50%) and randomly changed individual phenotypes to unknown. I evaluated removing 5% to 60% of the phenotype data at 5% intervals.

To evaluate the robustness of SOLAR and SOLAR*Strap* to biases, specifically non-random missingness, pedigrees were removed from the heritability estimation with a probability determined by a beta distribution. The beta distribution is a continuous probability distribution bounded by 0 and 1 and parameterized alpha and beta. Each family can be assigned a probability by sampling this distribution. Most families will have the same probability of missing data with a small number of families have a much lower probability. By varying the beta and alpha parameters I can change the proportion of families with a much lower probability of missing data. I varied the value of the beta parameter from 0.001, 0.01, 0.1, 1.0, 10.0, to 100.0 and I set the alpha parameter such that the average probability of missingness across all families was constant at 50%.

Using the simulation results, I evaluated the effect of increasing the sample size (or the number of families being sampled in each iteration when running SOLAR*Strap*). I

hypothesized that as the number of families approaches the number of available families the heritability estimate of SOLAR*Strap* would converge to the heritability estimate of SOLAR. I expected that the number of families sampled would not have an effect on the heritability estimate produced by SOLAR or SOLAR*Strap*. I evaluated this relationship using linear regression of the simulation results. One of the primary quality control metrics for SOLAR*Strap* is the Proportion of Significant Attempts (or POSA). I evaluated the relationship between the POSA score (which ranges from 0 to 1) and the accuracy of the heritability estimates produced.

Computational and statistical software

Statistical analysis, data preparation, and figure creation were performed using Python 2.7. Relationship inferences were implemented in Julia 0.4.3. All correlations were reported as Pearson correlation coefficients unless otherwise noted.

Literature review

For validation purposes, I compared the heritability estimates from this to the ones reported in the most recent meta-analysis of twin correlations and heritability (MaTCH) (Polderman et al. 2015). Using the ICD-10 hierarchy, I grouped our ICD codes to match the main chapters and subchapters reported in the MaTCH database. Since the meta-analysis grouped all traits into higher-level traits, losing a lot of granularity, I also performed a literature review on heritability estimates on 128 traits. I started by analyzing studies that were included in the table available at http://www.snpedia.com/index.php/Heritability (accessed on March 2016). In total, I reviewed heritability estimates with confidence intervals from

		UK Biobank		MaTCH dataset		Observational heritability			
Code	Trait	h^2	SE	h^2	SE	site	model	h_o^2	h_o^2 SE
VI	Diseases of the nervous system	0.0246	0.0216	0.5221	0.0302	Weill Cornell	AE	0.1505	0.0431
X	Diseases of the respiratory system	0.0506	0.0191	0.6215	0.0385	Mount Sinai	AE	0.1556	0.0422
X	Diseases of the respiratory system	0.0506	0.0191	0.6215	0.0385	Weill Cornell	AE	0.3111	0.0592
XI	Diseases of the digestive system	0.0354	0.0092	0.4390	0.0193	Weill Cornell	AE	0.3098	0.0345
XII	Diseases of the skin and subcutaneous tissue	0.0204	0.0180	0.7877	0.0204	Weill Cornell	AE	0.2276	0.1193

Table 5.2: Comparison of heritability estimates from the UK Biobank, the MaTCH database and observational heritability.

61 published reports.

Additionally, I compared our heritability estimates to those reported using the UK Biobank dataset (Ge et al. 2017). I used the estimates reported with ICD 10 codes to match the heritability estimates reported by Ge et al. to our estimates. Overall, I observed that the estimates from the UK Biobank were significantly lower than those computed using EHR data (Figure 5.2). I also compared the heritability estimates from this set of traits to the MaTCH database. Table 5.2 contained the traits along with heritability estimates from the UK Biobank, the MaTCH database, and our estimates using EHR data.

Results

To validate the accuracy and robustness of SOLAR*Strap*, I used simulations of quantitative and dichotomous traits with heritability ranging from 5-95%. SOLAR was precise in estimating the heritability of both quantitative (r^2 = 0.999) and dichotomous (r^2 = 0.994) traits (Figure 5.1A). I ran SOLAR*Strap* in the simulated quantitative traits, and it accurately estimated the heritabilities regardless of the sampling size (Figure 5.1B, r^2 = 0.986, p = 3.22e-15). For dichotomous traits, I ran SOLAR*Strap* in two scenarios: 1) including all

families regardless of the number of cases in the family and 2) including only families with at least one case. In the latter scenario, I randomly chose one of the cases in each family to be the proband. SOLAR*Strap* accurately recapitulated the heritability estimates regardless of the number of families sampled in both cases, with lower accuracy when a proband was assigned than the complete ascertainment (r^2 = 0.988, p = 7.57e-15 without proband and r^2 = 0.930, p = 2.85e-11 with proband; Figure 5.1C and 5.1D). I found that both SOLAR and SOLAR*Strap* produced accurate estimates given complete data and in the presence of random missingness (Figure 5.1E). However, SOLAR*Strap* produced more accurate estimates in the presence of ascertainment biases that vary from family to family (Figure 5.1F).

As expected, SOLAR*Strap* produced estimates with larger confidence intervals than SOLAR. SOLAR*Strap* becomes more sensitive to bias as the number of families sampled increased towards the total number of families available (Figure 5.1G); however, the estimate of heritability is not dependent on the number of families sampled (Figure 5.1H, r=0.02, p=4.1e-8). I used the Proportion of Significant Attempts (POSA) as a quality score for the heritability estimates generated by SOLAR*Strap*. A higher POSA score represents a more accurate heritability estimate from SOLAR*Strap* (Figure 5.1I). I injected noise into the data by randomly shuffling a subset of the patient diagnoses, simulating misclassification (misdiagnosis or missed diagnosis) in the medical records. Injection of 5% noise reduced the estimate 13% (from h_o^2 =0.77 to h_o^2 =0.67) and 10% noise reduced the estimate 30% (from h_o^2 =0.77 to h_o^2 =0.53, Figure 5.1J). Misclassification was one explanation of lower than expected estimates compared to a carefully ascertained study.

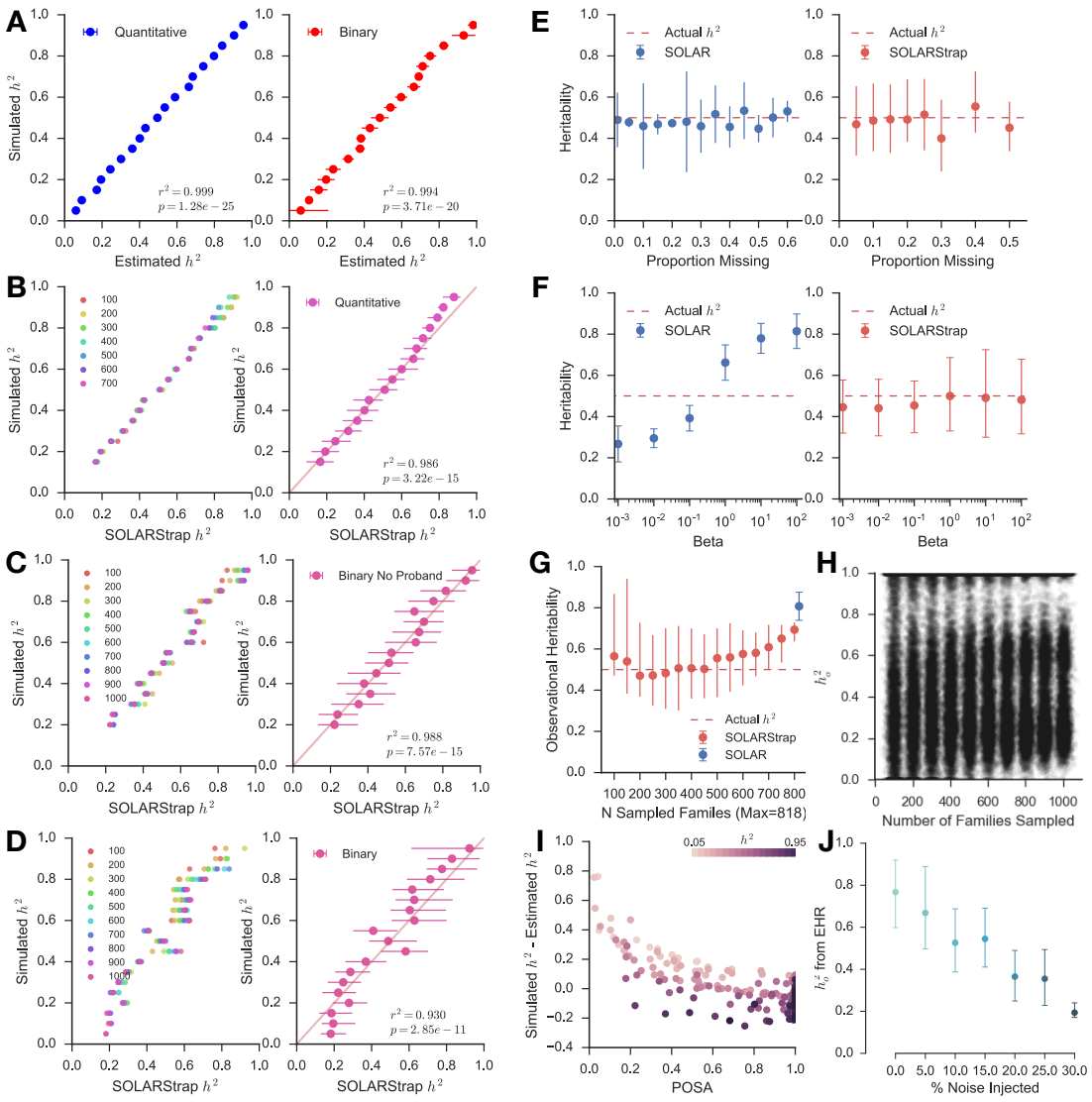

Figure 5.1: Validation of SOLAR*Strap* accuracy and robustness using simulated data. (A) Traits with heritability ranging from 5% to 95% were generated using the SOLAR. We used actual family structures extracted from the EHR by RIFTEHR to generate the simulated traits. We then created dichotomous (binary) versions of the trait by choosing a threshold that would yield a trait with 15% prevalence. SOLAR was very accurate at recapitulating the correct heritability for both quantitative (r^2 = 0.999) and binary (r^2 = 0.994) traits. In (B), (C) and (D), the number of families varied from 100 to 1000, being represented by different colors. (B) SOLAR*Strap* was run on each of the simulated quantitative traits and was accurate at estimating the true heritability (r^2 = 0.986). SOLAR*Strap* was accurate regardless of the number of families that was used in the sampling procedure (left). (C) SOLAR*Strap* was run on each of the binary traits in the setting of complete ascertainment. SOLAR*Strap* achieved equal accuracy as in the quantitative case (r^2 = 0.988). (D) SOLAR*Strap* was run on each of the binary traits in the setting of incomplete ascertainment. In this case, families without any cases were dropped and a proband was randomly assigned in each family. The accuracy is lower than the case of complete ascertainment (r^2 = 0.930). (E) In the presence of randomly missing information, both SOLAR and SOLAR*Strap* produce accurate estimates of the true heritability

even when up to 60% of the data are removed. However, in four cases where the proportion removed was 35%, 45%, and above 50% SOLAR*Strap* estimates did not pass our internal quality control criteria. (F) SOLAR is sensitive to this bias and produces inaccurate results as the strength of the bias increases. SOLAR*Strap* is robust to these biases and produces accurate estimates of heritability even in the most extreme case of bias. (G) As the number of families sampled increases toward the total number of available families SOLAR*Strap* becomes more sensitive to bias – in the most extreme case where the number of sampled families is equal to the total number of available families SOLAR*Strap* reduces to simply running SOLAR. (H) The estimate of heritability is not dependent on the number of families sampled (r=0.02, p=4.1e-8). (I) The Proportion of Significant Attempts (POSA) is a primary estimate of quality for heritability estimates produced by SOLAR*Strap*. The accuracy of SOLAR*Strap* increases as the POSA increases (shown as error here). (J) The effect of noise injection on the estimate of observational heritability of rhinitis. We injected noise into the data by randomly shuffling a subset of the patient diagnoses. This simulates misclassification (misdiagnosis or missed diagnosis) in the medical records. When no noise is injected the estimate is 0.77 (0.60-0.92). As noise is introduced the estimate of the heritability decreases to 0.36 (0.23-0.49) once one quarter of the data are randomized.

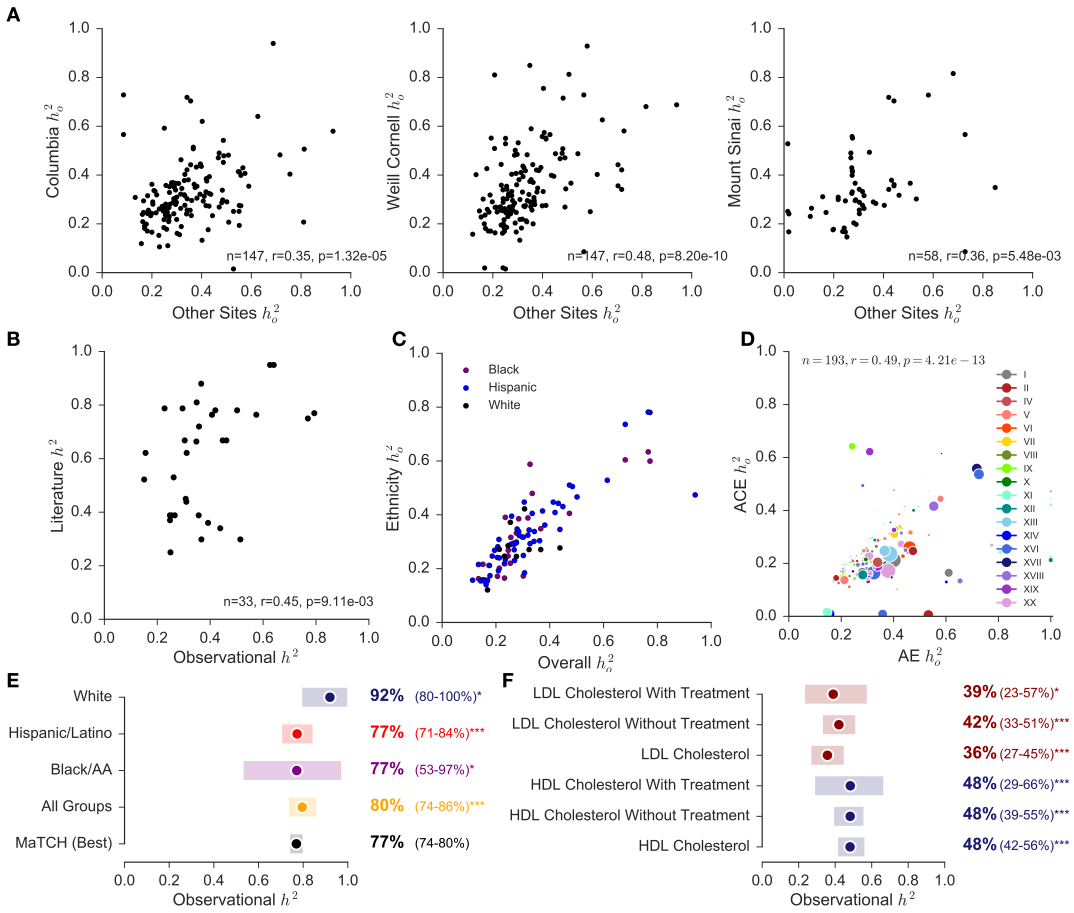

Figure 5.2: Estimating heritability of disease using electronic health records. We designed a method, called SOLAR*Strap*, for estimating the heritability of traits where the phenotype is derived under unknown ascertainment biases, the h_o^2. (A) We found that performance was consistent across sites and (B) that h_o^2 is significantly correlated with literature estimates of h^2. (C) Heritability estimates stratified by race and ethnicity using the AE model are correlated with estimates of h_o^2. (D) These models are also correlated when computing heritability estimates for ICD10 codes alone. (E) Heritability of traits that have been studied before, such as height, have been recapitulated by our study. We also stratified heritability of height by self-reported race and ethnicity as available in EHR. (F) Observational heritability of HDL cholesterol (blue) is significantly higher than heritability of LDL cholesterol (red). This difference is still observed after stratifying patients by the presence or absence of HMG-CoA reductase inhibitors as treatment for hypercholesterolemia.

I found that heritability estimates are significantly correlated across sites (Figure 5.2A). I identified traits with heritability estimates and then computed the correlation between the estimates found in each one of the study sites to the other two sites. Columbia had 147 traits

that overlapped with traits from the other two sites, with correlation r=0.35, p=1.32e-05. Similarly, Weill Cornell had 147 traits, with correlation r=0.48, p=8.20e-10, and Mount Sinai had 58 traits, r=0.36, p=5.48e-03. I mined the literature for heritability estimates and found 91 phenotypes that mapped to phenotypes I curated from the EHR. I also included all traits reported in the latest meta-analysis (Polderman et al. 2015). I used simulations to set the quality control parameters of the SOLAR*Strap* procedure. Thirty-three traits passed these quality control criteria. I found that they were significantly correlated with literature estimates for these traits (r=0.45, p=9.11e-03, Figure 5.2B), and 16 (48%) had overlapping confidence intervals (Table 5.3). On average, observational heritability estimates were 27% lower than those reported in the literature. I also stratified the heritability estimates by race and ethnicity. The estimates stratified by race and ethnicity are significantly correlated with the overall heritability estimates (Figures 5.2C and 5.3).

Name	Site	h_o^2 (95% CI)	h^2 (95% CI)
Acne	Columbia	0.35 (0.22-0.55)	0.81 (0.73-0.89)
Allergy, Unspecified	Columbia	0.30 (0.18-0.42)	0.67 (0.61-0.72)
Asthma*	Weill Cornell	0.37 (0.21-0.58)	0.30 (0.22-0.37)
Asthma*	Columbia	0.51 (0.30-0.63)	0.30 (0.22-0.37)
Asthma with status asthmaticus	Columbia	0.45 (0.27-0.56)	0.67 (0.61-0.72)
Atopic dermatitis	Columbia	0.42 (0.25-0.62)	0.78 (0.73-0.83)
Atopic dermatitis and related conditions*	Columbia	0.50 (0.34-0.78)	0.78 (0.73-0.83)
Attention deficit hyperactivity disorder	Columbia	0.36 (0.22-0.50)	0.72 (0.56-0.85)
Celiac disease*	Columbia	0.77 (0.41-0.98)	0.75 (0.55-0.96)
Depression	Columbia	0.25 (0.17-0.30)	0.37 (0.31-0.42)
Depressive disorder*	Weill Cornell	0.25 (0.15-0.41)	0.39 (0.36-0.42)
Depressive disorder	Columbia	0.27 (0.13-0.35)	0.39 (0.36-0.42)
Disease of skin and subcutaneous tissue	Columbia	0.30 (0.17-0.46)	0.79 (0.75-0.83)
Diseases of the digestive system*	Weill Cornell	0.31 (0.22-0.40)	0.44 (0.40-0.48)
Diseases of the nervous system	Weill Cornell	0.15 (0.08-0.23)	0.52 (0.46-0.58)
Diseases of the respiratory system	Weill Cornell	0.31 (0.22-0.40)	0.62 (0.55-0.70)
Diseases of the respiratory system	Mount Sinai	0.16 (0.10-0.25)	0.62 (0.55-0.70)
Diseases of the skin and subcutaneous tissue	Weill Cornell	0.23 (0.13-0.32)	0.79 (0.75-0.83)
Eczema (adults)*	Columbia	0.44 (0.30-0.57)	0.34 (0.02-0.66)
Exacerbation of asthma*	Columbia	0.46 (0.26-0.63)	0.67 (0.61-0.72)
Glaucoma*	Columbia	0.39 (0.24-0.65)	0.36 (0.18-0.54)
Height*	Columbia	0.79 (0.67-0.94)	0.77 (0.74-0.80)
Major depressive disorder, recurrent*	Columbia	0.36 (0.21-0.51)	0.39 (0.36-0.42)
Major depressive disorder, single episode	Columbia	0.25 (0.17-0.31)	0.39 (0.36-0.42)
Migraine*	Columbia	0.31 (0.17-0.48)	0.45 (0.41-0.49)
Obesity*	Weill Cornell	0.57 (0.40-0.82)	0.76 (0.67-0.85)
Obesity	Columbia	0.41 (0.31-0.49)	0.76 (0.67-0.85)
Osteoarthritis	Columbia	0.26 (0.15-0.38)	0.53 (0.44-0.62)
Rhinitis (children)*	Weill Cornell	0.63 (0.39-0.95)	0.95 (0.78-0.97)
Rhinitis (children)	Columbia	0.64 (0.47-0.77)	0.95 (0.78-0.97)
Type 1 diabetes mellitus*	Weill Cornell	0.35 (0.23-0.53)	0.66 (0.49-0.84)
Type 1 diabetes mellitus	Columbia	0.37 (0.20-0.70)	0.88 (0.78-0.94)
Type-2 diabetes*	Columbia	0.25 (0.15-0.32)	0.25 (0.15-0.35)

Table 5.3: Comparison between observational heritability (h_o^2) and heritability estimates (h^2) previously reported in the literature. Among the 33 traits, 16 (48%) have overlapping confidence intervals, highlighted with a star (*).

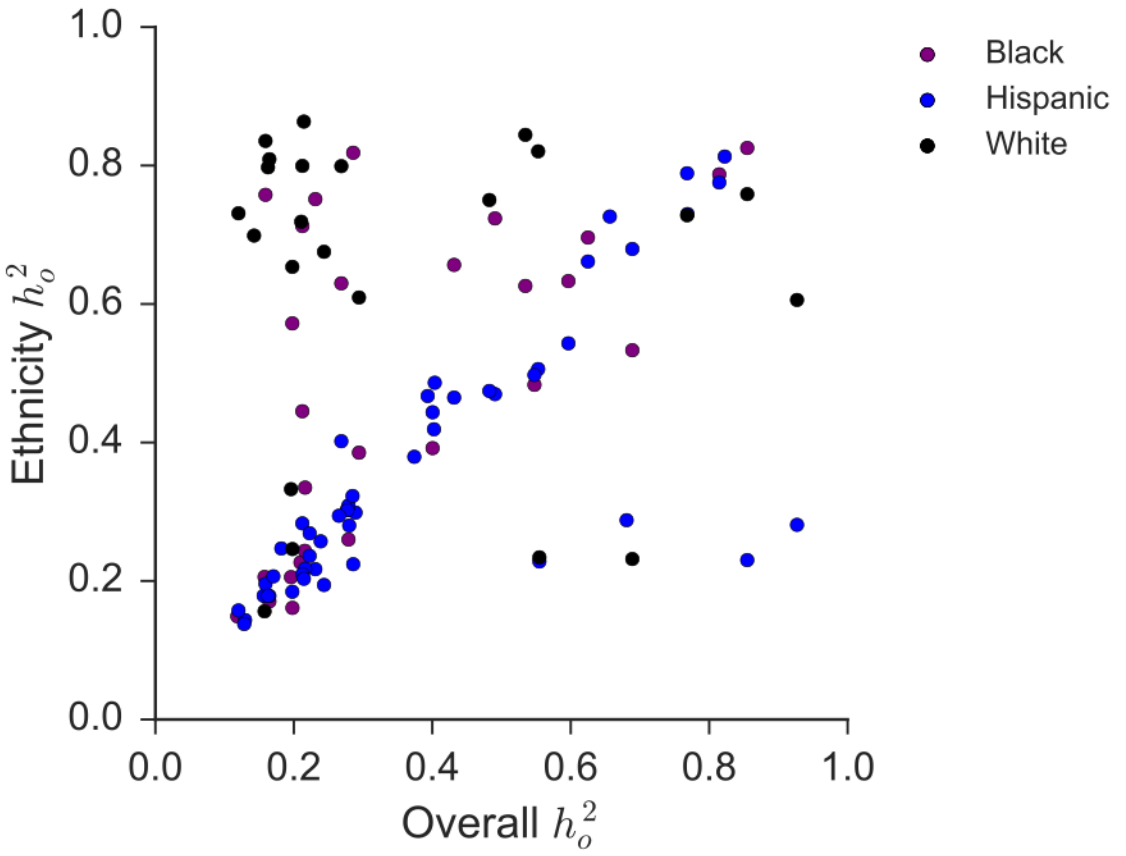

Figure 5.3: Correlation between the estimates stratified by race and ethnicity and the overall heritability estimates using the ACE model.

In addition to the additive genetic model (AE), I also modeled heritability with a term for common environment (ACE) using the mother ID as the household ID. ACE and AE models are overall significantly correlated (r=0.66, p=1.25e-34, Figure S2) and are also correlated when computing heritability estimates for ICD10 codes alone (r=0.49, p=4.21e-13, Figure 5.2D).

I found that phenotypes from the EHR could increase sample size and recapitulate heritability estimates that are well known. For example, the most heritable trait I found was for sickle cell disease, h_o^2=0.97 (0.75-1.00), N=857 (Table 5.1). I also computed heritability of height and stratified the estimates based on self-reported race and ethnicity as captured

in the EHR. The latest meta-analysis reported heritability of height to be 0.77 (CI=0.74-0.80) (Polderman et al. 2015). Using EHR data, I obtained observational heritability of 0.80 (CI=0.74-0.86). The heritability of height among whites had a lower quality control score and is higher than the other groups. (Figure 5.2E).

Using phenotypes from the EHR for heritability can provide clarity for poorly studied traits, revealing subtle differences between closely related conditions, and open up new avenues of heritability research. For example, two previous studies had shown conflicting evidence for the relative heritability of HDL cholesterol and LDL cholesterol (Pietiläinen et al. 2009; Souren et al. 2007). The larger of these two studies (*N*=378) found no difference in the heritability of these two traits when adjusting for age and sex, while the other found a slightly higher heritability for HDL, but was underpowered to detect significance. In this study, I presented evidence that HDL is more heritable than LDL (h_o^2 =0.48 95% CI: 0.42 - 0.56 vs 0.36 95% CI: 0.27 - 0.45 at Columbia; h_o^2 =0.51 95% CI: 0.35 - 0.67 vs 0.26 95% CI: 0.15 - 0.38 at Weill Cornell). This finding held when accounting for the use of HMG-CoA reductase inhibitors as treatment for hypercholesterolemia (Figure 5.2F). At 96,241 patients in the Columbia cohort and 33,239 patients in the Weill Cornell cohort, this study was the largest heritability study of cholesterol ever conducted, to my knowledge.

		Trait with Highest Heritability			Trait with Lowest Heritability		
Dichotomous Disease Category	Median h_2^o	ICD9 Code	Name	Median h_2^o (95% CI)	ICD9 Code	Name	Median h_2^o (95% CI)
Hematologic Diseases	0.50	287.31	Immune thrombocytopenic purpura	0.71 (0.33-0.96)	285.9	Anemia	0.20 (0.15-0.36)
Mental Health Diseases	0.41	309.28	Adjustment disorder with mixed anxiety and depressed mood	0.95 (0.36-1.00)	315.39	Other developmental speech or language disorder	0.11 (0.09-0.15)
Sense Organs Diseases	0.41	365.11	Primary open angle glaucoma	0.93 (0.52-1.00)	382.9	Unspecified otitis media	0.10 (0.06-0.16)
Endocrine and Metabolic Diseases	0.40	278.02	Overweight	0.71 (0.54-0.88)	272.4	Other and unspecified hyperlipidemia	0.23 (0.15-0.37)
Gastrointestinal Diseases	0.39	579	Celiac disease	0.78 (0.55-0.97)	521	Dental caries	0.12 (0.07-0.18)
Infectious Diseases	0.34	111	Pityriasis versicolor	0.85 (0.50-0.94)	780.6	Fever	0.11 (0.05-0.23)
Respiratory Diseases	0.34	477.9	Allergic rhinitis, cause unspecified	0.72 (0.25-0.93)	464.4	Croup	0.09 (0.05-0.12)
Cardiovascular Diseases	0.33	785.2	Undiagnosed cardiac murmurs	0.59 (0.42-0.84)	786.59	Other chest pain	0.18 (0.11-0.25)

		Trait with Highest Heritability			Trait with Lowest Heritability		
Dichotomous Disease Category	Median h_2^o	ICD10 Code	Name	Median h_2^o (95% CI)	ICD10 Code	Name	Median h_2^o (95% CI)
Pregnancy, Childbirth and Puerperium	0.54	O30	Multiple gestation	0.76 (0.36-1.00)	O30-O48	Maternal care related to the fetus and amniotic cavity and possible delivery problems	0.41 (0.19-0.61)
Hematologic Diseases	0.45	D57	Sickle-cell disorders	0.97 (0.75-1.00)	D64	Other anemias	0.18 (0.11-0.30)
Injury and Poisoning	0.40	T59	Toxic effect of other gases, fumes and vapors	0.81 (0.49-0.98)	S01	Open wound of head	0.18 (0.10-0.36)
Infectious Diseases	0.40	B35	Dermatophytosis	0.81 (0.41-0.98)	B80	Enterobiasis	0.11 (0.04-0.13)
Genitourinary Diseases	0.37	N92	Excessive, frequent and irregular menstruation	0.85 (0.62-0.99)	N80-N98	Noninflammatory disorders of female genital tract	0.15 (0.09-0.20)
Respiratory Diseases	0.35	J01	Acute sinusitis	0.85 (0.61-0.98)	J02	Acute pharyngitis	0.02 (0.01-0.03)
Eye Diseases	0.34	H35	Other retinal disorders	0.55 (0.33-0.77)	H10	Conjunctivitis	0.18 (0.10-0.22)
Gastrointestinal Diseases	0.34	K90	Intestinal malabsorption	0.84 (0.69-0.98)	K02	Dental caries	0.14 (0.09-0.20)
Endocrine and Metabolic Diseases	0.34	E20-E35	Disorders of other endocrine glands	0.60 (0.28-0.89)	E84	Cystic fibrosis	0.01 (0.01-0.02)
Cardiovascular Diseases	0.33	I15	Secondary hypertension	0.50 (0.31-0.89)	IX	Diseases of the Circulatory System	0.18 (0.10-0.28)
Skin Diseases	0.32	L70	Acne	0.72 (0.20-0.91)	L80-L99	Other disorders of the skin and subcutaneous tissue	0.17 (0.11-0.29)
Ear and Mastoid Diseases	0.31	H61	Other disorders of external ear	0.82 (0.68-0.93)	H66	Suppurative and unspecified otitis media	0.11 (0.06-0.22)
Mental Health Diseases	0.31	F93	Emotional disorders with onset specific to childhood	0.78 (0.27-1.00)	F40-F48	Anxiety	0.02 (0.01-0.03)
External Causes of Morbidity and Mortality	0.31	V49	Car occupant injured in other and unspecified transport accidents	0.94 (0.87-0.99)	V04	Pedestrian injured in collision with heavy transport vehicle or bus	0.01 (0.00-0.01)
Signs and Symptoms	0.30	R92	Abnormal findings on diagnostic imaging of breast	0.48 (0.26-0.65)	R62	Lack of expected normal physiological development	0.07 (0.05-0.10)
Musculoskeletal Diseases	0.27	M71	Other bursopathies	0.61 (0.25-0.99)	M00-M25	Arthropathies	0.18 (0.11-0.25)
Congenital malformations	0.27	XVII	Congenital Malformations	0.73 (0.50-0.96)	Q85	Phakomatoses	0.05 (0.00-0.09)
Neoplasms	0.25	D23	Other benign neoplasms of skin	0.35 (0.20-0.53)	II	Neoplasms	0.17 (0.08-0.27)
Perinatal Diseases	0.22	XVI	Certain Conditions Originating In the Perinatal Period	0.62 (0.45-0.84)	P00-P04	Newborn affected by maternal factors and by complications of pregnancy	0.05 (0.01-0.08)
Neurological Diseases	0.17	G47	Sleep disorders	0.31 (0.19-0.48)	G44	Other headache syndromes	0.02 (0.01-0.03)

		Trait with Highest Heritability			Trait with Lowest Heritability		
Quantitative Disease Category	Median h_2^o	LOINC Code	Name	Median h_2^o (95% CI)	LOINC Code	Name	Median h_2^o (95% CI)
Endocrine Disorders	0.30	3016-3	Thyrotropin [Units/volume] in Serum or Plasma	0.37 (0.23-0.49)	3026-2	Thyroxine (T4) [Mass/volume] in Serum or Plasma	0.26 (0.16-0.36)
Gastrointestinal Disorders	0.30	2324-2	Gamma glutamyl transferase [Enzymatic activity/volume] in Serum or Plasma	0.45 (0.35-0.56)	1975-2	Total Bilirubin serum/plasma	0.11 (0.08-0.16)
Hemorrhage	0.18	5902-2	Prothrombin time - patient	0.25 (0.16-0.35)	718-7	Hemoglobin	0.14 (0.08-0.19)
Metabolic and Nutritional Disorders	0.41	2573-4	Lipoprotein.alpha [Mass/volume] in Serum or Plasma	0.49 (0.41-0.58)	2498-4	Iron [Mass/volume] in Serum or Plasma	0.25 (0.14-0.35)
Metabolic Disorders	0.38	2085-9	Cholesterol in HDL [Mass/volume] in Serum or Plasma	0.51 (0.35-0.67)	2089-1	Cholesterol in LDL [Mass/volume] in Serum or Plasma	0.26 (0.15-0.38)
Reticuloendothelial Disorders	0.29	4679-7	Reticulocytes %	0.93 (0.77-1.00)	26450-7	Eosinophils %	0.12 (0.07-0.18)

Table 5.4: Heritability Ranges for Dichotomous and Quantitative Trait Categories. The median observational heritability and ranges are shown for dichotomous trait categories, both ICD9 and ICD10 codes, and for quantitative trait categories, LOINC codes. Within each category, the trait with the highest heritability and the trait with the lowest heritability are shown. Mendelian conditions are annotated with (*) and traits with literature heritability estimates are marked with.

Heritability is used to estimate genetic contribution to complex, polygenic, or quantitative traits rather than classic Mendelian disorders in which the presence or absence of a single genetic mutation determines the development of the disease. Interestingly, our algorithm was able to provide estimates of heritability for Mendelian traits without genetic information based only on EHR data. For example, I observed high heritability estimates for common highly penetrant Mendelian diseases with autosomal transmission, such as sickle cell disease (h_o^2 = 0.97, 95% CI 0.75-1.00, *N*=857 families), but low heritability estimates for other rare recessive Mendelian traits, such as cystic fibrosis (h_o^2 = 0.01 95% CI: 0.01-0.02 *N*=7,682 families). Recovering a heritability estimate of almost 1 for sickle cell is reassuring since that is exactly what would be expected in the presence of a highly penetrant mutation and when carriers are also frequently correctly identified in the EHR. However, the heritability of cystic fibrosis was very low. This is likely because the additive model used for heritability estimation is clearly misspecified for a rare disease with a known recessive pattern of inheritance and asymptomatic carrier status. Moreover, because of the

	Cystic Fibrosis (CF)		Sickle Cell Disorder (SCD)				
Relationships	Relationships count	Relationships ratio	Relationships count	Relationships ratio	p-value	Difference (SCD - CF)	Ratio (SCD/CF)
Aunt/Uncle	189	0.0608	2526	0.0699	3.08e-01	0.0092	1.1506
Child	316	0.1016	6312	0.1748	2.18e-14***	0.0731	1.7196
Cousin	166	0.0534	1997	0.0553	8.35e-01	0.0019	1.0357
First cousin once removed	78	0.0251	1330	0.0368	1.57e-03**	0.0117	1.4679
Grandaunt/Granduncle	65	0.0209	1239	0.0343	1.10e-04***	0.0134	1.6410
Grandchild	60	0.0193	2024	0.0560	1.63e-20***	0.0367	2.9041
Grandnephew/Grandniece	141	0.0454	1673	0.0463	8.21e-01	0.0010	1.0215
Grandparent	123	0.0396	1486	0.0411	1.00e+00	0.0016	1.0401
Great-grandaunt/Great-granduncle	30	0.0096	570	0.0158	8.9e-03**	0.0061	1.6357
Great-grandchild	34	0.0109	860	0.0238	1.35e-06***	0.0129	2.1775
Great-grandnephew/Great-grandniece	89	0.0286	786	0.0218	1.06e-02*	-0.0069	0.7603
Great-grandparent	22	0.0071	598	0.0166	1.52e-05***	0.0095	2.3400
Great-great-grandchild	24	0.0077	388	0.0107	1.40e-01	0.0030	1.3918
Great-great-grandparent	13	0.0042	272	0.0075	3.55e-02*	0.0034	1.8012
Nephew/Niece	213	0.0685	2866	0.0794	2.84e-01	0.0108	1.1584
Parent	920	0.2959	4324	0.1197	2.00e-159***	-0.1762	0.4046
Sibling	532	0.1711	4925	0.1364	1.34e-14***	-0.0347	0.7970
Spouse	94	0.0302	1938	0.0537	5.37e-08***	0.0234	1.7749

Table 5.5: Distribution of relationship types among families with cystic fibrosis and sickle cell disease.

availability of carrier screening and prenatal diagnosis, cystic fibrosis families are nowadays typically small (Castellani et al. 2009; Dupuis et al. 2005; Scotet et al. 2012; Slieker et al. 2005); affected cases also frequently suffer from infertility limiting the number of observed disease transmissions per family. Indeed, in our dataset families with cystic fibrosis were smaller (average family size 3.0 for cystic fibrosis vs 4.6 for sickle cell disease, p=8.8e-14), had more advanced average age (average 40 years old vs 36 years old for sickle cell disease, p=4.1e-17), had fewer "child" and "grandchild" relationships (p=2.18e-14 and p=1.63e-20, respectively), and included more parental relationships (p=2.00e-159) when compared to the sickle cell disease cohort (Table 5.5).

In addition, subtle phenotypical variations that are routinely collected clinically can be studied. For example, analysis of the highest and lowest heritability estimates by category provides us with interesting findings. Among neurological diseases, I observed that sleep disorders are highly heritable (h_o^2=0.31 95% CI: 0.19-0.48); whereas headache syndromes are not (h_o^2=0.02 95% CI: 0.01-0.03). A comprehensive list of heritability estimates for

multiple diseases' categories is available in Table 5.1. Finally, this study demonstrated that the EHR can identify novel traits for future genetic studies. Of the 500 traits I computed heritability estimates for, only 33 of which had been previously studied as part of the latest meta-analysis or identified by our literature review.

Discussion

Heritability is a key component in precision medicine and is typically estimated based on family history. Collection of comprehensive and accurate family history is time-consuming and does not occur during the vast majority of clinical encounters (Polubriaginof, Tatonetti, and Vawdrey 2015). The construction of pedigrees by inference of relatedness from administrative records, as described in Aim 2.2, allows for rapidly assessing family history and heritability at scales that were previously impossible to achieve. I used EHR-inferred relationships to calculate heritability estimates among individuals with defined relationships.

Previous research in this area has focused on family studies of known relatives, primarily twins. Mayer and colleagues used EHR data to create a cohort of 2,000 twins/multiple births and measured concordance among identified twins for two highly heritable diseases, muscular dystrophy and fragile-X syndrome (Mayer et al. 2014). This study looked not only at twins, but entire families across several generations.

Importantly, most previous studies have predominantly involved White Europeans and may not be representative of other populations. However, the results presented reflect the diverse, multiethnic population of New York City – the majority of our patient population

is not self-reported as "white." For example, I stratified patients that had height available in the EHR by self-reported race and ethnicity and used these cohorts of patients to compute heritability of height. I observed that the heritability estimate was higher among whites in comparison to other race and ethnicity groups. Bias might explain this difference since this group had a lower quality control score than the others. I also investigated income as a possible confounder using patient ZIP codes and Census data. Overall, the population self-identified as white has twice the average income than other populations – one possible explanation for this difference given that heritability estimates increase in more homogenous environments. This could create a difference in heritability of height both across ethnicities and across income levels. In other cases, traits have been shown to be more heritable in high socioeconomic strata than in lower strata (Bronfenbrenner and Ceci 1994; Harden, Turkheimer, and Loehlin 2007; Turkheimer et al. 2003).

However, the stratification by race and ethnicity was not feasible for all traits. Over 68% of the families have a single race and ethnicity reported and over 29% of the families have two distinct race and ethnicity groups reported. Estimates of traits that had a large enough sample size to stratify by race and ethnicity are available at http://riftehr.tatonettilab.org. For traits that were stratified by race and ethnicity, heritability estimates were significantly correlated with the overall heritability estimate.

The primary challenge when using traits defined from an observational resource, like the EHR, is incomplete phenotype information resulting in ascertainment bias. In a heritability study, the phenotype of each study participant is, ideally, carefully evaluated and quantified. This is not feasible, however, when the cohort contains millions of patients with thousands of phenotypes. The bias may depend on many latent factors, including the trait

being studied, the trait status of relatives, the proximity to the hospital, and an individual's ethnicity and cultural identification, among others. The consequence of this uncontrolled ascertainment bias is that heritability estimates will be highly dependent on the particular individuals in the study cohort. I observed that a small number of highly biased families could significantly sway the heritability estimate. Repeated sub-sampling will be robust to these types of biases. EHR-based heritability estimates are particularly well-suited for complex traits that require large numbers of patients (e.g., Type 2 Diabetes Mellitus and Obesity).

The unique nature of the relationships and phenotypes derived from the EHR may necessitate novel methods for estimating heritability. I used a mixed linear model implemented in SOLAR (Almasy and Blangero 1998) to estimate heritability and used repeated sampling, which I call SOLAR*Strap*, for efficiency and to correct for ascertainment heterogeneities. I evaluated the impact of bias and missingness on SOLAR*Strap*p by comparing the heritability estimates with simulated data and demonstrated that SOLAR*Strap* is robust to bias. Overall, quantitative traits perform better than dichotomous traits, and traits commonly documented in EHRs perform better than rare and poorly documented conditions (e.g. mental health disorders). There may be more accurate ways to estimate heritability from this unique data source. Future work should focus on using only certain types or relationships or use alternative modeling strategies. Fragmentation of care is an additional limitation when using EHR data for genetic research. Patients often go to multiple healthcare systems, and therefore, the information available in a single institution is incomplete. Future implementations may address this limitation by accounting for the number of visits and documentation of primary care physician in the healthcare system or by integrating

records across a regional healthcare network.

There are significant bioethical considerations regarding the use of the RIFTEHR method, including how best to balance the competing demands of protecting patients' privacy with clinicians' duty to warn relatives of potential genetic risks. The method could readily be applied in EHR systems, such that clinicians could easily access the health information of a patient's family members. In the United States, accessing a family member's health information in this manner may be considered a violation of the 1996 Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule (United States, 1996). On the other hand, case law in the United States has established that healthcare providers have a responsibility to inform a patient's relatives about heritable conditions that may reasonably put the relatives "at risk of harm" (Suarez 2011). These conflicts may need to be resolved before automatic relationship inference can be used clinically. It is worth noting there is a risk of reidentification of family structures, even when de-identified according to the HIPAA Safe Harbor. For example, unique family structures could be identified by cross-referencing obituaries and other online tools. Extra safeguards are necessary to mitigate these risks when releasing these data.

Conclusion

I have described and validated a novel method for identifying familial relationships in patient medical records in Aim 2.2, and used 7.4 million relationships inferred from the EHRs at three academic medical centers to estimate heritability of 500 traits without genetic testing. I found that heritability estimates were concordant across the three centers, and are

broadly consistent with published studies, suggesting that the method may have broad applicability. Genetic information is valuable but expensive and not always available. In this case, familial relationships extracted from emergency contact information can personalize disease risk prediction and facilitate heritability determination for phenotypes that were not previously investigated in family-based or twin studies. The correspondence the heritability estimates presented in this study with family-based estimates provides a direct and novel validation of the value of electronic health records for generating inferences about disease, making RIFTEHR a valuable tool for the advancement of precision medicine.

5.2 Aim 3.2 - Estimating disease screening rates using electronic health records data

Background

that are at a high risk for disease development, and therefore promote disease prevention, screening, and early diagnosis and treatment. Current clinical guidelines, such as those from the U.S. Preventative Services Task Force (USPSTF) recommend additional or early disease screening for patients at higher risk for developing certain diseases, such as cancer, cardiovascular, and gastrointestinal conditions. Despite these lofty goals, there remains a gap in effectively screening patients. For instance, previous research has shown that breast cancer screening takes up valuable time during patient care visits to conduct accurately (Owens et al. 2011). Furthermore, there has been little research on clinician adherence to the recommendation of early screening among high-risk patients (Jemal and Fedewa 2017; Solbak et al. 2018).

The lack of clarity in understanding adherence to guidelines has a large potential impact on how care is delivered. Adherence to clinical guidelines is important, particularly for chronic diseases such as diabetes mellitus. Previous studies have shown that adherence to treatment guidelines that include assessing various physiological and social determinant information is generally low (Oude Wesselink et al. 2015). Prior work to evaluate clinical guideline adherence in diabetes mellitus has focused on disease management, either by determining the frequency of testing for disease outcomes, such as retinopathy, or process measures, such as measuring for hemoglobin A1c (An et al. 2018; Khunti et al. 2018).

There has been less focus in the literature on early or preventative screening for diabetes mellitus, particularly in relation to family history. Fortuitously, the widespread use of EHRs in combination with the method previously described in Aim 2.2, which identifies family medical history from existing clinical databases, allowed us to study whether patients have been properly screened for conditions in a comprehensive way.

Objectives

The purpose of this study was to use EHR data to determine the rates of screening among patients known to be at high risk for a prevelant condition, diabetes mellitus, and for a rare condiation, celiac disease.

Research Questions

- *Can EHR data be used to measure disease screening and adherence to clinical guidelines?*

Methods

As a proof-of-concept to determine the usefulness of EHR data in assessing disease screening rates, I applied similar methodology focusing on two distinct conditions: diabetes and celiac disease. These conditions were determined because both conditions have additional screening recommendations for patients with a known family history of disease and both conditions have additional screening recommendations for patients with a known family history of disease, with diabetes being highly prevalent affecting nearly 1 in 10

Americans while celiac disease is considered a rare condition.

With approval of the Institutional Review Board of Columbia University Medical Center, I conducted a retrospective analysis of family members of patients diagnosed with diabetes mellitus visiting NewYork-Presbyterian Hospital/Columbia University Medical Center from 2007 to 2017. Patients with a diagnosis of diabetes mellitus were identified using a validated and previously implemented EHR phenotype, available at PheKB (*Phenotype KnowledgeBase*). This EHR phenotype used a combination of diagnosis codes, medications and laboratory test results to identify patients with diabetes mellitus in our institution. Because an EHR phenotype for celiac disease had not been developed and validated, I examined relatives of patients (N=2,081) with biopsy-diagnosed celiac disease in a prospectively maintained database at NewYork-Presbyterian Hospital/Columbia University Medical Center.

To identify family history in electronic health records (EHRs), patients' relatives were identified using RIFTEHR (Relationship Inference from the Electronic Health Record), as described in Aim 2.2, a novel validated method that used the first name, last name, phone number and ZIP code of patients' emergency contacts to identify familial relationships. Once the relationships were identified, RIFTEHR inferred additional relationships according to family structure. The identified relationships were previously validated using both clinical and genetic data, as previously described in Aim 2.2 (Polubriaginof et al. 2017).

Once the cohort of family members was identified, I extracted demographic information, such as sex, age, race, and ethnicity from the EHR. While race and ethnicity were stored as distinct fields in our database, I found that transforming the two fields into a single field addressed many cases of missing data. Therefore, regardless of race, patients with

a reported ethnicity of "Hispanic" are reported in this study as "Hispanic." Patients with ethnicity recorded as "non-Hispanic" or "Unknown" were reported using the race information available (e.g., "White," "Black or African American," "Asian"). Patients without race and ethnicity information were reported as "Uninformative."

Diabetes mellitus

I measured diabetes screening by identifying individuals that had at least one of the following laboratory tests after the index case diagnosis date: fasting glucose (LOINC code 1558-6), random glucose (LOINC codes 2339-0, 2345-7), or hemoglobin A1C (LOINC codes 4548-4, 17856-6, 4549-2, 17855-8). I included all family members over 18 years of age. I calculated descriptive statistics of the identified cohort, along with the rate of screening among family members. Additionally, I performed a multivariate analysis to determine factors that increase the likelihood of receiving a screening test. To determine the influence of each parameter in the logistic regression model, I computed the standardized coefficients (β) by multiplying the beta coefficient (B) to the standard deviation of the corresponding parameter in the data. Python 2.7 was used to perform these analyses.

Celiac disease

The EHR was queried and each patients' records were manually reviewed to extract celiac disease testing information. The manual review included extraction of the following elements: 1) serology results, 2) duodenal biopsy results, 3) occurrence of a visit with a gastroenterologist, 4) presence of signs or symptoms of celiac disease in clinical notes and/or ICD codes, and 5) documentation of family history of celiac disease. Demographic infor-

mation such as gender, age, and race and ethnicity were queried from the EHR's database. Celiac disease screening was defined as either antibody testing or endoscopic evaluation with duodenal biopsy. SAS (Cary, NC) version 9.4 was used to perform both univariate and multivariate analyses to identify predictors of celiac disease screening. I tested the following variables *a priori* and included all variables in the multivariable analysis. All reported p values are 2-sided.

Results

Diabetes mellitus

Overall, I identified 13,086 patients with diabetes mellitus that also had familial relationships extracted by RIFTEHR. These patients had 56,794 family members in our database, distributed across 12,613 families. Familial relationships spanned up to four generations, including relationships such as great-great-grandparents. Of those, 45,778 family members (12,181 families) were over 18 years of age, and 27,757 (8,188 families) had a clinical visit after the index case had been diagnosed with diabetes mellitus; this was the population deemed eligible for diabetes screening (Figure 5.4).

The cohort of patients eligible for diabetes screening was represented by 18,406 (66.3%) females, with an average age of 46 years old, with the majority being self-reported as Hispanic (72.7%). Table 5.6 summarizes the demographic information of the study cohort. Among the eligible-for-screening cohort, 19,264 (69.4%) received diabetes screening, and 8,493 (30.6%) patients did not. Among first-degree relatives of the index cases, 71.6% received at least one diabetes screening test. The cohort of individuals that received screening

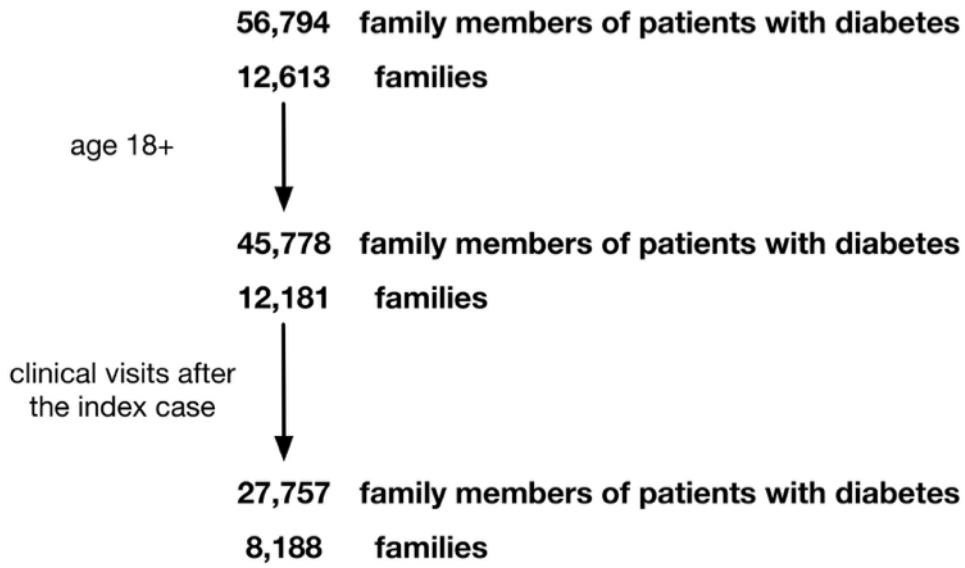

Figure 5.4: Cohort of individuals eligible for early diabetes screening. Eligibility criteria included being 18 years of age and having a clinical visit after the family member was diagnosed with diabetes mellitus.

was significantly older than the group that did not receive screening (average age 50 vs. 38, $p < 0.0001$).

The multivariate analysis found that age ($\beta = 0.67$, $p < 0.0001$), having more than one family member affected ($\beta = 0.11$, $p < 0.0001$), and being a female ($\beta = 0.08$, $p < 0.0001$) were the most important contributors to being screened for diabetes mellitus. Results for all features are shown in Table 5.7, and screening rates for these features are shown in Table 5.8.

Characteristics	Count	Percentage
Females	18,406	66.3%
Age	46 ± 20.4	
Race and Ethnicity		
American Indian	21	0.1%
Asian	171	0.6%
Black	3,706	13.4%
Hispanic	15,128	54.5%
Pacific Islander	76	0.3%
Uninformative	7,731	27.9%
White	10,212	36.8%

Table 5.6: The traditional process of collecting patient-provided information.

Features	Beta coefficient (B)	Standardized coefficient (β)	p-value
Age	0.03	0.67	<0.0001
More than one family member diagnosed	0.26	0.11	<0.0001
Female	0.17	0.08	<0.0001
First degree relative diagnosed	0.08	0.04	<0.0001
Second degree relative diagnosed	0.05	0.02	<0.0001
Third degree relative diagnosed	0.14	0.02	<0.0001
Fourth degree relative diagnosed	-0.09	-0.01	<0.0001

Table 5.7: Results of a multivariate analysis. To determine the influence of each parameter in the logistic regression model, I computed the standardized coefficients (β) by multiplying the beta coefficient (B) to the standard deviation of the corresponding parameter in the data.

Features	Screened	Total	Percentage Screened
Females	13,178	18,406	71.6%
Males	6,086	9,351	65.1%
More than one family member diagnosed	5,467	19,264	28.4%
First degree relative diagnosed	7,394	10,332	71.6%
Second degree relative diagnosed	2,520	3,993	63.1%
Third degree relative diagnosed	169	435	38.9%
Fourth degree relative diagnosed	288	511	56.4%

Table 5.8: Screening rates stratified by features.

Celiac disease

I applied the RIFTEHR algorithm to identify family members of the 2,081 index cases of celiac disease, yielding 379 distinct families and 852 relatives. The inclusion criteria included only relatives seen at our institution *after* the index case was diagnosed, which resulted in a total of 272 distinct families and 539 relatives (Table 5.9).

There was a relatively even distribution of men (47.1%) and women (52.9%), and of those 18 years and older (52.5%) as compared to those under 18 years (47.5%). The majority of individuals identified were first-degree relatives (71.1%) of patients with celiac disease and had been seen more than once (88.3%) at our institution after their relative was diagnosed. Non-Hispanic White (58.6%) and Hispanic (28.9%) were the two most commonly documented ethnicities in our study population. From manual review of the EHR, 316 of the 529 total relatives (58.6%) did not have any associated symptoms or conditions related to celiac disease.

I found that 193 of the 383 (50.4%) first-degree relatives had been screened for celiac disease (Table 5.10). When restricting this analysis to first-degree relatives with associated symptoms or conditions related to celiac disease, I found that 71.5% (118/165) were tested. Since screening practices are largely influenced by the available data at the time of the visit, each patient's record was reviewed to determine if a family history of celiac disease had been documented anywhere within the record. Of all 539 relatives, only 120 (22.3%) had a family history of celiac disease documented. When subcategorized by degree of relative, I found that 30.3% of first-degree relatives had documentation of family history of celiac disease, as compared to only 2.6% for all other degrees of relatives.

	N (%)
Age Group	
< 18	256 (47.5%)
18-39	114 (21.2%)
40-69	133 (24.7%)
70+	36 (6.7%)
Gender	
Male	254 (47.1%)
Female	285 (52.9%)
Race	
Non-Hispanic White	316 (58.6%)
African American	14 (2.6%)
Hispanic	156 (28.9%)
Other/Unknown	53 (9.8%)
Relative	
First	383 (71.1%)
All other	156 (28.9%)
Number of times seen at CUMC	
Once	63 (11.7%)
2-5	206 (38.2%)
>5	270 (50.1%)
CD Signs/Symptoms during any visit	
Diarrhea	54 (10.0%)
Bloating	18 (3.3%)
Abdominal Pain	136 (25.2%)
Fatigue	2 (0.4%)
Fe. Def. anemia	14 (2.6%)
Osteoporosis/OA	29 (5.4%)
GERD	62 (11.6%)
DM1/Autoimmune thyroid/IgA Def./PBC	11 (2.0%)
None of above	316 (58.6%)

Table 5.9: Demographics of relatives (*N*=539).

Variable	Total	First Degree Relative	All Other Relatives	p value
Screened for CD	212/539 (39.3%)	193/383 (50.4%)	19/156 (12.2%)	p <0.0001
Documented family history of CD	120/539 (22.3%)	116/383 (30.3%)	4/156 (2.6%)	p <0.0001

Table 5.10: Screening and charting practices based upon degree of relative. Percent of symptomatic first degree relatives tested for CD 118/165 = 71.52%.

On univariate analysis, there were several factors that were associated with a higher likelihood of being screened (Table 5.11). Only 5.6% of relatives over the age of 69 were screened, a far lower rate compared to all other age categories, which ranged from 35.3% - 44.1%. Screening practices also varied by race, with 58.6% of Non-Hispanic Whites, 25% of Hispanics, and 0% of African Americans tested. Additionally, the presence of symptoms (59.2% vs. 25.3%, $p < 0.0001$), whether the relative was seen by a gastroenterologist (87.1% vs. 20.1%, $p < 0.0001$), whether there was documentation of a family history of celiac disease in the EHR (89.2% vs. 25.1%, $p < 0.0001$), and the degree of relative (first-degree 50.4% vs. all other degrees 12.2%, $p < 0.0001$), were associated with testing for celiac disease. Notably, neither sex (Male 39% vs. Female 39.7%, p=0.87) nor the number of times a relative had been seen at our institution after the initial family member had been diagnosed (once 36.5% vs. 2-5 times 45.2% vs. > 5 times 35.6%, p=0.09) affected the likelihood of celiac disease testing.

On multivariate analysis (Table 5.12), I found that age, number of visits to our institution, being seen by a gastroenterologist, the presence of symptoms or conditions associated with celiac disease, a documented family history of celiac disease, and the degree of relative, to be significant predictors of screening. Specifically, I found that relatives aged 18-39 were more than two times more likely to be screened than relatives under the age of 18 years old (OR 2.27, 95% CI: 1.12-4.58, p=0.02). When the number of visits was considered as a binary variable, those seen more than five times were less likely to be screened as compared to those seen one to five times, though this was of borderline significance (OR 0.57, 95% CI: 0.32-1.00, p=0.05). Other significant predictors included the presence of any condition or symptom related to celiac disease (OR 3.69, 95% CI 2.11-6.47, $p < 0.0001$)

Variable	Screened (%)	p-value
Age		
< 18	113/256 (44.1%)	
18-39	50/114 (43.9%)	p < 0.0001
40-69	47/133 (35.3%)	
70+	2/36 (5.6%)	
Gender		
Male	99/254 (39.0%)	p = 0.873
Female	113/285 (39.7%)	
Race		
Non-Hispanic White	149/316 (58.6%)	
African American	0/14 (0%)	p < 0.0001
Hispanic	39/156 (25%)	
Other/Unknown	24/53 (45.3%)	
CD Signs/Symptoms		
Symptomatic	132/223 (59.2%)	p < 0.0001
Asymptomatic	80/316 (25.3%)	
Number of visits		
Once	23/63 (36.5%)	p = 0.093
2-5	93/206 (45.2%)	
>5	96/270 (35.6%)	
Seen by GI		
Yes	35/155 (87.1%)	p < 0.0001
No	77/384 (20.1%)	
Documented family history of CD		
Yes	107/120 (89.2%)	p < 0.0001
No	105/419 (25.1%)	
Degree of Relative		
First	193/383 (50.4%)	p < 0.0001
Other	19/156 (12.2%)	

Table 5.11: Factors associated with screening: univariate analysis.

and being a first-degree relative (OR 4.90, 95% CI: 2.34-10.25, $p < 0.0001$). The two factors most strongly associated with screening were whether the relative had been seen by a gastroenterologist (OR 15.16, 95% CI: 7.72-29.80, $p < 0.0001$) and whether there was documentation in the EHR of a family history of celiac disease (OR: 11.9, 95% CI: 5.56-25.48, $p < 0.0001$). Race and sex were not associated with celiac disease testing on multivariate analysis.

A total of 79 of the 539 relatives (14.7%) had biopsies consistent with celiac disease. Fourteen individuals had biopsy-proven celiac disease but no record of antibody testing recorded within the EHR. Of the 82 patients who tested positive for celiac antibodies (endomysial, transglutaminase, and/or gliadin peptide), 80 (97.6%) were first-degree relatives, and a total of 65 (79.3%) had a biopsy confirming the diagnosis (Table 5.13).

Variable	Adjusted* Odds Ratio	95% Confidence Interval	p-value
Age in years			
<18	1.0	[ref]	[ref]
18-39	2.27	1.12 - 4.58	0.02
40-69	1.03	0.53 – 2.02	0.93
70+	0.27	0.05 – 1.43	0.12
Sex			
Female	1.0	[ref]	[ref]
Male	0.882	0.52 – 1.51	0.65
Race/Ethnicity			
Non-Hispanic White	1.0	[ref]	[ref]
Hispanic	0.75	0.39 – 1.46	0.40
Other/Unknown	1.16	0.52 – 2.57	0.72
Number of visits to CUMC			
1 - 5 visits	1.0	[ref]	[ref]
> 5 visits	0.57	0.32 – 0.999	0.0495
Seen by a gastroenterologist			
No	1.0	[ref]	[ref]
Yes	15.16	7.72 – 29.80	<.0001
Any symptom/sign of celiac disease			
No	1.0	[ref]	[ref]
Yes	3.69	2.11 – 6.47	<.0001
Documented family history of CD			
No	1.0	[ref]	[ref]
Yes	11.9	5.56 – 25.48	<.0001
Degree of relative			
Other	1.0	[ref]	[ref]
First	4.90	2.34 – 10.25	<.0001

Table 5.12: Multivariable analysis examining patient factors associated with screening in all relatives. *Adjusted for all variables listed in the table. Symptoms/signs of celiac disease include diarrhea, bloating, abdominal pain, fatigue, osteoporosis, osteoarthritis, GERD, Type 1 diabetes, autoimmune thyroid disease, IgA deficiency, and primary biliary cholangitis. [ref] refers to the group of reference.

Screened Relatives	Biopsy consistent with CD
First Degree Relative:	
Positive Antibody	63/80
N = 80	(78.8%)
All Other Relatives:	
Positive Antibody	2/2 (100%)
N = 2	

Table 5.13: Pathology results of screened relatives.

Discussion

In this study, I measured clinician adherence to diabetes mellitus and celiac disease screening among high-risk patients. Because current screening guidelines include family history of diabetes or celiac disease as a risk factor, respectively, I used a novel method for identifying families and gathering corresponding medical histories through patient-provided emergency contact information stored in the EHR. This method, along with an EHR phenotyping algorithm, was used to identify patients at risk for disease development that were eligible for additional testing. I found that 30.6% of patients at high risk for diabetes and 49.6% of patients at high risk for celiac disease were not appropriately screened for their respective diseases, even though early diagnosis is known to decrease morbidity. Previous research suggests similar findings for a myriad of different diseases, most notably in relation to cancer screening (Jemal and Fedewa 2017; Solbak et al. 2018).

Given that fewer resources are required to carry out proper screening for diabetes mellitus as compared to cancer screening, further studies should be conducted to understand the challenges preventing recommended diabetes mellitus screening. While cancer screen-

ing is often costly and requires more complex tests (e.g., MRI, genetic testing), diabetes screening is relatively simple and inexpensive. In celiac disease, screening can be initiated with serology, which is less invasive and less complex than a biopsy. Interestingly, I observed that 28.5% of patients presenting symptoms who had a first-degree relative affected by celiac disease were not screened.

This study found that there were several factors associated with increased screening for diabetes mellitus. These factors included being female, age, and having more than one family member diagnosed with diabetes. One of the major differences between the subpopulations who received screening vs. no screening was age, where the screened subpopulation was far more elderly on average (50 years old vs. 38 years old, $p < 0.0001$). One possible interpretation of this finding is that individual clinicians are not adhering to clinician guidelines in favor of considering patient's individual factors—in this case age—in determining the necessity of screening.

For celiac disease, there were multiple contributing factors to the overall low adherence to screening rates. As previously described in other conditions, being seen by a specialist in that discipline is associated with a higher likelihood of being screened (Patwardhan et al. 2011). In our study, only 39% of relatives were seen by a gastroenterologist, but those who did were significantly more likely to be screened. Additionally, the American College of Gastroenterology (ACG) guidelines recommend screening for first-degree symptomatic relatives (Rubio-Tapia et al. 2013), but I found that both being a first-degree relative and being symptomatic were independently associated with an increased likelihood of being screened. Those patients seen more than five times without being screened were overall less likely to be screened. This may be due to a significant number of acute conditions

that dictated numerous visits and took precedence over celiac disease screening, or reflect that after several visits, provider and patient may no longer be as cognizant of the family member who was previously diagnosed with celiac disease, and as a result, were less likely to be tested.

While clinicians may consciously consider but choose to not screen certain patients, I believe this to be an improbable explanation of the low rate of screening adherence. Because our institution is a tertiary facility, patients are often seeking for specialized care, where disease screening may not be a primary focus. Further, prior research provides a few additional potential explanations for the low adherence to screening guidelines, including lack of family history documentation and lack of patient and physician awareness (Sequist et al. 2009; Wee, McCarthy, and Phillips 2005). Even though family history has always been considered "a core element of clinical care" (Berg et al. 2009), it has been found to be poorly captured in the EHR. Lack of time to obtain, organize, and analyze family history data is perhaps one of the most important challenges in the quality of family history documentation (Green 2007; Guttmacher, Collins, and Carmona 2004; Rich et al. 2004; Scheuner et al. 2009; Sussner, Jandorf, and Valdimarsdottir 2011; Wilson et al. 2012a). Additionally, uncertainty about the medical history of family members, as well as inaccuracies in patient recall, compound the challenge of obtaining accurate family history data (Green 2007; Peace, Valdez, and Lutz 2012; Sussner, Jandorf, and Valdimarsdottir 2011). When it is captured in the EHR, family medical history information is frequently stored in clinical notes, which cannot be easily abstracted during a patient visit (Chen et al. 2014; Polubriaginof, Tatonetti, and Vawdrey 2015), and may ultimately result in poor screening rates. The results of the logistic regression demonstrated that patients who had multiple

family members diagnosed with diabetes was one of the most likely factors leading to a patient being screened for diabetes. This result points to the importance of family history data availability during the clinical encounter.

While the use of RIFTEHR for identifying family history in research has tremendous benefits, there are privacy issues regarding the use of this method for clinical practice, and tradeoffs must be made between providing optimal care and safeguarding the privacy and confidentiality of family members' health information (United States 1996). Notwithstanding the ethical considerations, the use of RIFTEHR for identifying familial relationships using EHR data unlocks new opportunities for secondary use of data to facilitate identification patients at high risk for disease development and support appropriate monitoring of prevention strategies such as disease screening.

Conclusion

In summary, 30.6% and 49.6% of patients that were eligible for early diabetes and celiac disease screening, respectively, did not receive the appropriate testing that could lead to early diagnosis, and therefore, decrease patient morbidity. In this study, I demonstrated that electronic health record data along with novel and innovative informatics methods can increase availability of data, and therefore utility of large electronic clinical databases, ultimately resulting in improvements in clinical care.

This page intentionally left blank.

Chapter 6

Conclusions and Future Work

6.1 Summary of Work

The body of research represented in this dissertation investigated the quality of patient-provided databases, the impact of interventions targeting these data, and the usefulness of these data in assessing disease risk.

In the first Aim, I focused on determining the quality of patient-provided data stored in clinical databases. To accomplish this goal, I assessed the data quality of three patient-provided data types: race and ethnicity, family history, and smoking status. When assessing the quality of race and ethnicity, I identified that data completeness, correctness, and concordance were all issues for this type of information. When assessing the quality of family history, my results showed that patients' family history records were rarely complete in the EHR. Smoking status data suffered similar problems with concordance and completeness. Furthermore, I found that changes to a patient's smoking status had plausibility issues, in that not all changes to smoking status could have been logically possible (e.g., a "current smoker" becoming a "never smoker"). Overall, the results of these studies demonstrated that such patient-provided information is currently poorly captured in the EHR.

The results from the three studies conducted in Aim 1 demonstrated that while the im-

portance of patient-provided data is well-known, these data are not being optimally captured during clinical encounters. The unavailability of patient-provided data at the point of care can negatively impact clinical decisions, such as limiting the assessment of disease risk. Disease risk assessment is a key factor in determining whether patients may benefit from additional disease screening and modified treatment. The fact that these data are not readily available to clinicians can result in disease comorbidity that might have been prevented or mitigated.

Based on the results of Aim 1, I investigated the impact of various intervention types on the quality of patient-provided data in Aim 2. Several types of interventions exist, including 1) high-level policy changes, such as the Meaningful Use program, 2) local health information technology initiatives, such as deploying patient-facing tools that collect and share information with patients, and 3) the use of informatics methods that leverage existing datasets to facilitate the identification of high-risk patients.

I found that each of the three types of interventions had a different effect on the quality of patient-provided data. While policy changes seemed to encourage the collection of patient-provided data using pre-determined categories, they did not necessarily translate into better data quality. I found that with patient-facing tools, patients were willing to provide even sensitive information, such as race and ethnicity, and that by doing so, they were able to enhance the data quality of the information contained in their medical records. In my studies, two forms of patient-facing tools, HCAHPS surveys and patient portals, enabled these changes. Using informatics methods, I demonstrated how issues of incomplete family history information can be overcome, in some cases, by accurately and automatically deducing certain family history information based on inferred relationships between

patients using other patient-provided information.

In Aim 3, I applied the informatics method developed in Aim 2 to assess patients' disease risk. This work analyzed disease risk at two levels: first at a population level, by measuring disease heritability, and second, at the individual level, by assessing disease screening rates among high-risk patients. In the first study on heritability, I successfully estimated disease heritability for 500 distinct traits, some of which had not previously been reported in the literature. Further, I showed how this method could be readily applied to diverse racial and ethnic groups, which overcomes a significant limitation of most genetic studies, which are based on a population of European descent.

In the second study from Aim 3, which focused on patient screening, I leveraged inferred familial relationships to determine screening rates for two conditions: diabetes mellitus, a prevalent condition that affects 1 in 10 Americans, and celiac disease, an autoimmune condition that affects approximately 1% of the population. For both conditions, screening rates among family members that are considered to be at high-risk for disease development were very low. In sum, the studies I carried out in Aim 3 highlight the difficulty associated with identification of high-risk individuals in the clinical setting. The results of these two studies demonstrate the impact of informatics methods utilizing patient-provided information in both genetics and clinical practice.

6.2 Contributions

My research is a novel contribution to understanding how to use EHR data to assess disease risk. The contributions of this thesis include: 1) an overview of the quality of patient-provided information in clinical databases, 2) an assessment of the impact of different intervention types on the quality of patient-provided data, 3) the development and evaluation of a novel method that uses patient-provided information to generate a unique data set that can support biomedical research, and 4) the use of clinical data to understand disease risk and assess disease screening rates among high-risk individuals. Each study I conducted provided insight into new areas of exploration that had not been previously reported in the biomedical literature. A summary of the publications and presentations generated during the course of my research are shown in the following Table 6.1.

Aim 1 explored the data quality of patient-provided data, including race and ethnicity, family history, and smoking status. The studies included in this chapter demonstrated that patient-provided data suffers from the same data quality issues as clinical data when stored in the EHR system.

Aim 2 assessed the impact of different intervention types on the quality of patient-provided information. The results showed that patient-facing tools were a superior method for capturing high-quality patient-provided data, compared with policy changes, which were most effective for driving the collection of the data in a structured format. Further, Aim 2 also introduced a novel and validated method to extract familial relationships from clinical databases, enabling the inference of family history.

And lastly, studies from Aim 3 used the familial relationships inferred in Aim 2 to

Title	Journal/ Conference	Aim	Co-authors	Status
Quality of Race and Ethnicity Data in Electronic Health Records.	CRI 2016	1.1	**Polubriaginof F**, Boland MR, Perotte A, Vawdrey DK	Published
Challenges with quality of race and ethnicity data in observational databases.	JAMIA	1.1 and 2.1	**Polubriaginof F**, Patrick Ryan, Salmasian H, Perotte A, Safford MM, Hripcsak G, Tatonetti NP, Vawdrey DK	Submitted
An Assessment of Family History Information Captured in an Electronic Health Record.	AMIA 2015	1.2	**Polubriaginof F**, Tatonetti NP, Vawdrey DK	Published
Challenges with Collecting Smoking Status in Electronic Health Records.	AMIA 2017	1.3	**Polubriaginof F**, Salmasian H, Albert DA, Vawdrey DK	Published
Patient-provided Data Improves Race and Ethnicity Data Quality in Electronic Health Records.	AMIA 2016	2.1	**Polubriaginof F**, Salmasian H, Shapiro AW, Prey J, Hripcsak G, Perotte A, Tatonetti NP, Vawdrey DK	Published
Engaging Hospital Patients in the Medication Reconciliation Process Using Tablet Computers.	JAMIA	2.1	Prey JE, **Polubriaginof F**, Grossman LV, Creber RM, Perotte R, Qian M, Restaino S, Bakken S, Hripcsak G, Underwood J, Vawdrey DK	Submitted
An automated method to identify familial relationships in electronic health records.	ASHG 2016	2.2	**Polubriaginof F**, (Quinnies K, Vanguri R), Yahi A, Simmerling M, Ionita-Laza I, Salmasian H, Bakken S, Kiryluk K, Goldstein D, (Vawdrey DK, Tatonetti NP)	Published
Automated Identification of Families in Electronic Health Records to Support Clinical Research.	CRI 2017	2.2	**Polubriaginof F**, (Quinnies K, Vanguri R), Yahi A, Simmerling M, Ionita-Laza I, Salmasian H, Bakken S, Hripcsak G, Goldstein D, Kiryluk K, (Tatonetti NP, Vawdrey DK)	Published
Mining Electronic Health Records to Uncover Drugs that Result in Adverse Fetal Outcomes Following Maternal Exposure.	Scientific Report	2.2	Boland MR, **Polubriaginof F**, Tatonetti NP	Published
Has Meaningful Use improved collection of smoking status information?	CRI 2018	2.3	**Polubriaginof F**, Tariq AA, Tatonetti NP, Vawdrey DK	Published
Using relationships inferred from electronic health records to conduct genetic studies.	ASHG 2017	2.2 and 3.1	**Polubriaginof F**, (Vanguri R, Quinnies K, Belbin G), Yahi A, Salmasian H, Lorberbaum T, Nwankwo V, Li L, Shervey M, Glowe P, Ionita-Laza I, Simmerling M, Hripcsak G, Bakken S, Goldstein D, Kiryluk K, Kenny E, Dudley J, (Vawdrey DK, Tatonetti NP)	Published
Disease heritability estimates using the electronic health records of 9 million patients.	ASHG 2016	2.2 and 3.1	Tatonetti NP, **Polubriaginof F**, Quinnies K, Vanguri R, Yahi A, Simmerling M, Ionita-Laza I, Salmasian H, Bakken S, Kiryluk K, Goldstein D, Vawdrey DK	Published
Estimate of disease heritability using 7.4 million familial relationships inferred from electronic health records.	Cell	2.2 and 3.1	**Polubriaginof F**, (Vanguri R, Quinnies K, Belbin G), Yahi A, Salmasian H, Lorberbaum T, Nwankwo V, Li L, Shervey M, Glowe P, Ionita-Laza I, Simmerling M, Hripcsak G, Bakken S, Goldstein D, Kiryluk K, Kenny E, Dudley J, (Vawdrey DK, Tatonetti NP)	Published
Systematic estimation of the environmental contribution to disease using the electronic health records.	ASHG 2016	3.1	Vanguri RS, **Polubriaginof F**, Quinnies K, Vanguri R, Yahi A, Simmerling M, Ionita-Laza I, Salmasian H, Bakken S, Kiryluk K, Goldstein D, (Vawdrey DK, Tatonetti NP)	Published
Low rates of screening for celiac disease among family members.	DDW 2018	3.2	(Faye AS, **Polubriaginof F**), Green PHR, Vawdrey DK, Tatonetti NP, Lebwohl B	Published
Low Rates of Screening for Celiac Disease Among Family Members; Analysis of Algorithm-Identified Familial Relationships.	Clinical Gastroenterology and Hepatology	3.2	(Faye AS, **Polubriaginof F**), Green PHR, Vawdrey DK, Tatonetti NP, Lebwohl B	Submitted
Low Screening Rates for Diabetes Mellitus Among Family Members of Affected Relatives.	AMIA 2018	3.2	**Polubriaginof F**, Shang N, Hripcsak G, Tatonetti NP, Vawdrey DK	Submitted
Open Data: A review of current methods used for discovery in biomedical research.	BMC Medical Informatics and Decision Making	Background	(**Polubriaginof F**, Romano JD, Yahi A), Vawdrey DK, Tatonetti NP	Submitted

Table 6.1: Summary of publications. Authors in parenthesis contributed equally.

support both genetic and clinical research. These studies demonstrated that the availability of familial data along with clinical data can have a significant impact in multiple research areas.

6.3 Implications for Biomedical Informatics

The work in this dissertation contributed to the field of biomedical informatics in multiple areas, summarized in three main implications for the field.

First, I have shown that while patient-provided data is critical to accomplish multiple clinical tasks, there are data quality concerns that should be addressed when utilizing these data. Data available in the EHR are often incomplete or inaccurate, posing challenges for reuse.

Second, I have shown that the implementation of different intervention types had different impact in the collection and quality of patient-provided data. In general, policy changes resulted in increased data collection of these data types, and patient-facing tools resulted in higher data quality. These results suggest that there should be greater focus on using patient-facing tools when the objective is to increase the quality of this information.

Third, the availability of family history through the use of familial relationships in addition to clinical data can open up new avenues of research, support knowledge discovery, and facilitate the identification of clinical phenotypes.

6.4 Implications for Genetics Research

In this dissertation, I demonstrated the usefulness of utilizing EHR data to conduct large genetic studies in a diverse patient population. The use of EHR data can be used to empower genetic studies by significantly increasing the sample sizes available, with minor costs. Genetic data is a valuable but expensive and not always available resource. The use of EHR data in genetics can expedite research while decreasing cost. The RIFTEHR method can be used to personalize disease risk prediction and facilitate heritability estimation for phenotypes not previously studied in family-based or twin studies.

Further, the use of these data allowed for genetic research in multiple racial and ethnic groups, demonstrating the utility of using EHR data in conjunction with traditional genetic research data. Traditional genetic studies often focus on a single racial group, limiting the generalizability of its findings to other populations. The use of EHR data allows for studies to include other racial and ethnic groups, without impacting the research cost. The ability to conduct genetic research on multiple racial and ethnic groups at once will help us achieve the goals of precision medicine.

6.5 Implications for Clinical Care

In addition to the contributions and implications described above, my dissertation also impacts clinical care. First, my results indicate that patients are willing to participate in their care by reviewing or providing information, suggesting that providers need to encourage and engage their patients in sharing relevant information. Second, using RIFTEHR, I identified that disease screening rates among high-risk individuals were low. Future efforts should focus on ways to improve the identification of high-risk individuals by incorporating family history and other patient-provided information at the point of care.

6.6 Limitations

The work presented has many limitations. Many of the studies were conducted in a single institution, a large urban academic medical center. As such, the research findings may not be generalizable to other institutions. Additionally, one of the studies involved patient recruitment, and only included English-speaking participants, in addition to small sample size due to recruitment constraints. Overall, my studies focused on just three types of patient-provided data. However, there are many other types of patient-provided information. Other types of information may pose different data quality challenges compared with the data types included in this dissertation. Some of the studies focusing on data quality were not able to assess correctness of the data due to unavailability of data from the reference standard. Further, all reported studies heavily relied on EHR data, and therefore faced challenges related to fragmentation of care. Patients often seek care in multiple healthcare systems, resulting in data missingness which may impact the results of the studies reported.

6.7 Future Work

This dissertation can be expanded into several areas of future work. First, this dissertation built the foundational work to better inform how to capture patient-provided data. Additional work should be conducted to improve the quality of these data in clinical databases. These data currently face considerable data quality issues, as shown in Aim 1. Future interventions could use the work presented here to identify approaches that can potentially improve the quality of these data. Based on this work, patient-facing tools could greatly improve the quality of patient-provided data in EHRs, while decreasing clinician burden during clinical encounters. Future work could build on this finding by developing and deploying patient-facing tools to capture a collection of patient-provided information relevant to clinical care. Additionally, the work presented in Aim 2 exhibited multiple methods to assess the impact of different types of interventions. Future work could use similar methods to measure outcomes after the implementation of an intervention, allowing for rigorous evaluation of the intervention at hand, and directly assessing the impact of the data quality.

Second, future work should leverage the RIFTEHR method to power numerous research studies. Availability of family data in conjunction with rich clinical data is a powerful combination to support not only clinical studies but also clinical care. As demonstrated in Aim 3, the use of this data can generate new knowledge, such as estimation of disease heritability for diseases that have not previously explored and in populations that had not been studied. This work could be used to generate new hypotheses that could subsequently be tested using a traditional study design. While this dissertation has demonstrated one use case of these data in genetic research, there are many more opportunities in other fields as well. For

example, in this dissertation screening rates among high-risk individuals was explored for two conditions, and it demonstrated that there are challenges to identifying these patients during a clinical visit. Future work should investigate methods to use informatics solutions to facilitate the identification of high-risk individuals at the point of care to mitigate barriers to identifying high-risk patients. Such efforts could potentially increase adherence to clinical guidelines, provide a more individualized disease management plan, and ultimately decrease patient morbidity.

Third, in this dissertation, the identification of familial relationships was performed in three institutions, independent from each other. Patients often receive medical care at more than one institution. Future work should take advantage of health information exchange efforts, to identify familial relationships broadly. This approach will potentially reduce the challenges we currently face with the fragmentation of care, enabling robust and complete population studies. One of the major biases that was accounted for in Aim 3 was ascertainment bias, which affected the estimated heritability of disease. Linking the medical histories of patients from several different institutions (for example, through health information exchange processes) may provide a more holistic assessment of the patient's diseases and disease states.

6.8 Conclusion

Patient-provided information is needed to advance precision medicine, by enabling clinicians to provide more individualized disease screening and diagnosis as well as care management. The goal of this dissertation was to develop a better understanding of how patient-provided information in the EHR could facilitate the identification of patients at increased risk for developing disease. The studies included in this dissertation found data quality concerns for patient-provided information, and that different interventions could lead to increased collection and/or increased quality of patient-provided information. In general, allowing patients to review or directly supply patient-provided information resulted in the most complete and highest quality information. In the absence of patients providing information themselves, informatics methods, such as RIFTEHR, can be utilized to infer certain patient-provided information, such as family medical history. The use of inference methods unlocks new knowledge, such as disease heritability for multiple races and ethnicities, and enables assessment of adherence to guidelines for high-risk patients, such as those for diabetes mellitus or celiac disease. In conclusion, this dissertation outlines the data quality issues that exists for patient-provided information, how to overcome these data quality issues, and how to apply patient-provided information to generate new knowledge.

This page intentionally left blank.

Bibliography

Adams, Samantha A and Carolyn Petersen (2016). "Precision medicine: opportunities, possibilities, and challenges for patients and providers." In: *Journal of the American Medical Informatics Association*, ocv215–4.

Adler, Nancy E and William W Stead (2015). "Patients in context—EHR capture of social and behavioral determinants of health." In: *New England Journal of Medicine* 372.8, pp. 698–701.

Ahmad, Faraz S et al. (2017). "Validity of Cardiovascular Data From Electronic Sources:The Multi-Ethnic Study of Atherosclerosis and HealthLNK." In: *Circulation*, CIRCULATIONAHA.117.027436–52.

Almasy, L and J Blangero (1998). "Multipoint quantitative-trait linkage analysis in general pedigrees." In: *American journal of human genetics* 62.5, pp. 1198–1211.

Almgren, P et al. (2011). "Heritability and familiality of type 2 diabetes and related quantitative traits in the Botnia Study." In: *Diabetologia* 54.11, pp. 2811–2819.

An, JaeJin et al. (2018). "Adherence to the American Diabetes Association retinal screening guidelines for population with diabetes in the United States." In: *Ophthalmic epidemiology* 25.3, pp. 257–265.

Arar, Nedal et al. (2011). "Veterans' experience in using the online Surgeon General's family health history tool." In: *Personalized medicine* 8.5, pp. 523–532.

Aronsky, D and P J Haug (2000). "Assessing the quality of clinical data in a computer-based record for calculating the pneumonia severity index." In: *Journal of the American Medical Informatics Association* 7.1, pp. 55–65.

Aronson, Samuel J and Heidi L Rehm (2015). "Building the foundation for genomics in precision medicine." In: *Nature* 526.7573, pp. 336–342.

Arsoniadis, E G et al. (2015). "Characterizing Patient-Generated Clinical Data and Associated Implications for Electronic Health Records." In: *Stud Health Technol ...*.

Arts, Daniëlle et al. (2002). "Quality of data collected for severity of illness scores in the Dutch National Intensive Care Evaluation (NICE) registry." In: *Intensive care medicine* 28.5, pp. 656–659.

Bach, Peter B et al. (2004). "Primary care physicians who treat blacks and whites." In: *The New England journal of medicine* 351.6, pp. 575–584.

Baker, David W et al. (2007). "Attitudes toward health care providers, collecting information about patients' race, ethnicity, and language." In: *Med Care* 45.11, pp. 1034–1042.

Ball, M J and J Lillis (2001). "E-health: transforming the physician/patient relationship." In: *International Journal of Medical Informatics* 61.1, pp. 1–10.

Bardes, C L (2012). "Defining "patient-centered medicine"." In: *New England Journal of Medicine* 366.9, pp. 782–783.

Basch, Ethan (2010). "The Missing Voice of Patients in Drug-Safety Reporting." In: *New England Journal of Medicine* 362.10, pp. 865–869.

Basch, Ethan et al. (2009). "Adverse symptom event reporting by patients vs clinicians: relationships with clinical outcomes." In: *Journal of the National Cancer Institute* 101.23, pp. 1624–1632.

Basch, Ethan et al. (2017). "Overall survival results of a trial assessing patient-reported outcomes for symptom monitoring during routine cancer treatment." In: *Jama* 318.2, pp. 197–198.

Baumgart, Leigh A, Kristen J Vogel Postula, and William A Knaus (2015). "Initial clinical validation of Health Heritage, a patient-facing tool for personal and family history collection and cancer risk assessment." In: *Familial Cancer* 15.2, pp. 331–339.

Bederson, J B et al. (2000). "Recommendations for the management of patients with unruptured intracranial aneurysms - A statement for healthcare professionals from the Stroke Council of the American Heart Association." In: *Circulation* 102.18, pp. 2300–2308.

Benson, K and A J Hartz (2000). "A comparison of observational studies and randomized, controlled trials." In: *New England Journal of Medicine* 342.25, pp. 1878–1886.

Berg, Alfred O et al. (2009). "National Institutes of Health State-of-the-Science Conference Statement: Family History and Improving Health." In: *Annals of internal medicine*. University of Washington, Group Health Research Institute, Seattle, Washington, USA., pp. 872–877.

Berger, Marc L et al. (2009). "Good research practices for comparative effectiveness research: defining, reporting and interpreting nonrandomized studies of treatment effects

using secondary data sources: the ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report–Part I." In: *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research* 12.8, pp. 1044–1052.

Bernhardt, B A and R E Pyeritz (1989). "The economics of clinical genetics services. III. Cognitive genetics services are not self-supporting." In: *American journal of human genetics* 44.2, pp. 288–293.

Berry, Carolyn et al. (2014). "Moving to patient reported collection of race and ethnicity data: implementation and impact in ten hospitals." In: *International Journal of Health Care Quality Assurance* 27.4, pp. 271–283.

Berry, D A et al. (1997). "Probability of carrying a mutation of breast-ovarian cancer gene BRCA1 based on family history." In: *Journal of the National Cancer Institute* 89.3, pp. 227–238.

Bhalla, Rohit, Brandon G Yongue, and Brian P Currie (2012). "Standardizing race, ethnicity, and preferred language data collection in hospital information systems: results and implications for healthcare delivery and policy." In: *Journal for healthcare quality : official publication of the National Association for Healthcare Quality* 34.2, pp. 44–52.

Bill, Robert et al. (2014). "Automated extraction of family history information from clinical notes." In: *AMIA Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* 2014, pp. 1709–1717.

Birkhead, Guthrie S, Michael Klompas, and Nirav R Shah (2015). "Uses of electronic health records for public health surveillance to advance public health." In: *Annual review of public health* 36, pp. 345–359.

Blumenthal, D and M Tavenner (2010). "The "Meaningful Use" Regulation for Electronic Health Records." In: *New England Journal of Medicine* 363.6, pp. 501–504.

Blumenthal, David (2009). "Stimulating the adoption of health information technology." In: *The New England journal of medicine* 360.15, pp. 1477–1479.

Blustein, J (1994). "The reliability of racial classifications in hospital discharge abstract data." In: *American journal of public health* 84.6, pp. 1018–1021.

Boland, Mary Regina et al. (2015). "Birth month affects lifetime disease risk: a phenome-wide method." In: *J Am Med Inform Assoc* 22.5, pp. 1042–1053.

Booth, Helen P, A Toby Prevost, and Martin C Gulliford (2013). "Validity of smoking prevalence estimates from primary care electronic health records compared with na-

tional population survey data for England, 2007 to 2011." In: *Pharmacoepidemiology and drug safety* 22.12, pp. 1357–1361.

Boyle, Raymond, Leif Solberg, and Michael Fiore (2014). "Use of electronic health records to support smoking cessation." In: *Cochrane database of systematic reviews* 12.12, pp. CD008743–CD008743.

Brennan, P F and W W Stead (2000). "Assessing data quality: from concordance, through correctness and completeness, to valid manipulatable representations." In: *Journal of the American Medical Informatics Association* 7.1, pp. 106–107.

Bronfenbrenner, U and S J Ceci (1994). "Nature-nurture reconceptualized in developmental perspective: a bioecological model." In: *Psychological review* 101.4, pp. 568–586.

Brown, Jeffrey S, Michael Kahn, and Sengwee Toh (2013). "Data quality assessment for comparative effectiveness research in distributed data networks." In: *Med Care* 51.8 Suppl 3, S22–9.

Buntin, Melinda B and John Z Ayanian (2017). "Social Risk Factors and Equity in Medicare Payment." In: *The New England journal of medicine* 376.6, pp. 507–510.

Caligtan, Christine A et al. (2012). "Bedside information technology to support patient-centered care." In: *Int J Med Inform* 81.7, pp. 442–451.

Caplan, Lee, Charlotte Stout, and Daniel S Blumenthal (2011). "Training physicians to do office-based smoking cessation increases adherence to PHS guidelines." In: *Journal of community health* 36.2, pp. 238–243.

Castellani, Carlo et al. (2009). "Association between carrier screening and incidence of cystic fibrosis." In: *JAMA* 302.23, pp. 2573–2579.

Centers for Medicare & Medicaid Services. *CMS Electronic Health Records Incentive Programs*. URL: `https : / / www . cms . gov / Regulations - and - Guidance / Legislation / EHRIncentivePrograms / index . html ? redirect = /ehrincentiveprograms`.

— (2010). *Eligible Professional Meaningful Use Core Measures Measure 9 of 15*. Tech. rep. URL: `https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/downloads/9RecordSmokingStatus.pdf`.

— (2014a). *Eligible Hospital and Critical Access Hospital Meaningful Use Menu Set Measures Measure 4 of 6, Family Health History*. Tech. rep.

— (2014b). *Eligible Professional Meaningful Use Core Measures Measure 7 of 13*. Tech. rep.

Chakkalakal, Rosette J et al. (2015). "Standardized Data Collection Practices and the Racial/Ethnic Distribution of Hospitalized Patients." In: *Med Care* 53.8, pp. 666–672.

Chatterjee, Nilanjan, Jianxin Shi, and Montserrat García-Closas (2016). "Developing and evaluating polygenic risk prediction models for stratified disease prevention." In: *Nature Reviews Genetics* 17.7, pp. 392–406.

Chaudhry, Basit et al. (2006). "Systematic Review: Impact of Health Information Technology on Quality, Efficiency, and Costs of Medical Care." In: *Annals of Internal Medicine* 144.10, pp. 742–752.

Chen, Elizabeth S et al. (2012). "Characterizing the use and contents of free-text family history comments in the Electronic Health Record." In: *AMIA Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* 2012, pp. 85–92.

Chen, Elizabeth S et al. (2014). "Multi-source development of an integrated model for family health history." In: *J Am Med Inform Assoc* 22.e1, e67–80.

Chin, Marshall H (2015). "Using patient race, ethnicity, and language data to achieve health equity." In: *Journal of General Internal Medicine* 30.6, pp. 703–705.

Cimino, J J, V L Patel, and A W Kushniruk (2001). "What do patients do with access to their medical records?" In: *Studies in health technology and informatics* 84.Pt 2, pp. 1440–1444.

Claus, E B, N Risch, and W D Thompson (1994). "Autosomal dominant inheritance of early-onset breast cancer. Implications for risk prediction." In: *Cancer.* 73.3, pp. 643–651.

Cohn, W F et al. (2010). "Health Heritage© a web-based tool for the collection and assessment of family health history: initial user experience and analytic validity." In: *Public health genomics* 13.7-8, pp. 477–491.

Collins, Francis S and Harold Varmus (2015). "A New Initiative on Precision Medicine." In: *dx.doi.org* 372.9, pp. 793–795.

Collins, Sarah A et al. (2011). "Policies for patient access to clinical data via PHRs: current state and recommendations." In: *Journal of the American Medical Informatics Association* 18.Supplement$_1$, pp. i2–i7.

Committee on Accounting for Socioeconomic Status in Medicare Payment Programs et al. (2016). "Accounting for Social Risk Factors in Medicare Payment: Identifying Social Risk Factors." In:

Committee on the Recommended Social and Behavioral Domains and Measures for Electronic Health Records, Board on Population Health and Public Health Practice, and Institute of, Medicine (2015). "Capturing Social and Behavioral Domains and Measures in Electronic Health Records: Phase 2." In:

Concato, J, N Shah, and R I Horwitz (2000). "Randomized, controlled trials, observational studies, and the hierarchy of research designs." In: *New England Journal of Medicine* 342.25, pp. 1887–1892.

Conn, Joseph (2016). "Hospitals achieve 96% EHR adoption rate; data exchange still needs work." In: *Modern Healthcare*.

Coopey, Suzanne B et al. (2012). "The role of chemoprevention in modifying the risk of breast cancer in women with atypical breast lesions." In: *Breast cancer research and treatment* 136.3, pp. 627–633.

Cusack, C M et al. (2013). "The future state of clinical data capture and documentation: a report from AMIA's 2011 Policy Meeting." In: *Journal of the American Medical Informatics Association* 20.1, pp. 134–140.

Cutting, Garry R (2010). "Modifier genes in Mendelian disorders: the example of cystic fibrosis." In: *Annals of the New York Academy of Sciences* 1214.1, pp. 57–69.

Davis Giardina, Traber et al. (2014). "Patient access to medical records and healthcare outcomes: a systematic review." In: *J Am Med Inform Assoc* 21.4, pp. 737–741.

Dawber, T R, G F Meadors, and F E Moore (1951). "Epidemiological approaches to heart disease: the Framingham Study." In: *American journal of public health and the nation's health* 41.3, pp. 279–281.

Delbanco, Tom et al. (2010). "Open notes: doctors and patients signing on." In: *Ann Intern Med* 153.2, pp. 121–125.

Disease Control, Centers for and Prevention (2007). *National Health and Nutrition Examination Survey (NHANES).* Vol. 2007. Hyattsville, MD: ….

Dorsey, Rashida et al. (2014). "Implementing Health Reform: Improved Data Collection and the Monitoring of Health Disparities." In: *Annual review of public health* 35.1, pp. 123–138.

Douglas, Megan Daugherty et al. (2015). "Missed policy opportunities to advance health equity by recording demographic data in electronic health records." In: *American journal of public health* 105 Suppl 3.S3, S380–8.

Drennan, Kathleen B (2002). "Patient recruitment: the costly and growing bottleneck in drug development." In: *Drug discovery today* 7.3, pp. 167–170.

Dudley, Joel T, Tarangini Deshpande, and Atul J Butte (2011). "Exploiting drug-disease relationships for computational drug repositioning." In: *Briefings in bioinformatics* 12.4, pp. 303–311.

Dullabh, Prashila et al. (2014). "How Patients Can Improve the Accuracy of their Medical Records." In: *eGEMs (Generating Evidence & Methods to improve patient outcomes)* 2.3, pp. 1–15.

Dupuis, Annie et al. (2005). "Cystic Fibrosis Birth Rates in Canada: A Decreasing Trend since the Onset of Genetic Testing." In: *The Journal of pediatrics* 147.3, pp. 312–315.

Dwamena, F and M Holmes Rovner (2012). "Interventions for providers to promote a patient-centred approach in clinical consultations." In: *The Cochrane ...*.

Eisen, S, W True, and J Goldberg (1987). "The Vietnam era twin (VET) registry: method of construction." In: *Acta geneticae ...* 36.01, pp. 61–66.

Emery, J (1999). "Computer support for genetic advice in primary care." In: *The British journal of general practice : the journal of the Royal College of General Practitioners* 49.444, pp. 572–575.

Emery, J et al. (1999). "Computer support for recording and interpreting family histories of breast and ovarian cancer in primary care (RAGs): qualitative evaluation with simulated patients." In: *BMJ (Clinical research ed.)* 319.7201, pp. 32–36.

Engle, R L (1991). "The evolution, uses, and present problems of the patient's medical record as exemplified by the records of the New York Hospital from 1793 to the present." In: *Transactions of the American Clinical and Climatological Association* 102, 182–9–discussion 189–92.

Epstein, R M et al. (2010). "Why The Nation Needs A Policy Push On Patient-Centered Health Care." In: *Health Affairs* 29.8, pp. 1489–1495.

Faber, T et al. (2017). "Effect of tobacco control policies on perinatal and child health: a systematic review and meta-analysis." In: *The Lancet Public ...*.

Facio, Flavia M et al. (2010). "Validation of My Family Health Portrait for six common heritable conditions." In: *Genetics in Medicine* 12.6, pp. 370–375.

Feero, W Gregory (2013). "Connecting the Dots Between Patient-Completed Family Health History and the Electronic Health Record." In: *Journal of General Internal Medicine* 28.12, pp. 1547–1548.

Finkelstein, Joel B (2006). "E-prescribing first step to improved safety." In: *Journal of the National Cancer Institute* 98.24, pp. 1763–1765.

Frezzo, Theresa M et al. (2003). "The genetic family history as a risk assessment tool in internal medicine." In: *Genetics in Medicine* 5.2, pp. 84–91.

Friedman, C, G Hripcsak, and L Shagina (1999). "Representing information in patient reports using natural language processing and the extensible markup language." In: *Journal of the American ...*.

Gail, M H et al. (1989). "Projecting individualized probabilities of developing breast cancer for white females who are being examined annually." In: *Journal of the National Cancer Institute* 81.24, pp. 1879–1886.

Ge, Tian et al. (2017). "Phenome-wide heritability analysis of the UK Biobank." In: *PLoS Genetics* 13.4, e1006711–21.

Ginsburg, G S and H F Willard (2009). "Genomic and personalized medicine: foundations and applications." In: *Translational research : the journal of laboratory and clinical medicine* 154.6, pp. 277–287.

Giovanni, Monica A and Michael F Murray (2010). "The application of computer-based tools in obtaining the genetic family history." In: *Current protocols in human genetics* Chapter 9, Unit 9.21.

Gomez, Scarlett L and Sally L Glaser (2006). "Misclassification of race/ethnicity in a Population-based Cancer Registry (United States)." In: *Cancer Causes & Control* 17.6, pp. 771–781.

Gorber, Sarah Connor et al. (2009). "The accuracy of self-reported smoking: a systematic review of the relationship between self-reported and cotinine-assessed smoking status." In: *Nicotine & tobacco research* 11.1, pp. 12–24.

Green, R F (2007). "Summary of workgroup meeting on use of family history information in pediatric primary care and public health." In: *Pediatrics* 120 Suppl 2.Supplement, 100–S100.

Greenfield, S et al. (1988). "Patients' participation in medical care: effects on blood sugar control and quality of life in diabetes." In: *Journal of General Internal Medicine* 3.5, pp. 448–457.

Greenhalgh, Trisha et al. (2008). "Patients' attitudes to the summary care record and HealthSpace: qualitative study." In: *BMJ (Clinical research ed.)* 336.7656, pp. 1290–1295.

Grossman, Lisa V et al. (2017). “Implementation of acute care patient portals: recommendations on utility and use from six early adopters.” In: *J Am Med Inform Assoc* 136.3, p. 327.

Guttmacher, Alan E and Francis S Collins (2003). “Welcome to the genomic era.” In: *The New England journal of medicine* 349.10, pp. 996–998.

Guttmacher, Alan E, Francis S Collins, and Richard H Carmona (2004). “The family history–more important than ever.” In: *The New England journal of medicine* 351.22, pp. 2333–2336.

HL7 Version 3 Implementation Guide: Family History/Pedigree Interoperability, Release 1 - US Realm.

Hack, T F, L F Degner, and D G Dyck (1994). “Relationship between preferences for decisional control and illness information among women with breast cancer: a quantitative and qualitative analysis.” In: *Soc Sci Med* 39.2, pp. 279–289.

Hagland, M (2011). “Balancing act: can CMIOs and CIOs make physician documentation work for everyone?” In: *Healthcare informatics: the business magazine for ...*.

Halamka, John, Kenneth Mandl, and Paul Tang (2008). “Early experiences with personal health records.” In: *Journal of the American Medical Informatics Association* 15.1, pp. 1–7.

Hamilton, Natia S et al. (2009). “Concordance between self-reported race/ethnicity and that recorded in a Veteran Affairs electronic medical record.” In: *North Carolina medical journal* 70.4, pp. 296–300.

Hammond, W E et al. (1980). “Functional characteristics of a computerized medical record.” In: *Methods Inf Med* 19.3, pp. 157–162.

Harden, K Paige, Eric Turkheimer, and John C Loehlin (2007). “Genotype by environment interaction in adolescents’ cognitive aptitude.” In: *Behavior Genetics* 37.2, pp. 273–283.

Hasan, Sharique and Rema Padman (2006). “Analyzing the effect of data quality on the accuracy of clinical decision support systems: a computer simulation approach.” In: *AMIA Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pp. 324–328.

Hasnain-Wynia, R, D Pierce, and M A Pittman (2004). *Who, What, When, Where: The Current State of Data of Collection on Race and Ethnicity in Hospitals*. New York.

Hassol, Andrea et al. (2004). "Patient experiences and attitudes about access to a patient electronic health care record and linked web messaging." In: *Journal of the American Medical Informatics Association* 11.6, pp. 505–513.

Health Information Technology, The Office of the National Coordinator for (2017). *Health IT Quick Stats*. URL: https://dashboard.healthit.gov/quickstats/quickstats.php.

HealthData.gov. URL: https://www.healthdata.gov.

Healthcare Cost and Utilization Project (HCUP). URL: https://www.hcup-us.ahrq.gov.

Helfand, Mark and Susan Carson (2008). "Screening for Lipid Disorders in Adults: Selective Update of 2001 US Preventive Services Task Force Review." In:

Hemani, Gibran et al. (2013). "Inference of the genetic architecture underlying BMI and height with the use of 20,240 sibling pairs." In: *American journal of human genetics* 93.5, pp. 865–875.

Hersh, W R et al. (2013). "Caveats for the use of operational electronic health record data in comparative effectiveness research." In: *Med Care* 51.8 Suppl 3, S30–7.

Hirsch, Bradford R and Amy P Abernethy (2013). "Incorporating the Patient's Voice in the Continuum of Care." In: *Journal of the National Comprehensive Cancer Network : JNCCN* 11.1, pp. 116–118.

Hogan, W R and M M Wagner (1997). "Accuracy of data in computer-based patient records." In: *Journal of the American Medical Informatics Association* 4.5, pp. 342–355.

Hospital Consumer Assessment of Healthcare Providers and System. URL: http://hcahpsonline.org/home.asp.

Hoyt, Robert et al. (2013). "Digital family histories for data mining." In: *Perspectives in health information management* 10, 1a.

Hripcsak, G et al. (2011a). "Use of electronic clinical documentation: time spent and team interactions." In: *Journal of the American Medical Informatics Association* 18.2, pp. 112–117.

Hripcsak, George and David J Albers (2013). "Next-generation phenotyping of electronic health records." In: *J Am Med Inform Assoc* 20.1, pp. 117–121.

Hripcsak, George et al. (2011b). “Bias associated with mining electronic health records.” In: *Journal of biomedical discovery and collaboration* 6.0, pp. 48–52.

Hripcsak, George et al. (2015). “Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers.” In: *Studies in health technology and informatics* 216, pp. 574–578.

Hripcsak, George et al. (2016). “Characterizing treatment pathways at scale using the OHDSI network.” In: *Proceedings of the National Academy of Sciences of the United States of America* 113.27, pp. 7329–7336.

Hulse, Nathan C et al. (2010). “Deriving consumer-facing disease concepts for family health histories using multi-source sampling.” In: *Journal of Biomedical Informatics* 43.5, pp. 716–724.

Hulse, Nathan C et al. (2011). “Development and early usage patterns of a consumer-facing family health history tool.” In: *AMIA Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* 2011, pp. 578–587.

Institute of Medicine (US) Forum on Drug Discovery, Development, and Translation (2010). “Transforming Clinical Research in the United States: Challenges and Opportunities: Workshop Summary.” In:

Jamal, Ahmed et al. (2016). “Current Cigarette Smoking Among Adults - United States, 2005-2015.” In: *MMWR. Morbidity and mortality weekly report* 65.44, pp. 1205–1211.

Jemal, Ahmedin and Stacey A Fedewa (2017). “Lung Cancer Screening With Low-Dose Computed Tomography in the United States—2010 to 2015.” In: *JAMA oncology* 3.9, pp. 1278–1281.

Jones, R B, S M McGhee, and D McGhee (1992). “Patient on-line access to medical records in general practice.” In: *Health bulletin* 50.2, pp. 143–150.

Kaelber, David C et al. (2008). “A research agenda for personal health records (PHRs).” In: *Journal of the American Medical Informatics Association* 15.6, pp. 729–736.

Kahn, Michael G, Brian B Eliason, and Janet Bathurst (2010). “Quantifying clinical data quality using relative gold standards.” In: *AMIA Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* 2010, pp. 356–360.

Kaplan, Judith B (2014). “The Quality of Data on “Race” and “Ethnicity”: Implications for Health Researchers, Policy Makers, and Practitioners.” In: *Race and Social Problems* 6.3, pp. 214–236.

Kaplan, Robert M, David A Chambers, and Russell E Glasgow (2014). "Big Data and Large Sample Size: A Cautionary Note on the Potential for Bias." In: *Clinical and Translational Science* 7.4, pp. 342–346.

Kaplan, S H et al. (1995). "Patient and visit characteristics related to physicians' participatory decision-making style. Results from the Medical Outcomes Study." In: *Med Care* 33.12, pp. 1176–1187.

Khunti, Kamlesh et al. (2018). "Achievement of guideline targets for blood pressure, lipid, and glycaemic control in type 2 diabetes: A meta-analysis." In: *Diabetes research and clinical practice* 137, pp. 137–148.

Klinger, Elissa V et al. (2015). "Accuracy of race, ethnicity, and language preference in an electronic health record." In: *Journal of General Internal Medicine* 30.6, pp. 719–723.

Kogut, Stephen Jon et al. (2014). "Improving medication management after a hospitalization with pharmacist home visits and electronic personal health records: an observational study." In: *Drug, healthcare and patient safety* 6, pp. 1–6.

Kohane, Isaac S (2011). "Using electronic health records to drive discovery in disease genomics." In: *Nature Reviews Genetics* 12.6, pp. 417–428.

Kressin, Nancy R (2015). "Race/Ethnicity Identification: Vital for Disparities Research, Quality Improvement, and Much More Than "Meets the Eye"." In: *Med Care* 53.8, pp. 663–665.

LaVeist, Thomas A, Darrell Gaskin, and Patrick Richard (2011). "Estimating the economic burden of racial health inequalities in the United States." In: *International journal of health services : planning, administration, evaluation* 41.2, pp. 231–238.

Lee, Simon J Craddock, James E Grobe, and Jasmin A Tiro (2015). "Assessing race and ethnicity data quality across cancer registries and EMRs in two hospitals." In: *J Am Med Inform Assoc* 23.3, pp. 627–634.

Lei, J van der (1991). "Use and abuse of computer-stored medical records." In: *Methods Inf Med* 30.2, pp. 79–80.

Leveille, Suzanne G et al. (2012). "Evaluating the impact of patients' online access to doctors' visit notes: designing and executing the OpenNotes project." In: *BMC medical informatics and decision making* 12.1, p. 32.

Levey, Andrew S et al. (2009). "A New Equation to Estimate Glomerular Filtration Rate." In: *Annals of Internal Medicine* 150.9, pp. 604–612.

Li, Li et al. (2015). "Identification of type 2 diabetes subgroups through topological analysis of patient similarity." In: *Science translational medicine* 7.311, 311ra174–311ra174.

Lichtenstein, Paul et al. (2009). "Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study." In: *Lancet (London, England)* 373.9659, pp. 234–239.

Locatelli, Isabella et al. (2007). "A correlated frailty model with long-term survivors for estimating the heritability of breast cancer." In: *Statistics in medicine* 26.20, pp. 3722–3734.

Lorberbaum, Tal et al. (2016a). "An Integrative Data Science Pipeline to Identify Novel Drug Interactions that Prolong the QT Interval." In: *Drug safety : an international journal of medical toxicology and drug experience* 39.5, pp. 433–441.

Lorberbaum, Tal et al. (2016b). "Coupling Data Mining and Laboratory Experiments to Discover Drug Interactions Causing QT Prolongation." In: *Journal of the American College of Cardiology* 68.16, pp. 1756–1764.

Madigan, David et al. (2014). "A Systematic Statistical Approach to Evaluating Evidence from Observational Studies." In: *Annual Review of Statistics and Its Application* 1.1, pp. 11–39.

Maher, Brendan (2008). "Personal genomes: The case of the missing heritability." In: *Nature* 456.7218, pp. 18–21.

Maher, Molly et al. (2015). "A Novel Health Information Technology Communication System to Increase Caregiver Activation in the Context of Hospital-Based Pediatric Hematopoietic Cell Transplantation: A Pilot Study." In: *JMIR research protocols* 4.4, e119.

Maher, Molly et al. (2016). "User-Centered Design Groups to Engage Patients and Caregivers with a Personalized Health Information Technology Tool." In: *Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation* 22.2, pp. 349–358.

Markus, Hazel Rose (2008). "Pride, prejudice, and ambivalence: toward a unified theory of race and ethnicity." In: *The American psychologist* 63.8, pp. 651–670.

Martin, Erika G, Natalie Helbig, and Nirav R Shah (2014). "Liberating data to transform health care: New York's open data experience." In: *JAMA* 311.24, pp. 2481–2482.

Masterson Creber, Ruth et al. (2016). "Engaging hospitalized patients in clinical care: Study protocol for a pragmatic randomized controlled trial." In: *Contemporary clinical trials* 47, pp. 165–171.

Mayer, John et al. (2014). “Use of an electronic medical record to create the marshfield clinic twin/multiple birth cohort.” In: *Genetic epidemiology* 38.8, pp. 692–698.

Medicine, U S National Library of. *U S National Library of Medicine, Precision medicine*. URL: https://ghr.nlm.nih.gov/primer/precisionmedicine/definition.

Melton, Genevieve B et al. (2010). “Evaluation of family history information within clinical documents and adequacy of HL7 clinical statement and clinical genomics family history models for its representation: a case report.” In: *Journal of the American Medical Informatics Association* 17.3, pp. 337–340.

Moscou, Susan et al. (2003). “Validity of racial/ethnic classifications in medical records data: an exploratory study.” In: *American journal of public health* 93.7, pp. 1084–1086.

Moyer, Virginia A and U.S. Preventive Services Task Force (2014). *Risk assessment, genetic counseling, and genetic testing for BRCA-related cancer in women: U.S. Preventive Services Task Force recommendation statement.*

Mucci, Lorelei A et al. (2016). “Familial Risk and Heritability of Cancer Among Twins in Nordic Countries.” In: *JAMA* 315.1, pp. 68–76.

Murabito, J M et al. (2001). “Family breast cancer history and mammography: Framingham Offspring Study.” In: *American Journal of Epidemiology* 154.10, pp. 916–923.

Murphy, Shawn N et al. (2010). “Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2).” In: *J Am Med Inform Assoc* 17.2, pp. 124–130.

Murray, Michael F et al. (2013). “Comparing Electronic Health Record Portals to Obtain Patient-Entered Family Health History in Primary Care.” In: *Journal of General Internal Medicine* 28.12, pp. 1558–1564.

My Family Health Portrait. URL: https://familyhistory.hhs.gov/FHH/html/index.html.

National Center for Chronic Disease Prevention and Health Promotion Office on Smoking and Health (2014). “The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General.” In:

Nazi, K M et al. (2010). “Embracing a health services research perspective on personal health records: lessons learned from the VA My HealtheVet system.” In: *Journal of General Internal Medicine* 25 Suppl 1.S1, pp. 62–67.

Nazi, Kim M et al. (2015). “VA OpenNotes: exploring the experiences of early patient adopters with access to clinical notes.” In: *J Am Med Inform Assoc* 22.2, pp. 380–389.

Nelson, Nancy C et al. (2005). "Detection and Prevention of Medication Errors Using Real-Time Bedside Nurse Charting." In: *Journal of the American Medical Informatics Association* 12.4, pp. 390–397.

Ng, P C et al. (2008). "Individual Genomes Instead of Race for Personalized Medicine." In: *Clinical Pharmacology & Therapeutics* 84.3, pp. 306–309.

O'Leary, K J et al. (2015). "The effect of tablet computers with a mobile patient portal application on hospitalized patients' knowledge and activation." In: *Journal of the American Medical Informatics Association*, pp. 1–7.

OMB (1997). "Standards for Maintaining, Collecting, and Presenting Federal Data on Race and Ethnicity." In: *Statistical Policy Directive No. 15*.

Open Government Directive. URL: `https://www.fcc.gov/general/open-government-directive`.

Optum Data Assets. URL: `https://www.optum.com/content/dam/optum/resources/productSheets/5302_Data_Assets_Chart_Sheet_ISPOR.pdf`.

Orlando, Lori A et al. (2011). "Protocol for implementation of family health history collection and decision support into primary care using a computerized family health history system." In: *BMC health services research* 11.1, p. 264.

Orlando, Lori A et al. (2013). "Development and validation of a primary care-based family health history and decision support program (MeTree)." In: *North Carolina medical journal* 74.4, pp. 287–296.

Otte-Trojel, Terese et al. (2014). "How outcomes are achieved through patient portals: a realist review." In: *J Am Med Inform Assoc* 21.4, pp. 751–757.

Oude Wesselink, Sandra F et al. (2015). "Guideline adherence and health outcomes in diabetes mellitus type 2 patients: a cross-sectional study." In: *BMC health services research* 15, p. 22.

Owens, Kailey M et al. (2011). "Clinical use of the Surgeon General's "My Family Health Portrait" (MFHP) tool: opinions of future health care providers." In: *Journal of genetic counseling* 20.5, pp. 510–525.

Ozanne, E M et al. (2009). "Identification and management of women at high risk for hereditary breast/ovarian cancer syndrome." In: *Breast J* 15.2, pp. 155–162.

Ozanne, E M et al. (2013). "Which Risk Model to Use? Clinical Implications of the ACS MRI Screening Guidelines." In: *Cancer Epidemiology Biomarkers & Prevention* 22.1, pp. 146–149.

Pakhomov, Serguei V et al. (2008). "Agreement between patient-reported symptoms and their documentation in the medical record." In: *The American journal of managed care* 14.8, pp. 530–539.

Patrick, D L et al. (1994). "The validity of self-reported smoking: a review and meta-analysis." In: *American journal of public health* 84.7, pp. 1086–1093.

Patwardhan, Vilas et al. (2011). "Hepatocellular carcinoma screening rates vary by etiology of cirrhosis and involvement of gastrointestinal sub-specialists." In: *Digestive diseases and sciences* 56.11, pp. 3316–3322.

Peace, Jane, William Bisanar, and Nathan Licht (2012). "Will family health history tools work for complex families? Scenario-based testing of a web-based consumer application." In: *AMIA Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* 2012, pp. 709–715.

Peace, Jane, Rupa Sheth Valdez, and Kristin F Lutz (2012). "Data-based considerations for electronic family health history applications." In: *Computers, informatics, nursing : CIN* 30.1, pp. 37–45.

Phenotype KnowledgeBase. URL: `https://phekb.org`.

Pietiläinen, Kirsi H et al. (2009). "HDL subspecies in young adult twins: heritability and impact of overweight." In: *Obesity (Silver Spring, Md.)* 17.6, pp. 1208–1214.

Polderman, Tinca J C et al. (2015). "Meta-analysis of the heritability of human traits based on fifty years of twin studies." In: *Nature Publishing Group* 47.7, pp. 702–709.

Polubriaginof, Fernanda, Nicholas P Tatonetti, and David K Vawdrey (2015). "An Assessment of Family History Information Captured in an Electronic Health Record." In: *AMIA Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* 2015, pp. 2035–2042.

Polubriaginof, Fernanda et al. (2016). "**Patient-provided Data Improves Race and Ethnicity Data Quality in Electronic Health Records**." In: *AMIA Annual Symposium ...*.

Polubriaginof, Fernanda et al. (2017). "Estimate of disease heritability using 7.4 million familial relationships inferred from electronic health records." In: *bioRxiv*, p. 066068.

Porter, S C et al. (2000). "Parents as direct contributors to the medical record: validation of their electronic input." In: *Annals of Emergency Medicine* 35.4, pp. 346–352.

Powell, Karen P et al. (2013). "Collection of family health history for assessment of chronic disease risk in primary care." In: *North Carolina medical journal* 74.4, pp. 279–286.

Prey, Jennifer E, Susan Restaino, and David K Vawdrey (2014). "Providing hospital patients with access to their medical records." In: *AMIA Annual Symposium Proceedings*. Vol. 2014. American Medical Informatics Association, p. 1884.

Purcell, Shaun et al. (2007). "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." In: *The American Journal of Human Genetics* 81.3, pp. 559–575.

Pyper, Cecilia et al. (2004). "Patients' experiences when accessing their on-line electronic patient records in primary care." In: *The British journal of general practice : the journal of the Royal College of General Practitioners* 54.498, pp. 38–43.

Qureshi, N et al. (2009). *NIH State-of-the Science Conference: Family History and Improving Health 2009*. AHRQ.

Qureshi, Nadeem et al. (2007). "Collection and use of cancer family history in primary care." In: *Evidence report/technology assessment* 159, pp. 1–84.

Ralston, James D et al. (2007). "Patient web services integrated with a shared medical record: patient use and satisfaction." In: *Journal of the American Medical Informatics Association* 14.6, pp. 798–806.

Rao, S R et al. (2011). "Electronic health records in small physician practices: availability, use, and perceived benefits." In: *Journal of the American Medical Informatics Association* 18.3, pp. 271–275.

Reid, G T et al. (2009). "Family history questionnaires designed for clinical use: a systematic review." In: *Public health genomics* 12.2, pp. 73–83.

Reti, Shane R et al. (2010). "Improving personal health records for patient-centered care." In: *J Am Med Inform Assoc* 17.2, pp. 192–195.

Rich, Eugene C et al. (2004). "Reconsidering the family history in primary care." In: *Journal of General Internal Medicine* 19.3, pp. 273–280.

Ritchie, Marylyn D, Mariza de Andrade, and Helena Kuivaniemi (2015). "The foundation of precision medicine: integration of electronic health records with genomics through basic, clinical, and translational research." In: *Frontiers in genetics* 6, p. 104.

Robbin, A (1999). "The problematic status of US statistics on race and ethnicity: An "imperfect representation of reality"." In: *Journal of Government Information* 26.5, pp. 467–483.

Ronald, Angelica and Rosa A Hoekstra (2011). "Autism spectrum disorders and autistic traits: A decade of new twin studies." In: *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 156.3, pp. 255–274.

Rosenbloom, S T et al. (2011). "Data from clinical notes: a perspective on the tension between structure and flexible documentation." In: *Journal of the American Medical Informatics Association* 18.2, pp. 181–186.

Rubio-Tapia, Alberto et al. (2013). *ACG clinical guidelines: diagnosis and management of celiac disease.*

Rusanov, Alexander et al. (2014). "Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research." In: *BMC medical informatics and decision making* 14.1, p. 51.

Ryan, Patrick B et al. (2012). "Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the Observational Medical Outcomes Partnership." In: *Statistics in medicine* 31.30, pp. 4401–4415.

Sandin, Sven et al. (2014). "The Familial Risk of Autism." In: *JAMA* 311.17, pp. 1770–8.

Saslow, Debbie et al. (2007). *American Cancer Society guidelines for breast screening with MRI as an adjunct to mammography.*

Scheuner, Maren T, Pauline Sieverding, and Paul G Shekelle (2008). "Delivery of genomic medicine for common chronic adult diseases: a systematic review." In: *JAMA* 299.11, pp. 1320–1334.

Scheuner, Maren T et al. (2009). "Are electronic health records ready for genomic medicine?" In: *Genetics in Medicine* 11.7, pp. 510–517.

Schriger, D L et al. (1997). "Implementation of clinical guidelines using a computer charting system. Effect on the initial care of health care workers exposed to body fluids." In: *JAMA* 278.19, pp. 1585–1590.

Schriger, D L et al. (2000). "Implementation of clinical guidelines via a computer charting system: effect on the care of febrile children less than three years of age." In: *Journal of the American Medical Informatics Association* 7.2, pp. 186–195.

Scotet, Virginie et al. (2012). "Evidence for decline in the incidence of cystic fibrosis: a 35-year observational study in Brittany, France." In: *Orphanet Journal of Rare Diseases* 7.1, p. 14.

Selvachandran, S N et al. (2002). "Prediction of colorectal cancer by a patient consultation questionnaire and scoring system: a prospective study." In: *The Lancet* 360.9329, pp. 278–283.

Sequist, Thomas D et al. (2009). "Patient and physician reminders to promote colorectal cancer screening: a randomized controlled trial." In: *Arch Intern Med* 169.4, pp. 364–371.

Siegler, Eugenia L (2010). "The Evolving Medical Record." In: *Annals of Internal Medicine* 153.10, p. 671.

Slieker, Martijn G et al. (2005). "Birth Prevalence and Survival in Cystic Fibrosis: A National Cohort Study in the Netherlands." In: *Chest* 128.4, pp. 2309–2315.

Smith, R A, V Cokkinides, and O W Brawley (2012). "Cancer screening in the United States, 2012: A review of current American Cancer Society guidelines and current issues in cancer screening." In: *CA Cancer J Clin.* 62.2, pp. 129–142.

Solbak, Nathan M et al. (2018). "Patterns and predictors of adherence to colorectal cancer screening recommendations in Alberta's Tomorrow Project participants stratified by risk." In: *BMC public health* 18.1, p. 177.

Souren, N Y et al. (2007). "Anthropometry, carbohydrate and lipid metabolism in the East Flanders Prospective Twin Survey: heritabilities." In: *Diabetologia* 50.10, pp. 2107–2116.

Staroselsky, Maria et al. (2006). "Improving electronic health record (EHR) accuracy and increasing compliance with health maintenance clinical guidelines through patient access and input." In: *International Journal of Medical Informatics* 75.10-11, pp. 693–700.

Staroselsky, Maria et al. (2008). "An effort to improve electronic health record medication list accuracy between visits: Patients' and physicians' response." In: *International Journal of Medical Informatics* 77.3, pp. 153–160.

Stevens, Lesley A et al. (2006). "Assessing kidney function–measured and estimated glomerular filtration rate." In: *The New England journal of medicine* 354.23, pp. 2473–2483.

Suarez, R (2011). "Breaching Doctor-Patient Confidentiality: Confusion among Physicians about Involuntary Disclosure of Genetic Information." In: *S Cal Interdisc LJ*.

Sullivan, Patrick F, Mark J Daly, and Michael O'Donovan (2012). "Genetic architectures of psychiatric disorders: the emerging picture and its implications." In: *Nature Reviews Genetics* 13.8, pp. 537–551.

Sullivan, Patrick F, Kenneth S Kendler, and Michael C Neale (2003). "Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies." In: *Archives of general psychiatry* 60.12, pp. 1187–1192.

Sussner, Katarina M, Lina Jandorf, and Heiddis B Valdimarsdottir (2011). "Educational needs about cancer family history and genetic counseling for cancer risk among front-line healthcare clinicians in New York City." In: *Genetics in Medicine* 13.9, pp. 785–793.

Sweet, K M, T L Bradley, and J A Westman (2002). "Identification and referral of families at high risk for cancer susceptibility." In: *Journal of Clinical Oncology* 20.2, pp. 528–537.

Sweet, Kevin et al. (2014). "Clinically relevant lessons from Family HealthLink: a cancer and coronary heart disease familial risk assessment tool." In: *Genetics in Medicine* 17.6, pp. 493–500.

Tang, Paul C and Thomas H Lee (2009). "Your Doctor's Office or the Internet? Two Paths to Personal Health Records." In: *dx.doi.org* 360.13, pp. 1276–1278.

Tatonetti, Nicholas P et al. (2012). "Data-driven prediction of drug effects and interactions." In: *Science translational medicine* 4.125, 125ra31–125ra31.

Tenesa, Albert and Chris S Haley (2013). "The heritability of human disease: estimation, uses and abuses." In: *Nature Reviews Genetics* 14.2, pp. 139–149.

Thiru, Krish, Alan Hassey, and Frank Sullivan (2003). "Systematic review of scope and quality of electronic patient record data in primary care." In: *BMJ* 326.7398, pp. 1070–0.

Turkheimer, Eric et al. (2003). "Socioeconomic status modifies heritability of IQ in young children." In: *Psychological science* 14.6, pp. 623–628.

Tyrer, Jonathan, Stephen W Duffy, and Jack Cuzick (2004). "A breast cancer prediction model incorporating familial and personal risk factors." In: *Statistics in medicine* 23.7, pp. 1111–1130.

U.S. Preventive Services Task Force (2011). *Screening for osteoporosis: U.S. preventive services task force recommendation statement.*

United States (1996). "Health Insurance Portability and Accountability Act of 1996. Public Law 104-191." In: *United States statutes at large* 110, pp. 1936–2103.

Visscher, Peter M, William G Hill, and Naomi R Wray (2008). "Heritability in the genomics era–concepts and misconceptions." In: *Nature Reviews Genetics* 9.4, pp. 255–266.

Visscher, Peter M et al. (2007). “Genome partitioning of genetic variation for height from 11,214 sibling pairs.” In: *American journal of human genetics* 81.5, pp. 1104–1110.

Volk, Lynn A et al. (2007). “Do physicians take action on high risk family history information provided by patients outside of a clinic visit?” In: *Studies in health technology and informatics* 129.Pt 1, pp. 13–17.

Wagenknecht, L E et al. (2011). “Misclassification of smoking status in the CARDIA study: a comparison of self-report with serum cotinine levels.” In: *American journal of public health* 82.1, pp. 33–36.

Walker, Jan et al. (2011). “Inviting patients to read their doctors’ notes: patients and doctors look ahead: patient and physician surveys.” In: *Ann Intern Med* 155.12, pp. 811–819.

Wallace, Paul J et al. (2014). “Optum Labs: building a novel node in the learning health care system.” In: *Health Aff (Millwood)* 33.7, pp. 1187–1194.

Wang, Kanix et al. (2017). “Classification of common human diseases derived from shared genetic and environmental determinants.” In: *Nature Publishing Group* 49.9, pp. 1319–1325.

Wang, Yan et al. (2016). “Investigating Longitudinal Tobacco Use Information from Social History and Clinical Notes in the Electronic Health Record.” In: *AMIA Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* 2016, pp. 1209–1218.

Warner, Diana (2010). “Managing patient-provided information in EHRs.” In: *J AHIMA* 81.7, pp. 44–45.

Wasserman, Richard C (2011). “Electronic medical records (EMRs), epidemiology, and epistemology: reflections on EMRs and future pediatric clinical research.” In: *Academic pediatrics* 11.4, pp. 280–287.

Wee, Christina C, Ellen P McCarthy, and Russell S Phillips (2005). “Factors associated with colon cancer screening: the role of patient factors and physician counseling.” In: *Preventive Medicine* 41.1, pp. 23–29.

Wei, Wei-Qi and Joshua C Denny (2015). “Extracting research-quality phenotypes from electronic health records to support precision medicine.” In: *Genome medicine* 7.1, p. 41.

Weiner, Mark G and Peter J Embi (2009). “Toward Reuse of Clinical Data for Research and Quality Improvement: The End of the Beginning?” In: *Annals of Internal Medicine* 151.5, pp. 359–360.

Weingart, Saul N et al. (2005). “Patient-reported medication symptoms in primary care.” In: *Arch Intern Med* 165.2, pp. 234–240.

Weingart, Saul N et al. (2008). “Medication safety messages for patients via the web portal: The MedCheck intervention.” In: *International Journal of Medical Informatics* 77.3, pp. 161–168.

Weiskopf, Nicole Gray and Chunhua Weng (2013). “Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research.” In: *J Am Med Inform Assoc* 20.1, pp. 144–151.

Welch, Brandon M, Willard Dere, and Joshua D Schiffman (2015). “Family Health History.” In: *JAMA* 313.17, pp. 1711–2.

Welch, Brandon M and Kensaku Kawamoto (2013). “Clinical decision support for genetically guided personalized medicine: a systematic review.” In: *J Am Med Inform Assoc* 20.2, pp. 388–400.

Wilcox, Lauren et al. (2016). “Interactive tools for inpatient medication tracking: a multiphase study with cardiothoracic surgery patients.” In: *J Am Med Inform Assoc* 23.1, pp. 144–158.

Wilson, B J et al. (2012a). “Family history tools in primary care: does one size fit all?” In: *Public health genomics* 15.3-4, pp. 181–188.

— (2012b). “Family history tools in primary care: does one size fit all?” In: *Public health genomics* 15.3-4, pp. 181–188.

Winden, Tamara J et al. (2015). “Towards the Standardized Documentation of E-Cigarette Use in the Electronic Health Record for Population Health Surveillance and Research.” In: *AMIA Joint Summits on Translational Science proceedings AMIA Summit on Translational Science* 2015, pp. 199–203.

Woods, Susan S, Neil C Evans, and Kathleen L Frisbee (2016). “Integrating patient voices into health information for self-care and patient-clinician partnerships: Veterans Affairs design recommendations for patient-generated data applications.” In: *Journal of the American Medical Informatics Association* 23.3, pp. 491–495.

Wu, R Ryanne and Lori A Orlando (2015). “Implementation of health risk assessments with family health history: barriers and benefits.” In: *Postgraduate Medical Journal* 91.1079, pp. 508–513.

Wu, R Ryanne et al. (2013). “Patient and primary care provider experience using a family health history collection, risk stratification, and clinical decision support tool: a type 2

hybrid controlled implementation- effectiveness trial." In: *BMC Family Practice* 14.1, pp. 1–1.

Wu, R Ryanne et al. (2014). "Quality of family history collection with use of a patient facing family history assessment tool." In: *BMC Family Practice* 15.1, pp. 1–8.

Wu, R Ryanne et al. (2015). "Protocol for the "Implementation, adoption, and utility of family history in diverse care settings" study." In: *Implementation Science* 10.1, pp. 1–10.

Wuerdeman, Lisa et al. (2005). "How accurate is information that patients contribute to their Electronic Health Record?" In: *AMIA Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* 2005, pp. 834–838.

Yoon, Paula W, Maren T Scheuner, and Muin J Khoury (2003). "Research priorities for evaluating family history in the prevention of common chronic diseases." In: *American Journal of Preventive Medicine* 24.2, pp. 128–135.

Yoon, Paula W et al. (2009). "Developing Family Healthware, a family history screening tool to prevent common chronic diseases." In: *Preventing chronic disease* 6.1, A33.

Yuan, Jiawei et al. (2017). "Towards a privacy preserving cohort discovery framework for clinical research networks." In: *Journal of Biomedical Informatics* 66, pp. 42–51.