RISK PREDICTION OF MULTIPLE SELECTED CHRONIC DISEASES USING SELF- AND PROXY-REPORTED FAMILY HEALTH HISTORY AND LIFESTYLE RISK FACTORS

Yuling Jiang

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Biomedical Informatics

The University of Utah

December 2013

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Yuling Jiang**

has been approved by the following supervisory committee members:

Catherine Staes	, Chair	**10/31/2013** Date Approved
Julio Facelli	, Member	**10/31/2013** Date Approved
Nathan Hulse	, Member	**10/31/2013** Date Approved
Steven Hunt	, Member	**10/31/2013** Date Approved
Scott Narus	, Member	**10/31/2013** Date Approved

And by **Wendy Chapman**, Chair of

the Department of **Biomedical Informatics**

and by David Kieda, Dean of The Graduate School.

ABSTRACT

Family health history (FHH) is an independent risk factor for predicting an individual's chance of developing selected chronic diseases. Though various FHH tools have been developed, many research questions remain to be addressed. Before FHH can be used as an effective risk assessment tool in public health screenings or population-based research, it is important to understand the quality of collected data and evaluate risk prediction models. No literature has been identified whereby risks are predicted by applying machine learning solely on FHH. This dissertation addressed several questions. First, using mixed methods, we defined 50 requirements for documenting FHH for a population-based study. Second, we examined the accuracy of self- and proxy-reported FHH data in the *Health Family Tree* database, by comparing the disease and risk factor rates generated from this database with rates recorded in a cancer registry and standard public health surveys. The rates generated from the *Health Family Tree* were statistically lower than those from public sources (exceptions: stroke rates were the same, exercise rates were higher). Third, we validated the *Health Family Tree* risk predictive algorithm. The very high risk (≥2) predicted the risk of all concerned diseases for adult population (20 ~ 99 years of age), and the predictability remained when using disease rates from public sources as the reference in the relative risk model. The referent population used to establish the expected rate of disease impacted risk classification: the lower expected disease rates generated by the *Health Family Tree*, in comparison to the rates from public

sources, caused more persons to be classified at high risk. Finally, we constructed and evaluated new predictive models using three machine learning classifiers (logistic regression, Bayesian networks, and support vector machine). A limited set of information about first-degree relatives was used to predict future disease. In summary, combining FHH with valid risk algorithms provide a low cost tool for identifying persons at risk for common diseases. These findings may be especially useful when developing strategies to screen populations for common diseases and identifying those at highest risk for public health interventions or population-based research.

To my family.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

AHIC – American Health Information Community

BMI – body mass index

BRFSS – Behavioral Risk Factor Surveillance System

CHD – coronary heart disease

CMH – Cochran-Mantel-Haenszel

EHR – electronic health records

FHH – family health history

FHS – family history score

HFT – Health Family Tree

HL7 – Health Level Seven

MI – myocardial infarction

NCHS – National Center for Health Statistics

NCS – National Children's Study

PHR – personal health records

SVM – support vector machine

UCR – Utah Cancer Registry

UHAS – Utah Health Access Survey

WEKA – Waikato Environment for Knowledge Analysis

ACKNOWLEDGEMENTS

I would like especially to thank my committee chair, Catherine Staes, for her guidance and support for my dissertation. I am also very grateful for other members of my committee, Julio Facelli, Steven Hunt, Nathan Hulse, and Scott Narus for their key contributions and crucial feedback. I thank faculty members Nancy Staggers, Ted Adams, Sean Firth, James Vanderslice, and Gang Luo who also gave me guidance with the data analysis. I would like to express many thanks to the students and staff at the Biomedical Informatics Department, University of Utah. Finally, I express thanks to my parents who supported me through graduate school.

CHAPTER 1

INTRODUCTION

Chronic diseases such as heart disease, stroke, cancer, and diabetes are common, expensive, but preventable health problems in the United States.[1] A positive family health history is often considered an independent risk factor for these diseases.[2] This dissertation focuses on using family health history information and personal lifestyle risk factors to assess a healthy individual's risk of developing selected chronic diseases. Preventing the onset of chronic disease is an important strategy for population health.

The following hypotheses were tested by the research described in this dissertation: 1) The requirements for documenting family health history for population-based studies and public health will differ from the published requirements for integrating family health history in an electronic health record system; 2) The disease rates generated from self-reported and proxy-reported family history data are similar to the rates recorded in public databases; 3) A current risk algorithm can predict a healthy individual's risk for developing certain chronic diseases, based on self- and proxy- reported family health history information; and 4) A new prediction model based on machine learning can predict a healthy individual's risk for developing chronic diseases, using diabetes as an example.

Each chapter of this dissertation addresses a different component of the research. Chapter 2 provides background information about chronic diseases, family health history

and the rationale for conducting this research. Chapter 3 provides a description of requirements for documenting family health history for longitudinal population-based studies and public health, and an evaluation of a national tool for meeting the requirements. Chapter 4 provides an examination of the accuracy of family health history data by comparing disease rates generated from self- and proxy-reported data with rates generated from a cancer registry and public health standardized surveys. Chapter 5 provides a validation of the current risk algorithm for predicting a healthy individual's risk for developing certain chronic diseases by using the individual's family health history. Chapter 6 includes a description and evaluation of new risk prediction models for diabetes using three machine learning methods. Finally, Chapter 7 includes a discussion of the implications and potential future work associated with the research presented in this dissertation.

References

1. The Lancet. Tackling the burden of chronic diseases in the USA. *The Lancet.* Jan 17 2009;373(9659):185.

2. CDC. Awareness of family health history as a risk factor for disease: United States, 2004 *MMWR Morb Mortal Wkly Rep.* 2004;53:1044-1047.

CHAPTER 2

BACKGROUND

Chronic Diseases in the United States

Chronic diseases are ongoing, often incurable illnesses or conditions, such as heart disease, cancer, and diabetes. Approximately 133 million Americans, 45% of the population, are affected by at least one chronic disease.[1] By 2020, the number is projected to grow to 157 million, with 81 million persons having multiple conditions.[2] Among persons age 65 and older, half (51%) have hypertension and approximately a third have arthritis (37%) or heart disease (29%).[3] These three conditions are the most common chronic diseases among persons over 65 years of age. [3] Chronic diseases are the leading cause of death and disability in the United States. Each year, 7 out of 10 deaths in the United States are due to chronic diseases. Heart disease, cancer, and stroke are responsible for more than 50% of all deaths.[4] About 25% of people with chronic diseases have some type of activity limitations.[3] Chronic diseases also account for the majority of health spending in the United States: more than 75% of the health care costs, about $2 trillion, are due to chronic conditions.[5] The most expensive chronic diseases, heart disease and stroke, cost Americans $432 billion per year.[6]

Many chronic diseases are preventable if individuals address the myriad of risk factors that contribute to the onset of disease. For example, tobacco use in the United States since the 1950s has declined greatly from 57% to 23% among men, and from 34%

to 18% among women.[7] This decrease happened after the 1964 report of the surgeon general, which linked smoking and lung cancer.[8] A report by the CDC also showed that one year after quitting smoking, excess risk for heart disease can drop 50%[9]; exercise at moderate intensity and lowered intake of fat and calories can reduce the risk for diabetes by 58%.[9] The World Health Organization has estimated that 80% of heart disease, stroke, and type 2 diabetes and 40% of cancers would be prevented if the following three risk factors for chronic diseases were eliminated: physical inactivity, poor diet, and smoking.[10] The goal of public health is to prevent chronic diseases through primary prevention (i.e., health promotion efforts that encourages healthy living), secondary prevention (i.e., screening efforts for early detection among at-risk populations), and tertiary prevention (i.e., management of existing diseases).[9] Given that half of all chronic disease is caused by unhealthy behaviors, there is a need to develop tools to identify those at risk and encourage healthy behaviors.[7]

Family Health History and Its Value for Chronic Disease Management

There are four major determinants that affect a person's health: biology, environment (including physical and social), lifestyle, and healthcare.[11] The description of genetic relationships and medical history of a family, known as family health history (FHH), reflects all of the determinants contextualized within the family, such as genetic predispositions, shared environmental factors, common lifestyle factors, and shared healthcare.[12] Among medical practitioners, FHH is traditionally considered one of the major components for a complete medical history in the official medical record. FHH has been used by clinicians for chronic disease diagnosis, treatment, prevention, and patient

education.[13,14] FHH has also been used by public health professionals to identify high risk populations and then, based on the assessed level of risk, to apply screening strategies for chronic diseases.[12,15] For example, positive FHH of diabetes,[16-20] coronary heart disease (CHD) or myocardial infarction (MI),[21-27] stroke,[28,29] or various cancers[30-34] are all considered risk factors for these diseases. In the more recent era of genomic and personalized medicine, FHH still retains its importance. Being a noninvasive, low-cost, and proven tool, FHH is given new meaning and power to interpret the complex interactions between gene and environment that cause different levels of health and disease.[35] Though the cost of sequencing a person's genome has dropped significantly, the interpretation of the genetics and the interaction of the genetic and environmental factors are far from being completely understood and may still be best represented by the FHH. In addition, FHH may be used to determine the likelihood of whether or not genetic variations will be pathologic for a specific individual. Furthermore, FHH also reflects the shared behavioral factors such as diet and exercise within the family. Thus, FHH is used as an important tool in different health related areas: in clinical medicine, FHH is used to diagnose and manage affected patients[7]; in public health, FHH is used to stratify a healthy population to identify high risk subpopulations and to prevent chronic diseases through health education and/or health screening[15,36]; and finally, in population-based research, FHH is used to stratify risk within the study population for descriptive analysis or to select cases and controls for future genetic and epidemiologic analysis. This dissertation seeks to apply FHH to risk assessment for longitudinal population-based research and public health screening strategies.

Family Health History Tools

To better use FHH for risk assessment either in clinical or public health settings, many tools have been developed. Traditional paper-based tools such as the American Medical Association's *Prenatal Genetic Screening Questionnaire*,[37] *Pediatric Genetic Screening Questionnaire*,[38] and *Adult Family History Form*[39] are used to collect information for screening, diagnosis, and treatment. The last decade has seen rapid development of informatics tools, including computerized and web-based tools. For example, in 2004, the Surgeon General initiated a national public health campaign[19] that released a web-based tool, *My Family Health Portrait*,[40] for the public to collect, save, and share the family's medical history of multiple diseases and conditions with their healthcare providers and family members. This tool is publicly available and uses standard vocabulary (including LOINC® and SNOMED-CT) and the HL7 family history data model to allow interoperability with electronic health records.[41]

Other than *My Family Health Portrait*, various universities, health organizations, and research institutes have also developed tools to meet their needs for collecting, interpreting, and applying FHH. *Family Healthware*[42] is a web-based research tool developed by the Centers for Disease Control and Prevention to assess a person's familial risk for six chronic diseases, including diabetes, CHD, stroke, and colon, breast, and ovarian cancer. Alternatively, *MyGenerations*[43] collects family history on cancers and provides risk assessment. *Your Disease Risk*[44] assesses the risks for diabetes, heart diseases, stroke, osteoporosis, and cancer using family history and lifestyle risk factors information. *Hughes RiskApps*[45] allows family history and other risk data to be entered

and calculates risks for breast and ovarian cancer. Similar tools for different cancer or chronic disease risk assessment include *Hereditary Cancer Quiz*,[46] *Family Healthlink*,[47] and *Be Ready Quiz*.[48]

Utah's tool: *Health Family Tree*

Health Family Tree is an additional tool that uses FHH to predict risk. The *Health Family Tree* program was developed by researchers in Utah and Texas in the early 1980s.[49] The tool included a paper-based questionnaire and a computer-based database and algorithm. The questionnaire was distributed to high school students in Utah through their required Health Education class. From 1983 to 2001, 57,238 students in 55 high schools collected their family history by documenting information about common diseases and general lifestyle risk factors about their family members.[50] Their family members included the student's siblings, parents, aunts, and uncles and grandparents. The common diseases included: diabetes, MI, CHD, stroke, high blood pressure, high blood cholesterol, breast cancer, lung cancer, and colon cancer. The lifestyle risk factors assessed included: smoking, drinking, being overweight/obese, and exercise. The information was collected with consent by each student as assigned homework, with help from their parents, on a 36 x 23 inch folding paper that was designed to fit on a kitchen table. The collected information was transcribed and stored in a database, and an algorithm was developed to automatically predict the risk for the above diseases.[51] The algorithm predicted risk based on comparing the observed number of disease events to the expected number of disease events within the family. The expected number of events was calculated by multiplying the age- and sex-specific person-years for each person in the family by the age- and sex-specific incidence rates generated from the entire database

population.[51] The large volume of persons represented in the data and the systematic collection of information from a large number of high school students throughout the state led to the assumption that the expected number of events could be derived from the database itself.

Besides the database, the *Health Family Tree* program also developed an algorithm to calculate the risks of developing the above diseases. This risk algorithm was validated in 1986 for predicting heart diseases.[51] The researchers found that the definition of elevated risk from the *Health Family Tree* algorithm successfully predicted unaffected family member's risk of developing future CHD. In addition, preliminary analysis of the tools ability to predict MI and diabetes was reported in 2009.[52] The *Health Family Tree* tool has been used by the Utah Department of Health for the purpose of screening populations in the community since it was developed. A web-based version of the *Health Family Tree* (http://healthfamilytree.utah.edu/) was developed in 2005 and could continue to serve similar purposes on a larger scale.

Requirements for Documenting Family Health History

Before expanding data collection about FHH using any tools, it is essential to understand the requirements for collecting FHH. Various tools collect different family history information based on their own needs; thus, the collected data vary greatly in terms of required data elements and the degree of relatives that are included. In 2008, the American Health Information Community (AHIC), a federal group formed to advise the Secretary of the Department of Health Human Services on methods of increasing electronic health record adoption in healthcare facilities, published the data requirements (i.e., core data set) for representing FHH in an Electronic Health Record and Personal

Health Record.[53] This core data set may or may not be adequate to develop or select informatics tools for documenting FHH for longitudinal population-based research or screening. Furthermore, in addition to the data requirements, other functional and nonfunctional requirements must be defined for documenting FHH. No current literature addresses these other requirements, such as functional requirements for using and maintaining the collected FHH data or the usability requirements for collecting FHH. When developing or adopting any computer-based systems, it is critical to incorporate usability into the process.[54] One study showed that the ultimate acceptance or rejection of a healthcare information system is largely dependent on the system's usability.[55] The usability issue is even more important for developing or adopting tools to collect family history information from the general public, because the system must accommodate various languages, levels of education, and computer skills. A well-designed, user friendly computerized tool may reduce the burden and save the time needed for entering FHH, thereby potentially increasing the completeness of the information. A tool with good usability may also reduce human errors during data input and hence improve the quality of the information collected.

Accuracy of Family Health History

Although the literature has shown that FHH information can be used for risk assessment, the quality of data collected needs to be examined. Using the *Health Family Tree* as an example, the information about the students and their siblings and parents was self-reported while information about the student's aunts, uncles, and grandparents was classified as proxy-reported. These methods, both self-report and proxy-report, have been widely used to collect FHH data. To collect an individual's family history, informants

(students and their parents in this case) need to not only report on their own medical history and risk factors (self-report) but also report on the medical history and risk factors of their first, second, or even third degree relatives (proxy-report). According to a systematic review by the National Institute of Health in 2009, the accuracy of FHH varies based on the disease being studied.[56] In general, correct reporting of the absence of disease in relatives was better than correct reporting of the existence of disease.[56] The results from multiple studies showed that the specificities of reporting family history of cancer were high, ranging from 0.91 to 1.00. In contrast, the sensitivities reported in these studies varied by the type of cancer: breast, 0.72 to 0.95; colon, 0.33 to 0.90; ovarian, 0.38 to 0.42; and prostate, 0.47 to 0.79.[56] Similar patterns were observed for reporting family history of other diseases such as diabetes, hypertension, and cardiovascular disease: the specificities were high, ranging from 0.76 to 0.98, and the sensitivities varied greatly from 0.18 to 0.89.[56] No clear association was observed between accuracy and informant age, sex, or educational level. In 1986, a data accuracy study was also performed to assess the quality of the information reported by the student and their parents for their relatives in the *Health Family Tree* program. A subset of the families was selected and the family members were contacted by mailing a questionnaire with additional questions, phone calling, or personal interviews to confirm the reported disease status. Results showed the sensitivity of capturing disease events was 0.67 and the specificity was 0.96.[51]

The accuracy studies described above used similar methods to verify the relatives' actual disease status. The methods require locating each relative's medical records, records in disease or death registries, or contacting the relative directly. While these

methods are advantageous because they directly capture the history of the person, there are several disadvantages. For example, the methods require finding the existing records or contacting multiple relatives, which makes the accuracy evaluation very resource- and time-consuming. Alternatively, if the rates of disease and risk factors generated from one data source are similar to the rates available from publicly available sources, the accuracy of this data source may be validated on the population level and fewer resources will be used. These strategies address the accuracy of the counts of events that comprise the numerator in the rate calculation.

When assessing the *Health Family Tree* risk assessment algorithm, it is also important to evaluate the accuracy of the expected rates being used to define the expected occurrence of disease. The *Health Family Tree* risk assessment algorithm relies on the disease rates generated from the database itself. There are two major reasons to describe the disease rates generated by the Health Family Tree and compare the rates with general population disease rates: 1) to test the assumption that it is valid to generate expected rates using the *Health Family Tree* database, and 2) to generalize the risk algorithm used in this program and possibly implement the algorithm in a standalone decision support system for risk prediction and disease prevention. Other decision systems could use the publicly available data in the risk predictive algorithm and assess the risks for medical or public health decision support without the need of a database that contains a large amount of records. Currently, the relationship between the self- and proxy- reported information in the *Health Family Tree* and information from public sources is unknown.

New Predictive Model Using Family Health History

The algorithm used in the *Health Family Tree* program is based on classical statistical methods and was implemented in 1983. Since 1983, new methods have been developed to discover knowledge in large databases. For example, machine learning methods use a discovery process to describe structural patterns. These patterns may be used to predict outcomes from any similar data. While both statistical and machine learning methods have evolved in parallel, there are important differences. Statistical methods are more concerned with testing hypotheses, whereas machine learning methods are more concerned with formulating the process of generalization.[57]

Recently, data mining and machine learning methods have been used to analyze healthcare data, which is known to be complex and voluminous. For example, these methods have been applied to control hospital infections, to rank hospitals, and to identify high risk patients.[58,59] Studies applying machine learning methods to multiple risk factors including FHH to predict the risks of developing coronary heart events and diabetes showed promising results.[60-63] No literature has been identified whereby risks are predicted by applying machine learning solely on FHH information.

The *Health Family Tree* program created a rich database that contains more than one million individuals' self- and proxy- reported demographic, family relationship, medical history, and lifestyle risk factor information. Machine learning may help to develop effective predictive models for classifying high risk from low risk individuals in order to implement screening and population-based interventions. Compared with statistical models based on relative risk, the machine learning methods may have practical benefits for implementation. The *Health Family Tree* statistical algorithm

requires a reference population to generate the expected incidence rates. Very limited information about incidence rates is available for most chronic diseases. Cancer incidence rates are the only exception due to nationally instituted cancer registries. Machine learning techniques build predictive models by training on the features collected by the Health Family Tree tool. Ideally, these trained and validated models can be applied to predict disease risks in any tool that collects the required set of features.

Filling a Knowledge Gap

"Documentation of family medical history" is one of the "Meaningful Use" objectives defined by the US Health and Human Services.[64] Even so, family health history is an important, yet underused tool in both clinical and public health settings and population-based studies. Possible barriers to this effort include lack of awareness of the importance of FHH, lack of time, lack of accurate, detailed information, and lack of validated risk assessment,[13-15] and possibly poor tool support partly due to loosely defined requirements. Currently, there is no literature that defines requirements beyond data requirements, and there is no consensus on data and function requirements; therefore, though multiple tools exist, it is unknown how the tools can be used for longitudinal population-based research and public health screening. Literature has reported on the accuracy of self- and proxy-reported family history, but the results of those studies vary greatly, indicating that more research is needed. The *Health Family Tree* program was developed and has been used in the community as a screening tool for over 20 years, but has only been validated for heart diseases. An online version of this tool is developed and could be implemented in a larger community. Validating the risk algorithm for more diseases and health conditions may increase acceptance of the tool as

a population screening tool. Finally, recently developed machine learning techniques have not been applied in predictive models that use only FHH. How the new models built by machine learning techniques compare with traditionally developed predictive models still remains unknown. This dissertation addresses the gaps by: 1) defining the requirements for a tool to document FHH for longitudinal population-based studies, and evaluating if a current national tool meets these requirements; 2) examining the accuracy of self- and proxy-reported data by comparing disease rates generated from the *Health Family Tree* database with rates from public data sources; 3) validating whether the risk score derived from the *Health Family Tree* algorithm can predict an individual's future risk for multiple diseases or conditions (diabetes, MI, CHD, stroke, high blood pressure, high blood cholesterol, breast cancer, lung cancer, and colon cancer); and 4) building and evaluating a diabetes predictive model through machine learning that uses only FHH and lifestyle risk factors to classify healthy individuals.

References

1. Wu S, Green A. *Projection of Chronic Illness Prevalence and Cost Inflation*. Santa Monica, CA: RAND Health; 2000.

2. The Lancet. Tackling the burden of chronic diseases in the USA. *The Lancet.* Jan 17 2009;373(9659):185.

3. Anderson G. *Chronic Conditions: Making the Case of Ongoing Care*. Baltimore, MD: John Hopkins University; 2004.

4. Kung HC, Hoyert DL, Xu J, Murphy SL. Deaths: final data for 2005. *Natl Vital Stat Rep.* Apr 24 2008;56(10):1-120.

5. CDC. The Power to Prevent, The Call to Control: At A Glance 2009. 2009; http://www.cdc.gov/chronicdisease/resources/publications/aag/chronic.htm. Accessed April 1, 2013.

6. Mensah GA, Brown DW. An overview of cardiovascular disease burden in the United States. *Health Aff (Millwood).* Jan-Feb 2007;26(1):38-48.

7. Schroeder SA. Shattuck Lecture. We can do better—improving the health of the American people. *N Engl J Med.* Sep 20 2007;357(12):1221-1228.

8. Bayne-Jones S, Burdette W, Cochran W. *Smoking and Health: Report of the Advisory Committee to the Surgeon General of the Public Health Service*. Washington, DC: Public Health Service; 1964.

9. CDC. *The Power of Prevention: Chronic Disease...the Public Health Challenge of the 21st Century.* Atlanta: Centers for Disease Control and Prevention;2009.

10. WHO. *Preventing Chronic Diseases: A Vital Investment*. Geneva: World Health Organization;2005.

11. Evans RG, Stoddart GL. Producing health, consuming health care. *Soc Sci Med.* 1990;31(12):1347-1363.

12. CDC. Awareness of family health history as a risk factor for disease: United States, 2004. *MMWR Morb Mortal Wkly Rep.* 2004;53:1044-1047.

13. Trotter TL, Martin HM. Family history in pediatric primary care. *Pediatrics.* Sep 2007;120 Suppl 2:S60-65.

14. Hinton RB, Jr. The family history: reemergence of an established tool. *Crit Care Nurs Clin North Am.* Jun 2008;20(2):149-158, v.

15. Yoon PW, Scheuner MT, Peterson-Oehlke KL, Gwinn M, Faucett A, Khoury MJ. Can family history be used as a tool for public health and preventive medicine? *Genet Med.* Jul-Aug 2002;4(4):304-310.

16. Harrison TA, Hindorff LA, Kim H, et al. Family history of diabetes as a potential public health tool. *Am J Prev Med.* Feb 2003;24(2):152-159.

17. Bjornholt JV, Erikssen G, Liestol K, Jervell J, Thaulow E, Erikssen J. Type 2 diabetes and maternal family history: an impact beyond slow glucose removal rate and fasting hyperglycemia in low-risk individuals? Results from 22.5 years of follow-up of healthy nondiabetic men. *Diabetes Care.* Sep 2000;23(9):1255-1259.

18. Nakanishi S, Yamane K, Kamei N, Okubo M, Kohno N. Relationship between development of diabetes and family history by gender in Japanese-Americans. *Diabetes Res Clin Pract.* Aug 2003;61(2):109-115.

19. Rahman M, Simmons RK, Harding AH, Wareham NJ, Griffin SJ. A simple risk score identifies individuals at high risk of developing Type 2 diabetes: a prospective cohort study. *Fam Pract.* Jun 2008;25(3):191-196.

20. Meigs JB, Cupples LA, Wilson PW. Parental transmission of type 2 diabetes: the Framingham Offspring Study. *Diabetes.* Dec 2000;49(12):2201-2207.

21. Sesso HD, Lee IM, Gaziano JM, Rexrode KM, Glynn RJ, Buring JE. Maternal and paternal history of myocardial infarction and risk of cardiovascular disease in men and women. *Circulation.* Jul 24 2001;104(4):393-398.

22. Jousilahti P, Puska P, Vartiainen E, Pekkanen J, Tuomilehto J. Parental history of premature coronary heart disease: an independent risk factor of myocardial infarction. *J Clin Epidemiol.* May 1996;49(5):497-503.

23. Hopkins PN, Williams RR, Kuida H, et al. Family history as an independent risk factor for incident coronary artery disease in a high-risk cohort in Utah. *Am J Cardiol.* Oct 1 1988;62(10 Pt 1):703-707.

24. Myers RH, Kiely DK, Cupples LA, Kannel WB. Parental history is an independent risk factor for coronary artery disease: the Framingham Study. *Am Heart J.* Oct 1990;120(4):963-969.

25. Roncaglioni MC, Santoro L, D'Avanzo B, et al. Role of family history in patients with myocardial infarction. An Italian case-control study. GISSI-EFRIM Investigators. *Circulation.* Jun 1992;85(6):2065-2072.

26. Djousse L, Gaziano JM. Parental history of myocardial infarction and risk of heart failure in male physicians. *Eur J Clin Invest.* Dec 2008;38(12):896-901.

27. Friedlander Y, Arbogast P, Schwartz SM, et al. Family history as a risk factor for early onset myocardial infarction in young women. *Atherosclerosis.* May 2001;156(1):201-207.

28. Jousilahti P, Rastenyte D, Tuomilehto J, Sarti C, Vartiainen E. Parental history of cardiovascular disease and risk of stroke. A prospective follow-up of 14371 middle-aged men and women in Finland. *Stroke.* Jul 1997;28(7):1361-1366.

29. Kadota A, Okamura T, Hozawa A, et al. Relationships between family histories of stroke and of hypertension and stroke mortality: NIPPON DATA80, 1980-1999. *Hypertens Res.* Aug 2008;31(8):1525-1531.

30. Keku TO, Millikan RC, Martin C, Rahkra-Burris TK, Sandler RS. Family history of colon cancer: what does it mean and how is it useful? *Am J Prev Med.* Feb 2003;24(2):170-176.

31. Cauley JA, Song J, Dowsett SA, Mershon JL, Cummings SR. Risk factors for breast cancer in older women: the relative contribution of bone mineral density and other established risk factors. *Breast Cancer Res Treat.* Apr 2007;102(2):181-188.

32. Wei EK, Giovannucci E, Wu K, et al. Comparison of risk factors for colon and rectal cancer. *Int J Cancer.* Jan 20 2004;108(3):433-442.

33. Rodriguez C, Calle EE, Miracle-McMahill HL, et al. Family history and risk of fatal prostate cancer. *Epidemiology.* Nov 1997;8(6):653-657.

34. Gao Y, Goldstein AM, Consonni D, et al. Family history of cancer and nonmalignant lung diseases as risk factors for lung cancer. *Int J Cancer.* Jul 1 2009;125(1):146-152.

35. Guttmacher AE, Collins FS, Carmona RH. The family history—more important than ever. *N Engl J Med.* Nov 25 2004;351(22):2333-2336.

36. Chowdhury S, Dent T, Pashayan N, et al. Incorporating genomics into breast and prostate cancer screening: assessing the implications. *Genet Med.* Jun 2013;15(6):423-432.

37. AMA. Prenatal Screening Questionnaire. http://www.ama-assn.org/resources/doc/genetics/ped_screening.pdf. Accessed Dec 31, 2012.

38. AMA. Pediatric Clinical Genetics Questionnaire. http://www.ama-assn.org/resources/doc/genetics/ped_clinical_genetic.pdf. Accessed Dec 31, 2012.

39. AMA. Adult Family History Form. http://www.ama-assn.org/resources/doc/genetics/adult_history.pdf. Accessed Dec 31, 2012.

40. My family health portrait. 2005; http://familyhistory.hhs.gov. Accessed March 18, 2013.

41. Learn More About My Family Health Portrait. https://familyhistory.hhs.gov/fhh-web/popup/getHelp/helpDetailsLearnMore.action. Accessed Feb 27, 2011.

42. Yoon PW, Scheuner MT, Jorgensen C, Khoury MJ. Developing Family Healthware, a family history screening tool to prevent common chronic diseases. *Prev Chronic Dis.* Jan 2009;6(1):A33.

43. NorthShore. MyGenerations. http://www.northshore.org/genetics/mygenerations/. Accessed April 1, 2012.

44. Washington University. Your disease risk: the source on prevention http://www.yourdiseaserisk.wustl.edu/. Accessed April 1, 2012.

45. Screening (medicine). http://en.wikipedia.org/wiki/Screening_(medicine). Accessed October 12, 2011.

46. Hereditary Cancer Quiz. https://www.hereditarycancerquiz.com/. Accessed April 1, 2012.

47. Welcome to Family HealthLink. https://familyhealthlink.osumc.edu/Notice.aspx. Accessed April 1, 2012.

48. A Quiz for HBOC Testing. http://www.bracnow.com/considering-testing/check-inherited-cancer-risk.php#brq. Accessed April 1, 2012.

49. Williams RR, Hunt SC, Barlow GK, et al. Health family trees: a tool for finding and helping young family members of coronary and cancer prone pedigrees in Texas and Utah. *Am J Public Health.* Oct 1988;78(10):1283-1286.

50. Johnson J, Giles RT, Larsen L, Ware J, Adams T, Hunt SC. Utah's Family High Risk Program: bridging the gap between genomics and public health. *Prev Chronic Dis.* Apr 2005;2(2):A24.

51. Hunt SC, Williams RR, Barlow GK. A comparison of positive family history definitions for defining risk of future disease. *J Chronic Dis.* 1986;39(10):809-821.

52. Jiang Y, Staes CJ, Adams TD, Hunt SC. Evaluation of risk scores derived from the health family tree program. *AMIA Annu Symp Proc.* 2009;2009:286-290.

53. Feero WG, Bigley MB, Brinner KM. New standards and enhanced utility for family health history information in the electronic health record: an update from the American Health Information Community's Family Health History

Multi-Stakeholder Workgroup. *J Am Med Inform Assoc.* Nov-Dec 2008;15(6):723-728.

54. Nielsen J. *Usability Engineering*. New York: AP professional, Academic Press; 1993.

55. Kushniruk AW, Patel VL, Cimino JJ. Usability testing in medical informatics: cognitive approaches to evaluation of information systems and user interfaces. *Proc AMIA Annu Fall Symp.* 1997:218-222.

56. Wilson BJ, Qureshi N, Santaguida P, et al. Systematic review: family history in risk assessment for common diseases. *Ann Intern Med.* Dec 15 2009;151(12):878-885.

57. Witten I, Frank E. *Data mining: practical machine learning tools and techniques*. 2 ed: Morgan Kaufmann; 2005.

58. Koh HC, Tan G. Data mining applications in healthcare. *J Healthc Inf Manag.* Spring 2005;19(2):64-72.

59. Obenshain MK. Application of data mining techniques to healthcare data. *Infect Control Hosp Epidemiol.* Aug 2004;25(8):690-695.

60. Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC Med Inform Decis Mak.* 2010;10:16.

61. Karaolis MA, Moutiris JA, Hadjipanayi D, Pattichis CS. Assessment of the risk factors of coronary heart events based on data mining with decision trees. *IEEE Trans Inf Technol Biomed.* May 2010;14(3):559-566.

62. Karaolis M, Moutiris JA, Papaconstantinou L, Pattichis CS. Association rule analysis for the assessment of the risk of coronary heart events. *Conf Proc IEEE Eng Med Biol Soc.* 2009;2009:6238-6241.

63. Barakat NH, Bradley AP, Barakat MN. Intelligible support vector machines for diagnosis of diabetes mellitus. *IEEE Trans Inf Technol Biomed.* Jul 2010;14(4):1114-1120.

64. Meaningful use matrix: Health IT policy council recommendations to national coordinator for defining meaningful use final—August 2009. http://healthit.hhs.gov/portal/server.pt?open=512&objID=1815&parentname=CommunityPage&parentid=7&mode=2&in_hi_userid=11113&cached=true. Accessed March 23, 2011.

CHAPTER 3

DOCUMENTING FAMILY HEALTH HISTORY FOR LONGITUDINAL STUDIES: WILL CURRENT REQUIREMENTS AND A NATIONAL TOOL MEET THE NEEDS?

Background

Family health history (FHH) is the description of the genetic relationships and medical history of a family.[1] Since families tend to live close to each other and share many lifestyle choices, such as diet and physical activity habits, FHH reflects the shared genetic susceptibility, environmental factors, and common behaviors among family members.[2] These factors interact with each other and are related to health status, so FHH can be used as a useful proxy for factors that contribute to disease. A positive FHH for cancer, diabetes, or heart diseases is considered a risk factor for these diseases.[3-7] Thus, FHH can be used to target prevention strategies towards individuals and populations at greater risk of certain diseases.[8]

FHH is useful in longitudinal population-based research. The National Children's Study (NCS) is a good example of a longitudinal study that examines the effects of natural or man-made environmental, biological, genetic, and psychosocial factors on the health and development of 100,000 children (from before birth until 21 years of age across the United States.[9] The goal of the NCS was to improve the health of children in

future generations by identifying the genetic and environmental contributors to disease. Understanding FHH may be particularly valuable for some of the NCS priority health conditions. FHH can be used as a proxy for factors that mediate exposure-disease relationships to a) stratify risk within the study population during descriptive analysis, and b) select cases and controls for future genetic analysis.

The collection of FHH has been enabled through the use of informatics methods and tools. In 2004, the Surgeon General initiated a national public health campaign[10] and released a web-based tool, *My Family Health Portrait,*[11] for the public to collect, save, and share their family history about multiple diseases and conditions with their healthcare providers and family members.[12] The tool is publicly available and uses standard vocabulary (including LOINC®, SNOMED-CT, and HL7 Vocabulary) and the HL7 family history data model to allow interoperability with electronic health records.[13] *My Family Health Portrait* may be useful for population-based longitudinal studies, but its adequacy has yet to be evaluated.

Currently, the collection of family health history information in the NCS is limited. Before expanding data collection about family health history within the NCS, it is important to understand the requirements for collecting family health history for longitudinal studies. In 2008, the American Health Information Community (AHIC) published data requirements (i.e., core data set) for representing FHH in an Electronic Health Record and Personal Health Record.[14] This core data set may or may not adequately address the data needs for a longitudinal study such as the NCS. Furthermore, the literature does not provide sufficient detail to define functional and nonfunctional requirements needed for a tool that documents FHH for longitudinal studies. Finally,

before adapting an existing tool for a population-based study to use, it is important to analyze the usability of the tool for use by participants in a study.

Therefore, the objectives of this research were to: 1) develop requirements for documenting FHH for a longitudinal population-based study and determine whether the requirements differ from the published requirements for integrating FHH into an electronic health record; and 2) evaluate whether the publicly available tool, *My Family Health Portrait*, meets the usability and other requirements identified.

Methods

Requirement analysis

The requirement analysis involved the development of new requirements and a comparison with published requirements for documenting FHH in an electronic health record. A variety of methods were used to explore the breadth and depth of requirements needed for the NCS. First, the technical requirements were defined through conversations with one of the NCS study directors and evaluating current standards related to representing family history.[15-17] Second, the data requirements were defined and confirmed using a variety of methods:

a) we reviewed the 2008 AHIC publication[14] and the design document used to develop a FHH tool for Intermountain Healthcare;

b) we used two existing tools and identified the required data elements to collect a FHH and use the information for risk assessment.[18] The two tools were: *My Family Health Portrait*,[11] the publicly-available web-based tool developed by the US Surgeon General's Office, and the *Health Family Tree,*[19,20] a paper and web-based tool developed by the University of Utah; and

c) we interviewed five domain experts to understand the clinical and patient-focused operational definition of clinical terms and to determine the minimum degree of relatives required for data collection. The five domain experts included: a pediatrician, and experts in preterm birth, obesity, asthma, and autism.

Third, the functional requirements were defined by reviewing the two tools listed above and the Intermountain Healthcare design document and considering requirements that emerged from the other methods. Fourth, the initial usability and security requirements were defined by considering a) usability standards,[21,22] b) usability issues that were brought up during interviews with clinicians, and c) usability issues identified during the evaluation of *My Family Health Portrait*. All the feedback from the usability evaluations were analyzed and iteratively incorporated into the usability requirements document. Finally, the ethical requirements were defined by examining the literature.[23,24]

The supersets of requirements we developed were classified into their logical groupings. For each requirement, we provided justification and assessed the relevance of the requirement from three perspectives: 1) an NCS study field manager who needs to collect and manage the information, 2) an NCS study participant who needs to collect and update the information periodically during at least 20 years of follow-up, and 3) an NCS researcher who may need to stratify the study population by levels of risk during analysis or follow-up. We compared the requirements we established for a longitudinal study such as the NCS with the published requirements for using FHH in an electronic health record or personal health record.[14]

Evaluation of *My Family Health Portrait*

We evaluated the Surgeon General's tool, *My Family Health Portrait*, in light of the requirements we developed. The tool was assessed to determine the domain of diseases it addresses, the data collection format, and the severity of unmet criteria. The primary author (YJ), who is a PhD student and a previous Preventive Medicine and Public Health Practitioner, explored the use of *My Family Health Portrait* from the perspective of a researcher and an NCS participant, determining whether each requirement was "met" or "not met."

Next, we evaluated *My Family Health Portrait* to determine its usability for the target NCS population and to elicit usability requirements. We performed the evaluation with five mothers of children under two years of age who resided in Salt Lake County. Three mothers were selected by convenience. Two of the mothers were selected using the Utah Population Database and Resource for Genetic and Epidemiologic Research, University of Utah. The latter two mothers were selected because their children had at least three first, second, and third degree relatives that had a preterm birth documented in their birth certificate record.

Two researchers conducted the usability evaluation in each participant's home setting. After obtaining consent, the participants were given tasks that required them to use *My Family Health Portrait* to collect their child's FHH information. We used a naturalistic observation with a "think aloud" technique whereby users talked about what they were doing as they interacted with the tool.[21] The observer provided minimal instructions or interruptions and took field notes about the experience.

After the visit, the researchers used a heuristic evaluation method[25] to determine the severity and types of heuristic violations associated with the problems the mothers encountered. Both researchers discussed and agreed on the heuristic violation categories and assigned severity ratings. If the participant had difficulty recalling particular information about relatives, we scheduled a follow-up home visit at the end of the initial visit to allow the participant to contact relatives about missing information. This study was reviewed and approved by the Institutional Review Board of University of Utah.

Results

Requirement analysis

We identified 50 requirements for a tool that may be used to document FHH for a longitudinal population-based study such as the NCS (Table 3.1). The requirements were classified into six categories: 1) technical requirements; 2) functional requirements for collecting, displaying, and storing FHH, including the data to be collected; 3) functional requirements for exporting and maintaining FHH in the context of the NCS; 4) usability requirements; 5) security requirements; and 6) ethical and legal requirements. The relevance of the requirements varied among the three perspectives assessed: a study manager, a study participant, and a researcher who would need to use the data to assess risk.

While most of the requirements are self-explanatory, selected requirements may need to be highlighted and explained. To meet the needs of a large-scale population-based study, the tool:

- should be a self-administered, web-based application to save time and resources.
- should include definitions of medical terms or only use nonmedical terms.

- should allow the user to periodically update their FHH, which will evolve during the course of the longitudinal study. This requires that the tool be able to record the version used and time stamp each update.
- should support a context for parents to proxy for their child (the proband) without confusion. In most situations involving NCS, a parent will need to fill in the information and use the tool for their child. The text and relationships displayed to the user should reflect this proxy context. When a family health history tool is designed for adult users to collect their own FHH, the user interface will reflect the adult's perspective. For example, a typical user interface may include questions such as “How many brothers do YOU have?” or “What is this person’s relationship to YOU?” If a mother uses the tool from her child’s perspective, the current *My Family Health Portrait* user interface could be confusing if the context is not clarified before data collection. If she uses the tool from her own perspective, the health history for the paternal side of the child’s family will not be collected, though the child’s cousins (i.e., the mother’s nieces and nephews) from both the maternal and paternal sides will probably be collected. If both parents collect their own history, their nieces’ and nephews’ (e.g., the child’s cousins’) information could be duplicated.
- should address how to store personal identifiers of the family members to enable data entry and updates while meeting privacy considerations. Though parents of the child consent to participate, the child’s FHH may contain health information about the child’s first, second, and even third degree relatives. Obtaining consent

from all the relatives is not practical nor may it be necessary because the history is being reported for the participants who have consented to participate in the study.

- should collect sufficient information to generate risk scores for analysis, and policies need to be developed to address whether, when, and how this information should be shared with study participants. When clear, actionable information is obtained about significant health risks, there is an ethical obligation to notify study participants.[26,27]

In general, the data requirements we defined for documenting FHH for a longitudinal population-based study are similar to the data requirements published by AHIC for integrating FHH into an electronic health record,[14] Both sets of requirements include data related to the individual's identification, relationship to the proband, gender, age, adoptive status, and year of death. While the AHIC requirements included "date of death" and "cause of death" in the core data set, we determined that it is also important to clearly document the "living status" at the time the history information is gathered. The "living status" is especially useful to avoid confusion when the "date of death" and "cause of death" are not available. In addition to the data requirements, we defined technical, functional, usability, security, and ethical requirements. These additional types of requirements were not clearly stated in the requirements published by AHIC.

Evaluation of *My Family Health Portrait*

The *My Family Health Portrait* tool met 36 (72%) of the 50 requirements we defined (see Table 3.1). All the data requirements we identified were met, while the functional requirements were less likely to be fulfilled. For example, the tool did not meet the following important functional and other requirements:

- Requirement 2.6: create a time stamp when the FHH is saved (useful for updating family history over time);
- Requirement 2.7: allow the user to define data sensitivity and data sharing status (useful for protecting the participants' privacy);
- Requirement 4.4: each health condition has a surface form that represents the common language used for explaining the concept (useful for the participants to understand the concepts and record accurate information);
- Requirement 4.9: allow researchers to annotate a participant's record (useful for assessing the risk for participants);
- Requirement 4.7: the user interface should support a context whereby a mother can proxy for her child, and the text and relationships should reflect this proxy context (useful for collecting complete family history for the child and avoiding confusion for the parent).

The usability of *My Family Health Portrait* was analyzed by observing use of the program by five participants: the biological mothers of a child living in the same household with the child's father. The mothers were 28 to 42 years of age, had a Bachelor's degree or above, and had regular access to computers and the Internet and confidence in their computer and Internet skills. The five participants took 30 to 90 minutes to enter, view, and print their FHH.

While entering, viewing, and printing the FHH, we observed 21 usability problems (see Table 3.2). Three participants were not able to find a specific disease they wanted to document in the drop-down list of diseases provided in the tool. For example, to record that a relative has asthma, the user had to open the dropdown list labeled "select

disease," then select "lung disease (more options…)," then open the dropdown list of "please specify (lung disease)," then select "asthma." The users could not find the disease because they expected an alphabetic list of diseases or they did not understand that asthma was located under the category of lung disease. This problem violated the usability heuristics of "*Match* between the system and the real world: the image of the system perceived by users should match the model that the users have about the system" and "*Visibility* of system: users should know what is going on with the application."[25]

In total, the problems we identified violated seven usability heuristics. The most frequently violated heuristics concerned the *match* between the user's expectation and the function of the tool. The second most frequently violated heuristic concerned the *language* used in the tool, which was not always presented in a form understandable to the user. None of the violations were catastrophic, requiring that the tool be fixed before it could be implemented; however, we identified major and minor violations that should be fixed and several cosmetic problems.

Discussion

FHH reflects shared genetic, environmental and behavioral risks within a family that may be important for guiding analyses and interpreting findings from longitudinal population-based studies such as the NCS. The limited FHH information collected during the pilot phase of the NCS could be enhanced using informatics strategies and tools. Our investigation identified requirements for an application to collect, store, and analyze FHH data for a longitudinal population-based study and evaluated an existing publicly available tool against the requirements. We found that the requirements for documenting FHH in an electronic health record[14] are similar to those required by a longitudinal

population-based study, and the Surgeon General's tool meets many but not all of the requirements we identified. The recently developed HL7 standards[15] for modeling and messaging family history information and the recent inclusion of "Documentation of family medical history" as one of the "Meaningful Use" objectives[28] further increase opportunities to collect family health histories and reuse already-collected data or build upon already-developed tools. To facilitate interoperability between different FHH applications, the HL7 Clinical Genomic Work Group is developing a family history model that will structure different data elements as attributes of multiples classes for the purpose of exchanging family history and genetic tests data, which is beyond the scope of this specific study.[15] But the data requirements we identified are consistent with the model that is currently being developed by the HL7 workgroup. All the data elements we identified are included in the HL7 family history model, with the exception of the "adoptive status" of relatives. We included the adoptive status of the relatives in the data requirements for a FHH because it is useful for risk analysis, given that FHH incorporates the effects of shared environmental and behavioral factors as well as genetics factors.[2] With an added data element of "adoptive status," any application that complies with the HL7 FHH standard model should be able to be used to document FHH for risk analysis for a longitudinal population-based study.

Our Study has several strengths. We identified requirements for documenting FHH for a longitudinal population-based study, a critical step before expanding data collection about FHH among the thousands of study participants that were expected in the National Children's Study. We expanded the data requirements defined by AHIC for representing FHH in an electronic health record and personal health record and included

requirements that will ensure that risk stratification can be performed.[18,29] The information gathered from the clinical domain experts was used to define the minimum degree of relatives and the functional needs for providing an operational definition of a health condition. We did not attempt to define the text that should be used, as this would require more extensive domain expertise and validation than was possible during this study. We identified, however, that a tool must allow for commonly used terms to be displayed to the user when they are being asked to consider whether they have family members with a particular condition. Finally, we highlighted the relevance of the requirements for different users: the study manager, the study participant involved with data collection and entry, and the researcher involved in data analysis. This strategy will allow the NCS stakeholders to evaluate and validate the requirements we defined in the context of their use.

The requirement analysis may have limitations. In particular, the analysis is partially subjective and based on the authors' experience. Even so, the mixed method approach to requirements development and iterative development process has strengths.

For a longitudinal population-based study, the timing of the collection of FHH information needs to be considered. No previous publication was identified that addresses the effect of the timing of collection of FHH on the quality of the data. FHH evolves over time, so the collection of a history is not necessarily required at the beginning of a longitudinal study. A complete and up-to-date history should be collected at the end of a study. There is a risk that study participants and their family members may not recall events or may not be alive to give a history, which will lead to the loss of some health history information. It may be acceptable to initiate the collection of FHH information

after the start of a study and update the information periodically and when new health events occur. The impact of the timing of collection on the quality of the health history needs further research.

The evaluation of the Surgeon General's *My Family Health Portrait* tool against the 50 requirements revealed strengths and weaknesses in this existing data collection tool. The tool did not meet all the requirements, and the major unmet requirements are valuable for administrators of longitudinal studies to consider when they select or modify tools for their own use. Additionally, the issue of instrumentation biases must be further explored and understood.

The usability assessment uncovered issues that should be addressed by NCS before selecting a web-based tool for collecting family history data. First and foremost, a tool should clearly define the context for data collection and documentation in the situation when a parent is entering information as a proxy for their child. Second, the tool should allow a user to easily select a health condition and review consumer-friendly descriptions for the conditions that are relevant for NCS. Studies have shown that the design of a survey instrument plays an important role in self-administered tools,[30,31] and there are two common response formats (radio buttons and drop-down boxes) that are often used by online surveys.[32] *My Family Health Portrait* used a drop-down box to collect a relative's health conditions. While Dillman et al.[33] suggest that radio buttons are favorable as they present questions in a similar way to a conventional paper form, no significant differences were found in the completion rate or time to completion between radio buttons and drop-down boxes.[32,34] Other research showed that radio buttons could be an initial barrier because they require longer downloading time than a drop-down

box.[35] The same study also used the number of answers that have nonsubstantial values as an indicator of data quality and found no significant difference between the data quality produced by these two response formats.[35] Furthermore, no literature was identified that addresses how the different formats would affect the accuracy of the self- or proxy-reported FHH data. In the context of proxy-reporting of a relative's health history, both formats have strengths and weaknesses. The radio button format explicitly steps a user through all the conditions of interest to force them to recall and evaluate the status of their relatives. This strategy may improve the sensitivity and specificity for documenting conditions among relatives. It may be time-consuming however, especially when there are many conditions of interest. The drop-down format is efficient and may not need to be updated for different studies, but it may result in users only recording the health events that are well known by the user while missing the less "memorable" conditions that would need to be sought in a pick list.

The usability assessment we performed had strengths and limitations. The assessment was performed in the participant's home, which is where most NCS participants would input their FHH. On the other hand, the participants may not be representative of the NCS study population. The participants were relatively young and well educated, with good computer literacy and skills. As a result, we may have missed usability problems that would be encountered by those with lower computer literacy and skills. Also, the heuristic evaluation method identified heuristic problems but did not identify design strengths that should be replicated. In addition, the heuristics evaluation relied on the evaluator's subjective judgment; however, in our evaluation, the two observers agreed on the results initially or after discussion. Finally, our enrollment for the

usability assessment was lower than expected. We attempted to enroll families from the community without an established relationship with the family. Our experience of unsuccessful enrollment is not unique as similar challenges face researchers conducting clinical trials[36] and previous studies have examined recruiting barriers.[37] To alleviate recruitment issues in the future, we recommend recruiting participants through health care facilities with pre-existing patient relationships and providing incentives to the participants to compensate for their time and effort. Despite the limitations, our study defined functional and nonfunctional requirements not yet available in the literature, evaluated how the national tool met the requirements, and the usability of the national tool. In addition, although we only had five participants for the usability assessment, past studies have shown that as few as three to six users can detect 80 percent of design problems and provide a maximal benefit to cost ratio.[21,38]

Conclusion

FHH is an important independent risk factor for many conditions and should be systematically collected when performing a longitudinal, population based study of health outcomes for diseases that may have a genetic component. The requirements for a tool to gather this information are similar to those for gathering FHH for electronic health records. We also identified features that should be considered when selecting a tool for the NCS. The Surgeon General's *My Family Health Portrait* tool does not currently meet all the requirements, but the developing FHH standards and tools built to those standards will meet most of the requirements for tools needed to collect data for longitudinal studies of health outcomes for diseases with a potential genetic component.

Table 3.1. Requirements for a tool to gather family health histories for a longitudinal study, using the National Children's Study as an example

#	Requirement description	Study manager	Participant	Researcher	Primary justification for the requirement
1.Technical requirements					
1.1	The tool should be self-administered, with user instructions	√	√	N/A	To save time and cost for population study
1.2	The tool should be a computerized, web-based application	√	√	√	Easy access for everyone involved
1.3	Comply with current terminology (such as SNOMED-CT), use GEDCOM as family tree structure, use HL7 for information model and data structure	√	N/A	√	To facilitate family health history data exchange
2.Functional requirements for collecting, displaying, and storing family health history					
2.1	Allow the user to create a family record	√	√	√	To build the family structure
2.2	Once a family record is created, automatically create an individual record for the proband (the child in NCS), the proband's father, and the proband's mother	√	√	√	To provide the user with the family members of a nuclear family to start collecting family health history
2.3	Allow the user to add additional individual records to the family record by adding relationships	√	√	√	To provide users the function of adding more family members
2.4	Allow both the family record and each individual record to be opened, edited, saved, closed, and updated when needed	√	√	√	To provide the functions for recording and editing the history throughout the duration of the longitudinal study
2.5	Allow family health history data to be stored in the web-based central location	√	√	√	To save the initial and follow-up data
2.6	Create a time stamp when the family health history is saved	√	√	√	To track the version of the family history and compare date of documentation with patient enrollment in NCS
Family record					
2.7	Allow the user to define data sensitivity and sharing status	N/A	√	N/A	To protect the privacy of the user
2.8	Allow data entry for the child's 1st, 2nd, and 3rd degree relatives	√	√	√	To cover the required degree of relationship by the four NCS diseases
2.8.1	Allow data entry for the child's 1st degree relatives	√	√	√	Parents and siblings are required by all four diseases history collection
2.8.2	Allow data entry for the child's 2nd degree relatives	√	√	√	Grandparents, aunts, uncles, nieces and nephews are required by obesity, asthma, and autism history collection
2.8.3	Allow data entry for the child's 3rd degree relatives	√	√	√	Cousins are required for asthma and autism history collection; great grandparents and cousins are required for obesity history collection
2.9	Display the whole family health history, either as a spreadsheet or pedigree	√	√	√	To view the big picture of the family health history
2.10	Allow the user to print the spreadsheet or pedigree	√	√	√	To enable the user to view and share family health history
2.11	Indicate the completeness of required data fields	√	√	√	To save time for continuous collection to complete the family record

Table 3.1 Continued

Individual's record					
2.12	Allow the user to indicate consanguinity	√	√	√	To help interpret the risk evaluation
2.13	Allow unknown values for each data field except "name" and "year of birth"	√	√	√	Name is needed for the user to identify family members, year of birth is needed for generating risk assessments
2.14	Allow text for explaining medical terms for lay people	N/A	√	N/A	To avoid user confusions
2.15	Allow the user to indicate the level of certainty for each health condition field	√	√	√	To record users' different level of certainty for recalled information
2.16	Each person's data sheet should contain a textbox for annotations	√	√	√	To record additional information
2.17	Allow the user to print a paper version with all fields whether they are filled or not	N/A	√	N/A	To enable users to use the unfilled paper version for data collection, the filled version for sharing
2.18	Each individual's data collection sheet should include the following individual's identifiers and demographic data	√	√	√	These are the basic data fields that should be included for family health history collection and use it for risk assessment
2.18.1	Individual's record identifier automatically assigned by the system	√	√	√	To uniquely identify individuals included in the family
2.18.2	Name	N/A	√	N/A	For the users to identify the relative for data entry
2.18.3	Relationship to the proband	√	√	√	To build the family pedigree
2.18.4	Gender	√	√	√	To identify the individual and to be used for the risk assessment
2.18.5	Year of birth	√	√	√	To identify the individual and to be used for the risk assessment
2.18.6	Living status	√	√	√	To be used for the risk assessment
2.18.7	Year of death	√	√	√	To be used for the risk assessment
2.18.8	Adoptive status	√	√	√	To be used for the risk assessment
2.19	Each individual data record should allow users to document the presence of none to many health conditions	√	√	√	To collect health condition information for the risk assessment
2.19.1	Health condition (including pre-term birth, obesity, asthma, and autism)	√	√	√	To be used for the risk assessment *Clinicians may collect the history of both mother giving birth to a preterm baby and the person born prematurely. We here only document the history of the person who was born prematurely
2.19.2	Year of onset	√	√	√	To be used for the risk assessment
3.Functional requirements for exporting, using, and maintaining of family health history in the context of NCS data management					
3.1	Export file in a structured and coded format that can be translated into HL7 standard code sets when it is stored for the NCS	√	N/A	√	NCS needs standardized data for data storage and exchange
3.2	Allow data to be extracted for risk assessment	√	N/A	√	To enable risk assessment based on the family health history data collected
3.3	Allow the user to update histories periodically over time	√	√	√	To enable data collection over years
3.4	Allow the user to save different versions of family history data	√	√	√	To enable view and compare different versions of family history data over years

Table 3.1 Continued

4.Usability requirements					
For participants					
4.1	Friendly interface that complies with the ISO usability standard (ISO 9241)	N/A	√	N/A	To facilitate the human-computer interaction
4.2	All text and instructions in the application should be written to a 6th grade reading level or lower	N/A	√	N/A	To meet the needs of users who have different levels of education, literacy…
4.3	When the user logs in, the system should display a welcome message with the user's name, and bring the user to the "family record" by default	N/A	√	N/A	To ease data input
4.4	Each health condition has a surface form that represents the common language used for the concept	N/A	√	N/A	To use user's language and avoid confusion
4.5	Display additional text explaining each health condition that is the interest of the study.	N/A	√	N/A	To help users understand medical terms. For example, the text explaining preterm birth may be "Birth less than 37 weeks of gestational age or 21 days or more earlier than the due date"; the text explaining autism may be "autism spectrum disorder, including autism, Asperger, and POD-NOS")
4.6	Display error messages properly	N/A	√	N/A	To ease the data input
4.7	The user interface should support a context whereby a mother can proxy for her child (the proband), and the text and relationships should reflect this proxy context. i.e, ask "relationship to child" rather than "relationship to me"	N/A	√	N/A	To avoid user confusions which could lead to incomplete data collection
For study managers and researchers					
4.8	Replicate participants' view	√	N/A	√	To facilitate the risk assessment process
4.9	Allow users to annotate a participant's record	√	N/A	√	To facilitate the risk assessment process
5.Security requirements					
5.1	Store family health history data securely and confidentially	√	√	√	To protect the participants' privacy and data security
5.2	User authentication and authorization, only authorized users have access	√	√	√	To protect the participants' privacy and data security
5.3	Log users out after a pre-specified period of inactivity (e.g., ~30 minutes)	N/A	√	N/A	To protect the participants' privacy and data security
6.Ethical and legal requirements					
6.1	Each individual has an unique identifier so when records are exported to the NCS for analysis, non-consenting relatives can be de-identified by excluding the individual names from the records	√	N/A	√	To protect the user's privacy. There are 18 specific identifiers need to be removed according to HIPAA "safe harbor" method. Name is one of the 18 identifiers being required to collect by NCS
6.2	Requirement for observational study: do not report risk assessment to the users unless actionable information is obtained about significant health risks	N/A	√	N/A	NCS is a pure observational study; communication risk to the users introduces interventions.

N/A = not applicable

Table 3.2. Summary of usability violations observed during use of *My Family Health Portrait* to collect family history for the NCS, 2010

Sub-tasks	Usability problem description	Heuristic violated	Violation severity
Adding first relative	User was not sure from whose (her self's or her child's) perspective she should collect family history data	Match	Major
Entering relative's living status	User skipped the "living status" question; thus, the tool did not capture any birth date information	Match	Major
Entering relative's living status	User indicated that a person was dead, then no birth date information would be asked and recorded	Match	Major
Entering relative's demographic information	After the user made a mistake of using the wrong format of birth date, an error message displayed and previous entered data of race and ethnicity were lost.	Undo	Major
Looking at the disease list	User commented that none of the 4 diseases was on the list	Match Visibility	Major Major
Entering relative's asthma history	User did not find asthma in the disease list, so she entered "asthma" as free text under "other diseases"	Match Visibility	Major Major
Entering relative's preterm history	User could not locate preterm birth from the dropdown list and did not know there was an option to enter free text	Match	Major
Saving family history	The process took longer than 30 seconds; user had to use another option (click "select the link") to save history.	Feedback	Major
Printing family history	"Preterm birth" and "Preterm Birth" were printed as different conditions	Match	Major
Entering number of relatives	User did not know whether to input just living relatives or all relatives	Language	Minor
Entering relative's birth date	User had to choose only one of the three ways (full birth date, age, or estimated age) to enter birth date. The user was trying to enter birth month, day and estimated age.	Match Flexibility	Minor Minor
Entering child's weight & height	User was not sure whether the 'weight & height' was referring to birth weight & height or current weight & height	Language	Minor
Entering relative's death cause	The tool only allowed one death cause and the user wanted to input more than one.	Match	Minor
Entering relative's preterm history	User was not sure about the definition of preterm birth	Language	Minor
Entering relative's asthma history	After a user typed "asthma", an error message showed up: "Stack over flow: 39"	Message Feedback Match	Minor Minor Minor
Viewing family history	User did not see the horizontal scroll bar for viewing the whole family tree	Visibility	Minor
Saving relative's condition	A button marked "add" needs to be clicked to save a relative's condition. The user was not sure about the function of the button but she tried it anyway.	Language	Cosmetic
Adding a cousin	User could not find how to add a cousin	Visibility	Cosmetic
Entering relative's estimated age	User was not sure about the definition of "in infancy" so defined it using her own definition as "less than 2 years old"	Language	Cosmetic
Opening saved family history	The "browse" and "open" buttons must be clicked in specific order. A user clicked "open" first and could not open the application	Match Language	Cosmetic Cosmetic
Entering relative's race and ethnicity	User was not aware of the difference between race and ethnicity	Language	Cosmetic

Key: Description of usability heuristics (Zhang, et al., 2003):

- **Match** between the system and the real world: the image of the system perceived by users should match the model the user have about the system. Also, assess the fit of the device with the kind of work that is being done;
- Reversible actions (**undo**): users should be allowed to recover from errors;
- **Visibility** of System State: users should know what is going on with the device or application;
- Informative **Feedback:** prompt, informative feedback about actions;
- Use the Users' **Language:** the language should be always presented in a form understandable by the users;
- **Flexibility** and efficiency: users need flexibility for shortcuts, customization;
- Goode error **message**s: users need to understand the errors and how to recover from the errors.

References

1. Bennett R. *The Practical Guide to the Genetic Family History*. New York: Wiley-Liss; 1999.

2. CDC. Awareness of family health history as a risk factor for disease: United States, 2004 *MMWR Morb Mortal Wkly Rep.* 2004;53:1044-1047.

3. Hunt SC, Gwinn M, Adams TD. Family history assessment: strategies for prevention of cardiovascular disease. *Am J Prev Med.* Feb 2003;24(2):136-142.

4. Harrison TA, Hindorff LA, Kim H, et al. Family history of diabetes as a potential public health tool. *Am J Prev Med.* Feb 2003;24(2):152-159.

5. Burke W, Fesinmeyer M, Reed K, Hampson L, Carlsten C. Family history as a predictor of asthma risk. *Am J Prev Med.* Feb 2003;24(2):160-169.

6. Keku TO, Millikan RC, Martin C, Rahkra-Burris TK, Sandler RS. Family history of colon cancer: what does it mean and how is it useful? *Am J Prev Med.* Feb 2003;24(2):170-176.

7. Ziogas A, Anton-Culver H. Validation of family history data in cancer family registries. *Am J Prev Med.* Feb 2003;24(2):190-198.

8. Trotter T, Martin H. Family history in pediatric primary care. *Pediatrics.* 2007;120 Suppl 2:S60-65.

9. AMA. Prenatal Screening Questionnaire. http://www.ama-assn.org/resources/doc/genetics/ped_screening.pdf. Accessed Dec 31, 2012.

10. Roncaglioni MC, Santoro L, D'Avanzo B, et al. Role of family history in patients with myocardial infarction. An Italian case-control study. GISSI-EFRIM Investigators. *Circulation.* Jun 1992;85(6):2065-2072.

11. My family health portrait. 2005; http://familyhistory.hhs.gov. Accessed March 18, 2013.

12. Guttmacher AE, Collins FS, Carmona RH. The family history—more important than ever. *N Engl J Med.* Nov 25 2004;351(22):2333-2336.

13. Huang HL, Chang FL. ESVM: evolutionary support vector machine for automatic feature selection and classification of microarray data. *Biosystems.* Sep-Oct 2007;90(2):516-528.

14. Feero WG, Bigley MB, Brinner KM. New standards and enhanced utility for family health history information in the electronic health record: an update from the American Health Information Community's Family Health History Multi-Stakeholder Workgroup. *J Am Med Inform Assoc.* Nov-Dec 2008;15(6):723-728.

15. HL7 Clinical Genomics. http://www.hl7.org/Special/committees/clingenomics/index.cfm. Accessed April 1, 2011.

16. UDOH. Data and confidence limits for percentage of adults who reported current cigarette smoking, adults aged 18 and older, Utah and U.S., 1989-2011. http://ibis.health.utah.gov/indicator/view_numbers/CigSmokAdlt.Ut_US.html. Accessed October 29, 2013.

17. UDOH. Data and confidence limits for percentage of adults who reported binge drinking in the past 30 days, Utah and U.S., 2005-2011. http://ibis.health.utah.gov/indicator/view_numbers/AlcConBinDri.UT_US.html. Accessed October 29, 2013.

18. Hunt SC, Williams RR, Barlow GK. A comparison of positive family history definitions for defining risk of future disease. *J Chronic Dis.* 1986;39(10):809-821.

19. Health Family Tree. 2005; http://healthfamilytree.utah.edu Accessed April 1, 2012.

20. Johnson J, Giles RT, Larsen L, Ware J, Adams T, Hunt SC. Utah's Family High Risk Program: bridging the gap between genomics and public health. *Prev Chronic Dis.* Apr 2005;2(2):A24.

21. Nielsen J. *Usability Engineering*. New York: AP Professional, Academic Press; 1993.

22. ISO. ISO 9241. International Organization for Standardization. http://www.iso.org/iso/catalogue_detail.htm?csnumber=52075. Accessed June 5, 2011.

23. Lounsbury DW, Reynolds TC, Rapkin BD, Robson ME, Ostroff J. Protecting the privacy of third-party information: recommendations for social and behavioral health researchers. *Soc Sci Med.* Jan 2007;64(1):213-222.

24. Botkin J. Protecting the privacy of family members in survey and pedigree research. *JAMA.* Jan 10 2001;285(2):207-211.

25. Zhang J, Johnson TR, Patel VL, Paige DL, Kubose T. Using usability heuristics to evaluate patient safety of medical devices. *J Biomed Inform.* Feb-Apr 2003;36(1-2):23-30.

26. Affleck P. Is it ethical to deny genetic research participants individualised results? *J Med Ethics.* Apr 2009;35(4):209-213.

27. Ravitsky V, Wilfond BS. Disclosing individual genetic results to research participants. *Am J Bioeth.* Nov-Dec 2006;6(6):8-17.

28. Meaningful use matrix: Health IT policy council recommendations to national coordinator for defining meaningful use final—August 2009.

http://healthit.hhs.gov/portal/server.pt?open=512&objID=1815&parentname=CommunityPage&parentid=7&mode=2&in_hi_userid=11113&cached=true. Accessed March 23, 2011.

29. Jiang Y, Staes CJ, Adams TD, Hunt SC. Evaluation of risk scores derived from the health family tree program. *AMIA Annu Symp Proc.* 2009;2009:286-290.

30. Couper M. Web surveys: a review of issues and approaches. *Public Opin Q.* Winter 2000;64(4):464-494.

31. Smith TW. Little things matter: a sampler of how differences in questionnaire format can affect survey responses. *Proc Am Stat Assoc.* 1995:1046-1051.

32. Healey B. Drop downs and scroll mice: the effect of response option format and input mechanism employed on data quality in Web surveys. *Social Science Computer Review.* Spring 2007 2007;25(1):111-128.

33. Dillman DA, Tortora RD, Bowker D. Principle for constructing web surveys (SESRC technical report). 1998; http://survey.sesrc.wsu.edu/dillman/papers/1998/principlesforconstructingwebsurveys.pdf. Accessed July 8, 2012.

34. Couper MP, Tourangeau R, Conrad FG, Crawford SD. What they see is what we get. *Social Science Computer Review.* February 1, 2004;22(1):111-127.

35. Heervegh D, Loosveldt G. An evaluation of the effect of response formats in data quality in Web surveys. *Social Science Computer Review.* 2002;20(4):471-484.

36. Caldwell PH, Hamilton S, Tan A, Craig JC. Strategies for increasing recruitment to randomised controlled trials: systematic review. *PLoS Med.* 2010;7(11):e1000368.

37. Cuggia M, Besana P, Glasspool D. Comparing semi-automatic systems for recruitment of patients to clinical trials. *Int J Med Inform.* Jun 2011;80(6):371-388.

38. Yao P, Gorman PN. Discount usability engineering applied to an interface for Web-based medical knowledge resources. *Proc AMIA Symp.* 2000:928-932.

CHAPTER 4

QUALITY ASSESSMENT OF THE DISEASE AND RISK FACTOR DATA AND RATES GENERATED FROM THE *HEALTH FAMILY TREE*

Background

Self-administered questionnaires are frequently used in population-based studies to obtain information about study subjects.[1] In current public health practice and epidemiology studies, questionnaires are often used to collect health related information[2] including family health history (FHH). FHH uses genetic relationships and the medical history of a family to assess each family member's risk of disease.[3] To record an individual's FHH, the informant needs to not only report on his or her own medical history (self-report) but also report on his or her first, second, or even third degree relative's medical history (proxy-report).[4,5] The accuracy of self- and proxy- reported FHH information remains unclear. According to a systematic review from a National Institute of Health conference in 2009, the accuracy of reporting FHH varies based on the disease being studied.[6] In general, correct reporting of the absence of disease in relatives was better than correct reporting of the existence of disease.[6] The results from multiple studies showed that the specificities of reporting family history of cancer were high, ranging from 0.91 to 1.00. In contrast, the sensitivities reported in these studies varied by the type of cancer: breast, 0.72 to 0.95; colon: 0.33 to 0.90; ovarian, 0.42 to 0.38; and

prostate, 0.47 to 0.79.[6] Similar patterns were observed when reporting family history of other diseases such as diabetes, hypertension, and cardiovascular disease: the specificities were high, ranging from 0.76 to 0.98, while the sensitivities varied from 0.18 to 0.89.[6] No clear association was observed between accuracy and informant age, sex, or educational level.[6]

The *Health Family Tree* program was initially developed by researchers in Utah and Texas in the early 1980s.[7] From 1983 to 2001, researchers collaborated with schools and distributed a paper questionnaire as a take-home assignment for high school students in Utah through their required Health Education class. The questionnaire (Figures 4.1 and 4.2) was used to document information about common diseases and general lifestyle risk factors about family members.[8] Students were instructed to finish the assignment with help from their parents. Then the content of the paper questionnaire was transferred onto a scan form, scanned, and stored in the computer-based *Health Family Tree* database. The family members included the student's siblings, parents, aunts and uncles, and grandparents.

The information collected about each individual by the *Health Family Tree* questionnaire can be categorized into three sets of information. The first set of information concerned demographic characteristics, including year of birth, sex, age (now or at death), living status causes of death, and blood relationship with the student or not. The second set of information concerned diseases and health conditions, including diabetes, myocardial infarction (MI), coronary heart disease (CHD), stroke, high blood pressure, high blood cholesterol, breast cancer, lung cancer, and colon cancer. The third set of information concerned lifestyle risk factors, including smoking, overweight/obese,

drinking, and exercise. The questions used to gather the disease and risk factor information are described in Table 4.1.

In 1986, an accuracy study was performed to assess the quality of the information reported about CHD by the students and their parents.[9] A subset of the families was selected and the family members were contacted to confirm the reported disease status using a mailed questionnaire with additional questions, a phone call, or a personal interview. Results showed the sensitivity of capturing CHD events was 67% and the specificity was 96%.[9] These performance measures are high but within the range of performance measures reported in other studies.[6]

The above accuracy study and other accuracy studies included in the National Institute of Health review used similar methods to verify a relative's actual disease status. They either directly contacted the individual associated with the information, reviewed medical records, or reviewed disease or death registries. These methods require locating each relative's records or interviewing the relative. While these methods directly capture the relative's history from an authoritative source or the person themselves (i.e., self report), there are several disadvantages. In particular, these methods are resource- and time-intensive because the verification process requires finding existing records or contacting multiple relatives. These methods may be justified when validating an individual's outcome and risk information needed to document 'observed' events within a single family; however, they may not be needed to determine the validity of self- and proxy-reported data to generate 'expected' rates in a prediction algorithm. Population-based rates of disease and risk factors may be derived from cancer registries (which are highly sensitive and specific) and from standardized public health survey interviews of

individuals who self-report their health status. The relationship between the self- and proxy- reported information in the *Health Family Tree* and information from these public sources is unknown.

There are several reasons to assess the quality of the self- and proxy-reported family health history information reported in the *Health Family Tree*. First, biases are known to be present when reporting FHH,[6] but the magnitude and direction of differences between self- and proxy-reported information is not known. Second, these differences may impact the interpretation of risk algorithms based on self- and proxy-reported data. Therefore, the objective of this study was to compare the disease and risk factor rates generated from the *Health Family Tree* database of self- and proxy-reported family health history data with rates available from authoritative public data sources.

Methods

Study population

Between 1983 and 2001, 57,238 high school students distributed throughout northern Utah completed a family health history. All records in the *Health Family Tree* database were candidates for this analysis, including 1,195,599 individuals' self- or proxy- reported medical history and lifestyle risk factors. The individuals may be represented twice, if more than one student from a single family participated in the class assignment. Prior to analyzing the data, we systematically checked all variables for errors and missing values. After cleaning, 1,021,909 (85.5%) valid records remained. We defined and handled errors and missing values in the following manner:

- When "age" was missing for "living" relatives, we calculated age using the reported year of birth.

- When there was a mismatch between the reported "sex" of the family member and the type of relative (i.e., grandmother should be a female), we used the relationship defined on the *Health Family Tree* (i.e., used the "relative number") to assign a value for 'sex.' For example, relative number 4 and number 6 are grandmothers of the high school student and their "sex" should be "female," and never "male."
- Records with more than four reasons of death were removed from the analysis assuming that they were errors.
- Records for parents of a high school student that report an age less than 25 years were removed from the analysis.
- Records with an invalid or uncorrectable "sex," "age," "relative number," or "year of birth" were removed from the analysis.

Comparison of rates generated from *Health Family Tree* and public sources

To compare the rates generated by the *Health Family Tree* with rates available from public data sources, we calculated rates from the *Health Family Tree* database. All diseases and risk factors were included except MI, drinking, and being overweight/obese. We did not identify any public data source that used definitions of drinking and overweight/obesity that were similar to those used by the *Health Family Tree*, and we were unable to find public data sources concerning MI in the time periods addressed by *Health Family Tree* (Table 4.2). We limited the *Health Family Tree* data to time periods that matched or were similar to the time frames of the available public data (Table 4.2). For diabetes, high blood pressure, high blood cholesterol, breast cancer, lung cancer,

colon cancer, and smoking, we selected the time period from 1990 to 1999. All records that were collected between 1990 and 1999 for persons that were alive at the time of data collection constituted the sample for the rate comparison for these diseases or risk factor. For CHD and stroke, we calculated rates based on information collected in 1996 because the public population-based data for Utah were only available for that year. For exercise, we generated rates of exercise using the latest five years (1997-2001) of data in the *Health Family Tree* database. Public data for Utah about exercise was only available for 2001, 2003, 2005, 2009, and 2010. The 2001 data alone was an insufficient sample, so we used data for all the years in the comparison.

We compared either incidence or prevalence rates depending on the available rates reported from public sources. For example, the public data available for cancers allows for the calculation of incidence rates, thus incidence rates were used for comparing cancer rates. All other diseases and risk factors were only reported as prevalence rates in the public data sets, thus prevalence rates were used for comparing noncancer diseases and risk factors. The following public data sources were used for the analysis:

- We obtained the prevalence rates from 1990-1999 of diabetes, high blood pressure, high blood cholesterol, smoking, and exercise from the Behavioral Risk Factor Surveillance System (BRFSS)[10] using the online portal of Utah's Indicator-Based Information System for Public Health.[11]
- We obtained prevalence rates for 1996 of CHD and stroke from the Utah Healthcare Access Survey (UHAS)[12] with assistance from analysts at the Utah Department of Health.

- We obtained incidence rates from 1990-1999 of breast, lung, and colon cancer from the Utah Cancer Registration (UCR)[13] using the SEER*Stat tool.[14]

We graphed the rates from each data source for each disease and risk factor to compare the magnitude of the rates and the pattern by sex and age group. We used the Cochran-Mantel-Haenszel statistic to test the significance of differences in the rates, as the Cochran-Mantel-Haenszel statistics are a collection of tests analyzing categorical data while controlling for covariates. We also used the chi-square test to examine disagreement for each sex- and age-subgroup. We also compared the rates reported by public data sources with the rates generated by the *Health Family Tree* stratified by proxy- versus self-report. Records concerning the students or their siblings and parents were classified as self-reported, as the parents were supposed to help the student fill out the questionnaire. Records concerning the student's aunts and uncles and grandparents were classified as proxy-reported.

The *Health Family Tree* computer-based program and SAS 9.2[15] were used to calculate and compare the rates. This study was reviewed and approved by the Institutional Review Board of University of Utah.

Results

Study population

A total of 1,021,909 valid records were available from students that participated in the *Health Family Tree* school assignment from 1983 to 2001. The demographic distribution of the records is shown in Table 4.3. About one-third (32.8%) of the records were for the students or their siblings and parents (which we classified as self-reported). The remaining two-thirds of the records were for the student's aunts and uncles and

grandparents (which we classified as proxy-reported). The proportion of records for students, mothers, fathers, maternal grandmothers, maternal grandfathers, paternal grandmothers, and paternal grandfathers were similar (5%-6% each). The proportion of records for student's siblings, maternal aunts and uncles, and paternal aunts and uncles were higher (14%-22%).

Comparison of rates generated from *Health Family Tree* and public sources

The *Health Family Tree* and public data sources have similar patterns of rates by age and sex groups (Figure 4.4). For example, the rates of all diseases increased with age, smoking rates are higher in the middle age groups, and exercise rates decrease with age. The agreement between rates generated by subjects in the *Health Family Tree* and public data varied across disease categories (Table 4.4). There was no significant difference in the rates reported for stroke (Cochran-Mantel-Haenszel (CMH) $p = 0.18$ overall, and $p = 0.1$ for self- and proxy-reported subjects), for self-reported breast cancer ($p = 0.08$) and lung cancer ($p = 0.25$), and for proxy-reported diabetes ($p = 0.05$)). However, for all other comparisons shown in Table 4.4, there was a significant difference in the rates reported in the *Health Family Tree* and the public data sources ($p < 0.05$). While there are exceptions, the low agreement was primarily due to underreporting of events in the *Health Family Tree* compared with events reported in public data sources (Tables 4.5-4.14; subgroup chi-square $p < 0.05$). Of the underreported diseases, high blood pressure and high blood cholesterol were severely underreported. Only a few sex- and age-specific groups reporting CHD (male 60-69 age group: mixed and proxy-reported), stroke (male 60-69 age group: mixed and proxy-reported), breast cancer (female 20-29 age group:

mixed, self-, and proxy-reported), lung cancer (male 20-29 age group: mixed, self-, and proxy-reported), and colon cancer (male and female 20-29 age groups: mixed, self-, and proxy-reported) in *Health Family Tree* yielded significantly higher rates than public sources (Tables 4.5-4.14; subgroup chi-square p values < 0.05). These over-reported subgroups were either the oldest (60-69) or youngest (20-29) age groups.

A comparison of self-reported information and proxy-reported information again showed variation. Both self-reported and proxy-reported rates, in comparison with public data, a) underreported diabetes, CHD, high blood pressure, high blood cholesterol, colon cancer, and smoking; b) reported stroke at a rate similar to the public data; and c) over-reported exercise rates. When comparing reported and public data for breast and lung cancer, the rates for self-reporting a history of breast cancer and lung cancer were not significantly different from the rates recorded in the public data. Proxy-reported rates of these two cancers were significantly lower than rates from public data sources, which caused the combined self- and proxy-reported rates to be significantly lower. Furthermore, though both self- and proxy-rates of high blood pressure and high blood cholesterol were significantly lower than public data, self-reported rates were higher than proxy-reported rates and were closer to the public rates. Conversely, self-reported rates for smoking were lower than proxy-reported rates and were more underreported in comparison with rates from public data sources. Self-reported rates of exercise were higher and more overreported than proxy-reported rates in comparison with public data rates.

Discussion

Family health history is an important independent risk factor that may be used in the prediction of certain chronic diseases. Most family health history data are self- or proxy-reported; therefore, it is valuable to assess the accuracy of this type of data prior to its use in clinical and public health decision-making. Previous research focused on comparison of patient self-reports with medical records at the individual level. In the current analysis, we used a different approach to examine the data accuracy, comparing population prevalence and incidence rates generated from the self- and proxy-reported *Health Family Tree* database with the corresponding rates generated from authoritative public data sources, such as cancer registries and standardized public health surveys.

We used the unique *Health Family Tree* database and extracted rates for different diseases and risk factors, stratified by sex and age groups from the database. We then compared the extracted rates with the public rates. We found that the disease and risk factor rates have similar patterns by sex and age as population rates reported in public data sources but were statistically significantly different, and generally lower than the population rates. One exception was reported exercise rates, which were higher than exercise rates reported in the public data set.

The comparison also showed that for different diseases and risk factors, the age- and sex-distribution showed similar patterns but agreement about the rates of the health events varied: the rates reported were similar for stroke but were significantly different for other diseases. Most of the diseases (and smoking behavior) included in the comparison study were underreported in the *Health Family Tree* database when compared to public data. On the other hand, exercise was overreported in the *Health*

Family Tree database. This analysis helped us to pinpoint the accuracy of FHH in terms of particular diseases and related disease behaviors. For example, more salient events such as stroke may be recalled better by the informants than other chronic conditions such as high blood pressure and high blood cholesterol, while protective factors (i.e., negative risk factors) such as exercise were estimated to be higher than the rates found in public data sources. Furthermore, self- and proxy-reported data were similar when reporting most of the diseases and the two lifestyle factors; however, they differed when reporting breast cancer and lung cancer (self-reported is accurate while proxy-reported is low).

There may be limitations in this study. The *Health Family Tree* population represents the northern Utah population while the standardized public health surveys and cancer registry represent the state population. The impact of this difference is likely minimal because approximately 80% of Utah residence live along the Wasatch Front in the northern part of the state.[16] The *Health Family Tree* data were collected from 1983 to 2001, so the major disease events in the database could have occurred during the early to late part of 20th century. We tried to address this issue by using the disease events that happened within time periods comparable to the available public population data sources. Finally, and most importantly, the questions used to ascertain information about diseases and risk factors in the *Health Family Tree* are not worded exactly the same as the questions used to gather information for the public data sources. For example, in the *Health Family Tree*, CHD was defined as answering "yes" to "Has he/she ever been told by a doctor that he/she **suffers** from" either "heart attack" or "coronary bypass surgery." In the UHAS, CHD was defined as answering "yes" to "Has a medical doctor or other

medical professional ever told you (him/her) that you (they) have heart disease, such as angina, congestive heart failure, or heart attack?" The difference in wording could lead to a significant difference in reporting. The term "suffer" may lead respondents to believe that the question is about their (dis)ability to respond to a medical condition (in other words, whether the condition is causing them to suffer or not). Furthermore, UHAS queried for more conditions (i.e., angina) when asking about CHD than were used by the *Health Family Tree*. This may explain the underreported rates of CHD in the *Health Family Tree* database. Similarly, vigorous exercise was defined more strictly in BRFSS than the *Health Family Tree*: "Vigorous exercise at least 3 times **for 20 minutes** per week" (BRFSS) vs. "Vigorous routine exercise at least 3 times per week" (*Health Family Tree*). This difference may explain the overreported rates of exercise in the *Health Family Tree* database. Though we only included the diseases and risk factors for which we could find similar definitions in the public data sources and excluded the ones that did not match (overweight/obesity, drinking), to what extent the rate of under- or overreporting was due to the different questions remains unknown. Despite existing limitations, the accuracy examination provides information about the quality of the self- and proxy-reported FHH data and hence should be considered when using these data for risk assessment.

Besides the questionnaire and database, another component of the *Health Family Tree* program is an algorithm to predict risk. This algorithm predicts risks based on the comparison between the number of observed events and expected events within each family. The number of expected events was calculated from the disease rates generated from the database itself. The expected events will be underestimated when the population

is underreporting events. The underestimated expected events will lead to an overestimated risk as the risk algorithm compares observed events (numerator) to expected events (denominator). The overestimated risk will categorize more people in the higher risk group and increase the chance of applying more strict screening strategies. In most public health screening strategies, this outcome is acceptable since the cost of screening healthy individuals is much lower than the treatment of affected patients. Persons identified as higher risk by a FHH-based tool may be provided with health education messages and may be directed to seek further recommended screening.

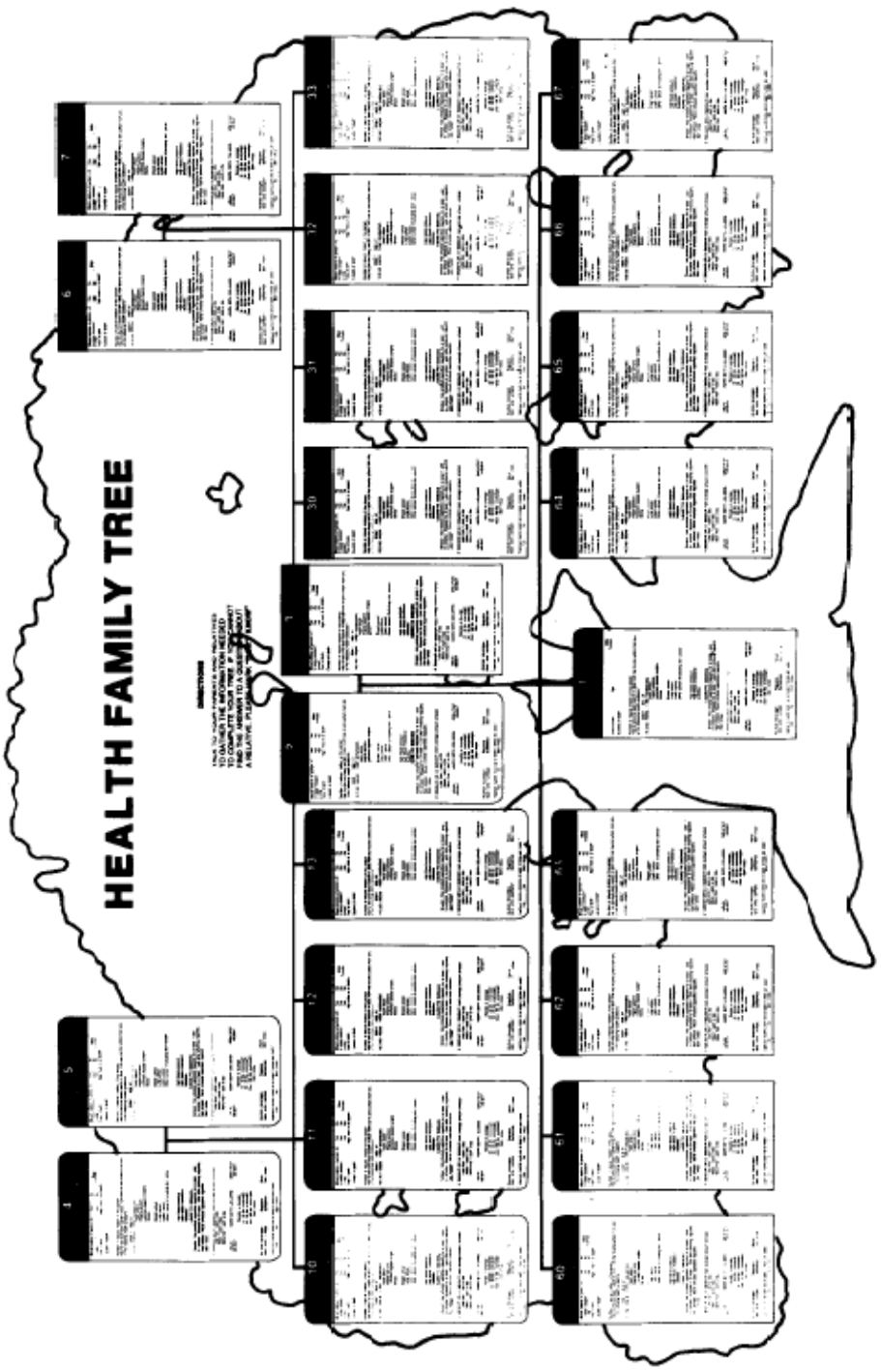
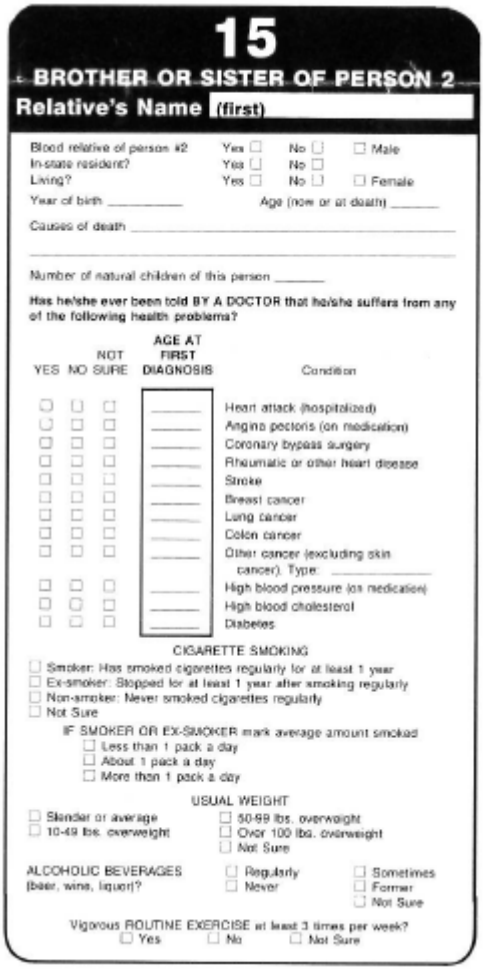


Figure 4.1. The *Health Family Tree* given to each family with a questionnaire box for each first degree relative of the student's parents.



15

BROTHER OR SISTER OF PERSON 2

Relative's Name (first)

Blood relative of person #2 Yes ☐ No ☐ ☐ Male
In-state resident? Yes ☐ No ☐
Living? Yes ☐ No ☐ ☐ Female

Year of birth _____ Age (now or at death) _____

Causes of death _____

_____

Number of natural children of this person _____

Has he/she ever been told BY A DOCTOR that he/she suffers from any of the following health problems?

YES	NO	NOT SURE	AGE AT FIRST DIAGNOSIS	Condition
☐	☐	☐	_____	Heart attack (hospitalized)
☐	☐	☐	_____	Angina pectoris (on medication)
☐	☐	☐	_____	Coronary bypass surgery
☐	☐	☐	_____	Rheumatic or other heart disease
☐	☐	☐	_____	Stroke
☐	☐	☐	_____	Breast cancer
☐	☐	☐	_____	Lung cancer
☐	☐	☐	_____	Colon cancer
☐	☐	☐	_____	Other cancer (excluding skin cancer). Type: _____
☐	☐	☐	_____	High blood pressure (on medication)
☐	☐	☐	_____	High blood cholesterol
☐	☐	☐	_____	Diabetes

CIGARETTE SMOKING

☐ Smoker: Has smoked cigarettes regularly for at least 1 year
☐ Ex-smoker: Stopped for at least 1 year after smoking regularly
☐ Non-smoker: Never smoked cigarettes regularly
☐ Not Sure

IF SMOKER OR EX-SMOKER mark average amount smoked
☐ Less than 1 pack a day
☐ About 1 pack a day
☐ More than 1 pack a day

USUAL WEIGHT

☐ Slender or average ☐ 50-99 lbs. overweight
☐ 10-49 lbs. overweight ☐ Over 100 lbs. overweight
☐ Not Sure

ALCOHOLIC BEVERAGES (beer, wine, liquor)? ☐ Regularly ☐ Sometimes ☐ Never ☐ Former ☐ Not Sure

Vigorous ROUTINE EXERCISE at least 3 times per week?
☐ Yes ☐ No ☐ Not Sure

Figure 4.2. The questionnaire box containing questions asked for each family member in the *Health Family Tree* program.

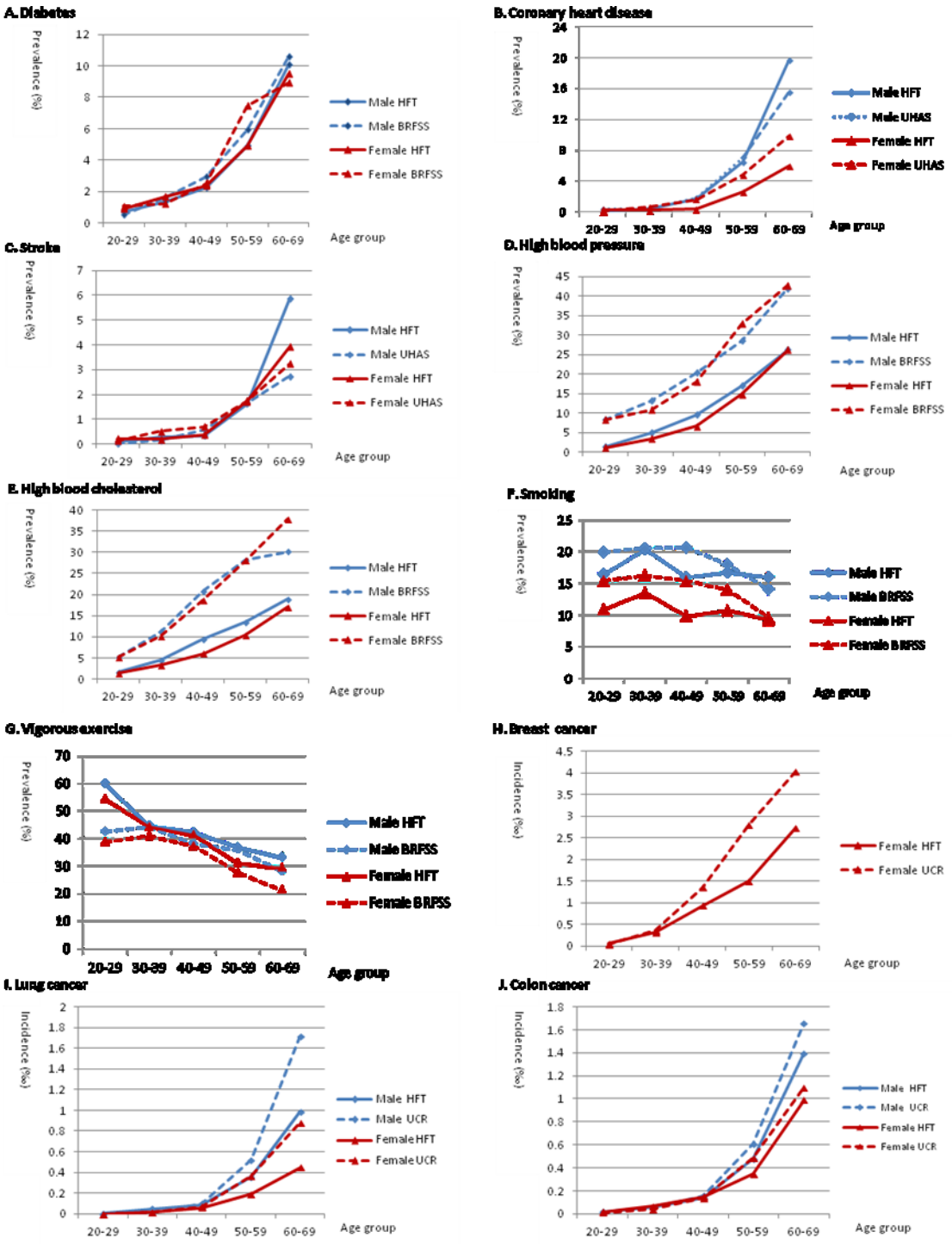


Figure 4.3 Disease and risk factor prevalence (%) or incidence rates (‰) generated from the *Health Family Tree* (HFT) database and the rates reported by public data sources including Utah Cancer Registry (UCR), Utah Behavioral Risk Factor Surveillance System (BRFSS), and Utah Healthcare Access Survey (UHAS).

Table 4.1. Diseases and lifestyle risk factors collected by the *Health Family Tree*

Disease/Risk factor	Question/Instruction to the informants	Value
Diabetes	Has he/she ever been told BY A DOCTOR that he/she suffers from **diabetes**?	Yes/No/Not sure
Myocardial infarction	Has he/she ever been told BY A DOCTOR that he/she suffers from **heart attack** (hospitalized)?	Yes/No/Not sure
Coronary heart disease	Has he/she ever been told BY A DOCTOR that he/she suffers from **heart attack** or **coronary bypass surgery**?	Yes/No/Not sure
Stroke	Has he/she ever been told BY A DOCTOR that he/she suffers from **stroke**?	Yes/No/Not sure
High blood pressure	Has he/she ever been told BY A DOCTOR that he/she suffers from **high blood pressure** (on medication)?	Yes/No/Not sure
High blood cholesterol	Has he/she ever been told BY A DOCTOR that he/she suffers from **high blood cholesterol**?	Yes/No/Not sure
Breast cancer	Has he/she ever been told BY A DOCTOR that he/she suffers from **breast cancer**?	Yes/No/Not sure
Lung cancer	Has he/she ever been told BY A DOCTOR that he/she suffers from **lung cancer**?	Yes/No/Not sure
Colon cancer	Has he/she ever been told BY A DOCTOR that he/she suffers from **colon cancer**?	Yes/No/Not sure
Cigarette smoking	Has smoked cigarettes regularly for at least 1 year	Smoker
	Stopped for at least one year after smoking regularly	Ex-smoker
	Never smoke cigarettes regularly	Non-smoker
Usual weight	Your opinion based on the person's usual weight	Slender or average/10-49 lbs. overweight/50-99 lbs. overweight/ Over 100 lbs overweight/Not sure
Alcoholic beverages (beer,	Drinking some type of alcohol (beer, wine, liquor) 3 or more times a week on the average	Regularly/Sometimes/Never /Former/Not sure
Routine exercise	Vigorous routine exercise at least 3 times per week. "Vigorous routine exercise" means the exercise that raises your heart rate and increases breathing for about half an hour or more without interruption. Jogging, aerobic dancing, and swimming are examples.	Yes/No/Not sure

Table 4.2. Diseases and risk factors that were included in the rates comparison between the *Health Family Tree* program and public data sources

	***Health Family Tree* (HFT)**		**Public data sources**	
	Data source & year	**Definition**	**Data source & year**	**Definition**
Diseases				
Diabetes	HFT 1990-1999	"Yes" to "Has he/she ever been told by a doctor that he/she suffers from ... [health problem]*?" All diseases were queried in the term as they are in the first column except CHD. CHD was defined as "yes" to either "heart attack" or "coronary bypass surgery"	BRFSS 1990-1999	"Yes" to "Have you ever been told by a doctor that you have diabetes?"
Coronary heart disease (CHD)	HFT 1996		UHAS 1996	"Yes" to "Has a medical doctor or other medical professional ever told you (him/her) that you (they) have heart disease, such as angina, congestive heart failure, or heart attack?
Stroke	HFT 1996		UHAS 1996	"Yes" to "Has a medical doctor or other medical professional ever told you (him/her) that you (they) have had a stroke?
High blood pressure	HFT 1990-1999		BRFSS 1990-1999	"Yes" to "Have you ever been told by a doctor, nurse, or other professional that you have high blood pressure?"
High blood cholesterol	HFT 1990-1999		BRFSS 1990-1999	"Yes" to "Have you ever been told by a doctor, nurse, or other professional that you r blood cholesterol is high?"
Breast cancer	HFT 1990-1999		UCR 1990-1999	Indicated as "breast cancer" in the registry
Lung cancer	HFT 1990-1999		UCR 1990-1999	Indicated as "lung cancer" in the registry
Colon cancer	HFT 1990-1999		UCR 1990-1999	Indicated as "colon cancer" in the registry
Lifestyle factors				
Smoking	HFT 1990-1999	Select "Has smoked cigarettes regularly for at least 1 year."	BRFSS 1990-1999	"Every day" or "Some days" to "Do you now smoke cigarettes every day, some days, or not at all?" AND who had smoked ≥100 cigarettes during their lifetime
Exercise	HFT 1997-2001	"Yes" to "Vigorous routine exercise at least 3 times per week."	BRFSS 2001, 2003, 2005, 2009, 2010	Vigorous exercise at least 3 times for 20 minutes per week

* The HFT program included a description of the diseases queried by the tool.
BRFSS: Behavioral Risk Factor Surveillance System; UHAS: Utah Healthcare Access Survey; UCR: Utah Cancer Registry

Table 4.3. Demographic distribution of records in the *Health Family Tree* database collected from 1983 to 2001

	Male	Female
Age in years	Number (%)	Number (%)
<20	71,764(7.0)	69,700(6.8)
20-29	49,283(4.8)	45,348(4.5)
30-39	89,053(8.7)	96,903(9.5)
40-49	126,349(12.4)	118,347(11.6)
50-59	62,949(6.2)	57,716(5.7)
60-69	54,819(5.4)	60,866(6.0)
>70	59,708(5.8)	58,104(5.7)
All	513,925(50.3)	507,984(49.7)

Table 4.4. Significance of the difference in disease and risk factor rates from the *Health Family Tree* and selected public data sources

	P values from Cochran-Mantel-Haenszel test		
	All subjects	**Self-reported subjects (students, siblings, and parents)**	**Proxy-reported subjects (grandparents, aunts and uncles)**
Diabetes	0.02	<0.01	0.05
Coronary heart disease	0.01	<0.01	0.03
Stroke	0.18	0.11	0.10
High blood pressure	<0.01	<0.01	<0.01
High blood cholesterol	<0.01	<0.01	<0.01
Breast cancer	<0.01	0.08	<0.01
Lung cancer	<0.01	0.25	<0.01
Colon cancer	<0.01	<0.01	<0.01
Smoking	<0.01	<0.01	<0.01
Vigorous exercising	<0.01	<0.01	<0.01

Table 4.5. Diabetes (DM) counts and prevalence from two data sources: Behavioral Risk Factor Surveillance Survey (BRFSS); *Health Family Tree* (HFT)

Sex	Age	DM	UHAS		HFT					
			Self- and proxy-reported		Mixed self+proxy-reported (everyone)		Self-reported (students, parents, siblings)		Proxy-reported (grandparents, uncles&aunts)	
			Counts	Prevalence	Counts	Prevalence	Counts	Prevalence	Counts	Prevalence
Male	20-29	Yes	11	0.54	140	0.70	79	0.59	61	0.92
		No	2152		19833		13253		6580	
	30-39	Yes	36	1.59	515	1.34	105	1.54	410	1.30
		No	2432		37938		5728		31210	
	40-49	Yes	57	2.95	1289	2.22	440	2.31	849	2.18**
		No	1943		56725		18580		38145	
	50-59	Yes	70	5.93	1256	4.94	252	4.65	1004	5.02
		No	1179		24144		5163		18981	
	60-69	Yes	108	10.58	2150	10.09	50	9.14	2100	10.12
		No	902		19151		497		18654	
Female	20-29	Yes	31	1.06	168	0.88	97	0.76**	71	1.14
		No	2663		18878		12708		6170	
	30-39	Yes	44	1.21	688	1.65	200	1.79	488	1.60
		No	3081		41057		10956		30101	
	40-49	Yes	71	2.51	1312	2.35	404	2.23**	908	2.42
		No	2371		54404		17746		36658	
	50-59	Yes	103	7.48	1185	4.92**	101	3.32**	1084	5.15**
		No	1476		22895		2945		19950	
	60-69	Yes	121	8.90	2410	9.49	13	7.26	2397	9.51
		No	1208		22977		166		22811	
Mantel Haenszel p value						0.02		<0.01		0.05

**P value <.05 when compared with BRFSS

Table 4.6. Coronary heart disease (CHD) counts and prevalence from two data sources: Utah Healthcare Access Survey (UHAS); *Health Family Tree* (HFT)

Sex	Age	CHD	UHAS		HFT					
			Self- and proxy-reported		Mixed self+proxy-reported (everyone)		Self-reported (students, parents, siblings)		Proxy-reported (grandparents, uncles&aunts)	
			Counts	Prevalence	Counts	Prevalence	Counts	Prevalence	Counts	Prevalence
Male	20-29	Yes	5	0.41	3	0.14	2	0.13	1	0.16
		No	1218		2126		1510		616	
	30-39	Yes	5	0.36	17	0.46	2	0.30	15	0.49
		No	1374		3700		672		3028	
	40-49	Yes	23	1.83	106	1.70	37	1.88	69	1.61
		No	1231		6144		1926		4218	
	50-59	Yes	56	7.09	179	6.48	29	5.01	150	6.86
		No	734		2585		550		2035	
	60-69	Yes	97	15.59	418	19.75**	10	15.38	408	19.89**
		No	525		1698		55		1643	
Female	20-29	Yes	2	0.14	7	0.35	5	0.35	2	0.34
		No	1405		2000		1421		579	
	30-39	Yes	9	0.65	11	0.26	6	0.53	5	0.17**
		No	1377		4154		1135		3019	
	40-49	Yes	21	1.63	25	0.42**	7	0.39**	18	0.44**
		No	1270		5905		1804		4101	
	50-59	Yes	39	4.81	68	2.60**	5	1.39**	63	2.80**
		No	771		2546		356		2190	
	60-69	Yes	67	9.83	151	5.96**	1	5.88**	150	5.96**
		No	615		2383		16		2367	
Mantel Haenszel p value						0.01		<0.01		0.03

**P value <.05 when compared with UHAS

Table 4.7. Stroke counts and prevalence from two data sources: Utah Healthcare Access Survey (UHAS); *Health Family Tree* (HFT)

Sex	Age	Stroke	UHAS		HFT					
			Self- and proxy-reported		Mixed self+proxy-reported (everyone)		Self-reported (students, parents, siblings)		Proxy-reported (grandparents, uncles&aunts)	
			Counts	Prevalence	Counts	Prevalence	Counts	Prevalence	Counts	Prevalence
Male	20-29	Yes	0	0	2	0.09	0	0.00	2	0.32
		No	1222		2127		1512		615	
	30-39	Yes	2	0.15	10	0.27	1	0.15	9	0.30
		No	1377		3707		673		3034	
	40-49	Yes	7	0.56	20	0.32	5	0.25	15	0.35
		No	1249		6230		1958		4272	
	50-59	Yes	13	1.65	45	1.63	8	1.38	37	1.69
		No	777		2719		571		2148	
	60-69	Yes	17	2.73	124	5.86**	3	4.62	121	5.90**
		No	606		1992		62		1930	
Female	20-29	Yes	2	0.14	4	0.20	2	0.14	2	0.34
		No	1406		2003		1424		579	
	30-39	Yes	7	0.51	8	0.19	2	0.18	6	0.20
		No	1379		4157		1139		3018	
	40-49	Yes	9	0.70	22	0.37	8	0.44	14	0.34
		No	1284		5908		1803		4105	
	50-59	Yes	14	1.73	45	1.72	5	1.39	40	1.78
		No	796		2569		356		2213	
	60-69	Yes	22	3.23	99	3.91	0	0.00	99	3.93
		No	659		2435		17		2418	
Mantel Haenszel p value						0.18		0.11		0.10

**P value <.05 when compared with BRFSS

Table 4.8. High blood pressure (HBP) counts and prevalence from two data sources: Behavioral Risk Factor Surveillance Survey (BRFSS); *Health Family Tree* (HFT)

Sex	Age	HBP	UHAS		HFT					
			Self- and proxy-reported		Mixed self+proxy-reported (everyone)		Self-reported (students, parents, siblings)		Proxy-reported (grandparents, uncles&aunts)	
			Counts	Prevalence	Counts	Prevalence	Counts	Prevalence	Counts	Prevalence
Male	20-29	Yes	123	8.31	276	1.38	157	1.18	119	1.79
		No	1358		19697		13175		6522	
	30-39	Yes	219	13.08	1862	4.84	487	7.13	1375	4.35
		No	1455		36591		6346		30245	
	40-49	Yes	275	20.24	5578	9.61	2257	11.87	3321	8.52
		No	1084		52436		16763		35673	
	50-59	Yes	237	28.52	4300	16.93	1122	20.72	3178	15.90
		No	594		21100		4293		16807	
	60-69	Yes	294	42.00	5586	26.22	148	27.06	5438	26.20
		No	406		15715		399		15316	
Female	20-29	Yes	153	8.33	207	1.09	124	0.97	83	1.33
		No	1682		18839		12681		6158	
	30-39	Yes	235	10.74	1400	3.35	460	4.12	940	3.07
		No	1954		40345		10696		29649	
	40-49	Yes	294	18.04	3689	6.62	1273	7.01	2416	6.43
		No	1336		52028		16877		35150	
	50-59	Yes	352	32.84	3572	14.83	474	15.56	3098	14.73
		No	720		20508		2572		17936	
	60-69	Yes	379	42.73	6678	26.30	41	22.91	6637	26.33
		No	508		18709		138		18571	
Mantel Haenszel p value						<0.01		<0.01		<0.01

All sub-group p values <.05 when compared with BRFSS

Table 4.9. High blood cholesterol (HBC) counts and prevalence from two data sources: Behavioral Risk Factor Surveillance Survey (BRFSS); *Health Family Tree* (HFT)

Sex	Age	HBC	UHAS		HFT					
			Self- and proxy-reported		Mixed self+proxy-reported (everyone)		Self-reported (students, parents, siblings)		Proxy-reported (grandparents, uncles&aunts)	
			Counts	Prevalence	Counts	Prevalence	Counts	Prevalence	Counts	Prevalence
Male	20-29	Yes	67	5.16	315	1.58	195	1.46	120	1.81
		No	1231		19658		13137		6521	
	30-39	Yes	165	11.23	1753	4.56	501	7.33	1252	3.96
		No	1304		36700		6332		30368	
	40-49	Yes	249	20.77	5496	9.47	2689	14.14	2807	7.20
		No	950		52518		16331		36187	
	50-59	Yes	210	28.15	3432	13.51	1024	18.91	2408	12.05
		No	536		21968		4391		17577	
	60-69	Yes	189	30.14	4018	18.86	115	21.02	3903	18.81
		No	438		17283		432		16851	
Female	20-29	Yes	84	5.20	255	1.34	164	1.28	91	1.46
		No	1530		18791		12641		6150	
	30-39	Yes	197	10.23	1377	3.30	526	4.71	851	2.78
		No	1728		40368		10630		29738	
	40-49	Yes	270	18.63	3363	6.04	1435	7.91	1928	5.13
		No	1179		52353		16715		35638	
	50-59	Yes	268	28.03	2520	10.47	410	13.46	2110	10.03
		No	688		21560		2636		18924	
	60-69	Yes	298	37.82	4341	17.10	31	17.32	4310	17.10
		No	490		21046		148		20898	
Mantel Haenszel p value						<0.01		<0.01		<0.01

All sub-group p values <.05 when compared with BRFSS

Table 4.10. Smoking counts and prevalence from two data sources: Behavioral Risk Factor Surveillance Survey (BRFSS); *Health Family Tree* (HFT)

Sex	Age	Smoker	UHAS		HFT					
			Self- and proxy-reported		Mixed self+proxy-reported (everyone)		Self-reported (students, parents, siblings)		Proxy-reported (grandparents, uncles&aunts)	
			Counts	Prevalence	Counts	Prevalence	Counts	Prevalence	Counts	Prevalence
Male	20-29	Yes	431	19.94	3157	16.48**	1854	14.40**	1303	20.74
		No	1731		16005		11024		4981	
	30-39	Yes	507	20.55	7517	20.43	1505	22.61**	6012	19.95
		No	1960		29270		5150		24120	
	40-49	Yes	414	20.68	8933	15.96**	2165	11.59**	6768	18.15**
		No	1588		47045		16515		30530	
	50-59	Yes	225	17.97	4095	16.75	462	8.69**	3633	19.00
		No	1027		20347		4855		15492	
	60-69	Yes	142	14.06	3287	15.99	47	8.77**	3240	16.18
		No	868		17273		489		16784	
Female	20-29	Yes	416	15.45	2004	10.89**	1160	9.33**	844	14.16
		No	2277		16397		11279		5118	
	30-39	Yes	509	16.30	5473	13.52**	1629	14.83**	3844	13.03**
		No	2614		35019		9359		25660	
	40-49	Yes	377	15.44	5321	9.82**	1139	6.36**	4182	11.53**
		No	2065		48860		16762		32093	
	50-59	Yes	222	14.07	2500	10.69**	95	3.15**	2405	11.80
		No	1356		20888		2918		17970	
	60-69	Yes	127	9.57	2300	9.28	15	8.62	2285	9.28
		No	1200		22497		159		22338	
Mantel Haenszel p value						<0.01		<0.01		<0.01

**P value <.05 when compared with BRFSS

Table 4.11. Exercise counts and prevalence from two data sources: Behavioral Risk Factor Surveillance Survey (BRFSS); *Health Family Tree* (HFT)

Sex	Age	Exercise	UHAS		HFT					
			Self- and proxy-reported		Mixed self+proxy-reported (everyone)		Self-reported (students, parents, siblings)		Proxy-reported (grandparents, uncles&aunts)	
			Counts	Prevalence	Counts	Prevalence	Counts	Prevalence	Counts	Prevalence
Male	20-29	Yes	576	42.51	2773	59.93**	2183	62.69**	590	51.53**
		No	779		1854		1299		555	
	30-39	Yes	824	44.21	3135	44.63	622	45.90	2513	44.33
		No	1040		3889		733		3156	
	40-49	Yes	745	38.11	5741	42.22**	2137	44.83**	3604	40.81**
		No	1210		7858		2630		5228	
	50-59	Yes	687	35.95	2314	36.79	713	43.40**	1601	34.45
		No	1224		3976		930		3046	
	60-69	Yes	390	28.10	1507	33.08**	35	34.31**	1472	33.05**
		No	998		3049		57		2982	
Female	20-29	Yes	679	38.93	2457	54.32**	1931	57.28**	526	45.66**
		No	1065		2066		1440		626	
	30-39	Yes	1033	40.93	3552	44.25**	1150	48.65**	2402	42.41
		No	1491		4476		1214		3262	
	40-49	Yes	908	37.30	5512	41.37**	2273	48.48**	3239	37.51
		No	1526		7812		2416		5396	
	50-59	Yes	652	27.56	1813	30.99**	398	40.95**	1415	29.01
		No	1714		4037		574		3463	
	60-69	Yes	373	21.24	1661	29.22**	10	26.32	1651	29.24**
		No	1383		4024		28		3996	
Mantel Haenszel p value						<0.01		<0.01		<0.01

**P value <.05 when compared with BRFSS

Table 4.12. Breast cancer counts and incidence rates from two data sources: Utah Cancer Registry (UCR); *Health Family Tree* (HFT)

Sex	Age	Breast cancer	UCR		HFT					
			Cancer registry		Mixed self+proxy-reported (everyone)		Self-reported (students, parents, siblings)		Proxy-reported (grandparents, uncles&aunts)	
			Counts	Incidence (‰)	Counts	Incidence (‰)	Counts	Incidence (‰)	Counts	Incidence (‰)
Female	20-29	Cases	58	0.04	119	0.06**	37	0.10**	82	0.06**
		Person-years	1626620		1830769		377551		1438596	
	30-39	Cases	522	0.36	504	0.32**	125	0.42	379	0.29**
		Person-years	1435745		1584906		295508		1289116	
	40-49	Cases	1574	1.35	964	0.93**	178	1.50	786	0.86**
		Person-years	1163077		1031016		118746		912892	
	50-59	Cases	2076	2.81	945	1.51**	25	1.82**	920	1.50**
		Person-years	738287		627074		13736		613333	
	60-69	Cases	2294	4.03	1084	2.74**	4	1.69**	1080	2.75**
		Person-years	568905		395909		2371		393443	
Mantel Haenszel p value						<0.01		0.08		<0.01

**P value <.05 when compared with UCR

Table 4.13. Lung cancer counts and incidence rates from two data sources: Utah Cancer Registry (UCR); *Health Family Tree* (HFT)

Sex	Age	Lung cancer	UCR		HFT					
			Cancer registry		Mixed self+proxy-reported (everyone)		Self-reported (students, parents, siblings)		Proxy-reported (grandparents, uncles&aunts)	
			Counts	Incidence (‰)	Counts	Incidence (‰)	Counts	Incidence (‰)	Counts	Incidence (‰)
Male	20-29	Cases	4	0.00	23	0.01**	7	0.02**	16	0.01**
		Person-years	1659401		1769231		368421		1454545	
	30-39	Cases	28	0.02	47	0.05	11	0.04	36	0.03
		Person-years	1473926		1566667		297297		1241379	
	40-49	Cases	122	0.10	96	0.09	18	0.11	78	0.09
		Person-years	1163520		1032258		159292		866667	
	50-59	Cases	376	0.52	215	0.36**	10	0.35**	205	0.36**
		Person-years	717142		593923		28736		566298	
	60-69	Cases	894	1.72	380	0.99**	4	0.90**	376	0.99**
		Person-years	518444		385396		4449		380952	
Female	20-29	Cases	7	0.00	6	0.00	2	0.01	4	0.00
		Person-years	1626620		2000000		400000		1333333	
	30-39	Cases	24	0.02	40	0.02	10	0.03	30	0.02
		Person-years	1435745		1600000		294118		1304348	
	40-49	Cases	82	0.07	61	0.06	8	0.07	53	0.06
		Person-years	1163077		1033898		119403		913793	
	50-59	Cases	270	0.37	121	0.19**	4	0.29	117	0.19**
		Person-years	738287		633508		13841		622340	
	60-69	Cases	498	0.88	183	0.45**	1	0.42**	182	0.45**
		Person-years	568905		404867		2375		402655	
Mantel Haenszel p value						<0.01		0.25		<0.01

**P value <.05 when compared with UCR

Table 4.14. Colon cancer counts and incidence rates from two data sources: Utah Cancer Registry (UCR); *Health Family Tree* (HFT)

Sex	Age	Colon cancer	UCR		HFT					
			Cancer registry		Mixed self+proxy-reported (everyone)		Self-reported (students, parents, siblings)		Proxy-reported (grandparents, uncles&aunts)	
			Counts	Incidence (‰)	Counts	Incidence (‰)	Counts	Incidence (‰)	Counts	Incidence (‰)
Male	20-29	Cases	15	0.01	43	0.02**	8	0.02**	35	0.02**
		Person-years	1659401		1791667		380952		1400000	
	30-39	Cases	75	0.05	90	0.06	21	0.07	69	0.06
		Person-years	1473926		1551724		300000		1254545	
	40-49	Cases	191	0.16	147	0.14	28	0.18	119	0.14
		Person-years	1163520		1027972		158192		868613	
	50-59	Cases	439	0.61	284	0.48**	18	0.63	266	0.47**
		Person-years	717142		594142		28662		565957	
	60-69	Cases	861	1.66	539	1.40**	8	1.81	531	1.40**
		Person-years	518444		384177		4425		379828	
Female	20-29	Cases	13	0.01	35	0.02**	9	0.02**	26	0.02**
		Person-years	1626620		1842105		375000		1444444	
	30-39	Cases	58	0.04	107	0.07**	22	0.07**	85	0.07**
		Person-years	1435745		1597015		297297		1287879	
	40-49	Cases	158	0.14	156	0.15	20	0.17	136	0.15
		Person-years	1163077		1033113		119048		918919	
	50-59	Cases	359	0.49	221	0.35**	13	0.94**	208	0.34**
		Person-years	738287		633238		13815		619048	
	60-69	Cases	625	1.10	401	0.99**	1	0.42**	400	1.00
		Person-years	568905		403421		2358		401204	
Mantel Haenszel p value						<0.01		<0.01		<0.01

**P value <.05 when compared with UCR

References

1. Enterline PE, Capt KG. A validation of information provided by household respondents in health surveys. *Am J Public Health Nations Health.* Feb 1959;49(2):205-212.

2. Huerta JM, Tormo MJ, Egea-Caparros JM, Ortola-Devesa JB, Navarro C. Accuracy of self-reported diabetes, hypertension and hyperlipidemia in the adult Spanish population. DINO study findings. *Rev Esp Cardiol.* Feb 2009;62(2):143-152.

3. Bennett R. *The Practical Guide to the Genetic Family History*. New York: Wiley-Liss; 1999.

4. Nelson LM, Longstreth WT, Jr., Koepsell TD, van Belle G. Proxy respondents in epidemiologic research. *Epidemiol Rev.* 1990;12:71-86.

5. Herrmann N. Retrospective information from questionnaires. I. Comparability of primary respondents and their next-of-kin. *Am J Epidemiol.* Jun 1985;121(6):937-947.

6. Wilson BJ, Qureshi N, Santaguida P, et al. Systematic review: family history in risk assessment for common diseases. *Ann Intern Med.* Dec 15 2009;151(12):878-885.

7. Williams RR, Hunt SC, Barlow GK, et al. Health family trees: a tool for finding and helping young family members of coronary and cancer prone pedigrees in Texas and Utah. *Am J Public Health.* Oct 1988;78(10):1283-1286.

8. Johnson J, Giles RT, Larsen L, Ware J, Adams T, Hunt SC. Utah's Family High Risk Program: bridging the gap between genomics and public health. *Prev Chronic Dis.* Apr 2005;2(2):A24.

9. Hunt SC, Williams RR, Barlow GK. A comparison of positive family history definitions for defining risk of future disease. *J Chronic Dis.* 1986;39(10):809-821.

10. CDC. Behavioral Risk Factor Suveillance System. http://www.cdc.gov/brfss/. Accessed April 1, 2012.

11. UDOH. Welcome to IBIS-PH: Utah's public health data resource. http://ibis.health.utah.gov/. Accessed October 29, 2013.

12. UDOH. Utah Healthcare Access Survey (UHAS). http://health.utah.gov/opha/OPHA_UHAS.htm. Accessed April 1, 2012.

13. UCR. Utah Cancer Registry. http://ucr.utah.edu/. Accessed April 1, 2012.

14. ADA. Diabetes statistics. http://www.diabetes.org/diabetes-basics/diabetes-statistics/. Accessed October 29, 2013.

15. SAS. About SAS. http://www.sas.com. Accessed April 1, 2012.

16. Utah State Government. *The Wasatch Fault*. Salt Lake City: Utah State Government; 1996.

CHAPTER 5

SUCCESSFUL RISK PREDICTION FOR COMMON DISEASES USING FAMILY HEALTH HISTORY

Background

A person's family health history (FHH) is a valuable, noninvasive, and relatively inexpensive tool for predicting his or her risk of disease. Family health histories have aided clinicians making diagnosis and treatment decisions[1] and have been used by public health professionals to identify high risk populations for disease prevention interventions, screening, and research.[2] FHH has been used to independently predict the risk of many prevalent chronic diseases. For example, a positive FHH has been associated with increased risk of coronary heart disease (CHD) and myocardial infarction (MI).[3-9] The association remained significant after adjusting for smoking, hypertension, high cholesterol, obesity, and socioeconomic status.[4-7] Similar predictive effects were found for stroke,[10,11] diabetes,[12-15] and several cancers.[16-19]

A variety of risk assessment tools based on FHH information have been developed,[20-29] including *My Family Health Portrait*,[29] a tool released in 2009 by the office of the US Surgeon General. Most of the tools are useful for collecting and displaying information for a clinician to review and identify potential risks. Besides collecting and displaying family history information, other tools also assess risks for

healthy individuals using rule-based logic or regression models. For example, *MyGenerations*[28] provides cancer risk assessments, and *Your Disease Risk*[23] assesses risk for cancer, diabetes, heart diseases, and other common diseases. In the early 1980s, a tool called *Health Family Tree* was developed at the University of Utah and was used to collect and use family health history and general lifestyle information to predict risks for chronic diseases.[22] From 1983 to 2001, family health history and lifestyle information were collected from 57,238 high school students in Utah during their required high school health class.[31] The information was collected by each student with help from their parents on a 36 x 23 inch folding paper that was designed to fit on a kitchen table. The information queried by the *Health Family Tree* about each individual can be categorized into three groups: demographic information, disease information, and lifestyle risk factor information. The collected information was then stored in a database and an algorithm was developed to automatically predict the risk for a variety of diseases.[31] The risk prediction algorithm compared the observed number of disease events to the expected number of disease events within each family. The expected number of events was calculated by multiplying the age- and sex-specific person-years for each person in the family by the age- and sex- specific incidence rates generated from the entire set of records in the database.[31] The large volume of persons represented in the data and the systematic collection of information from most high school students throughout the state led to the assumption that the expected rate in the general population could be derived from the database itself.

The risk algorithm developed for the *Health Family Tree* tool was validated in 1986 for predicting heart diseases.[31] The researchers found that the definition of elevated

risk from the *Health Family Tree* algorithm successfully predicted unaffected family member's risk of developing future CHD. The researchers also assessed the quality of the information reported by the student and their parents for their relatives. They selected a subset of the families, and contacted the family members by mailing a questionnaire with additional questions, phone calling, or personal interviews to confirm the reported disease status. The sensitivity of capturing disease events was 67% and the specificity was 96%.[31]

The risk algorithm has not been validated for predicting the other diseases included in the family health history captured by the *Health Family Tree,* although preliminary analyses concerning MI and diabetes were reported in 2009.[32] In addition, no validation has been performed to assess the impact of the lifestyle risk factors gathered by the tool, including smoking, drinking, overweight/obesity, and exercise. There are two major reasons for performing such a validation. First, the original paper-based *Health Family Tree* tool had been implemented as a web-based tool that would allow automated calculation of risk scores that may be presented to a user.[33] Validation of the risk scores for all the diseases included in the tool is important before general use and presentation of risk scores to users. Second, the algorithm used by *Health Family Tree* could be applied to other family history tools that document events observed in a family; however, to predict future risk, the tool would need to define expected rates of disease for the relatives included in the family. When a tool has no historic data that can be used to generate expected rates, the system would need to use reference rates in the risk algorithm and understand their impact on risk prediction if the rates are likely to under- or overestimate the true population rates.

Therefore, the objectives of this research were: 1) to describe the absolute rates generated from the *Health Family Tree* database and used in the risk prediction; 2) to validate the *Health Family Tree* risk algorithm by assessing the risk score's ability to predict an individual's future risk for selected common diseases; and 3) to validate the risk algorithm using publicly available data.

Methods

Study population

From 1983 to 2001, a total of 1,195,599 records were generated and stored in the *Health Family Tree* database. The records included individuals' self- or proxy- reported medical history and may have duplicates if more than one student from a single family participated in the class assignment. However, we do not expect this duplication will impact the expected rates used in the risk prediction algorithm because the algorithm predicts risks based on relative risk, which compares observed events with expected events. When duplication exists, both observed and expected events will be duplicated and the ratio will remain the same. Also, we do not expect that persons at higher risk for any disease included in the analysis will be disproportionately represented by families with more than one child participating in the *Health Family Tree* class exercise.

We systematically checked all variables in the database of 1,195,599 records for errors and missing values. After cleaning the "illegal" data, 1,021,909 (85.5%) "valid" records remained. We defined and handled errors and missing values in the following manner:

- When "age" was missing for "living" relatives, we calculated age using the reported year of birth.

- When there was a mismatch between the reported "sex" of the family member and the type of relative (i.e., grandmother should be a female), we used the relationship defined on the *Health Family Tree* (i.e., used the "relative number") to assign a value for "sex." For example, relative number 4 and number 6 are grandmothers of the high school student and their "sex" should be "female," and never "male."
- Records with more than four reasons of death were removed from the analysis assuming that they were errors.
- Records for parents of a high school student that report an age less than 25 years were removed from the analysis.
- Records with an invalid or uncorrectable "sex," "age," "relative number," or "year of birth" were removed from the analysis.

For each student, we treated their paternal and maternal family as separate families (Figure 5.1): one family includes the student, siblings, mother, maternal aunts and uncles, and maternal grandparents; a second family includes the student, siblings, father, paternal aunts and uncles, and paternal grandparents.

Generate rates from the *Health Family Tree*

To calculate the rates of diseases and lifestyle risk factors, we used all individuals between 20 and 99 years of age in the *Health Family Tree* database at the time of data collection. Incidence rates stratified by sex and age groups in 10-year-increments were calculated for all diseases collected in the tool (diabetes, MI, CHD, stroke, high blood pressure, high blood cholesterol, breast cancer, lung cancer, and colon cancer), while prevalence rates were calculated for all risk factors collected in the tool (smoking,

drinking, overweight/obesity, and exercise).

Calculate risk score

As described in the previous publication from 1986,[31] the FHS is calculated using the following equation:

If $| O - E | > ½$ then, $FHS = \frac{(|O-E|-1/2)}{\sqrt{E}} \times \frac{|O-E|}{O-E}$

Or,

if $| O - E | \leq ½$ then $FHS = 0$

The observed incidence of disease (O) was the observed number of events in the family; the expected number of events (E) was calculated by multiplying the age- and sex-specific person-years for each person in the family by the age- and sex- specific incidence rates generated from all records in the database.[31]

After calculating the risk score, the individuals were classified into the following groups based on the risk scores calculated by the algorithm:

- Very high risk group: (FHS ≥ 2.0)
- High risk group: (1.0 ≤ FHS < 2.0)
- Medium risk group: (0.5 ≤ FHS < 1.0)
- Low risk/Reference group: (FHS < 0.5)

Validate risk algorithm

The validation of the risk algorithm was performed for each disease and health condition collected. A retrospective cohort study design was used. From the cleaned original database, we created two datasets for analysis: a “baseline dataset” and a “follow-up dataset.” The “baseline dataset” was based on the family members’ statuses as

of a "cut-off" year, defined as 13 years prior to the year of data collection. This definition was used in the previous validation study in 1986 to balance the number of events in a family prior to and after the cut-off years.[31] The prior study also found that different cut-off years had little impact on the follow-up incidence rates.[31] The "baseline dataset" contained each individual's disease and vital status as of the cut-off year and was then used by the family history score (FHS) algorithm to calculate the risk. The "follow-up dataset" contained the disease events that occurred after the cut-off date and served to document outcomes. The calculated risk scores were merged with the outcomes in the "follow-up dataset" and then assessed using regression analysis. (Figure 5.2)

We analyzed the differences between the reference group and the very high, high, and medium risk groups using a Cox proportional hazards model. Follow-up time was defined as the time since the cut-off year until the onset of the condition (incidence), death, or the year of data collection, whichever was earliest. For every health condition, age was always included in the Cox proportional hazards model as a covariate. We grouped the individual's age at the cut-off year into five groups (20-39, 40-49, 50-59, 60-69, and 70-89), and used the age group as a covariate. To test the effect of behavioral risk factors including smoking, drinking, weight, and exercise, the model was evaluated twice: with and without risk factors as covariates. Before including lifestyle risk factors as covariates, interaction analysis was performed to examine the interaction effect between family disease history and each of the risk factors. Then the risk factors were added in the model all at once to test the effect of risk factors.

All the analyses were performed using SAS version 9.2.[34] A p-value < 0.05 was considered statistically significant. We obtained approval from the Institutional Review

Board at the University of Utah for this analysis.

Validate risk algorithm using public rates

To test the ability of the algorithm to predict future onset of disease using publicly available data, we used several different data sources and estimated age- and sex-specific "expected" rates from the available data. We excluded CHD, high blood pressure, and high blood cholesterol from this analysis because no public incidence rates were identified.

First, in order to assess prediction of future diabetes, we used the estimated national incidence of being diagnosed with diabetes in the year 1999 from the National Health Interview Survey of the National Center for Health Statistics (NCHS).[35] The NCHS data only provided estimated diabetes incidence rates for three age groups (18-44 years, 45-64 years, and 65-79 years), while the *Health Family Tree* algorithm used incidence rates for age groups ranging from 20 to 99 in 10 year increments (i.e., 20-29 years, 30-39 years, etc., up to 90-99 years). To estimate the incidence rates for the age groups used by the algorithm, we used piecewise linear interpolation for the age groups ranging from 20 to 79 years. We applied the incidence rate for the 70-79 age group to the 80-89 and 90-99 age groups. Second, to generate 'expected' rates of MI and stroke, we used incidence rates reported for the U.S. from the Atherosclerosis Risk in Communities Surveillance reported by the American Heart Association[36] and again performed piecewise linear interpolation to generate estimated rates for each age- and sex-specific subgroups. Third, to obtain incidence rates for breast, lung, and colon cancer, we queried the Utah Cancer Registration (UCR)[37] using the SEER*Stat[38] tool. Cancer rates in the UCR are available for every age ranging from 0 to 84 years old, and one rate is reported

for anyone 85 years and older. Again, we applied the rates for the age group 70-79 years to the age groups 80-89 and 90-99 years. Finally, we obtained the expected prevalence rates of smoking and exercise from the Behavioral Risk Factor Surveillance System (BRFSS)[39] using the web portal for the Utah Department of Health's Indicator-based Information System (IBIS-UT).[40]

Using the expected rates from the public data sources and the risk algorithm previously described, we calculated risk scores for diabetes, MI, stroke, breast cancer, lung cancer, and colon cancer. Then, we validated the risk algorithm using the same process described above. Finally, we compared the risk scores calculated using public data sources as reference with the scores calculated using *Health Family Tree* as reference. We used the Finn's r statistics to measure concordance between the two references. We also used Bowker's test of symmetry to analyze the direction of differences when there were disagreements in the risk categories assigned when using the two different reference populations.

Results

Study population

There were 1,021,909 records included in the analysis, which came from a total of 71,127 family units (each student's paternal and maternal family was treated as a separate family unit). After using the cut-off year to split the original database, 1,006,566 records were included in the "baseline" dataset and used to calculate risk scores, and from 956,169 to 981,418 records were included in the "follow-up" datasets and used to assess outcomes. The numbers varied by disease type because only those remaining "at risk" for that disease type were included in the "follow-up" dataset.

Rates from the *Health Family Tree*

The rates for disease and risk factors varied by sex and age group (Table 5.1). For all diseases, the incidence of disease increased with age. In general, males (especially males in older age groups) had higher incidence rates of CHD, MI, stroke, lung cancer, and colon cancer than females at the same age. However, females 60-69 years of age had higher incidence rates of high blood pressure than males. Sex differences in incidence are not observed for diabetes and high blood cholesterol. For risk factors, males had higher prevalence rates than females except for the category of overweight/obese. The rates of those reported to be overweight or obese increased with age while exercise rates decrease with age.

Validation of risk algorithm

Any elevated risk score (very high, high, or medium), with or without consideration for lifestyle risk factors of smoking, drinking, overweight/obesity, or lack of exercise, was predictive for future onset of diabetes, high blood pressure, and high blood cholesterol (all Cox proportional hazards model $p < 0.0001$). Similarly, very high risk scores (FHS ≥ 2.0) were predictive for all diseases included in the analysis, with or without considering the lifestyle risk factors (Cox proportional hazards model $p < 0.0001$ or $p = 0.0002$). The other risk scores (high and medium) were usually, but not always, predictive for the diseases analyzed (Table 5.2). When assessing risk factors (smoking, drinking, overweight/obesity, and exercise), we found no significant interaction between family disease history and each risk factor. So all lifestyle risk factors themselves were added as covariates one at a time in the hazards model, and the significance of family history was evaluated after every addition of a risk factor. The inclusion of any lifestyle

risk factors in the models for most diseases did not change the predictive ability of the family history risk scores reported without consideration of the lifestyle risk factors. Only the prediction of lung cancer was affected by adding lifestyle risk factors. When the risk factor of smoking was added to the hazard model, the family history risk score was no longer able to predict future onset of lung cancer for those families classified at high risk ($1 \leq FHS < 2$) for both males and females.

Validation of risk algorithm using public data

The predictive ability of risk scores using the rates from public data sources for expected rates (Table 5.3) were similar to the findings based on rates generated from the *Health Family Tree* database (Table 5.2). Only two subgroups showed a change in the statistical significance of the hazard ratio. The hazard ratio for females classified at high risk score for stroke became predictive using the public data. In contrast, the hazard ratio for males classified as high risk for lung cancer became not significant ($p = 0.3580$ for males; $p = 0.1719$ for females). Similarly, including both risk factor rates (smoking and exercise) queried from public data as covariates in the model did not change the significance level of the risk prediction for most diseases except for those in the high risk group for lung cancer.

When comparing risk scores calculated using the *Health Family Tree* population as a reference to generate expected rates with risk scores calculated using rates from public data sources, most risk assessment groups remained the same (Table 5.4-5.9). The proportion of individuals classified into the same risk groups using both reference populations is: diabetes 94.2%, MI 94.5%, stroke 97.6%, breast cancer 99.6%, lung cancer 99.9%, colon cancer 99.9% (Finn's r statistics = 0.9). The agreement was mostly

due to the fact that most individuals are at low risk for any of the above diseases. Bowker's test of symmetry ($p < 0.001$ for each disease) indicated that the risk scores calculated using the *Health Family Tree* as reference were disproportionately in one direction when they did not agree with the scores defined using the public source as reference. When the risk scores calculated from the two references were not in the same group (Table 5.4-5.9), more people were classified into the higher risk group when using the *Health Family Tree* as a reference than when using the public sources as a reference.

Discussion

Our validation study of the *Health Family Tree* risk algorithm indicated that the very high risk scores (FHS $\geq$ 2.0) derived from the algorithm can effectively predict the risks for all the concerned diseases and conditions for an adult population who is between 20 and 99 years of age. We also confirmed the predictability of the FHH after including lifestyle risk factors as covariates in the risk model. In addition, we demonstrated that the *Health Family Tree* risk algorithm could be applied to other systems that collect and store family history information. Risks can be predicted by comparing observed events collected by the system with expected events that may be calculated using disease rates from public data sources as expected rates.

We described the absolute rates (i.e., incidence for diseases and prevalence for lifestyle risk factors) generated from the *Health Family Tree* database and used in the risk prediction algorithm. In a separate previous analysis, the absolute rates were shown to have an age- and sex-specific pattern that is similar to public rates although many of the rates underestimated incidence and prevalence reported in public data sources. These

rates were obtained from authoritative public data sources such as the Utah Department of Health and the Utah Cancer Registry.

The family history risk algorithm validation demonstrated that when the risk score is high enough, all concerned diseases could be predicted based on merely the history reported by one or a few family members. For a subset of the diseases including diabetes, high blood pressure, and high blood cholesterol, any elevated risk scores predict the individual's risk for developing these diseases. When a single patient or a population that is not affected but categorized into one of the risk level (very high, high, or medium) groups for a specific disease, clinical and public health decision makers may choose an appropriate preventative strategy for this person or population based on their risk level.

Additionally, we applied publically available disease rates to validate the risk algorithm. This established the process to generalize the risk algorithm to other systems that collect FHH. A prerequisite to adopt the *Health Family Tree* algorithm is to generate the expected rates from a reference population. If the other system contains a large number of records representing the general population with similar information collected by the *Health Family Tree*, the same method of generating expected rates and the risk algorithm could be applied directly. However, if the other system does not collect enough records and information to generate expected population-based rates, the population incidence rates recorded in public data sources can be used as a reference to calculate the expected events for the algorithm. Then the predicted risks can be used for clinical or public health decision support, such as identifying high risk individuals and/or populations for further screening.

When making decisions based on this risk prediction, clinicians and public health professionals should be aware that the choice of reference rates will impact risk prediction by affecting the expected number of events. If a population with a lower than expected incidence of chronic disease is chosen as the reference to predict risks using this algorithm, the prediction scores will be higher than when using the expected risk population as reference. On the other hand, if a higher risk population serves as the reference, then the prediction scores will be lower than when using an average (i.e., expected) risk population as reference. These effects should be considered when researchers, clinicians, and public health professionals are making decisions based on the risk prediction generated from a referent population.

This study has many strengths. While many other risk prediction methods simply use the counts of affected relatives and the degree of relationships to estimate the risk, we used a quantitative risk score. To calculate the quantitative risk score, information about first-degree relatives of the student's parents was needed. Besides the student, the parents were also informants as they were actively involved in the data collection process. Furthermore, calculation of the risk score takes advantage of information about the family size, age of persons in the family, and the incidence of disease in the family and the reference population. The reference population could be either the population of the family history database itself or the general public population. The information included in this calculation is more comprehensive than the information often used to assess family history of common disease (e.g., the numbers of affected first or second degree relatives). Finally, using discrete cut-off risk score categories in the prediction instead of continuous risk scores is practical: the unaffected population will be categorized into

different risk levels after risk assessment, and high risk groups can be targeted for different screening and follow-up strategies.

The possible barriers associated with extensive use of FHH are lack of awareness of the importance of FHH, lack of time, lack of accurate, detailed information, and lack of validated risk assessment.[1,2,20,41,42] This study mainly addressed the challenge of lack of validated risk assessment. First, this study described the disease rates used in the *Health Family Tree* risk algorithm. Then the algorithm was validated by examining the relationship between the predicted relative risk (risk scores based on FHH prior to a certain year) and the reported outcome (the health events that occurred after a certain year). By substituting some disease and risk factor rates generated from the database with publically available data rates, this study confirmed the feasibility of adopting the *Health Family Tree* algorithm in other systems that collect family history information.

This study has limitations. First, the data collected by the *Health Family Tree* were either self-reported or proxy-reported, which may lead to recall bias. In a separate study, we assessed data quality by comparing the data reported by our participants with data collected and recorded by authoritative data sources, such as standardized public health surveys and the cancer registry. That analysis found most diseases were underreported in the *Health Family Tree* database when compared to authoritative data sources. When the reference population used in the analysis is underreporting the diseases incidence, the risk score became inflated. Inflated scores will increase the sensitivity and decrease the specificity of the risk prediction. In the context of using the risk scores for public health screening and recommending healthy behaviors, higher sensitivity with lower specificity is potentially acceptable. Another limitation concerns

generalizability. The study population was primarily from Utah, which has one of the highest proportions of white residents compared to many other states. Also, family sizes in Utah are larger than those in many other states. In addition, the levels of tobacco and alcohol use are lower. For example, in 2011, the percentage of adults aged 18 and older who reported current cigarette smoking was 11.3% for Utah, but 20.4% for US;[43] the percentage of adults who reported binge drinking in the past 30 days was 12.0% for Utah, but 18.3% for US.[44] However, a previous Texas study showed similar results when comparing their data from the HFT project with Utah.[45] Furthermore, the risk prediction is based on relative risk that compares observed with the expected incidence calculated from a reference population. As long as the appropriate reference population is selected, the risk score is valid. Despite the limitations, we recommend use of the *Health Family Tree* risk algorithm when family history data is available in order to identify persons in need of further assessment for risk. This strategy is a potentially low cost method for subsetting a population and finding those at high risk.

In conclusion, the *Health Family Tree* risk algorithm can effectively predict a healthy adult individual's future risk for developing a variety of common chronic diseases by using the individual's family health history, with or without considering lifestyle risk factors. Other family health history tools could use the *Health Family Tree* risk algorithm with the family health history data collected in their system and incidence data derived from public data sources to predict the future risk of common diseases for their population.

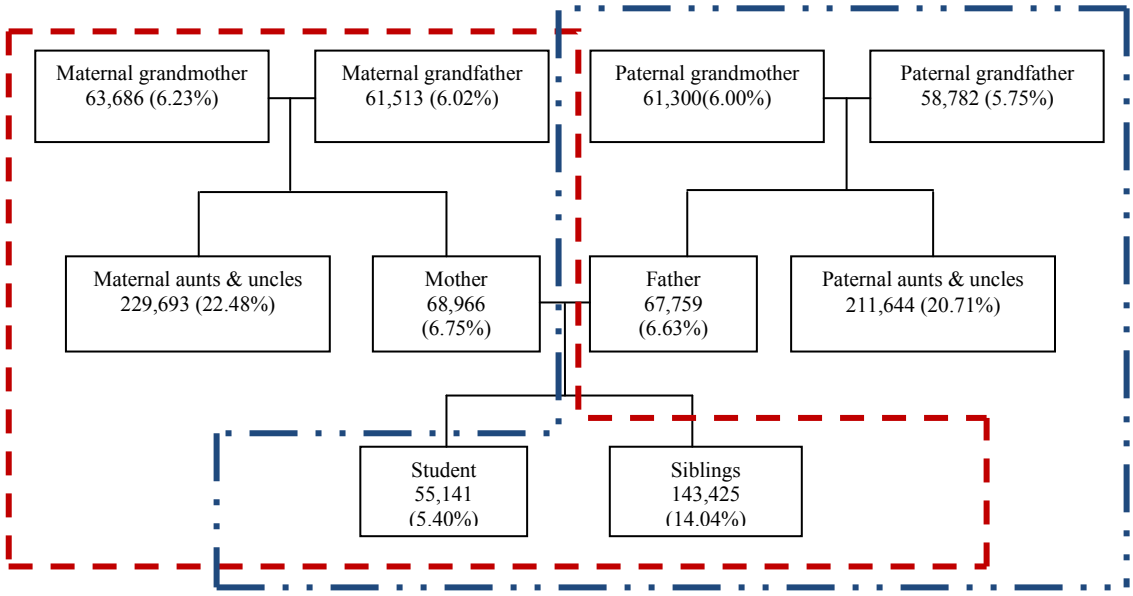


Figure 5.1. Family unit structure and number of records in the *Health Family Tree*. Student's paternal and maternal families were treated as separate family units: one family unit (within the dashed outline) includes the student, siblings, mother, maternal aunts and uncles, and grandparents; a second family unit (within the dashed-and-dotted outline) includes the student, siblings, father, paternal aunts and uncles, and grandparents.

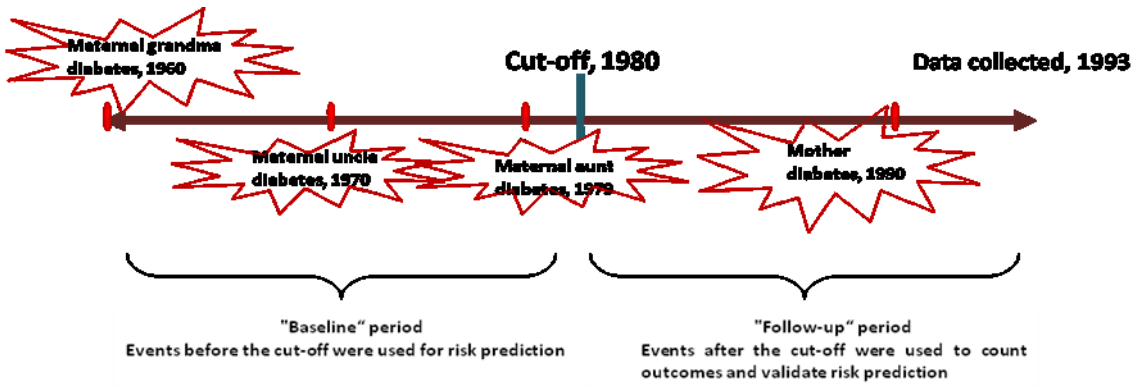


Figure 5.2. Retrospective cohort study design illustration. The figure represents one "family" used for analysis -the student's maternal side of the family. Diabetes events that happened during the "baseline" period before the cut-off year (1980 in this family's case) were used to calculate risk. Events that happened during the "follow-up" period after the cut-off were used to validate the prediction.

Table 5.1. Incidence rates of chronic diseases and prevalence rates of lifestyle risk factors extracted from the *Health Family Tree*, stratified by sex and age, reported 1983-2001

	Age groups							
	20-29	**30-39**	**40-49**	**50-59**	**60-69**	**70-79**	**80-89**	**90-99**
Incidence rates for diseases								
Diabetes (‰)								
Male	0.22	0.53	1.41	2.74	5.58	8.69	9.47	6.46
Female	0.27	0.69	1.47	2.84	5.71	8.86	11.54	7.17
Myocardial infarction (‰)								
Male	0.08	0.51	2.39	6.76	13.54	22.22	29.92	27.53
Female	0.05	0.20	0.64	1.90	4.87	10.16	17.70	17.37
Coronary heart diseases (‰)								
Male	0.09	0.52	2.48	7.06	14.80	25.82	34.44	29.94
Female	0.06	0.22	0.70	2.03	5.33	11.62	19.67	18.42
Stroke (‰)								
Male	0.03	0.11	0.43	1.42	4.30	11.23	23.37	27.10
Female	0.05	0.15	0.40	1.12	3.13	9.36	20.70	31.62
High blood pressure (‰)								
Male	0.63	2.65	6.45	9.13	14.92	19.54	21.71	16.74
Female	0.56	2.22	5.38	9.63	17.90	27.72	36.62	34.67
High blood cholesterol (‰)								
Male	0.29	1.62	4.26	5.22	8.34	10.21	9.36	5.38
Female	0.22	1.16	2.69	4.07	8.24	11.03	10.70	6.18
Breast cancer (‰)								
Male	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Female	0.07	0.37	1.03	1.55	2.64	3.96	5.63	3.74
Lung cancer (‰)								
Male	0.02	0.06	0.21	0.69	1.53	2.14	2.38	1.26
Female	0.01	0.05	0.13	0.32	0.62	0.74	0.97	1.65
Colon cancer (‰)								
Male	0.02	0.07	0.19	0.59	1.43	2.72	4.24	5.28
Female	0.03	0.09	0.20	0.46	1.10	1.88	2.66	2.34
Prevalence rates for lifestyle risk factors								
Smoking (%)								
Male	18.51	22.72	18.68	20.92	20.02	12.74	8.34	5.77
Female	12.16	14.62	11.11	12.08	10.22	4.97	2.20	2.12
Overweight/obese (%)								
Male	2.32	4.62	6.69	7.71	7.87	6.25	3.84	3.45
Female	4.12	7.96	9.77	11.70	11.64	9.61	6.93	4.63
Drinking (%)								
Male	35.00	40.41	34.57	34.67	36.65	28.70	22.20	18.58
Female	27.36	31.81	26.28	24.92	24.09	16.57	11.21	9.26
Exercise (%)								
Male	64.69	55.73	50.65	46.28	41.18	39.69	37.41	42.33
Female	56.78	50.11	46.21	40.95	35.82	31.16	25.59	29.71

Table 5.2. Hazard ratios and p-values of the Cox proportional hazards model, by gender and family history score category, with age as a covariate

	Very high (FHS≥2)		High (1≤FHS<2)		Medium (0.5≤FHS<1)	
	Male	**Female**	**Male**	**Female**	**Male**	**Female**
Diabetes						
Hazard ratio	3.6	3.4	2.1	2.0	2.0	1.8
95% CI	(3.2,4.1)	(3.0,3.8)	(1.9,2.3)	(1.8,2.1)	(1.8,2.2)	(1.6,2.0)
p value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
Myocardial infarction						
Hazard ratio	2.4	1.7	1.7	1.4	1.6	1.1
95% CI	(2.1,2.7)	(1.4,2.0)	(1.6,1.9)	(1.2,1.5)	(1.4,1.7)	(0.9,1.3)
p value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.2632
Coronary heart diseases						
Hazard ratio	2.3	1.7	1.6	1.2	1.6	1.1
95% CI	(2.0,2.5)	(1.4,2.0)	(1.5,1.8)	(1.1,1.4)	(1.4,1.7)	(1.0,1.3)
p value	<0.0001	<0.0001	<0.0001	0.0002	<0.0001	0.0735
Stroke						
Hazard ratio	1.5	1.7	0.9	1.1	0.7	1.0
95% CI	(1.2,2.0)	(1.4,2.1)	(0.7,1.2)	(0.9,1.3)	(0.5,1.0)	(0.8,1.3)
p value	0.0014	<0.0001	0.6651	0.5647	0.4465	0.9748
High blood pressure						
Hazard ratio	3.1	2.9	2.2	2.2	1.8	1.7
95% CI	(2.9,3.3)	(2.7,3.1)	(2.1,2.4)	(2.0,2.3)	(1.7,1.9)	(1.6,1.8)
p value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
High blood cholesterol						
Hazard ratio	4.8	4.6	3.1	2.4	2.6	2.3
95% CI	(4.5,5.2)	(4.2,5.0)	(2.9,3.3)	(2.2,2.6)	(2.5,2.8)	(2.2,2.5)
p value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
Breast cancer						
Hazard ratio	N/A	2.6	N/A	1.7	N/A	0.8
95% CI	N/A	(2.2,3.1)	N/A	(1.2,2.2)	N/A	(0.3,2.2)
p value	N/A	<0.0001	N/A	0.0009	N/A	0.6818
Lung cancer						
Hazard ratio	5.2	4.6	3.4	3.6	0.0	0.0
95% CI	(3.7,7.5)	(3.1,6.8)	(1.3,9.0)	(1.2,11.2)	(0.0,0.0)	(0.0,0.0)
p value	<0.0001	<0.0001	0.0149	0.0274	0.9516	0.9667
Colon cancer						
Hazard ratio	2.0	2.7	1.3	1.7	0.0	0.0
95% CI	(1.4,2.9)	(2.0,3.7)	(0.6,2.8)	(0.8,3.4)	(0.0,0.0)	(0.0,0.0)
p value	0.0002	<0.0001	0.4733	0.1384	0.9129	0.9380

Table 5.3. Hazard ratios and p-values of the Cox proportional hazards model, using public rates as reference in the algorithm, with age as a covariate, without risk factors

	Very high (FHS≥2)		High (1≤FHS<2)		Medium (0.5≤FHS<1)	
	Male	**Female**	**Male**	**Female**	**Male**	**Female**
Diabetes						
Hazard ratio	5.1	5.2	3.4	3.2	2.2	2.1
95% CI	(4.1,6.4))	(4.1,6.4)	(2.8,4.0)	(2.7,3.8)	(1.9,2.5)	(1.8,2.4)
p value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
Myocardial infarction						
Hazard ratio	2.1	1.5	1.6	1.3	1.5	1.0
95% CI	(1.9,2.3)	(1.3,1.8)	(1.4,1.7)	(1.2,1.4)	(1.4,1.7)	(0.8,1.1)
p value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.7115
Stroke						
Hazard ratio	1.7	1.5	1.2	1.6	0.9	0.9
95% CI	(1.2,2.4)	(1.1,2.2)	(0.9,1.6)	(1.3,2.0)	(0.6,1.2)	(0.7,1.2)
p value	0.0037	<0.0001	0.2750	<0.0001	0.4300	0.4450
Breast cancer						
Hazard ratio	N/A	2.7	N/A	1.8	N/A	1.4
95% CI	N/A	(2.2,3.3)	N/A	(1.3,2.3)	N/A	(0.8,2.4)
p value	N/A	<0.0001	N/A	<0.0001	N/A	0.2238
Lung cancer						
Hazard ratio	5.5	4.7	1.9	2.6	0.0	0.0
95% CI	(3.8,7.7)	(3.2,6.9)	(0.5,7.7)	(0.7,10.5)	(0.0,0.0)	(0.0,0.0)
p value	<0.0001	<0.0001	0.3580	0.1719	0.9518	0.9669
Colon cancer						
Hazard ratio	2.0	2.6	1.3	1.8	0.0	0.0
95% CI	(1.4,2.8)	(1.9,3.6)	(0.6,2.8)	(0.9,3.6)	(0.0,0.0)	(0.0,0.0)
p value	0.0003	<0.0001	0.4339	0.0918	0.9181	0.9393

Table 5.4. Comparison of family history scores (FHS) for diabetes generated by the *Health Family Tree* (HFT) algorithm based on two references

Percentages (%)		Rates from public source*				
		FHS<0.5	0.5≤FHS<1	1≤FHS<2	FHS>2	Total
Rates from *Health Family Tree*	FHS<0.5	94.0	0	0	0	94.0
	0.5≤FHS<1	2.5	0	0	0	2.5
	1≤FHS<2	1.8	0.6	0	0	2.4
	FHS>2	0	0.4	0.5	0.2	1.1
	Total	98.3	1.0	0.5	0.2	100

*Public source: incidence of diagnosed diabetes published by Centers for Disease Control and Prevention, available at: http://www.cdc.gov/diabetes/statistics/incidence/fig5.htm

Table 5.5. Comparison of family history scores (FHS) for myocardial infarction generated by the *Health Family Tree* (HFT) algorithm based on two references

Percentages (%)		Rates from public source*				
		FHS<0.5	0.5≤FHS<1	1≤FHS<2	FHS>2	Total
Rates from *Health Family Tree*	FHS<0.5	89.2	0	0	0	89.2
	0.5≤FHS<1	2.5	1.3	0	0	3.8
	1≤FHS<2	0	2.1	2.3	0.9	5.3
	FHS>2	0	0	0	1.7	1.7
	Total	91.7	3.4	2.3	2.6	100

*Public source: Heart disease and stroke statistics published by the American Heart Association, available at: http://my.americanheart.org/professional/General/AHA-Heart-Disease-and-Stroke-Statistics-2013-Update_UCM_445937_Article.jsp

Table 5.6. Comparison of family history scores (FHS) for stroke generated by the *Health Family Tree* (HFT) algorithm based on two references

Percentages (%)		Rates from public source*				
		FHS<0.5	0.5≤FHS<1	1≤FHS<2	FHS>2	Total
Rates from *Health Family Tree*	FHS<0.5	96.1	0	0	0	96.1
	0.5≤FHS<1	0.9	0.3	0	0	1.2
	1≤FHS<2	0.1	0.9	0.6	0	1.6
	FHS>2	0.2	0	0.4	0.6	1.2
	Total	97.3	1.2	1.0	0.6	100

*Public source: Heart disease and stroke statistics published by the American Heart Association, available at: http://my.americanheart.org/professional/General/AHA-Heart-Disease-and-Stroke-Statistics-2013-Update_UCM_445937_Article.jsp

Table 5.7. Comparison of family history scores (FHS) for breast cancer generated by the *Health Family Tree* (HFT) algorithm based on two references

Percentages (%)		Rates from public source*				
		FHS<0.5	0.5≤FHS<1	1≤FHS<2	FHS>2	Total
Rates from *Health Family Tree*	FHS<0.5	97.6	0	0	0	97.6
	0.5≤FHS<1	0.1	0.2	0	0	0.3
	1≤FHS<2	0	0.2	0.6	0	0.8
	FHS>2	0	0	0.2	1.2	1.4
	Total	97.7	0.4	0.8	1.2	100

*Public source: Utah Cancer registry, queried through SEER*Stat tool

Table 5.8. Comparison of family history scores (FHS) for lung cancer generated by the *Health Family Tree* (HFT) algorithm based on two references

Percentages (%)		Rates from public source*				
		FHS<0.5	0.5≤FHS<1	1≤FHS<2	FHS>2	Total
Rates from *Health Family Tree*	FHS<0.5	98.5	0	0	0	98.5
	0.5≤FHS<1	0	0.1	0	0	0.1
	1≤FHS<2	0	0	0.3	0.1	0.4
	FHS>2	0	0	0	1.0	1.0
	Total	98.5	0.1	0.3	1.1	100

*Public source: Utah Cancer registry, queried through SEER*Stat tool

Table 5.9. Comparison of family history scores (FHS) for colon cancer generated by the *Health Family Tree* (HFT) algorithm based on two references

Percentages (%)		Rates from public source*				
		FHS<0.5	0.5≤FHS<1	1≤FHS<2	FHS>2	Total
Rates from *Health Family Tree*	FHS<0.5	98.2	0	0	0	98.2
	0.5≤FHS<1	0	0.2	0	0	0.2
	1≤FHS<2	0	0	0.5	0.1	0.6
	FHS>2	0	0	0	1.0	1.0
	Total	98.2	0.2	0.5	1.1	100

*Public source: Utah Cancer registry, queried through SEER*Stat tool

References

1. Trotter TL, Martin HM. Family history in pediatric primary care. *Pediatrics.* Sep 2007;120 Suppl 2:S60-65.

2. Yoon PW, Scheuner MT, Peterson-Oehlke KL, Gwinn M, Faucett A, Khoury MJ. Can family history be used as a tool for public health and preventive medicine? *Genet Med.* Jul-Aug 2002;4(4):304-310.

3. Sesso HD, Lee IM, Gaziano JM, Rexrode KM, Glynn RJ, Buring JE. Maternal and paternal history of myocardial infarction and risk of cardiovascular disease in men and women. *Circulation.* Jul 24 2001;104(4):393-398.

4. Jousilahti P, Puska P, Vartiainen E, Pekkanen J, Tuomilehto J. Parental history of premature coronary heart disease: an independent risk factor of myocardial infarction. *J Clin Epidemiol.* May 1996;49(5):497-503.

5. Hopkins PN, Williams RR, Kuida H, et al. Family history as an independent risk factor for incident coronary artery disease in a high-risk cohort in Utah. *Am J Cardiol.* Oct 1 1988;62(10 Pt 1):703-707.

6. Myers RH, Kiely DK, Cupples LA, Kannel WB. Parental history is an independent risk factor for coronary artery disease: the Framingham Study. *Am Heart J.* Oct 1990;120(4):963-969.

7. Roncaglioni MC, Santoro L, D'Avanzo B, et al. Role of family history in patients with myocardial infarction. An Italian case-control study. GISSI-EFRIM Investigators. *Circulation.* Jun 1992;85(6):2065-2072.

8. Djousse L, Gaziano JM. Parental history of myocardial infarction and risk of heart failure in male physicians. *Eur J Clin Invest.* Dec 2008;38(12):896-901.

9. Friedlander Y, Arbogast P, Schwartz SM, et al. Family history as a risk factor for early onset myocardial infarction in young women. *Atherosclerosis.* May 2001;156(1):201-207.

10. Jousilahti P, Rastenyte D, Tuomilehto J, Sarti C, Vartiainen E. Parental history of cardiovascular disease and risk of stroke. A prospective follow-up of 14371 middle-aged men and women in Finland. *Stroke.* Jul 1997;28(7):1361-1366.

11. Kadota A, Okamura T, Hozawa A, et al. Relationships between family histories of stroke and of hypertension and stroke mortality: NIPPON DATA80, 1980-1999. *Hypertens Res.* Aug 2008;31(8):1525-1531.

12. Bjornholt JV, Erikssen G, Liestol K, Jervell J, Thaulow E, Erikssen J. Type 2 diabetes and maternal family history: an impact beyond slow glucose removal rate and fasting hyperglycemia in low-risk individuals? Results from 22.5 years of follow-up of healthy nondiabetic men. *Diabetes Care.* Sep 2000;23(9):1255-1259.

13. Nakanishi S, Yamane K, Kamei N, Okubo M, Kohno N. Relationship between development of diabetes and family history by gender in Japanese-Americans. *Diabetes Res Clin Pract.* Aug 2003;61(2):109-115.

14. Rahman M, Simmons RK, Harding AH, Wareham NJ, Griffin SJ. A simple risk score identifies individuals at high risk of developing Type 2 diabetes: a prospective cohort study. *Fam Pract.* Jun 2008;25(3):191-196.

15. Meigs JB, Cupples LA, Wilson PW. Parental transmission of type 2 diabetes: the Framingham Offspring Study. *Diabetes.* Dec 2000;49(12):2201-2207.

16. Cauley JA, Song J, Dowsett SA, Mershon JL, Cummings SR. Risk factors for breast cancer in older women: the relative contribution of bone mineral density and other established risk factors. *Breast Cancer Res Treat.* Apr 2007;102(2):181-188.

17. Wei EK, Giovannucci E, Wu K, et al. Comparison of risk factors for colon and rectal cancer. *Int J Cancer.* Jan 20 2004;108(3):433-442.

18. Rodriguez C, Calle EE, Miracle-McMahill HL, et al. Family history and risk of fatal prostate cancer. *Epidemiology.* Nov 1997;8(6):653-657.

19. Gao Y, Goldstein AM, Consonni D, et al. Family history of cancer and nonmalignant lung diseases as risk factors for lung cancer. *Int J Cancer.* Jul 1 2009;125(1):146-152.

20. Yoon PW, Scheuner MT, Jorgensen C, Khoury MJ. Developing Family Healthware, a family history screening tool to prevent common chronic diseases. *Prev Chronic Dis.* Jan 2009;6(1):A33.

21. Carmona RH, Wattendorf DJ. Personalizing prevention: the U.S. Surgeon General's Family History Initiative. *Am Fam Physician.* Jan 1 2005;71(1):36, 39.

22. Williams RR, Hunt SC, Barlow GK, et al. Health family trees: a tool for finding and helping young family members of coronary and cancer prone pedigrees in Texas and Utah. *Am J Public Health.* Oct 1988;78(10):1283-1286.

23. Washington University. Your disease risk: the source on prevention http://www.yourdiseaserisk.wustl.edu/. Accessed April 1, 2012.

24. OSU. Welcome to family healthlink. https://familyhealthlink.osumc.edu/. Accessed April 1, 2012.

25. BRACAnalysis. http://www.bracnow.com/. Accessed April 1, 2012.

26. Myriad. learn more about your family cancer history. http://www.myriadtests.com/index.php?page_id=227&usetemplate=whatisinherited&usetype=1. Accessed April 1, 2012.

27. Braithwaite D, Sutton S, Mackay J, Stein J, Emery J. Development of a risk assessment tool for women with a family history of breast cancer. *Cancer Detect Prev.* 2005;29(5):433-439.

28. NorthShore. MyGenerations. http://www.northshore.org/genetics/mygenerations/. Accessed April 1, 2012.

29. USSG. My family health portrait. https://familyhistory.hhs.gov/fhh-web/home.action. Accessed April 1, 2012.

30. Johnson J, Giles RT, Larsen L, Ware J, Adams T, Hunt SC. Utah's Family High Risk Program: bridging the gap between genomics and public health. *Prev Chronic Dis.* Apr 2005;2(2):A24.

31. Hunt SC, Williams RR, Barlow GK. A comparison of positive family history definitions for defining risk of future disease. *J Chronic Dis.* 1986;39(10):809-821.

32. Jiang Y, Staes CJ, Adams TD, Hunt SC. Evaluation of risk scores derived from the health family tree program. *AMIA Annu Symp Proc.* 2009;2009:286-290.

33. Health Family Tree. 2005; http://healthfamilytree.utah.edu Accessed April 1, 2012.

34. SAS. About SAS. http://www.sas.com. Accessed April 1, 2012.

35. CDC. Diabetes data & trends. http://www.cdc.gov/diabetes/statistics/incidence/fig5.htm. Accessed October 5, 2012.

36. Go AS, Mozaffarian D, Roger VL, et al. Heart disease and stroke statistics—2013 update: a report from the American Heart Association. *Circulation.* Jan 1 2013;127(1):e6-e245.

37. UCR. Utah Cancer Registry. http://ucr.utah.edu/. Accessed April 1, 2012.

38. ADA. Diabetes statistics. http://www.diabetes.org/diabetes-basics/diabetes-statistics/. Accessed October 29, 2013.

39. CDC. Behavioral Risk Factor Suveillance System. http://www.cdc.gov/brfss/. Accessed October 29, 2013.

40. UDOH. Welcome to IBIS-PH: Utah's public health data resource. http://ibis.health.utah.gov/. Accessed October 29, 2013.

41. Hinton RB, Jr. The family history: reemergence of an established tool. *Crit Care Nurs Clin North Am.* Jun 2008;20(2):149-158, v.

42. Yoon PW, Scheuner MT, Khoury MJ. Research priorities for evaluating family history in the prevention of common chronic diseases. *Am J Prev Med.* Feb 2003;24(2):128-135.

43. UDOH. Data and confidence limits for percentage of adults who reported current cigarette smoking, adults aged 18 and older, Utah and U.S., 1989-2011. http://ibis.health.utah.gov/indicator/view_numbers/CigSmokAdlt.Ut_US.html. Accessed October 29, 2013.

44. UDOH. Data and confidence limits for percentage of adults who reported binge drinking in the past 30 days, Utah and U.S., 2005-2011. http://ibis.health.utah.gov/indicator/view_numbers/AlcConBinDri.UT_US.html. Accessed October 29, 2013.

45. Johnson J, Giles RT, Larsen L, Ware J, Adams T, Hunt SC. Utah's Family High Risk Program: bridging the gap between genomics and public health. *Prev Chronic Dis.* 2005;2(2):A24.

CHAPTER 6

DEVELOPING NEW MODELS TO PREDICT DIABETES BY APPLYING MACHINE LEARNING METHODS TO THE *HEALTH FAMILY TREE* DATABASE

Background

Diabetes mellitus is a serious and very costly public health problem in the United States. According to data released in 2011 by Centers for Disease Control and Prevention, 25.8 million U.S. children and adults, 8.3% of the population, have diabetes.[1] In 2012, total health care costs of diagnosed diabetes in the United States were $245 billion, which was a 41% increase compared to 2007.[2] In 2007, diabetes was listed as the underlying cause or contributing factor for a total of 231,404 deaths.[1] Meanwhile, the incidence of diabetes continues to grow. The number of diagnosed diabetes cases is projected to reach 29 million by 2050.[3]

Numerous research methods including data mining and machine learning have been applied to diagnose or predict the development of diabetes. Data mining is the extraction of implicit, previously unknown, and potentially useful information from data.[4] Machine learning is the technical basis of data mining and was defined as the acquisition of structural descriptions from examples.[4] Besides their application in industrial fields such as retail and banking, data mining and machine learning techniques have been applied to healthcare data and research.[5,6] Various machine learning methods such as

Decision Trees, Naive Bayes/Bayesian Networks, and Support Vector Machine (SVM) have been applied to healthcare databases for diagnosis or prediction of the development of diabetes,[7-10] coronary heart diseases,[11,12] and other diseases.[13-16]

When compared to traditional statistical methods, machine learning methods have some advantages. For example, when predicting the development of diabetes, we treat the disease outcome (having diabetes or not) as the dependent variable, and a series of factors such as sex, age, weight, diet, and family health history of diabetes as independent variables. Traditional statistical methods assume the following a priori data model between the independent and dependent variables: response variable = f (predictor variable, random noise, parameters).[17] The assumed data model may not be true. In contrast, machine learning methods rely on the input and output data themselves rather than an assumed a priori data model. Machine learning methods allow a black box between independent and dependent data and aim to find models that predict outcomes (such as diagnosis of diabetes) based on inputs (various features present in the data). The approach is to find a function f(x)—an algorithm that operates on x to predict the responses y. These methods focus on the data themselves and the properties of algorithms, and their use has advanced rapidly in recent decades.[17]

Besides the theoretical advantages, machine learning methods may have additional practical benefits for the research questions addressed in this dissertation. The *Health Family Tree* algorithm introduced in previous chapters used traditional statistical methods to predict a healthy individual's risk for developing selected diseases. To implement this statistical model, the method requires a reference population to generate the expected incidence rates. In previous studies, we used the *Health Family Tree*

population (>1 million health records reported by the informants) to generate the incidence rates to calculate a family history risk score. For other systems that collect family health history (FHH) information without a large representative population to use the same algorithm, we also queried available public disease incidence rates and used interpolation methods to estimate the incidence rates that were needed by the algorithm. With the exception of cancer, very limited information about incidence rates is available for most chronic diseases. For many chronic diseases, the onset of the disease is not clear, so public health standardized surveys often query participants about their chronic disease status to generate prevalence rates. Public health surveys often do not ascertain an onset date, which is required for calculating incidence rates. Thus, the incidence rates of most chronic diseases are not available in public data sources. As a result, without interpolating incidence rates from prevalence data, the *Health Family Tree* algorithm may be difficult to implement in other systems. On the other hand, machine learning techniques build predictive models by training on the features collected by the system. Ideally, these trained and validated models can be applied to predict diabetes risk in other systems that collect these same features. Though there are many advantages of machine learning methods, there are disadvantages. In general, machine learning requires a large number of samples to train the classifiers. In addition, the algorithms used by machine learning methods may be very complicated and be a black box to the user.[17] For our purposes, we have a large data set with which to train the algorithms, and we are most interested in accurate prediction; therefore, these limitations may not apply to our research.

There are two key types of machine learning methods: supervised and unsupervised. Supervised machine learning operates under supervision by being provided with the actual outcome for each training example.[4] In contrast, unsupervised machine learning tries to find hidden structure in unlabeled data without knowing the outcomes.[4] The goal of this study was to develop models to predict the presence or risk of diabetes using supervised machine learning methods in order to identify high and low risk populations for population-based studies or public health screening. The specific objectives were to use the limited set of information about a subject (i.e., age and sex, comorbidities, lifestyle risk factors, and family health history) available in the *Health Family Tree* database and three different data mining algorithms to: a) predict the presence of diabetes among individuals in the sample, and b) predict the future development of diabetes.

Methods

Sample

The target population included the students' parents, aunts and uncles (n = 578,062 individuals) in the *Health Family Tree* database. This sample was drawn from the set of validated records used in Chapters 4 and 5. Considering there is no parental information for the students' grandparents, and the low incidence of diabetes in the students and their siblings, we chose the students' parents, aunts, and uncles as the study population for this analysis.

The target disease for this analysis was diabetes. Literature shows that the risk factors for diabetes include age, sex, weight/Body Mass Index (BMI), heart disease, stroke, family history of diabetes among first-degree relatives (parents and siblings),

smoking, and lack of physical activity.[18,19] The American Diabetes Association recommended a set of risk factors including but not limited to: body mass index, physical inactivity, first-degree relative with diabetes, race, hypertension, HDL cholesterol level, obesity, and history of cardiovascular diseases.[20] Multiple diabetes risk score tools such as the Cambridge diabetes risk score,[21] Danish diabetes risk score,[22] and Indian diabetes risk score[23] use a similar set of risk factors for their risk calculation.

Data preparation and feature examination

Supervised machine learning (whereby a desired target output is defined) was used to discover patterns in the *Health Family Tree* database and build models for classifying an individual's risk for developing diabetes. The outcome of interest was a binary classification: the individual is classified as having diabetes or not. The features shown in Table 6.1 were all included in the machine learning process because they either: 1) were collected by the *Health Family Tree* and are risk factors described in the literature and guidelines; or 2) were family health history related features that were generated from collected data on first-degree relatives and showed significant results (95% CI of odds ratio did not include one) from the univariate analysis.

The following procedures were conducted before training the classifiers:

- Feature preparation: five family history related features were created in the SAS database including: if the individual has a diabetic mother, if the individual has a diabetic father, if the individual has a diabetic sibling, the number of first-degree relatives with diabetes, and the ratio of first-degree relatives with diabetes to all first-degree relatives. When creating the family history related features, only events that happened before the subject's onset of diabetes were counted in the

family health history. Similarly, when classifying the subject's comorbidities, we only included comorbidities that were present when the subject developed diabetes. We made this assessment by comparing the year of onset for the diabetes and the other diseases reported by the subject.

- Feature selection: Given the limited set of 20 features available in the *Health Family Tree* database relative to the large number of instances (~578,000), no further feature selection was required. There were 14 features directly available in the dataset and an additional five features we derived from the data.

Classifier selection

Three classifiers including logistic regression, Bayesian network, and support vector machine (SVM) (see Table 6.2) were chosen to test prediction accuracy. The algorithms were selected for the following reasons:

a) The outcome is known in the *Health Family Tree* database; therefore the problem requires supervised machine learning methods. The selected algorithms all use supervised learning methods. Algorithms that use unsupervised machine learning such as clustering were excluded;

b) The outcome requires a binary classification, i.e., it predicts the disease or risk status as "yes" or "no." The selected algorithms can provide a binary output; and

c) The dataset contains a large number of instances (578,062 subjects: parents and uncles and aunts), and the choice of classifier has less effect on the machine learning results. In theory, all classifiers should give similar results.

Classifier training and evaluation

Waikato Environment for Knowledge Analysis (WEKA)[24] version 3.7.9 was used to train and evaluate the models. Ten-fold cross validation was used to evaluate the trained models.

The three classifiers were evaluated and the following performance metrics were compared: recall (i.e., sensitivity), precision (i.e., positive predictive value), and F-measure. The F-measure is calculated based on recall and precision.

Recall = TP/(TP + FN)

Precision = TP/(TP + FP)

F-measure = 2 * precision * recall/(precision + recall)

where TP, FP, TN, and FN represent the number of true positives, false positives, true negatives, and false negatives, respectively (Table 6.3).

The prevalence of positive diabetic individuals in the HFT database was relatively low (2.5%). This type of imbalanced dataset can affect the performance of various classifiers.[25] Therefore, we used the undersampling technique[26,27] to address the imbalanced classes, train, and evaluate the classifiers. The negative diabetic instances were undersampled randomly to create a subset of the population that was negative for the diabetes outcome of interest.

In addition to classifying an individual's diabetes at the time of data collection, we also trained and evaluated classifiers for predicting an individual's future risk to develop diabetes. This was done by using the status of the features at a point in time in the past to predict the presence or absence of diabetes in the future. For subjects that developed diabetes, we evaluated and used their age, status of comobidities, and their family health

history of diabetes at 5 years and 10 years before the year their diabetes was diagnosed. The 5- and 10-year gaps were arbitrarily chosen, considering that this dissertation mainly seeks to predict the future development of diabetes for public health prevention purposes. For subjects who did not develop diabetes by the time of data collection, we evaluated and used their age, status of comorbidities, and their family health history of diabetes at 5 years and 10 years before the year of data collection. Furthermore, to test how well FHH by itself can predict the presence or future development of diabetes, we also trained and evaluated the classifiers with only features related to family health history and the subject's age and sex.

This research was determined to be exempt from human subject research by the Institutional Review Board at the University of Utah because all data we used were de-identified.

Results

Study population

The target population for this analysis was the students' parents, aunts, and uncles, which are shown in the highlighted area and include a total of 578,062 records (Figure 6.1). After data cleaning and preparation for machine learning, a total of 564,485 (97.7%) records remained. These 564,485 records were used to train and evaluate the machine learning classifiers.

Features

The results of the univariate analysis (Table 6.4) showed that the distribution of almost all features were significantly different (Odds Ratio 95% CI did not include 1.0)

between the diabetic and nondiabetic groups. The proportion of males and prevalence of smoking were not significantly different between the diabetic and nondiabetic groups. Drinking and exercise reported by the subjects showed a protective effect on the presence of diabetes (Odds ratio CI <1), while the remaining features were associated with an increased risk of diabetes (Odds Ratio > 1.0). The prevalence of diabetes increased incrementally as the number of relatives with diabetes increased from one to seven.

Classifier evaluation

The performance of the models based on the three classifiers is presented in Table 6.5. All F-measures, as the weighted scores of both recall and precision, were greater than 0.50 (F-measure value by random guesses). The F-measures across the three models ranged from 0.64 to 0.70 when using all features (including the subject's sex and age group, comorbidities, lifestyle risk factors, and family health history) to predict the presence of diabetes. The F-measures ranged from 0.63 to 0.65 when using only the subject's age and sex and family health history to predict the presence of diabetes. For all three models, the F-measure decreased when removing comorbidities and lifestyle risk factor features. The F-measure was lower when using the model to predict future onset of diabetes in comparison to predicting the presence of diabetes.

Discussion

This study trained and evaluated classifiers to predict the future development of diabetes using a limited set of features that would be relatively easy to collect. These features included demographic characteristics (sex, age), family health history information about first-degree relatives (mother's diabetic history, father's diabetic

history, siblings' diabetic history, number and ratio of first-degree diabetic relatives), and optionally, other disease information and lifestyle risk factors.

Characteristics of the *Health Family Tree* dataset provides advantages and disadvantages for this analysis. The *Health Family Tree* database used in this study has a large number of instances, and relatively small number of features. Further feature selection is often used to select the most relevant features when there are many features but not many instances. When there are hundreds to thousands of features, a model trained on all the features will have high variance and tends to be overfitted to the training data.[23] When overfitting happens, the predictive model is too closely tied to the particular training data and will not apply well to fresh data. Feature selection is often needed to reduce the number of features to avoid the overfitting situation. The limited number of features in the HFT database gave us the advantage of analyzing the data without expending extra effort on feature selection. On the other hand, we were required to use an undersampling technique to address the imbalance created by the relatively rare presence of positive outcomes (diabetes) in the original dataset. The performance of all classifiers was improved after applying the undersampling technique. This improvement agreed with what other studies have used and found.[26,27]

The diabetes prediction models built by the machine learning methods have several strengths. First, instead of assuming a statistical model and a hypothesis based on the model, the machine learning method's approach is to build a classifier based on a subset of the actual data and then validate the classifier with the remaining data. The properties of the data were considered and included in the building process from the beginning. Second, the *Health Family Tree* database has a large number of instances and

a relatively small number of features. This characteristic avoids a potential problem often seen in machine learning applied to datasets with many features: the model overfits the data and is not generalizable.[28] Third, a model built by machine learning may be easier to implement than statistical models. To apply the statistical prediction models to the data collected by other systems, there is a need for a reference population to calculate the expected disease incidence rates because the model predicts risk based on a comparison of the observed and the expected events of disease. Aside from the practical issues that a reference population is not always available and disease incidence rates of reference population are usually not available, the choice of reference population will have a direct effect on the results of the prediction. The prediction models can be applied to public health or population-based research for prevention purposes. To apply the machine learning models using data collected by other systems (such as the electronic health records or personal health records), public health professionals or population health researchers would need to create an input data set including at least sex, age, and family health history for first degree relatives. The input data set can then be applied to the prediction models to obtain classification of the individual's diabetes status. While predicting the presence of diabetes is not useful when that information may already be available in the clinical record or from interviewing a person, the value of the classification is to predict the individuals that are currently nondiabetic but are likely to develop diabetes in the future (5 or 10 years). This population will be the target for public health education and interventions or risk stratification for population-based research.

There are limitations in this study. The prediction was built on the target instances that were included in the *Health Family Tree* database collected from the Utah population

from 1983 to 2002. Race information was not included in the *Health Family Tree* database, but it would be expected that most of the study population was Caucasian because the Utah population is 89% Caucasian based on 2000 census data.[29] The prevalence of diabetes differ among race/ethnicity groups after adjusting for age. From 2007 to 2009, the prevalence of people aged 20 years or older diagnosed with diabetes was 7.1% for non-Hispanic whites, 8.4% for Asian American, 12.6% for non-Hispanic blacks, and 11.8% for Hispanics.[30] Thus, the prediction may not predict diabetes for other populations representing different race groups. Similarly, the training and validation dataset used in this study did not include all known risk factors of diabetes such as body mass index and cholesterol level. Therefore, the model may be improved when adding these additional risk factors. Another limitation is the limited age range of the subjects. The target population used to train the classifiers were the high school students' parents, aunts, and uncles, who are mostly distributed in the middle-aged groups. The classifiers may perform differently in the younger and older age groups. Finally, the diabetes documented in the *Health Family Tree* was not specified as type 1 or type 2 diabetes. Diabetes was defined as answering "yes" to the question, "Has he/she ever been told by a doctor that he/she suffers from diabetes?," and then the age of diabetes diagnosis was recorded for those with diabetes. Since type 2 diabetes accounts for 90-95% of total diabetes cases[31] and most of the study target subjects (student's parents, aunts and uncles) were adults, the diabetes subtype reported in the target was more likely to be type 2 diabetes. The predictive ability may not be true for predicting the more rare events of type 1 diabetes.

Despite these limitations, we developed models that can accurately predict the presence or future development of diabetes in 5 or 10 years. The models were based on a limited set of self- and proxy-reported information and can be used to identify high and low risk persons within Caucasian, middle-aged adult populations for population-based studies or public health screening.

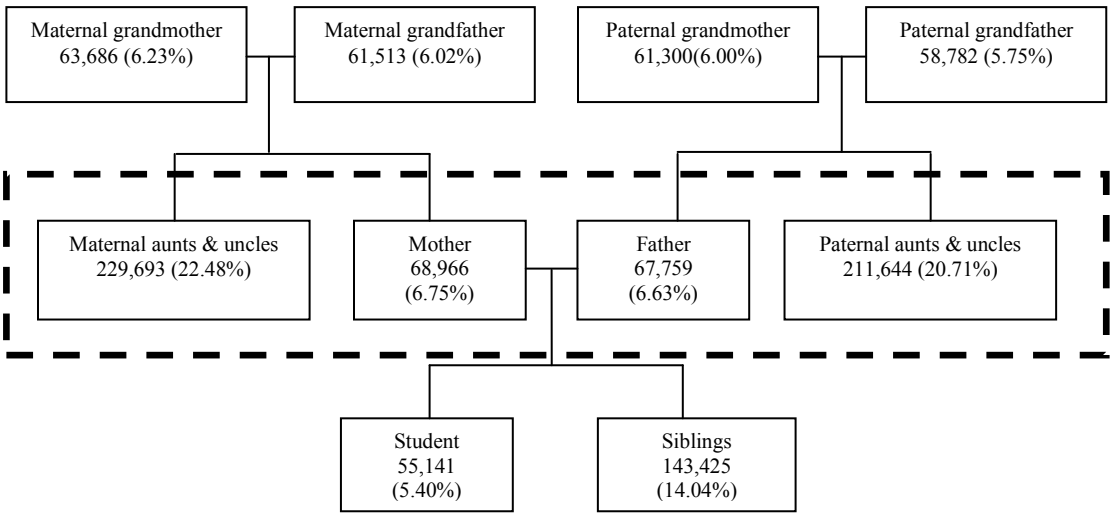


Figure 6.1. Family unit structure and number of records in the *Health Family Tree.* The student's parents, aunts, and uncles (highlighted) are the samples for machine learning.

Table 6.1. Features that were included in the machine learning

Features/Attributes	Data type	Values
Demographic		
Sex	Nominal	Male, female
Age group	Ordinal	<20, 20-29, 30-39, 40-49, 50-59, 60-69, ≥70
Comorbidity		
Myocardial Infarction	Nominal	Yes, no
Coronary heart disease	Nominal	Yes, no
Stroke	Nominal	Yes, no
Hypertension	Nominal	Yes, no
High blood cholesterol	Nominal	Yes, no
Breast cancer	Nominal	Yes, no
Lung cancer	Nominal	Yes, no
Colon cancer	Nominal	Yes, no
First-degree relatives' diabetes history		
Diabetic mother	Nominal	Yes, no
Diabetic father	Nominal	Yes, no
Diabetic sibling(s)	Nominal	Yes, no
Number of diabetic relatives	Numeric	0-9
Ratio of diabetic relatives	Numeric	0-1
Lifestyle risk factors		
Smoking	Nominal	Yes, no
Drinking	Nominal	Yes, no
Overweight	Nominal	Yes, no
Exercise	Nominal	Yes, no

Table 6.2. The advantages and disadvantages of the three chosen classifiers

	Logistic regression	**Bayesian network**	**Support vector machine**
Advantages	Based on the traditional statistic method, robust for categorical outcomes	Simple	Good for both linearly and nonlinearly separable data: Find optimal hyper plane for linear separable data; Fin kernels for data that are not linearly separable
	Features can be correlated because there are multiple ways to regularize	Fast	Nice theoretical guarantees regarding overfitting:
Disadvantages	Specified model ahead of time	Harder to handle continuous features	Slow Hard to interpret

Table 6.3. Confusion table of a two-class problem

		True classes	
		Diabetes	Non-diabetes
Prediction by the algorithm	Diabetes	True positive (TP)	False positive (FP)
	Non-diabetes	False negative (FN)	True negative (TN)

Table 6.4. Relationship between diabetes and individual features reported by parents, aunts, and uncles of students completing the *Health Family Tree*, 1983-2001

Feature and value	Diabetes prevalence (%)	Odds Ratio (95% CI)	Feature and value	Diabetes prevalence (%)	Odds Ratio (95% CI)
Sex			Smoking		
Male	2.5	1.0(0.9,1.0)	Yes	2.3	0.9(0.9,1.0)
Female	2.6	1.0 (ref)	No	2.5	1.0 (ref)
Age			Drinking		
<20	0.8	1.0(ref)	Yes	2.0	0.7(0.7,0.8)
20-29	1.3	1.6(1.3,2.0)	No	2.7	1.0 (ref)
30-39	1.6	2.0(1.6,2.4)	Overweight		
40-49	2.3	2.9(2.4,3.6)	Yes	3.5	2.1(2.0,2.1)
50-59	4.2	5.5(4.5,6.8)	No	1.7	1.0 (ref)
60-69	7.3	9.7(7.9,12.0)	Exercise		
>70	8.3	11.3(8.9,14.3)	Yes	1.8	0.6(0.6,0.7)
Myocardial Infarction			No	2.8	1.0 (ref)
Yes	6.7	2.9(2.6,3.1)	Diabetic mother		
No	2.4	1.0 (ref)	Yes	4.9	2.2(2.2,2.3)
Coronary heart disease			No	2.2	1.0 (ref)
Yes	6.4	2.7(2.5,3.0)	Diabetic father		
No	2.4	1.0 (ref)	Yes	4.4	1.9(1.8,2.0)
Stroke			No	2.3	1.0 (ref)
Yes	8.5	3.7(3.2,4.2)	Diabetic sibling(s)		
No	2.5	1.0 (ref)	Yes	11.1	6.8(6.6,7.1)
High blood pressure			No	1.8	
Yes	7.8	4.1(4.0,4.3)	Number of diabetic first degree relatives		
No	2.0	1.0 (ref)	0	1.4	1.0(ref)
High blood cholesterol			1	4.3	3.1(2.0,3.2)
Yes	7.2	3.4(3.3,3.6)	2	9.0	6.8(6.5,7.2)
No	2.2	1.0 (ref)	3	16.5	13.6(12.6,14.7)
Breast cancer			4	23.3	20.8(18.5,23.5)
Yes	4.6	1.9(1.6,2.2)	5	33.4	34.4(28.4,41.8)
No	2.5	1.0 (ref)	6	44.1	54.2(40.3,72.9)
Lung cancer			7	66.7	137(86,218)
Yes	8.2	3.5(2.8,4.3)	8	44.1	54.2(27.5,106)
No	2.5	1.0 (ref)	9	37.5	41.2(18.0,94.1)
Colon cancer					
Yes	7.8	3.3(2.8,4.0)			
No	2.5	1.0 (ref)			

Table 6.5. Evaluation of the three classifiers trained on the *Health Family Tree* data to predict diabetes (yes or no) at current time, in 5 years, and in 10 years

	Recall			Precision			F-measure		
	Current	In 5 years	In 10 years	Current	In 5 years	In 10 years	Current	In 5 years	In 10 years
Using all features*									
BN	0.63	0.55	0.58	0.73	0.72	0.73	0.68	0.63	0.65
LR	0.65	0.61	0.63	0.76	0.72	0.73	0.70	0.66	0.67
SVM	0.61	0.56	0.59	0.77	0.75	0.75	0.68	0.64	0.66
Using only age, sex and family health history									
BN	0.56	0.49	0.52	0.71	0.73	0.73	0.63	0.58	0.61
LR	0.59	0.58	0.58	0.73	0.71	0.72	0.65	0.64	0.65
SVM	0.58	0.55	0.58	0.74	0.73	0.73	0.65	0.63	0.64

BN= Bayesian network
LR= Logistic regression
SVM=Support vector machine
*Includes age, sex, family health history, comorbidities, and lifestyle risk factors.

References

1. CDC. *National Diabetes Fact Sheet: National Estimates and Genearl Information on Diabetes and Prediabetes in the United States*. Atlanta, GA: US Department of Health and Human Services; 2011.

2. ADA. Economic costs of diabetes in the U.S. in 2012. *Diabetes care.* 2013;36(4):1033-1046.

3. Boyle JP, Honeycutt AA, Narayan KM, et al. Projection of diabetes burden through 2050: impact of changing demography and disease prevalence in the U.S. *Diabetes Care.* Nov 2001;24(11):1936-1940.

4. Witten I, Frank E. *Data Mining: Practical Machine Learning Tools and Techniques*. 2 ed: Morgan Kaufmann; 2005.

5. Obenshain MK. Application of data mining techniques to healthcare data. *Infect Control Hosp Epidemiol.* Aug 2004;25(8):690-695.

6. Koh HC, Tan G. Data mining applications in healthcare. *J Healthc Inf Manag.* Spring 2005;19(2):64-72.

7. Breault JL, Goodall CR, Fos PJ. Data mining a diabetic data warehouse. *Artif Intell Med.* Sep-Oct 2002;26(1-2):37-54.

8. Huang Y, McCullagh P, Black N, Harper R. Feature selection and classification model construction on type 2 diabetic patients' data. *Artif Intell Med.* Nov 2007;41(3):251-262.

9. Barakat NH, Bradley AP, Barakat MN. Intelligible support vector machines for diagnosis of diabetes mellitus. *IEEE Trans Inf Technol Biomed.* Jul 2010;14(4):1114-1120.

10. Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC Med Inform Decis Mak.* 2010;10:16.

11. Karaolis M, Moutiris JA, Papaconstantinou L, Pattichis CS. Association rule analysis for the assessment of the risk of coronary heart events. *Conf Proc IEEE Eng Med Biol Soc.* 2009;2009:6238-6241.

12. Karaolis MA, Moutiris JA, Hadjipanayi D, Pattichis CS. Assessment of the risk factors of coronary heart events based on data mining with decision trees. *IEEE Trans Inf Technol Biomed.* May 2010;14(3):559-566.

13. Nassif H, Wu Y, Page D, Burnside E. Logical Differential Prediction Bayes Net, improving breast cancer diagnosis for older women. *AMIA Annu Symp Proc.* 2012;2012:1330-1339.

14. Saritas I. Prediction of breast cancer using artificial neural networks. *J Med Syst.* Oct 2012;36(5):2901-2907.

15. Biglarian A, Bakhshi E, Gohari MR, Khodabakhshi R. Artificial neural network for prediction of distant metastasis in colorectal cancer. *Asian Pac J Cancer Prev.* 2012;13(3):927-930.

16. Andersson B, Andersson R, Ohlsson M, Nilsson J. Prediction of severe acute pancreatitis at admission to hospital using artificial neural networks. *Pancreatology.* 2011;11(3):328-335.

17. Breiman L. Statistical modeling: the two cultures. *Statistical Science.* 2001;16(3):199-231.

18. Annis AM, Caulder MS, Cook ML, Duquette D. Family history, diabetes, and other demographic and risk factors among participants of the National Health and Nutrition Examination Survey 1999-2002. *Prev Chronic Dis.* Apr 2005;2(2):A19.

19. Dyck R, Karunanayake C, Pahwa P, et al. Prevalence, risk factors and co-morbidities of diabetes among adults in rural Saskatchewan: the influence of farm residence and agriculture-related exposures. *BMC Public Health.* 2013;13(1):7.

20. ADA. Standards of medical care in diabetes—2012. *Diabetes Care.* Jan 2012;35 Suppl 1:S11-63.

21. Thomas C, Hypponen E, Power C. Type 2 diabetes mellitus in midlife estimated from the Cambridge Risk Score and body mass index. *Arch Intern Med.* Mar 27 2006;166(6):682-688.

22. Glumer C, Carstensen B, Sandbaek A, Lauritzen T, Jorgensen T, Borch-Johnsen K. A Danish diabetes risk score for targeted screening: the Inter99 study. *Diabetes Care.* Mar 2004;27(3):727-733.

23. Joshi SR. Indian Diabetes Risk Score. *J Assoc Physicians India.* Sep 2005;53:755-757.

24. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten I. The WEKA data mining software: an update. *SIGKDD Explorations.* 2009;11(1).

25. Japkowicz N, S S. The class imbalance problem: a systematic study. *Intelligent Data Analysis.* 2002;6(5):429-449.

26. Weiss G. *Mining with Rarity: A Unifying Framework*: ACM;2004.

27. Drummond C, RC H. C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. Paper presented at: International Conference on Machine Learning 2003; Washington DC.

28. Guyon I, Elisseeff A. An introduction to variable and feature selection. *JMLR*. 2003;3:1157-1182.

29. US Census. Profile of general demographic characteristics: 2000 Census Summary file 1. http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk. Accessed June 1, 2013.

30. ADA. Diabetes statistics. http://www.diabetes.org/diabetes-basics/diabetes-statistics/. Accessed October 29, 2013.

31. CDC. What is diabetes? 2011; http://www.cdc.gov/chronicdisease/resources/publications/AAG/ddt.htm. Accessed June 25, 2013.

CHAPTER 7

DISCUSSION

Family Health History and Public Health or Population-based Research

Family health history (FHH) is an established disease risk assessment tool to stratify a healthy population and find individuals targeted for screening or health education for public health interventions or population-based research. Multiple advantages may exist when applying FHH for risk stratification and disease prevention. First, making good usage of FHH may reduce cost. One of the biggest problems related to healthcare in the United States is the cost. Since 1960, the US national health expenditure has been increasing rapidly.[1] The expenditure was $2.7 trillion in the year of 2011, which was 17.9% of the Gross Domestic Product.[1] Besides technology, insurance, administrative cost, changes in health care prices, and medical malpractice, the prevalence of chronic diseases was one of the important reasons that have lead to the high cost.[2] FHH may be used as a low cost, noninvasive screening tool for identifying populations at high risk for chronic diseases and implementing cost-effective primary and secondary prevention strategies. Another cost reducing benefit is through reducing time for the data collection. With the help of computerized and/or internet-based FHH tools, public health professionals and researchers could spend less or even no time to collect

FHH. Second, along with other information including genetic variations, biomarkers, environmental, and lifestyle factors, FHH can be included to develop comprehensive risk-prediction models used for genomic risk-stratified screenings.[3] Meanwhile, the quality in terms of both the completeness and the accuracy of the information may be improved through well-designed and carefully-controlled systems. Furthermore, with the integration of FHH into electronic health records (EHR) systems, the collected information could be structured and reused for public health programs and population-based research. The research performed for this dissertation addresses many questions that arise when suggesting the use of tools based on self-reported data and comparisons with population rates. For example, we addressed the following questions:

- *What is the Accuracy of self- or proxy-reported Family Health History Data?* Most current FHH data are patient self- and proxy-reported. This dissertation examined the accuracy of a database that contains more than one million self- and proxy-reported family medical history and lifestyle risk factor information. When compared to the Utah Cancer Registry and standardized public health survey data, the disease and lifestyle risk factor rates had similar patterns as compared to the rates from the public sources: all disease rates increased with age, smoking rates are higher in the middle age groups, and exercise rates decrease with age. Cochran-Mantel-Haenszel test results indicated that rates reported for stroke (overall, self- and proxy-reported), self-reported breast cancer, and self-reported lung cancer were not significantly different from the rates in the public data source, while the rates for other diseases and risk factors were significantly different. Chi-square tests by sex- and age-

subgroups indicated that most subgroup disease rates and smoking rates were underreported with a few exceptions in the extreme age groups; exercise rates were overreported compared to the rates in public data sources. The comparison between self- and proxy-reported data indicated that when reporting diseases, self-reported rates were closer to the rates in public data sources than proxy-reported, though both were underreported for most diseases; when reporting life style risk factors, self-reported rates were further away from the rates in public data sources, for both underreported smoking rates and overreported exercise rates.

- *Can the Health Family Tree Algorithm predict risk?* Using a retrospective cohort design, we validated the predictability of the *Health Family Tree* algorithm using both the Tree database itself and the public data sources as reference to generate the expected incidence. Both validations indicated similar results that the very high risk scores (FHS ≥2) derived from the algorithm predicted the future risk for all included diseases, with or without considering lifestyle risk factors.
- *Can Family Health History be used to predict risks without the use of population disease rates?* One factor that prevents the broader application of the traditional risk predictive model such as the *Health Family Tree* is its nonstraightforward implementation. To implement the algorithm to predict disease risk within other systems that collect family history information, reference population disease rates are required to calculate the expected disease events. This reference is not always available or straightforward to obtain. We

showed that machine learning methods can be used to train and evaluate another risk predictive model which is simpler to implement at any other health system.

Indications for Risk Assessments in Public Health and Population-based Research

For longitudinal, population-based research and public health screening, this research provided valuable indications about the effective use of FHH from data collection to the implementation of risk prediction models. For FHH data collection, we proposed 50 requirements including data, functional, and nonfunctional requirements to be considered. For risk prediction models, we validated a traditional statistical model and demonstrated how this model can be implemented in any system that collects FHH information. For longitudinal, population-based studies that contain a large amount of subjects' FHH information, the statistical risk prediction can be implemented similarly as the *Health Family Tree*, using the study population as a reference to generate the expected incidence. For other studies that do not have a large amount of subjects' FHH, or for a clinical system where the incidence of disease in the patient population does not reflect rates in the general population, the incidence from public data sources with interpolation can be applied to the risk prediction. The accuracy study of self- and proxy-reported FHH provided indications that most concerned diseases were underreported statistically, so the risk predictions may be overestimated. This effect was confirmed by the comparison of risks derived from using *Health Family Tree* as a reference vs. using public data sources as a reference. Finally, the new risk predictive models built by

machine learning methods provided effective prediction for a specific subpopulation (Caucasian, middle-age groups of adults) and a relatively straightforward method for implementation.

Contributions to Biomedical Informatics

Biomedical Informatics is a multidisciplinary science that includes a wide range of subfields. This dissertation addressed challenges related to two major domains of Biomedical Informatics: acquisition of quality data and implementation of decision support using predictive methods to identify populations at risk. First, the research examined health information acquisition, including requirements for data and functions that needed to be included, and the examination of the quality of the family health history data that were self- or proxy-reported. Second, the research evaluated an existing tool for predicting risk and developed new algorithms using different methods that may be easier to implement in current EHRs. Specifically, we validated an existing risk prediction algorithm based on classic statistical models and built a new predictive model based on machine learning for easier implementation. Finally, this research is unique in its focus on population health rather than individual clinical decision support in that we evaluated the use of a tool that has the potential to support public health strategies that allow populations to use a tool to identify those in need of further screening.

Future Directions

To further improve the use of FHH, there are research questions that need to be explored based on the work of this dissertation. To increase the interoperability of sharing FHH information (including the individual's medical information and family relationship)

between different systems and organizations, information modeling will be needed to structure how the information is organized for storage and transmissions. The Health Level 7 (HL7) community has been developing a clinical Genomics pedigree model.[4] This model is approved by the American National Standards Institute and is in the process of being accepted as an international standard. To improve the completeness and accuracy of FHH data, social networking may be used for data collection. With the advancement of the Internet and consumer technologies, more and more patients are involved in the decision process with their health and health care. Health2.0/medicine 2.0 is an analogy to Web 2.0 technology and it is developing quickly.[5] Multiple social networks/online communities such as Patient Like Me[6] have been created for patients to exchange information and interact with each other. Compared to the traditional method of collecting FHH through one member of the family, social networks allow multiple family members to participate, which may increase data quality and completeness but may introduce new problems such as how to resolve conflicting information. To further explore the application of supervised machine learning methods to the *Health Family Tree* data, predictive models of other diseases that were collected by the database should be built and evaluated. To streamline the process of using FHH for disease risk assessment, future work is needed to integrate FHH prediction models into the EHR or personal health records (PHR). Implementers need to consider the differences in the implementation requirements and the output provided to the user when choosing a risk predictive model. For example, a statistical risk model requires expected disease incidence rates from a reference population or data source and uses the relative risk of the individual to predict risk. In contrast, a machine learning model requires a validated

model and access to the data fields required by the model and can then output a disease classification prediction based on defined levels of recall and precision. Another future direction concerns how to communicate risk assessment results to the public and direct them to an accurate perception of their risks. According to the multiple theories such as the Health Belief Model,[7] Stages of Change Model,[8] and the Theory of Planned Behavior,[9] perceived risk and attitude are two of many important factors that lead to behavior change. Past studies have shown that an individual's attitude and perception of risk is affected by the manner in which information is presented to the user.[10] Also, patients may have different preferences between absolute risk and relative risk.[11] Thus, how to accurately deliver the risk assessment information to the public is another research question that needs to be studied in depth.

Conclusion

In conclusion, to better use family health history to predict an individual's or population's risk of developing selected chronic diseases, especially in the context of public health or population-based research, various requirements for family health history tools proposed herein should be considered when choosing an existing tool or building a new tool; risk prediction models may be built through various ways including statistics and machine learning; self- or proxy-reported family health history data collected by research projects such as *Health Family Tree* may generate lower disease prevalence or incidence rates compared to the rates generated from data collected by public health surveys or cancer registries. Disease predictive models built by the *Health Family Tree* program using the self- and proxy-reported data are still valid for predicting the future development of multiple chronic diseases for an unaffected adult between 20 and 99

years of age. These findings may be especially useful when developing strategies to screen populations for common chronic diseases and identifying those at highest risk for public health interventions or population-based research.

References

1. CMS. National health expenditures: aggregate and per capita amouts, annual percent change and percent distribution: selected calendar year 1960-2011. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/downloads/tables.pdf. Accessed May 4, 2013.

2. SSAB. *The Unsustainable Cost of Health Care*. Washington, DC: Social Security Advisory Board; 2009.

3. Chowdhury S, Dent T, Pashayan N, et al. Incorporating genomics into breast and prostate cancer screening: assessing the implications. *Genet Med.* Jun 2013;15(6):423-432.

4. HL7. HL7 version 3 standard: clinical genomics; pedigree, release 1. http://www.hl7.org/implement/standards/product_brief.cfm?product_id=8. Accessed May 4, 2013.

5. Van De Belt TH, Engelen LJ, Berben SA, Schoonhoven L. Definition of Health 2.0 and Medicine 2.0: a systematic review. *J Med Internet Res.* 2010;12(2):e18.

6. PatientLikeMe. Patient Like Me. http://www.patientslikeme.com/. Accessed May 4, 2013.

7. Becker MH, Maiman LA, Kirscht JP, Haefner DP, Drachman RH. The Health Belief Model and prediction of dietary compliance: a field experiment. *J Health Soc Behav.* Dec 1977;18(4):348-366.

8. Prochaska JO, DiClemente CC. Stages and processes of self-change of smoking: toward an integrative model of change. *J Consult Clin Psychol.* Jun 1983;51(3):390-395.

9. Ajzen I. The theory of planned behavior. *Organ Behav Hum Decis Process.* 1991;50:179-211.

10. Ancker JS, Kaufman D. Rethinking health numeracy: a multidisciplinary literature review. *J Am Med Inform Assoc.* Nov-Dec 2007;14(6):713-721.

11. Fortin JM, Hirota LK, Bond BE, O'Connor AM, Col NF. Identifying patient preferences for communicating risk estimates: a descriptive pilot study. *BMC Med Inform Decis Mak.* 2001;1:2.