Similarity of Medical Measures in Social Health Networks

by Lewis Evan Berman

B.S. in Computer Science, December 1985, University of Maryland
M.S. in Computer Science, February 1990, The George Washington University

A Dissertation submitted to

The Faculty of
The School of Engineering and Applied Science
of The George Washington University
in partial satisfaction of the requirements
for the degree of Doctor of Philosophy

May 15, 2011

Dissertation directed by:

Shmuel Rotenstreich
Associate Professor of Computer Science

The School of Engineering and Applied Science of The George Washington University certifies that Lewis Evan Berman has passed the Final Examination for the degree of Doctor of Philosophy as of March 25, 2011. This is the final and approved form of the dissertation.

Similarity of Medical Measures in Social Health Networks

Lewis E. Berman

Dissertation Research Committee:

Shmuel Rotenstreich, Associate Professor of Computer Science, Dissertation Director

Gabriel Parmer, Assistant Professor of Computer Science, Committee Member

James K. Hahn, Professor of Computer Science, Committee Member

Poorvi Vora, Associate Professor of Computer Science, Committee Member

George Thoma, Branch Chief, National Institutes of Health / National Library of Medicine, Committee Member

Copyright © 2011 by Lewis Evan Berman

All rights reserved

Dedication

To my parents Roz and Len Berman. I wish they were here to celebrate!

Acknowledgement

It is said that Ernest Shackleton renamed his ship the "Endurance" after his family motto, *fortitudine vincimus*, which translates to "by endurance we conquer." I am insufficiently brave to assume that I have conquered the art of research. However, I have certainly conquered my fear of attempting such a vocation, perhaps a bit later then preferred. Nevertheless, it is an effort for which humbleness is immediately made visceral and lasting pride in accomplishment is earned.

For this, I owe my deepest gratitude to Professor Shmuel Rotenstreich. When we first discussed the Ph.D. program, Professor Rotenstreich was forthright about the effort required. He assured me that hard work, a sprinkling of intelligence, and perseverance were the most important ingredients. Through this journey Professor Rotenstreich's keen insight, allowance for mistakes, and thoughtful criticism remained unwavering. He served as a mentor in research group meetings, on car rides home, on the metro, and over a meal. He routinely offered good cheer and humor. He taught me to think differently, be concise (excepting this acknowledgement), and to search for the essence of a problem. I am grateful for his contributions to my development, mentoring, and friendship.

I am also thankful to the faculty members who served on my doctoral committees: Professors James Hahn, Gabriel Parmer, and Poorvi Vora. Their interest, comments, and suggestions have been an invaluable aid in improving this dissertation. I would also like to thank Professor Murray Loew who served as my master's degree advisor.

Numerous students in our research group have been an endless source of help, time, and support. I want to acknowledge fellow students George Blankenship, Dave Christian, Amelia Estwick, Jae Hwang, Horatious Tanyi, and Brad Taylor. I am particularly grateful to George Dimitoglou for his support.

Over the years, I have had the privilege of working at the Naval Research Laboratory, National Library of Medicine, and National Center for Health Statistics. At these research institutions Drs. Tom ApRhys, Randy Curtin, Dave Kerr, Raynard Kington, Gerry McQuillan, Kathryn Porter, George Thoma, and Mr. Clifford Johnson offered their support, encouragement, and mentorship. They have made innumerable and invaluable contributions to my professional and academic achievements. I am particularly fortunate to have befriended Yechiam Ostchega, Ph.D., R.N., Capt. USPHS (Retired). He pushed me over the hump when I was uncertain about returning to school. His commitment to science and dedication to learning are inspiring. He was always available, he encouraged, and he mentored.

I am thankful to have two wonderful sisters, Robyn and Lori. They have been a constant source of love, humor, support, and strength throughout my life.

Lastly and most importantly, without my wife Sharon, and daughters Rachel and Danielle, this would not have been possible. They endured many meals, nights, and events alone while I pursued my studies. Their love, patience, understanding, cards and notes, and support have been a source of strength. They poked and prodded, encouraged, and asked endless questions. They always showed interest. I love you.

Abstract

Similarity of Medical Measures in Social Health Networks

The increasing use of Internet is making it possible for individuals to connect with one another for social networking and socially significant reasons. With the advent of electronic health records and the associated medical measurements, individuals will increasingly expect and gain access to their medical information. The availability of this information may serve as the foundation for automatically connecting similar individuals in social health networks (SHN).

Due to the potentially large number of online individuals and the number of medical measurements in an electronic health record, the record and feature space for connecting individuals is large. Therefore, this dissertation proposes a computational model to build SHNs using filtering, data reduction, and similarity coefficients. The filtering and data reduction steps reduce the scale of the problem. Two approaches for computing feature based similarity coefficients are explored.

A simulation of the computational model is demonstrated with two examples. The findings from the simulation demonstrate that the record and feature space can be significantly reduced. Furthermore, this research demonstrates that SHNs can be automatically formed with the defined computational model.

DEDICATION ..... **IV**
ACKNOWLEDGEMENT ..... **V**
ABSTRACT ..... **VII**
LIST OF FIGURES ..... **X**
LIST OF TABLES ..... **XI**
LIST OF ACRONYMS ..... **XII**
CHAPTER 1 INTRODUCTION **1**
1.1 PROBLEM STATEMENT 2
1.2 CHALLENGES AND SIGNIFICANCE 2
1.3 RESEARCH APPROACH 3
1.4 DISSERTATION ORGANIZATION 3
CHAPTER 2 LITERATURE REVIEW **4**
2.1 SOCIETAL SHIFT DUE TO INTERNET ACCESS 4
2.2 SOCIAL SUPPORT AND BRIDGING TO THE INTERNET 5
2.2.1 Current On-line Health Support *7*
2.3 MEDICAL MEASUREMENTS 7
2.3.1 Measurement Properties and Types *8*
2.3.2 Subjective and Objective Measurements *10*
2.3.3 Medical Data Coding *10*
2.3.4 Measurement Variation *11*
2.3.5 Medical Events *12*
2.3.6 Temporal Considerations *12*
2.4 DATA REDUCTION 13
2.5 SIMILARITY 14
2.5.1 Geometric Distances *15*
2.5.2 Ontological Models *17*
2.5.3 Feature-Based Models *18*
2.5.4 Gower's Model *20*

2.5.5 Feature Weighting *21*

2.6 SUMMARY 21

CHAPTER 3 SOCIAL HEALTH NETWORK COMPUTATIONAL MODEL 23

3.1 FILTERING 23

3.2 DATA REDUCTION 25

3.3 SIMILARITY 28

3.3.1 Tolerance Ranges *28*

3.3.2 Similarity Computation *31*

3.4 DATASET SELECTION CRITERION 32

3.4.1 Candidate Datasets *33*

3.4.2 Evaluation of Candidate Datasets *34*

3.4.3 Missing Data *35*

3.5 SHN SIMULATION 35

3.5.1 Example 1: Type 2 Diabetes *36*

3.5.2 Example 2: Chronic Kidney Disease *37*

3.5.3 Simulation Software *38*

CHAPTER 4 RESULTS 39

4.1 FILTERING 39

4.2 DATA REDUCTION 39

4.3 SIMILARITY 43

CHAPTER 5 DISCUSSION 53

CHAPTER 6 CONCLUSION 59

REFERENCES 60

APPENDIX A 71

List of Figures

Figure 1.Conceptualization of geometric distances. 16

Figure 2.Computing object distance by traversing an ontology. 17

Figure 3. Tversky Feature Contrast Model Coefficient. 20

Figure 4. SHN computational model. 23

Figure 5. Simple events with temporal overlap. 24

Figure 6. Ordinal measurement type. 30

Figure 7.Type 2 diabetes principal component values. 40

Figure 8.Chronic kidney disease principal component values. 40

Figure 9.Type 2 diabetes principal component unique and cumulative proportions. 41

Figure 10.Chronic kidney disease principal component unique and cumulative proportions. 41

Figure 11.Type 2 diabetes loadings. 42

Figure 12 Chronic kidney disease loadings. 42

Figure 13.Type 2 diabetes TFCMC descriptive similarity statistics. 49

Figure 14.Type 2 diabetes JC and GC descriptive similarity statistics. 49

Figure 15.Chronic kidney disease TFCMC descriptive similarity statistics. 50

Figure 16.Chronic kidney disease JC and GC descriptive similarity statistics. 51

Figure 17.Correlation coefficients associated with type 2 diabetes similarity coefficients. 52

Figure 18.Correlation coefficients associated with chronic kidney disease similarity coefficients. 52

List of Tables

Table 1.Feature-based similarity. 18

Table 2. Feature-based model variations. 19

Table 3. Temporal relationships in Allen's Algebra. 25

Table 4. Sample population correlation matrix for all measures. 27

Table 5. Example similarity matrix for five individuals. 32

Table 6. Dataset selection criteria. 32

Table 7.Application of dataset criterion to candidate datasets. 34

Table 8.Percentage-based tolerance ranges for type 2 diabetes measures. 44

Table 9.Percentage-based tolerance ranges for chronic kidney disease. 45

Table 10.Clinically relevant cut-point tolerance ranges for diabetes measures. 46

Table 11.Clinically relevant cut-point tolerance ranges for chronic kidney disease measures. 47

Table 12.Type 2 diabetes descriptive statistics. 48

Table 13.Chronic kidney disease descriptive statistics. 50

Table 14. NHANES 1999-2003 data items that correspond to diabetes. 71

Table 15. NHANES 1999-2003 data items that correspond to chronic kidney disease. 75

Table 16.NHANES 1999-2000, 2001-2002, 2003-2004 Data file URLs. 80

List of Acronyms

BCL - Breast Cancer List
BMI - Body Mass Index
BUN - Blood Urea Nitrogen
CDC - Centers for Disease Control and Prevention
CHI - Consolidated Health Informatics
CKD – Chronic Kidney Disease
CPT - Current Procedural Terminology
CPTR – Cut-point Tolerance Range
CVD – Cardiovascular Disease
EHR - Electronic Health Record
FA – Factor Analysis
FPG – Fasting Plasma Glucose Test
FBSC - Feature-based Similarity Coefficients
FHS - Framingham Heart Study
GC - Gower Coefficient
GFR - Glomeruler Filtration Rate
GO - Gene Ontology
HDL - High Density Lipoprotein
HIMSS - Healthcare Information and Management Systems Society
HIPAA - Health Insurance Portability and Accountability Act
HIT - Health Information Technology
ICD - International Classification of Disease
ICD 9 - International Classification of Disease 9th Edition
ICDF - Implantable Cardioverter Defibrillator
ICF - International Classification of Functioning, Disability, and Health
ICSG - Internet Cancer Support Group
IDDM – Insulin-dependent Diabetes Mellitus
JC - Jaccard Coefficient
LDL - Low Density Lipoprotein
LOINC - Logical Observation Identifiers Names and Codes
MDS – Multidimensional Scaling
NAMCS - National Ambulatory Medical Care Survey
NCHS - National Center for Health Statistics
NCVHS - National Committee on Vital and Health Statistics
NHANES - National Health and Nutrition Examination Survey
NHIS - National Health Interview Survey
NIH - National Institutes of Health
NN – Neural Network
NRA - National Research Act
OGTT – Oral Glucose Tolerance Test

PA - Privacy Act
PBTR – Percentage-based Tolerance Range
PC – Principal Component
PCA – Principal Component Analysis
PPMC - Pearson Product Moment Correlation
PHR - Personal Health Record
PIALP - Pew Internet & American Life Project
SHN – Social Health Network
SNOMED - Systematized Nomenclature of Medicine
SNOMED CT® - Systematized Nomenclature of Medicine Clinical Terms
SP - Survey Participant
TFCMC - Tversky Feature Contrast Model Coefficient
UMLS - Unified Medical Language System

Chapter 1 Introduction

Over the last 20 years, there has been an increase in the demand for information access due to the widespread use and ubiquitous nature of the Internet. Access to products and services is no longer a phone call or a drive away; instead, it is a link to a website. The increasing demand for online access has been satisfied with high-speed transmission through cellular service, Wi-Fi, and Internet service providers. This high-speed transmission ensures persistent access to information at home, in the car, and at work or school. Concurrently, medicine has undergone significant change in equipment, procedures, treatments, monitoring, and specialization. In this regard, the federal government of the United States (U.S.) is investing in health information technology (HIT) and electronic health records (EHR) with the hope that it will improve health [1].

With the increasing use of the Internet, online communities have developed for social collaboration to support connections between individuals. They tie individuals together by common interests, friendships, employment, and for other reasons. While online communities have been forming for pleasure and social interest, others are forming for socially significant purposes. There is evidence that social support buffers stress in terms of emotion, information sharing, and a feeling of being integrated with others [2].

Currently, individuals self-select into an online health community, or social health network (SHN), based on pre-defined groups. An alternative to self-selection is automated formation of health communities based on medical measurements. In essence,

medical measures from a personal health record may be used as a "matchmaking mechanism" between patients [3, page 6]. Moreover, since social support has a positive impact on health [2], automated grouping of individuals may contribute to improved health. It may have further benefits of sharing information, education, and contributing to future epidemiological research.

1.1 **Problem Statement**

A common problem across disparate disciplines is grouping objects based on feature similarity. This is a classification. In the biological sciences, classification has been emphasized to develop taxonomies such as the well-defined classification of the animal kingdom. Despite these accomplishments, to date, automated techniques have not been used to group similar individuals into SHNs. Therefore, this dissertation proposes to build a SHN computational model for grouping individuals similar to a *candidate* individual, based on their medical measurements. The computational model consists of filtering, data reduction, and similarity computation.

1.2 **Challenges and Significance**

There are several challenges in this dissertation. The first challenge is identifying the types of measures and associated characteristics, which can be used to determine whether a medical measurement from one person is the same as from another person. Second, there are myriad medical measurements. Thus, the problem space must be reduced so that the number of observations or measurements does not overwhelm the computational model. Third, similarity computation should support any type of medical

measurement whether it is raw data or a coded value. The similarity computation should be inexpensive for varied numbers of features. Fourth, the computational model should be inexpensive to compute when new individuals want to enter an existing group. Finally, the model should be assessed with a simulation.

This research is significant because to date no automated techniques have been developed for the purpose of grouping individuals into SHNs. A possible consequence of this research is an efficient computational model for SHN formation. Additionally, this work may be applicable to epidemiological studies.

1.3 Research Approach

The research approach in this dissertation is to develop the SHN computational model and simulate the model with a database of medical measures. The proposed SHN computational model will include several steps for filtering, data reduction, and similarity computation. The SHN is simulated with a publicly available database of medical measurements.

1.4 Dissertation Organization

Chapter 2 is review of the relevant literature. Chapter 3 concentrates on the computational model for grouping, comparison of similarity techniques, selection of a database, and a brief example. Chapter 4 presents the results of simulating the formation of SHNs using diabetes and chronic kidney disease as examples. Chapter 5 is a discussion of the results including some of the limitations of the model and simulation. The last chapter of this dissertation is the conclusion.

Chapter 2 Literature Review

Currently, existing SHNs form based on self-selection into pre-existing groups or by creating a new group. For example, a person interested in cancer attempts to find an online community that fosters conversation about this topic. While this approach has merit, it is limiting since there are many varieties of cancer and individuals may need groups that are better matches for their specific needs. Furthermore, individuals may also have challenges in finding the correct information or group [4].

This chapter is a review of the research and computational techniques related to the development of a computational model for SHNs. The topics cover societal shifts due to Internet access, the linkage of social support and the connection to health, existing on-line support groups, medical measurement types, data reduction, and similarity coefficients.

2.1 Societal Shift Due To Internet Access

In 2009, the Pew Research Center conducted the Pew Internet & American Life Project (PIALP) to monitor the online activities of American adults [5]. Findings from PIALP show that 79% of American adults use the Internet and 63% of households have broadband service. American adults, ages 18 and above, comprise 75% of the US population [6]. This suggests that there are roughly 180 million American adults using

the Internet[1]. This is consistent with polls showing 160 and 150 million people used the Internet for health information in 2007 and 2008 [7].

The top five Internet activities are sending or reading email (90%), using a search engine to find information (88%), searching for a map or driving directions (86%), looking for health medical information (83%), and looking for information on a hobby or interest (83%) [8]. Other findings show that adults are increasingly using online social networking sites, with usage at 8% in 2005 and 35% in 2009 [9]. This represents a sharp increase in usage of social networking sites.

2.2 **Social Support and Bridging to the Internet**

Social support has been defined "as the resources provided by other persons" that may have either a positive or negative effect on health [10, page 4]. Social support consists of four categories of behavior: emotional support, appraisal support, informational support, and instrumental support [11]. There is evidence that social support buffers stress and contributes to an individual's outlook and satisfaction in life. Additionally, some studies have shown that persons with weaker social networks are at greater risk for disease and death [12]. Thus, the ability to connect with others and strengthening social networks may offer people benefits.

A number of research studies have shown a positive association between Internet use and social support, particularly with respect to medical issues. For example,

---

1 This is calculated using the 2008 American Community Survey 1-Year Estimate of the US American population at 304 million and adults 18 and above comprising about 75% of the population. The 2008 American Community Survey 1-Year Estimates can be found at http://factfinder.census.gov.

implantable cardioverter defibrillators (ICDFs) are used by individuals at risk for a sudden cardiac death. In 2005, the Centers for Medicare and Medicaid Services started the National ICDF Registry, which requires all individuals with an ICDF to enroll in the registry [13]. One study investigated the use of Internet-based bulletin boards by individuals with ICDFs. This study found that bulletin board users were willing to share personal feelings and experiences, chronicle stories about themselves so they could connect to others, and offer support [14].

Another example is a 1997 evaluation of the on-line Breast Cancer List (BCL) discussion group, which was open and un-moderated. The BCL was created for people who had or were concerned about breast cancer topics, such as support, therapies, or surgery. Membership consisted of people mostly from the US and Canada and changed frequently. The BCL attracted a diverse group of people at different stages of breast cancer. Discussions concentrated on tumors, diagnoses, medication efficacy, and psychological support. An important finding of the BCL evaluation is that the discussion group has some advantages over traditional supportive communication. These advantages include persistent resource availability and membership of people from around the world [15].

Finally, one researcher investigated gender differences in Internet Cancer Support Groups (ICSG). This researcher focused on prostate cancer, breast cancer, and a mixed group. A review of the messages illustrated that in the ICSG, men in the prostate group were most interested in information exchange, followed by personal experience, and

personal opinion. Women in the breast cancer group were most interested in personal experience, followed by information exchange, and support. In the mixed group, individuals were most interested in information exchange, followed by personal experience, and support [16].

2.2.1 Current On-line Health Support

There are a number of existing web based health sites such as HopeCube, DailyStrength, and MyHealthExperience. They offer individuals emotional support, experience and information sharing, and medical advice. Some offer information on the latest drug treatment and alternative therapy research. There are a variety of pre-existing communities for conditions such as ALS/Motor Neuron Disease, Bipolar, and Parkinson's disease; in some cases, an individual may create a new community. These sites do not utilize self-reported information, objective health measurements, or health record data to connect individuals. Furthermore, in general, an individual enrolls and then self-selects an online community based on the name and description of a disease [17, 18, 19, 20].

2.3 Medical Measurements

Measurement is defined as the assignment of a number to an attribute of some instance of an object. An important consideration in measurement is that the "properties of the attribute are faithfully represented as numerical properties" as described by Krantz [21, page 1]. Krantz goes into detail about the systems that lead to measurement and the basic procedures for assigning numbers, including a rigorous discussion of the axioms and properties for this assignment. The measurement must characterize the object attribute or else the measurement does not convey useful information.

Medical measurements are the quantifiable result of tests, procedures, treatments, health history questions, or diagnoses, forming the foundation of an individual's health state. Examples of medical measurements are blood pressure, total cholesterol, among many others. Each measurement is constrained by a variety of factors such as the reliability and validity of the equipment used, consistency with other measures, procedures and methods, diurnal variation, and physiological processes.

2.3.1 Measurement Properties and Types

Measurements contain one or more of three basic properties of an object attribute. These properties are denoted as *magnitude*, *equal intervals*, and an *absolute zero point*. *Magnitude* refers to a measurement whereby the attribute of one instance of an object is determined to be greater than, less than, or equal to the attribute of different instance of that object. *Equal intervals* denote that the difference between any two adjacent points of an attribute on a measurement scale is equal. Thus, the difference between a weight of 150 pounds and 160 pounds is the same as the difference between a weight of 250 pounds and 260 pounds. Lastly, an *absolute zero point* indicates that nothing at all exists for the object attribute being measured [22].

In general, there are four measurement types. The first type is *nominal*[2] measurement, which separates data into discrete groups that are mutually exclusive. *Nominal* measurements do not possess the properties of *magnitude*, *equal intervals*, or *absolute zero point*. Thus, there is no ordering. In effect, *nominal* measurements are

---

[2] The literature uses n*ominal* and *categorical* interchangeably.

arbitrarily assigned labels for different objects. For example, gender is a *nominal* measurement for which males may be assigned a value of '1' and females the value of '2'.

The second type is *ordinal* measurement. *Ordinal* measurement assigns objects to categories such that these categories have a meaningful rank. *Ordinal* measurement possesses *magnitude*, but not *equal intervals* or *absolute zero point*. In epidemiological research, people may be pooled into different fitness groups such as poor, good, and outstanding based on an individual's perception of fitness level. While there is an ordering and a sense of the *magnitude* difference between fitness groups, it is not possible to determine the actual difference.

A third measurement type is *interval*, which possesses the *magnitude* and *equal intervals* attributes. An example of an *interval* measurement is Fahrenheit temperature. A temperature of 80° F is greater than a temperature of 60° F. However, temperature, like all *interval* measurements, has two interesting distinctions. First, a temperature of 0° F does not suggest the absence of temperature. Secondly, even though temperature measurements possesses *equal intervals* it is not the case that there is a true zero point and as a result, ratios between *interval* measures do not exist. Thus, 100° F is not twice as hot as 50° F.

The fourth measurement type is *ratio*, which possesses the attributes of *magnitude*, *equal intervals*, and *absolute zero point*. Thus, a person who weighs 200 pounds is twice as heavy as a person weighing 100 pounds (*magnitude*) and a 50 pound

difference between any two weights always has the same meaning (*equal interval*) [22, 23].

2.3.2 Subjective and Objective Measurements

Medical measurements may be either subjective or objective. Subjective measurements result from questions a health care provider asks an individual during an encounter or from self-reported information on a questionnaire. Questions may include topics such as weight history, smoking and alcohol use, illicit drug use, stress, use of prescription medications, pain, and physical activity. Objective measurements result from the judgment of a healthcare professional or a medical device. There are numerous categories of objective measurements including physical, biological specimens, image and signal data, and coding systems.

Physical measurements include items such as height, weight, blood pressure, vision, and hearing. Devices such as stadiometers, scales, blood pressure mercury manometers, and spirometers are used for physical measurement capture. Biological specimens are any material derived from an individual. Examples include tissue, hair, blood, nasal swabs, and urine. Many diagnostic medical measurements are the result of capturing an image or signal. For example, there are x-ray images, magnetic resonance images, and electro-cardiograph traces.

2.3.3 Medical Data Coding

Medical data may be represented as standardized codes from any of the more than 150 coding systems. These coding systems are used for procedures, measurements, diagnosis, disability, or treatment [24, 25]. For example, a blood pressure reading of

150/80 mmHg is represented in the Current Procedural Terminology (CPT®) with a code of '2000F' and in the International Classification of Disease (ICD) with a *code* of '401.9' [26, 27].

Some examples of medical coding systems are the CPT, ICD 9th Revision Clinical Modification (ICD-9-CM), the Logical Observation Identifiers Names and Codes (LOINC), and the Systematized Nomenclature of Medicine Clinical Terms® (SNOMED-CT). CPT describes medical, surgical, and diagnostic services under health insurance programs [28]. ICD-9-CM categorizes mortality statistics, contains a list of disease codes, and has entries for surgical and other procedures [29, 30]. LOINC is an extensive classification of laboratory measurements and other clinical observations [31, 32]. Lastly, SNOMED-CT® is an extensive ontology describing clinical data and relationships between data at a very granular level [33].

Two important aspects of a coding system are the amount of detail that can be coded and the complexity that can be represented by the standard. The greater the information and relationships the more likely it is that there will be greater meaning in the standard. Studies have shown that SNOMED-CT® contains the greatest amount of content coverage [34, 35, 36].

2.3.4 Measurement Variation

Measurements have several sources of variation. Some of these variations are due to methods, procedures, equipment, lifestyle sources, demographics, and diurnal variation. Methods variation may be arise due to machine calibration or quality assurance / quality control procedures. Lifestyle sources affecting measurements include

the consumption of food and drinks, smoking, fasting requirements, exercise, and medication use. Demographic variation is due to age, gender, or race / ethnicity differences. Lastly, measurements may change due to diurnal variations. For example, blood pressure may be higher prior to awakening and lower in the evening [37].

2.3.5 Medical Events

A health record consists of medical measurements at different points in time. A medical event consists of one or more medical measurements. A *simple event* is a single measurement. A *complex event* is a set of *simple events* for which there is a relationship among the measurements. *Simple events* may have linkages to other simple events or to complex events. *Complex events* may have linkages to other complex events. Events are marked with a timestamp, quantifying the time and date that the event occurred. An example of a complex event is a routine yearly physical exam with a primary care physician. All of the measurements recorded at the physical are related to one another and compose a complex event. Another example is diabetes where questions on numbness or pain in extremities, blood pressure, and several laboratory measures are important.

2.3.6 Temporal Considerations

Information processing research has long considered time in temporal logics, database design, query languages, event relationships, medical records, and knowledge-based decision-support systems [38, 39]. Temporality is defined as a time-frame over which a fact is true, establishing relationships among data or events [40]. A general temporal model consists of two elements: a time-line and the time unit. A time-line is

partitioned into chronons, which are the shortest meaningful time unit [41]. Chronons may be viewed in three different ways. First, there is a time point or time instance, which is a single chronon on a time line. Second, there is the time period, which is the difference between two time points. Lastly, there is the time interval that is a length of a segment along a time-line that is not fixed on a time-line [42]. The time unit defines the granularity of the chronon. The time-unit is either a continuous, rational, or integer number [43].

There are numerous pieces of temporal information related to medical measurements found in a medical record. This includes the time and date for collecting, processing (e.g., laboratory), shipping (e.g., specimens), and interpreting a measurement. In the 1970's, investigators pointed out the need to work with time more completely in patient's medical records and database systems [44]. Some of this research defined temporal query languages, precise definitions of temporal relationships, and overarching concepts [45, 46]. This work formed the basis for temporal research in medical reasoning, patient monitoring, clinical guidelines, internal medicine, hepatitis, and neonatology [39, 47, 48, 49, 50, 51].

2.4 **Data Reduction**

The definition of data reduction is the process of converting large sets of data into a smaller number of data points. Data reduction is useful because not all data points are important for understanding the concept of interest. Mathematically, data reduction is the transformation of an n-dimensional vector of observed data points or measurements, m =

$(m_1, m_2, \ldots, m_n)$, to a k-dimensional vector of variables $t = (t_1, t_2, \ldots, t_k)$ such that $k \leq n$. In addition, the transformation from m to t adheres to some criterion [52].

Data reduction methods fall into linear and non-linear methods. Some well-used linear methods include Principal Component Analysis (PCA) and Factor Analysis (FA). Non-linear methods include Principal Curves (PC), Multidimensional Scaling (MDS), and Neural Networks (NN), among other techniques. The linear methods are considered easier to implement than non-linear methods [52].

PCA is also known as singular value decomposition, the Karhunen-Loève transform, and the Hotelling transform. PCA has been applied in biology, medicine, chemistry, meteorology, and the social sciences. PCA is used specifically for data reduction. Conversely, FA is used to uncover 'latent' factors in a dataset and for exploratory research [52, 53, 54].

2.5 **Similarity**

Similarity is also known as communality, dissimilarity, distance, proximity, and resemblance [55, 56]. Similarity is the basis for classification and is defined to be the amount of resemblance between two individuals or objects based on the distinct information pertaining to the variables (i.e., features) of the individuals [57]. Similarity coefficients have been applied to several fields such as manufacturing systems, plant breeding, seed bank management, high throughput screening of chemical datasets, and determining the molecular markers of genetic relationships between individuals [58, 59, 60, 61].

One study investigated the cell formation (CF) problem in manufacturing systems. Cell formation is concerned with grouping parts or machines that are similar into product families to improve the manufacturing process. This study found similarity coefficients to be more flexible than other methods for solving the CF problem [58]. An early application of similarity coefficients to health problems was the classification of psychiatric patients. This study compared the clinical diagnosis of 46 patients to the groupings formed using a similarity coefficient. There was 70% agreement between the application of the similarity coefficient and the previous clinical diagnosis for 46 psychiatric patients [62]. Lastly, multimorbidity is the co-occurrence of two or more diseases in a single individual [63]. A study investigated the clustering of multimorbidity to determine if it improved the quality of chronic disease management and the implementation of evidence-based guidelines for patients with multiple chronic diseases. Their research utilized similarity to classify individuals with two or more chronic conditions into six clinically useful multimorbidity clusters [64].

The earliest known similarity coefficient was defined in 1901 by Jaccard who, investigated botany and plant physiology [65, 66]. There exist several other categories of techniques to determine similarity, or distance, between objects. These techniques, described in the next few sections, are geometric distances, ontological models, and feature-based models [57, 68, 69, 70, 71, 72, 73, 74, 75, 76].

2.5.1 Geometric Distances

Geometric distances assume that object features lie in a multi-dimensional geometric space, as illustrated in Figure 1. Thus, for two objects *A* and *B*, the judged

distance between these objects is defined as $\delta(A,B) = g[d[(A,B)]]$ where g is a "suitable monotonically non-decreasing function" [67, page 872].

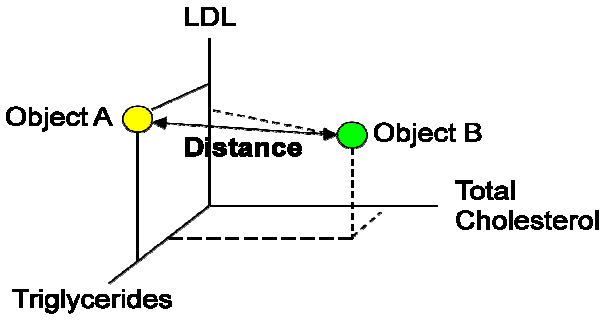

Figure 1.Conceptualization of geometric distances.

Geometric distance models have some limitations. First, it is assumed that object features are located in a multi-dimensional geometric space. This is not true however for nominal or ordinal categories and thus these types of measurements do not allow for meaningful distance computations. Secondly, some of the geometric distance computations may be adversely affected by very large differences in a single variable, even if that variable is not a highly important one. For example, if two sets of measurements were similar on all elements except for one measure that has a very large difference these two sets might be very computed to be far apart. Thus, this one medical measurement could skew the distance calculation even if the other medical measurements have very small differences.

2.5.2 Ontological Models

Medical ontologies, such as ICD-9 and SNOMED-CT®, are used to code medical measurements. Distances on an ontology are computed by traversing the ontology and counting the number of edges between two codes that are being compared. For example, as illustrated in Figure 2, if one person has a medical measurement that is represented by *node m* and another person has a medical measurement represented by *node l*, than the computed distance between the two nodes is seven. Several different examples for computing models have been developed for ontological distances and have been described in the literature [68, 69, 70, 71, 72].

Ontology based models also have limitations. First, medical record coding is susceptible to the coder and the precision used. Thus, coding errors introduced into the medical record might obfuscate path distances. Secondly, these models only consider a single measurement and no extension has been made for combining multiple concept distances together to compute an overall distance between objects.

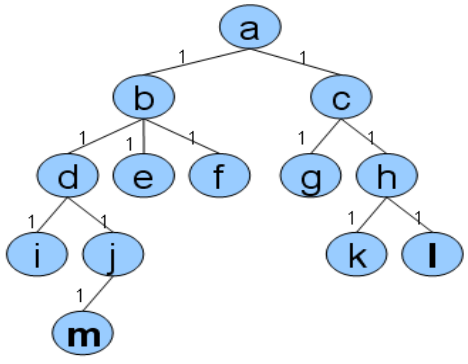

Figure 2.Computing object distance by traversing an ontology.

2.5.3 Feature-Based Models

Conceptually, feature-based models (FBM) consider the features in common and the features unique to an object. FBMs are computed using combinations of the set of common features, *a*, the unique features, *b* and *c*, and the features not contained by either object, *d*, as illustrated in Table 1 [73].

Table 1.Feature-based similarity.

		Object B		
		Contains Feature	Does not Contain Feature	Total
Object A	Contains Feature	*a*	*b*	*a* + *b*
	Does not contain feature	*c*	*d*	*c* + *d*
	Total	*a* + *c*	*b* + *d*	*a* + *b* + *c* +*d*

FBMs use binary attributes to compute similarity. As shown in Equation 1 the Jaccard coefficient (JC) computes the ratio of the number of features in common (i.e., intersection) between two objects and the total number features in common plus the number of features possessed uniquely by each of the two objects (i.e., union). The attributes of this coefficient, *a*, *b*, and *c* are illustrated in Table 1.

Jaccard Coefficient: $= \frac{a}{(a+b+c)}$ (1)

Where:
a = number of features in common
b = number of features possessed by 1st object
c = number of features possessed by 2nd object

Table 2 lists several FBMs described in the literature [74, 75]. There use does not appear to offer any substantive benefit for this application over JC. One disadvantage of JC is that it does not account for differences between objects. Thus, two objects with the same number of objects that have the same number of features in common are considered similar even though they may have several features, which are different.

Table 2. Feature-based model variations.

Similarity Coefficient	Definition
Jaccard	a / (a + b + c)
Russell and Rao	a / (a + b + c + d)
Hamann	((a + d) – (b + c)) / (a + b + c + d)
Dice	2a / (2a + b + c)
Sneath and Sokal	2(a + d) / (2(a + d) + (b + c))
Rogers and Tanimoto	(a + d) / ((a + d) + 2(b + c))
Ochiai	a / sqrt ((a + b) (a + c))
Yule	(ad – bc) / (ad +bc)
Kulczynski	(a / (a+b) + a/(a + c))/2
Pearson Phi	(ad-bc)/sqrt((a+b)(a+c)(d+b)(d+c))
Simple Matching	(a + d) / (a + b + c + d)

2.5.3.1 Tversky Feature Contrast Similarity Model

Tversky suggested using a set-theoretical approach known as the feature contrast model. The Tversky Feature Contrast Model Coefficient (TFCMC) computes similarity as a linear combination of the common and unique features of individual objects. Thus, for two objects *A* and *B*, there is a similarity function *S*; non-negative set functions *f* and *g* that define the weights of individual features and how they are combined; and two constants $\theta, \alpha, \beta \geq 0$ such that [76]:

$$S(A,B) = \emptyset g(A \cap B) - \left(\alpha f(A-B) + \beta f(B-A)\right) \quad (2)$$

In TFCMC if $\theta \neq 0$ and $\alpha = \beta = 0$, than the model reduces to a measure of the common attributes between *A* and *B*. Similarly, if $\theta = 0$, $\alpha = \beta$ TFCMC reduces to a measure of the distinction between *A* and *B*. TFCMC is illustrated in Figure 3.

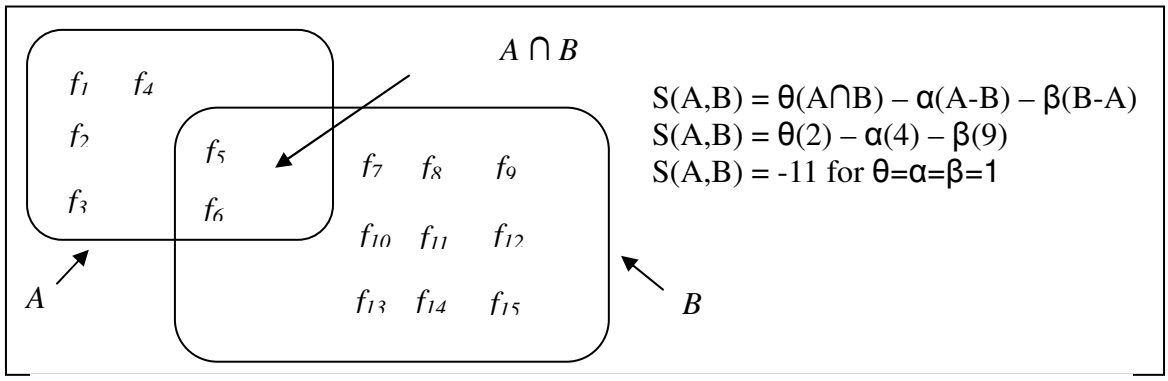

Figure 3. Tversky Feature Contrast Model Coefficient.

2.5.4 Gower's Model

In 1971, Gower proposed a different similarity coefficient that could simultaneously use variables of different measurement scales [57]. Gower computed the similarity between two objects, *A* and *B*, as follows:

$$S(A,B) = \frac{\left(\sum_{k=1}^{p} S(A,B)_k\right)}{\left(\sum_{k=1}^{p} W(A,B)_k\right)} \tag{3}$$

For nominal or ordinal data $S(A,B)_k = 1$ when the feature values are the same and 0 otherwise. For interval or ratio data $S(A,B)_k = 1 - | f_{Ak} - f_{Bk} | / R_k$ such that f_{Ak} and f_{Bk} are the values of the features for objects *A* and B; R_k equals the range for feature k across all objects (i.e., persons). In essence, this function scales the real valued features. A second feature of the Gower coefficient (GC) is the denominator, $W(A,B)_k$, which is a

type of binary weighting variable. It takes a value of 1 when the comparison between feature f_{Ak} and f_{Bk}, for objects *A* and B, is considered valid. Otherwise, it is equal to 0.

2.5.5 Feature Weighting

Weighting is defined as modifying a feature to make it more or less important in the computation of a similarity coefficient. JC and GC do not utilize weighting. TFCMC utilizes three different weighting factors, namely θ, α, and β. These factors can be used to emphasize common features or dissimilar features. For example, a larger value of θ relative to α and β will increase the effect of common features rather than dissimilar features.

Early work on weighting argued against feature weights, as there might not be an *a priori* reason to weight positive matches more heavily than non-matches. However, if there are good theoretical reasons for weighting a feature, than it can be utilized. The weight selected will depend upon the nature of the problem and a sense for the importance of matches [73, 77, 78].

2.6 Summary

This chapter contains a literature review of the societal shifts that have arisen due to the growth of broadband penetration and use of the Internet. These changes offer a mechanism to provide social support for individuals with health conditions. As individual medical measurements become available electronically and ubiquitously, the measurements can be utilized to compute similarity between individuals. Due to the type and myriad of measurements possible for grouping, techniques can be used to reduce the set of measurements to those that explain the most variation in a dataset. A related factor

is to filter the dataset to those individuals that align temporally or meet other criteria. In essence, such groupings result in SHNs. Chapter 3 describes the computational model for grouping individuals in SHNs and two example simulations that will be used to demonstrate the computational model.

Chapter 3 Social Health Network Computational Model

This dissertation proposes a computational model for creating SHNs from medical measures. The purpose of the SHN computational model is to group a *candidate* individual with others who are computed to be similar. As illustrated in Figure 4, the SHN computational model progresses linearly through filtering, data reduction, and similarity. This chapter will describe the steps in the computational model, criteria for selection of a simulation dataset, and a description of the simulation examples.

Figure 4. SHN computational model.

3.1 Filtering

The first step in computational model is to filter out individuals from the dataset to be matched with the *candidate*. Implementation of a filter is done with rules. The rules are declarative statements that in affect constrain the individuals that may be used for grouping with the *candidate*. A rule is a declarative statement as shown in equation 4. The predicates of *R,* (P_1,P_2, …,P_j), are operators used to express the logic of the filter. The operators are typically {>, <, ≠, =, ≥, ≤}. Filtering is *O(N)*, where *N* is the number of records in the dataset.

$$R: If\ \left(P_1 \wedge P_2 \ldots \wedge P_j\right)\ then\ \{Retain \mid Delete\} \quad (4)$$

Filtering is computed in two ways. First, the database is filtered according to demographic information such as age ranges, gender, and geography. Secondly, the database is filtered according to temporal criteria delineating when medical events or measurements must occur. For example, a *candidate* with elevated total cholesterol may be grouped with an individual having a similar diagnosis during the same time period. Total cholesterol measurements less than 200 are considered desirable [79]. Figure 5 illustrates this situation with a temporal overlap between the *candidate* and an individual who is a potential match based on a similar total cholesterol value.

A complex event consists of the co-occurrence or progression of multiple health factors. This introduces two temporal relationships. First, the health factors related to a condition may occur simultaneously or in some natural order. For example, a condition such as metabolic syndrome requires the co-occurrence of six different health conditions [80]. Secondly, there is the temporal relationship between individuals as described previously with the simple event.

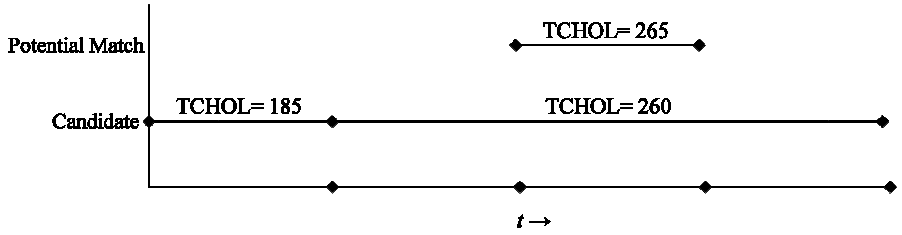

Figure 5. Simple events with temporal overlap.

Temporal ordering of measurements is characterized using a temporal interval logic such as that defined by Allen [81]. Between any two measurements or events there

are 13 relationships, as illustrated in Table 3. These relationships establish a set of temporal filtering rules to restrict individuals that can be matched with the *candidate*.

Table 3. Temporal relationships in Allen's Algebra.

Relation	Symbol	Inverse Symbol	*Example*
X *before* Y	<	>	XXX YYY
X *equal* Y	=	=	XXX YYY
X *meets* Y	m	mi	XXXYYY
X *overlaps* Y	o	oi	XXX YYY
X *during* Y	d	di	XXX YYYYYY
X *starts* Y	s	si	XXX YYYYYY
X *finishes* Y	f	fi	XXX YYYYYY

3.2 **Data Reduction**

The second step in the computational model is data reduction. Data reduction is used to improve the efficiency in computing the SHN by reducing the number of measurements used to compute similarity. PCA is used specifically for data reduction [52], has been used in health research [53], and is therefore used in this dissertation.

PCA takes independent measurements and reduces them to a smaller set of elements known as principal components (PC). The PCs are uncorrelated and represent most of the information in the original set of measurements [54]. The goal of PCA is to summarize the interrelationships for a set of measurements with a smaller set of uncorrelated orthogonal PCs that are linear combinations of the original measurements

[82]. The PCs explains the maximum amount of variance possible in the observed measurements with a smaller set of linearly transformed variables [52, 54]. If only a few principal components explain a high proportion of the variance in the observed variables and only a few of the measurements are highly correlated with these PCs, than the dataset may be reduced with a small loss of information. PCA is essentially based on the assumption that a smaller number of variables are responsible for the co-variation of the observed measurements.

Principal component analysis uses a correlation matrix from a set of data, in this case a set of observed measurements in a population. The aim of PCA is to reproduce this correlation matrix with the PCs, such that fewer PCs are needed to construct the correlation matrix. As illustrated in Table 4, each cell of a correlation matrix consists of the Pearson Product Moment Correlation (PPMC) between two different measurements. The PPMC is computed as shown in equation 5.

$$r_{xy} = \frac{\frac{\sum_1^N (x - \bar{x})(y - \bar{y})}{N - 1}}{\sqrt{var(x)var(y)}} = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}}, \quad var(x) = s_x^2 = \frac{\sum_1^N (x - \bar{x})^2}{N - 1} \qquad (5)$$

x = an individual observation of measurement X
$\bar{x}$ = mean for measurement X
y = individual observation for measurement Y
$\bar{y}$ = mean for measurement Y
N = number of paired observations for measurement X and Y

Table 4. Sample population correlation matrix for all measures.

	m_1	m_2	m_3	m_4
m_1	1.0	r_{m_1, m_2}	r_{m_1, m_3}	r_{m_1, m_4}
m_2	r_{m_2, m_1}	1.0	r_{m_2, m_3}	r_{m_2, m_4}
m_3	r_{m_3, m_1}	r_{m_3, m_2}	1.0	r_{m_3, m_4}
m_4	r_{m_4, m_1}	r_{m_4, m_2}	r_{m_4, m_3}	1.0

PPMC has a range of -1.0 to +1.0 since it is standardized by the variance of each measurement. When r_{xy} assumes a positive value, this indicates a direct relationship between two measurements such that the two measures are both increasing together; conversely, a negative value indicates an inverse relationship such that one measure is increasing while another measure is decreasing; and values close to zero indicate no relationship. The higher the absolute value of the correlation the stronger the relationship is between two types of measurements. An absolute value of r_{xy} between .50 - .69 is a moderate strength of relationship, between .70 - .89 is considered a strong relationship, and between .90 – 1.00 is considered a very strong relationship [53].

PCA finds a solution to the characteristic equation of the correlation matrix. Solving this equation results in eigenvalues, representing the variance in the measurements. Associated with the eigenvalues is an eigenvector, which contains weights, or loadings, associated with each item in the correlation matrix. The loadings represent the correlation of the item with a PC. The sum of the loadings is equal to the total variance that is explained by a PC. Similarly, since the total variance is known, the proportion of the total variance explained by a PC is equal to the sum of the loadings on a

PC divided by the total variance, where the total variance is equal to the number of measurements [53].

3.3 Similarity

The third element of the computation model is to compute the similarity between a *candidate* individual and other individuals. As discussed in Chapter 2, the geometric and ontological similarity coefficients are not suitable for computing similarity using medical measurements because they are restricted in the type of measurement data that can be used, may be adversely affected by single features, or the computational techniques have not been refined for multiple measurements. As a result, the Jaccard Coefficient (JC), the Tversky Feature Contrast Model Coefficient (TFCMC), and the Gower Coefficient (GC) are used and compared.

GC is appealing since it is computed on the raw data and can use all measurement categories directly. Conversely, a drawback of JC and TFCMC is that they operate on binary datasets. In order to accommodate the binary requirement, each measurement must be recoded to a binary value, which discerns whether the measurement associated with an individual is considered equivalent to the *candidate*.

3.3.1 Tolerance Ranges

Measurements of the same type from two different individuals can be an exact match. However, it may also be the case that these two values may differ but still be considered equivalent from a clinical perspective. For example, an individual with a blood pressure of 110/80 and another with 115/80 would both have normal blood pressure. However, if JC or TFCMC is used, than these two individuals would not be

considered a match unless some procedure is used to account for the blood pressure readings being essentially the same.

There are two approaches to this problem. The first approach is to define a percentage-based tolerance range (PBTR). A PBTR is determined by a tolerance level, τ, which is defined for the set of measurements. The tolerance level establishes a lower and upper value for each measurement. This establishes the range of values for a measurement that are considered equal to the *candidate's* measurement value. As shown in equation 5, the tolerance range for the j^{th} measurement is determined by the value of that measurement for the *candidate*, *c*, and the tolerance level τ.

$$T_j = \left(m_{cj} * (1 - \tau), m_{cj} * (1 + \tau)\right) \quad (5)$$

For example, assume a tolerance of 20% is used for body weight. If the *candidate's* body weight is 200 pounds, than the PBTR for body weight is $T_j = (180, 220)$. Thus, an individual with a body weight in this range is considered similar to the *candidate* for this feature. Conversely, someone with a body weight of 245 is not considered similar to the *candidate* for this feature.

The second approach is to set a cut point tolerance range (CPTR) for each of the medical measures. Often a medical measure has a clinically relevant cut-point, which establishes a threshold between healthy and un-healthy values. For example, the National Heart Lung and Blood Institute Obesity Education Initiative defined six classifications for body mass index (BMI). These classifications are cut points ranging from less than 18.5 kg/m^2 for underweight, 18.5 - 24.9 kg/m^2 for normal weight, to greater than or equal

to 40 kg/m^2 for extreme obesity [83]. Thus, for a *candidate* with a BMI value of 22 the CPTR, $T_j = (18.5, 24.9)$.

Both the PBTR and the CPTR approaches can be applied to interval and ratio data. For ordinal data, a tolerance range can be chosen as a range on the ordinal scale of potential values. For example, Figure 6 illustrates a question on mental health. The responses are ordered in ascending order of intensity. If the *candidate* person forming the social health network responds to this question with item response two, than they prefer to be grouped with people who have the same response or perhaps a subset of the possible categories. For instance, the tolerance set might be categories 2 and 3, represented as $T_j = (2, 3)$.

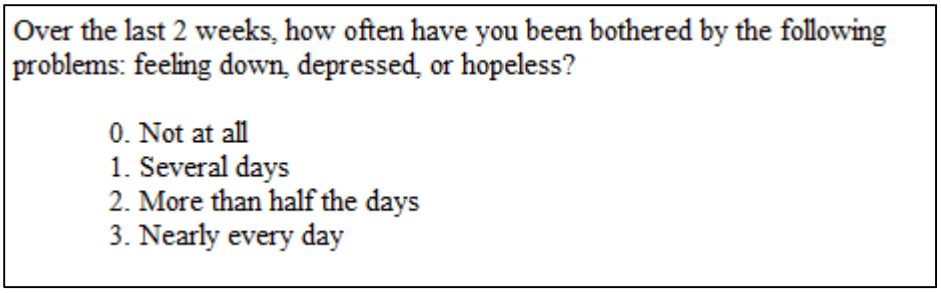

Figure 6. Ordinal measurement type.

For nominal data, there are two approaches. First, each response category of a nominal data item that the *candidate* possesses may be converted into an independent item. For example, if the nominal data item is a checklist of the prescription medications used by an individual this can be converted into 10 binary data items on the usage of each specific medication (e.g., using Lipitor / not using Lipitor, using aspirin / not using aspirin). Disuniting each element of a nominal data item in this manner has the

possibility of overwhelming the similarity computation. An alternative approach for nominal data is to associate a tolerance with this feature such as "X out of the Y nominal categories must be the same" for the binary data item to show agreement between the *candidate* and another individual. This would preclude the possibility of overwhelming the similarity computation by a disunited single nominal variable.

3.3.2 Similarity Computation

For a dataset of individuals $I = \{I_1, I_2, \ldots I_n\}$ each with a set of measurements $M = \{m_1, m_2, \ldots m_k\}$ an *NxN* similarity matrix can be computed between each pair of objects. This is $O(N^2)$. The computation can be simplified under three conditions. First, pair-wise computation of an object with itself is (i.e., on the diagonal) is not needed. Second, it is reasonable to assume that there is a symmetric relationship between two objects, thus *S(A,B) = S(B, A)*. Under these two conditions, the computation is reduced to the lower half of the matrix and thus there are $\frac{N^2 - N}{2}$ computations. Note, that the objective of this work is for a *candidate* individual to be matched with similar individuals. As such, the computation can be reduced to O(*N)* since only the similarity coefficient between the *candidate* and the other *N-1* individuals must be computed. An example similarity matrix is illustrated in Table 5. Assuming I_1 is the *candidate* than only the first column of the matrix needs to be computed.

Table 5. Example similarity matrix for five individuals.

	I_1	I_2	I_3	I_4	I_5
I_1	-	-	-	-	-
I_2	$S(I_2, I_1)$	-	-	-	-
I_3	$S(I_3, I_1)$	$S(I_3, I_2)$	-	-	-
I_4	$S(I_4, I_1)$	$S(I_4, I_2)$	$S(I_4, I_3)$	-	-
I_5	$S(I_5, I_1)$	$S(I_5, I_2)$	$S(I_5, I_3)$	$S(I_5, I_4)$	-

3.4 **Dataset Selection Criterion**

A dataset of medical measurements is necessary to simulate the computational model. Eight criteria, as outlined in Table 6, are used for the selection of a suitable simulation dataset. Meeting these criteria reduces the likelihood of bias resulting from limited scope, size, demographics, or geographic coverage. Furthermore, the dataset must be free, publicly available, and free of needing approval from institutional review board.

Table 6. Dataset selection criteria.

Criterion	Description
1. Scope	Includes laboratory, physical measures, and self-reported data.
2. Demographics	All ages and genders represented.
3. Sample Size	Large number of individuals represented.
4. Geographic Coverage	Broad geographic coverage.
5. Public Use	Publicly available dataset.
6. Institutional Review Board (IRB)	Necessity of IRB permission to use the dataset.
7. Documentation	Thorough documentation is available.
8. Recentness	Measurements are current.

3.4.1 Candidate Datasets

The United States National Institutes of Health (NIH) and the United States Centers for Disease Control and Prevention (CDC) operate clinical trials, cross-sectional studies, and surveillance activities either through intramural or extramural research. Three study datasets, from NIH and CDC, are evaluated against the dataset selection criteria in Table 6. These studies are the Framingham Heart Study (FHS), National Health and Nutrition Examination Survey (NHANES), and National Health Interview Survey (NHIS).

The FHS began in 1948 with the goal of determining the contributing factors in cardiovascular disease (CVD). There have been six different iterations of FHS. The initial cohort included 5,209 men and women ages 30 to 62 from Framingham, Massachusetts. This cohort enrolled individuals who had not yet developed symptoms of CVD, suffered a heart attack, or suffered a stroke. This study has identified significant CVD risk factors as well as relationships between phenotypic and genotypic data [84].

The NHANES is a nationally representative cross-sectional survey of the non-institutionalized population of the United States. Each year the NHANES enrolls approximately 5,000 individuals of all age ranges, genders, race, and ethnicities. Study participants participate in an interview in their home. After the home interview, a participant receives an extensive physical exam at one of three mobile examination centers. Content on the study includes cardiovascular disease, environmental exposures,

eye disease, kidney disease, obesity, physical fitness, physical functioning, and many other health indicators [85, 86].

The NHIS is a nationally representative cross-sectional survey of the non-institutionalized population of the United States. Each year the NHIS enrolls approximately 87,500 individuals of all age ranges, genders, race, and ethnicities. The study participants respond to an extensive household interview on basic health, health conditions, health utilization, chronic conditions, physician contacts, and hospitalization [87, 88].

3.4.2 Evaluation of Candidate Datasets

Table 7 illustrates the application of each criterion to the *candidate* datasets. Of the three *candidate* datasets, NHANES meets all eight criteria, NHIS meets seven of eight criteria, and Framingham meets five of eight criteria. Therefore, the SHN simulation uses the NHANES dataset.

Table 7.Application of dataset criterion to candidate datasets.

Criterion	NHANES	NHIS	Framingham
1. Scope	Yes	No	Yes
2. Demographics	Yes	Yes	Yes
3. Sample Size	Yes	Yes	Yes
4. Geographic Coverage	Yes	Yes	No
5. Public Use	Yes	Yes	No
6. Institutional Review Board	Yes	Yes	No
7. Documentation	Yes	Yes	Yes
8. Recentness	Yes	Yes	Yes

3.4.3 Missing Data

Surveys such as NHANES may have missing data for some individual's measurements. This can arise because individuals refuse to participate in the survey or because they refuse to participate in portions of the survey [89]. Missing data affects two elements of the computational model. First, it affects the data reduction piece, as PCA requires complete records for computation. However, PCA will automatically remove incomplete records to determine the variance structure.

Secondly, similarity computation needs to account for missing data. Conceptually, it is unknown if a measurement is missing because it was never observed or recorded, it is a feature that does not exist for an individual, or some other reason. The reasons for missing data are not encoded in the NHANES database and therefore it cannot be concluded that a person with a missing measurement has a value similar to the *candidate*. In this research, missing data is re-coded to *NULL* and is considered different from the *candidate's* measurement.

3.5 SHN Simulation

Publicly available data from NHANES 1999-2003 is used in the simulation. The dataset includes 31,124 individuals at birth age and older. This dataset comprises measures related to self-report questions on health, physical measures, and the results of laboratory tests [90, 91, 92]. The simulation is evaluated with two examples; the first is on type 2 diabetes and the second is on chronic kidney disease (CKD).

3.5.1 Example 1: Type 2 Diabetes

Type 2 diabetes (T2D) usually occurs in individuals who are older, obese, or lacking in physical activity. It occurs as insulin resistance such that the muscle, liver, and fat cells do not use insulin properly. As a result, the body needs additional insulin to get glucose into cells for energy [93]. T2D can be controlled with healthy eating habits, physical activity, weight loss, and for some individuals, with the use of medications [94].

A primary risk factor for type 2 diabetes is age, with those individuals over 45 being at increased risk. Some other risk factors associated with type 2 diabetes are abdominal obesity, ethnicity, HDL values lower than the normal range, history of gestational diabetes, hypertension, insulin resistance, overweight, physical inactivity, and a family history of diabetes [95, 96].

Symptoms of Type 2 diabetes include infections, blurry vision, and tingling or numbness in the hands and feet [97]. There are numerous health effects resulting from diabetes such as cataracts, glaucoma, or retinopathy; foot ulcers, amputations; hearing loss; heart disease, or hypertension; nervous system diseases; skin infections; or stroke [97, 98, 99].

Diabetes is diagnosed with a fasting plasma glucose (FPG) test, a regular plasma glucose test or an oral glucose tolerance test (OGTT). All three tests assess the level of glucose in the blood. A normal value is less than 100 mg/dL for people without diabetes. Values between 100 and 125 mg/dL is labeled as "impaired fasting glucose", while values greater 125 mg/dL are given a label of "provisional diagnosis of diabetes". A non-fasting plasma glucose test may also be used. If the value from this test is above 200 mg/dL,

than an individual may have diabetes. Confirmatory tests are usually required [100, 101, 102].

Type 2 diabetes is monitored with laboratory tests such as total cholesterol, HDL cholesterol, LDL cholesterol, triglycerides, and insulin [103]. Many of the type 2 diabetes related self-reported questions, physical measures, and laboratory tests are available in the NHANES dataset. Table 15 lists the measurements used for the diabetes simulation.

3.5.2 Example 2: Chronic Kidney Disease

The primary purpose of the kidneys is to remove and adjust waste products, water, and chemicals. The kidneys also help to control blood pressure, make red blood cells, and maintain bone health. Chronic kidney disease (CKD) is defined as damage to the kidneys from a physical injury or a disease such as diabetes. CKD is a progressive disease, which gets worse over time such that the kidney function decreases. Eventually, in the final stage the kidneys are failing and may either need dialysis to filter the blood or a kidney transplant [104].

There are numerous risk factors associated with CKD. These include illnesses such as diabetes, hypertension, cancer, hepatitis C, and urinary tract infections [104, 105, 106]. Symptoms associated with CKD are tiredness, loss of appetite, swelling in the feet or ankles, dry skin, and increased frequency of urination [107]. The major complications from CKD are dialysis and kidney transplant [104].

CKD is operationally defined as either kidney damage for at least three months or a glomerular filtration rate (GFR) less than 60 mL/min/1.73 m^2 for at least three months without kidney damage. In addition, the National Kidney Foundation® CKD delineates by five different CKD stages. Stages 1 through 4 have progressively lower GFR values and thus decreased levels of kidney function. Stage 5 is considered the worst stage as a person is in kidney failure and has a GFR value less than 15 [108].

While the primary measurement to diagnose and monitor CKD is GFR, there are numerous other laboratory tests using, blood or urine specimens, for detecting and diagnosing kidney damage. These tests include proteinuria, albuminuria, blood urea nitrogen, serum creatinine, hematocrit, hemglobin, phosphorus, and potassium [109, 110]. In addition, other risk factors for CKD are high blood pressure, cardiovascular disease, and diabetes [111].

3.5.3 Simulation Software

The computational model and software for the simulation runs on a Hewlett-Packard model p6210y personal computer with an AMD Athlon ™ II X4 620 Processor. The processor runs at 2.60 GHz and there is 6GB of installed RAM. Windows 7 64-bit operating system is installed on the personal computer. Filtering and data reduction is computed with software written in the SAS Statistical Software v9.1. Similarity is computed with software written in Java.

Chapter 4 Results

The SHN computational model is simulated with two examples, namely T2D and CKD. One dataset was prepared for each example by merging several datasets from NHANES 1999-200, NHANES 2001-2002, and NHANES 2003-2004. The T2D dataset includes 28 medical measurements while the CKD dataset includes 41 medical measurements. The simulation is examined from two different perspectives: 1) reduction in the record and feature space resulting from filtering and PCA, and 2) the correlation between the three similarity coefficients.

4.1 Filtering

T2D and CKD occur mostly in adults, thus the datasets were filtered in the first stage for individuals ages 20 and above. This resulted in the original dataset of 31,124 individuals being reduced to 49.2% of the original size. The dataset does not include temporal information due to confidentiality and disclosure concerns. Therefore, temporal matching is not utilized for this problem.

4.2 Data Reduction

The second step in the process is to conduct the principal component analysis (PCA) to reduce the scale of the feature space (i.e., medical measures). Figures 7 and 8 show the value of the principal components for T2D and CKD. For T2D the first 11 principal components (PC) are greater than 1.0. For CKD the first 12 PCs are greater than 1.0. Figures 9 and 10 show the unique and cumulative proportion that each PC contributes to the overall variance. For T2D the first 11 PCs uniquely contribute between

3.8% and 12.2% of the overall variance. For CKD the first 12 PCs uniquely contribute between 2.7% and 14.1% of the overall variance. In addition, the T2D and CKD PCs cumulatively contribute 70.7% and 78.4% to the overall variance. Thus, following the criteria for selection of PCs the first 11 T2D and first 12 CKD PCs are used for data reduction.

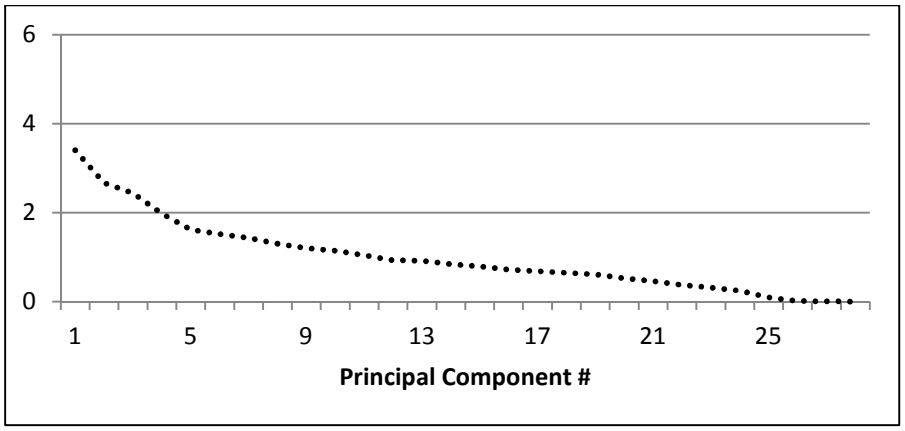

Figure 7.Type 2 diabetes principal component values.

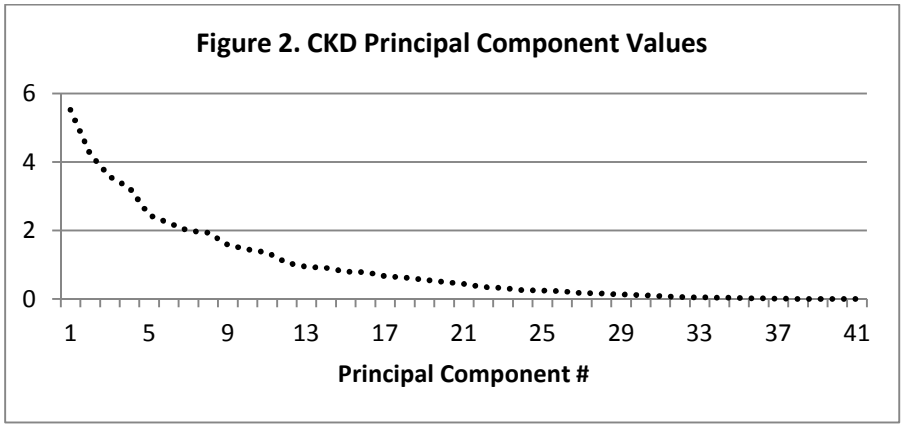

Figure 8.Chronic kidney disease principal component values.

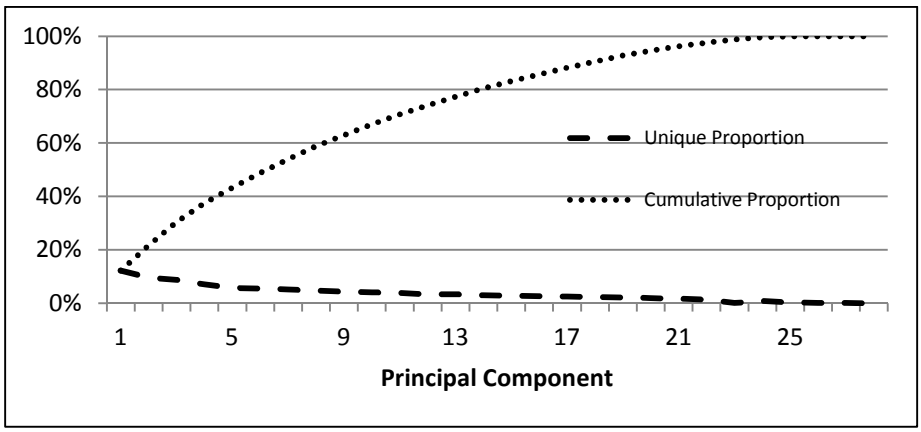

Figure 9.Type 2 diabetes principal component unique and cumulative proportions.

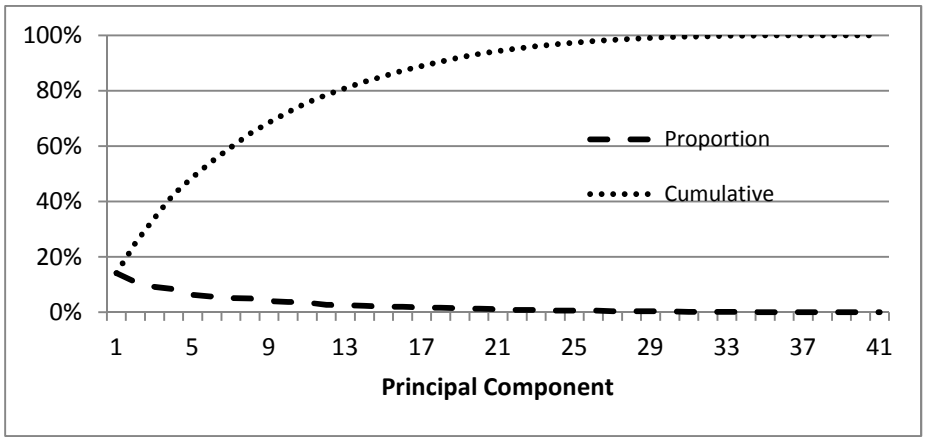

Figure 10.Chronic kidney disease principal component unique and cumulative proportions.

Figure 11 shows 18 of the original 28 measures related to T2D. Fourteen of these measures have a loading of 0.70 or greater on a PC. Four measures are loaded very close to 0.70 and are thus retained. Figure 12 shows 20 CKD measures where the absolute

value of the loading is 0.70 or greater. One measure, urine albumin (URXUMA) loaded very close to 0.70 and is considered in the analysis, as it is known to be related to CKD. In addition, although GFR has a load of 0.484 it is maintained in the next step of the computation since it is known to be an important marker for CKD. This results in 22 measurements being retained for the CKD similarity computation. Thus, PCA reduces the measurement space for T2D and CKD by 35.7% and 43.9%, respectively.

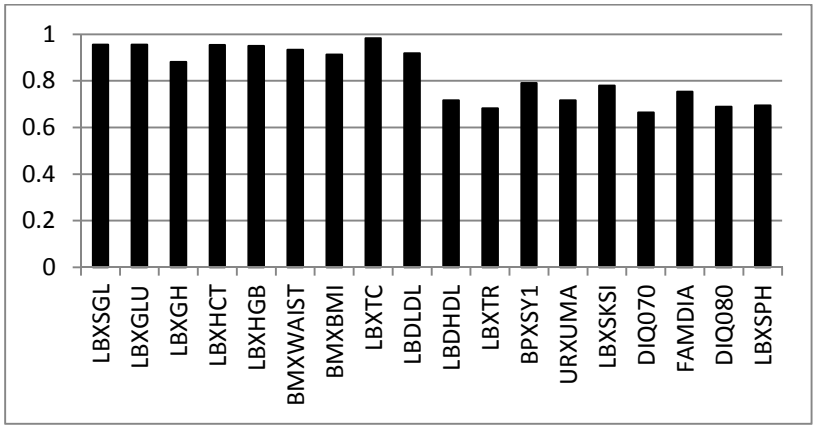

Figure 11.Type 2 diabetes loadings.

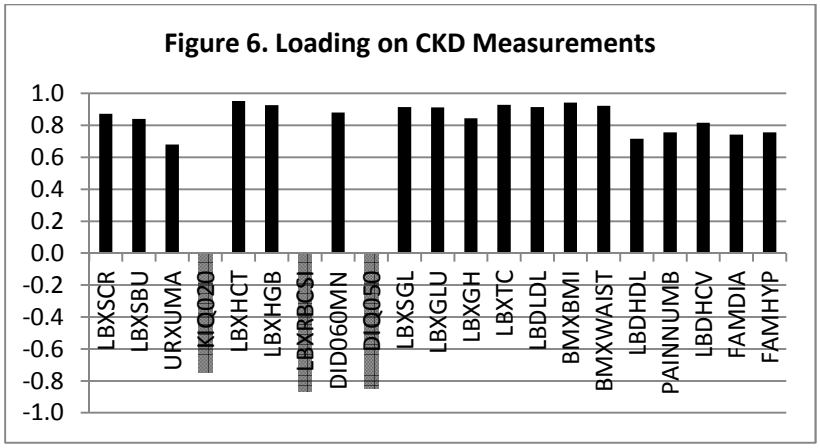

Figure 12 Chronic kidney disease loadings.

4.3 **Similarity**

Similarity coefficients are computed in the third step of the model. TFCMC, JC, and GC are used. For the TFCMC and JC the binary datasets are computed with PBTRs of 5%, 10%, 25%, and 50% as shown in Tables 8 and 9. The PBTRs for each measurement (i.e., variable) are calculated as described in equation 5. Thus, in the T2D example the *candidate* has a body mass index measurement (BMI) of 27 and a 5% PBTR of (26.201, 28.959). As the tolerance level increases the tolerance range around each measure becomes larger. For categorical data, individual categories may be selected; for ordinal data, ranges may be selected. FAMDIA is an example of a categorical measurement which can be coded with value of zero or one. A zero represents a *candidate* whom does not have a family history of diabetes. In T2D example, the FAMDIA PBTR range across all tolerance levels is essentially (0,0).

For the CPTR approach, the tolerance range used is one that is medically relevant. For example, the T2D *candidate* has a BMI measurement of 27.5 and a systolic blood pressure reading of 120. The literature describes a BMI of 27.5 to be in the overweight classification range of 25.0 – 29.9 [83]. Thus, the CPTR for BMI is (25, 29.9). Similarly, systolic blood pressure is considered normal if it is less than or equal to 120 mmHg. The *candidate's* blood pressure is exactly 120, so the CPTR can be set as less than or equal to 120 mmHg.

Tables 10 and 11 delineate the CPTRs for the T2D and CKD examples. For several measurements, the literature describes a CPTR delineating healthy and unhealthy

levels (refer to the references noted in Tables 10 and 11). Some measurements do not have a specific set of cut points for healthy and unhealthy values. Instead, these measurements have a reference range that denotes where the values of the measurement fall for a large percentage of the population. All reference ranges for these measurements are consistent with the age of the *candidate* and are inclusive of differences between males and females.

Table 8.Percentage-based tolerance ranges for type 2 diabetes measures.

		τ = 5%		τ = 10%		τ = 25%		τ = 50%	
Variable	Orig. Value	Min	Max	Min	Max	Min	Max	Min	Max
BMXBMI	27.5	26.2	28.9	24.8	30.3	20.6	34.4	13.7	41.3
BMXWAIST	101	95.9	106.0	90.9	111.1	75.7	126.2	50.5	151.5
BPXSY1	120	114	126	108	132	90	150	60	180
LBXGH	7.4	7.0	7.77	6.6	8.14	5.5	9.25	3.7	11.1
LBXGLU	178.4	169.4	187.3	160.5	196.2	133.8	223	89.2	267.6
LBXTC	167	158.6	175.3	150.3	183.7	125.2	208.7	83.5	250.5
LBDHDL	32	30.4	33.6	28.8	35.2	24	40	16	48
LBXTR	218	207.1	228.9	196.2	239.8	163.5	272.5	109	327
LBDLDL	92	87.4	96.6	82.8	101.2	69	115	46	138
URXUMA	26.4	25.0	27.7	23.7	29.0	19.8	33	13.2	39.6
LBXSGL	179	170.0	187.9	161.1	196.9	134.2	223.7	89.5	268.5
LBXSPH	2.6	2.4	2.7	2.34	2.8	1.95	3.25	1.3	3.9
LBXSKSI	3.8	3.6	4.0	3.5	4.27	2.91	4.86	1.9	5.8
LBXHGB	17	16.1	17.8	15.3	18.7	12.7	21.2	8.5	25.5
LBXHCT	51.1	48.5	53.6	45.9	56.2	38.3	63.8	25.5	76.6
DIQ080	1	1	1	1	1	1	1	1	1
DID060MN	0	0	0	0	0	0	0	0	0
FAMDIA	0	0	0	0	0	0	0	0	0

Table 9.Percentage-based tolerance ranges for chronic kidney disease.

		τ = 5%		τ = 10%		τ = 25%		τ = 50%	
Variable	Orig. Value	Min	Max	Min	Max	Min	Max	Min	Max
BMXBMI	37.4	35.5	39.3	33.7	41.1	28.0	46.8	18.7	56.1
BMXWAIST	120	114	126	108	132	90	150	60	180
DID060MN	0	0	0	0	0	0	0	0	0
DIQ050	2	2	2	2	2	2	2	2	2
FAMDIA	0	0	0	0	0	0	0	0	0
FAMHYP	0	0	0	0	0	0	0	0	0
GFR	58.0	55.1	60.9	52.2	63.8	43.5	72.5	29.0	87.1
KIQ020	2	2	2	2	2	2	2	2	2
LBDHCV	2	1.9	2.1	1.8	2.2	1.5	2.5	1	3
LBDHDL	52	49.4	54.6	46.8	57.2	39	65	26	78
LBDLDL	175	166.2	183.7	157.5	192.5	131.2	218.7	87.5	262.5
LBXGH	7.1	6.7	7.4	6.3	7.81	5.3	8.8	3.5	10.6
LBXGLU	205.9	195.6	216.1	185.3	226.4	154.4	257.3	102.9	308.8
LBXHCT	44.9	42.6	47.1	40.4	49.3	33.6	56.1	22.4	67.3
LBXHGB	14.6	13.8	15.3	13.1	16.0	10.9	18.2	7.3	21.9
LBXRBCSI	5.7	5.4	6.0	5.1	6.3	4.3	7.2	2.8	8.6
LBXSBU	23	21.8	24.1	20.7	25.3	17.2	28.7	11.5	34.5
LBXSCR	2	1.9	2.1	1.8	2.2	1.5	2.5	1	3
LBXSGL	199	189.0	208.9	179.1	218.9	149.2	248.7	99.5	298.5
LBXTC	272	258.4	285.6	244.8	299.2	204	340	136	408
PAINNUMB	1	1	1	1	1	1	1	1	1
URXUMA	31.7	30.1	33.2	28.5	34.8	23.7	39.6	15.8	47.5

Table 10.Clinically relevant cut-point tolerance ranges for diabetes measures.

Measure Variable Name	Original Value	Cut-point Tolerance Range
BMXBMI	27.58	$\geq$ 25 is overweight, thus 25 - 29.9 is used [83].
BMXWAIST	101	Higher risk category is $\geq$ 88 for women and $\geq$ 101 for men, thus $\geq$ 88 is used [112, 113].
BPXSY1	120	$\leq$ 120 normal [114]
LBXGH	7.4	> 5.2% [115]
LBXGLU	178.4	> 99 is abnormal [116]
LBXTC	167	< 200 is normal [115]
LBDHDL	32	< 35 is at risk [115]
LBXTR	218	< 250 is desirable [115]
LBDLDL	92	< 130 is desirable [115]
URXUMA	26.4	$\geq$ 20 is abnormal [117]
LBXSGL	179	> 180 is abnormal [118]
LBXSPH	2.6	Reference range is 2.8 - 4.1 for women and 2.3 - 3.7 for men. Thus, 2.3 - 4.1 is used [115].
LBXSKSI	3.89	Reference range is 3.5 - 5.1 [115].
LBXHGB	17	Reference range is 11.7-16.0 for women and 13.1-17.2 for men. Thus, 11.7 - 17.2 is used [115].
LBXHCT	51.1	Reference range is 35-47 for women and 39 - 50 for men. Thus, 35 - 50 is used [115].
DIQ080	1	1
DID060MN	0	0
FAMDIA	0	0

Table 11.Clinically relevant cut-point tolerance ranges for chronic kidney disease measures.

Measurement	Original Value	Cut-Point Tolerance Range
BMXBMI	37.450	Obese range is 30-39.9 and is thus used [83].
BMXWAIST	120	Higher risk category is ≥ 88 for females and ≥ 101 for males, thus ≥ 88 is used [112, 113].
DID060MN	0	0
DIQ050	2	2
FAMDIA	0	0
FAMHYP	0	0
GFR	58.069	30-59 is considered a moderate decrease in GFR [108]
KIQ020	2	2
LBDHCV	2	2
LBDHDL	52	>= 35 not at risk [115]
LBDLDL	175	>= 130 undesirable [115]
LBXGH	7.1	> 5.2% [115]
LBXGLU	205.900	> 99 is abnormal [116]
LBXHCT	44.900	Reference range is 36.1 - 44.3 for females and 40.7 - 50.3 for males. Thus, 36.1 - 50.3 is used [119].
LBXHGB	14.600	Reference range is 11.7 - 16.0 for females and 13.1 - 17.2 for males. Thus, 11.7 - 17.2 is used [115].
LBXRBCSI	5.760	Reference range is 4.2 - 5.4 for females and 4.7 - 6.1 for males. Thus, 4.2 - 6.1 used [120].
LBXSBU	23	>20 is abnormal [121]
LBXSCR	2	>1.4 is abnormal [122]
LBXSGL	199	> 180 is abnormal [118]
LBXTC	272	>= 200 abnormal [115]
PAINNUMB	1	1
URXUMA	31.700	≥ 20 abnormal [123]

Tables 12 and 13 and Figures 13 through 16 illustrate the descriptive statistics for the T2D and CKD examples. TFCMC can produce negative similarity scores when the majority of measurements between the *candidate* and another individual are dissimilar. In both examples, the similarity score at each percentile increases as the PBTR tolerance level increases. For example, in T2D at the 5% tolerance level and 95th percentile, the TFCMC similarity score results in a value of negative six; and at the 50% tolerance level and 95th percentile TFCMC has a similarity score of 14. Thus, higher similarity scores occur by increasing the tolerance level around a measurement. One must be careful in setting the tolerance level because high similarity scores can result between a candidate an individual who is in all likelihood dissimilar. In addition, the cut point tolerance ranges produce similarity scores at the different percentiles that fall between the 10% and 50% tolerance level in both examples.

Table 12.Type 2 diabetes descriptive statistics.

		Percentile						
Tolerance Level	Similarity Measure	Min	25	50	75	95	Max	Mean
5%	TFCMC	-18	-14	-12	-10	-6	2	-11.02
	JC	0	0.111	0.166	0.222	0.333	0.555	0.193
10%	TFCMC	-18	-10	-8	-6	-2	6	-8.365
	JC	0	0.222	0.277	0.333	0.444	0.667	0.267
25%	TFCMC	-18	-4	0	2	4	16	-1.942
	JC	0	0.388	0.5	0.555	0.611	0.944	0.446
50%	TFCMC	-18	2	6	8	14	18	4.432
	JC	0	.556	.667	.722	.889	1	.0623
Cut Points	TFCMC	-18	-6	-2	0	4	12	-2.940
	JC	0	0.333	0.444	0.5	0.611	0.833	0.418
	GC	0.056	0.673	0.717	0.841	0.881	0.963	0.689

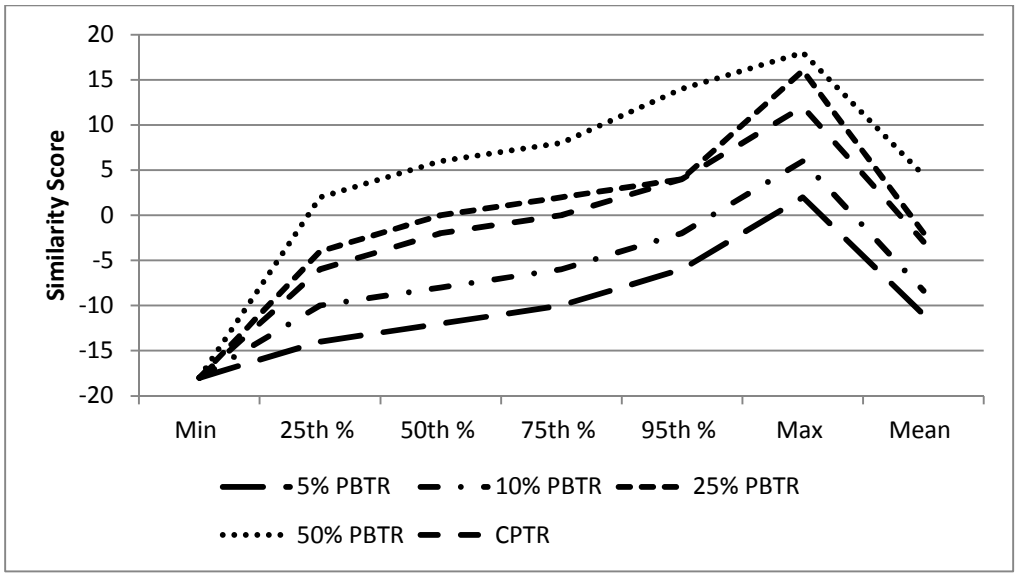

Figure 13.Type 2 diabetes TFCMC descriptive similarity statistics.

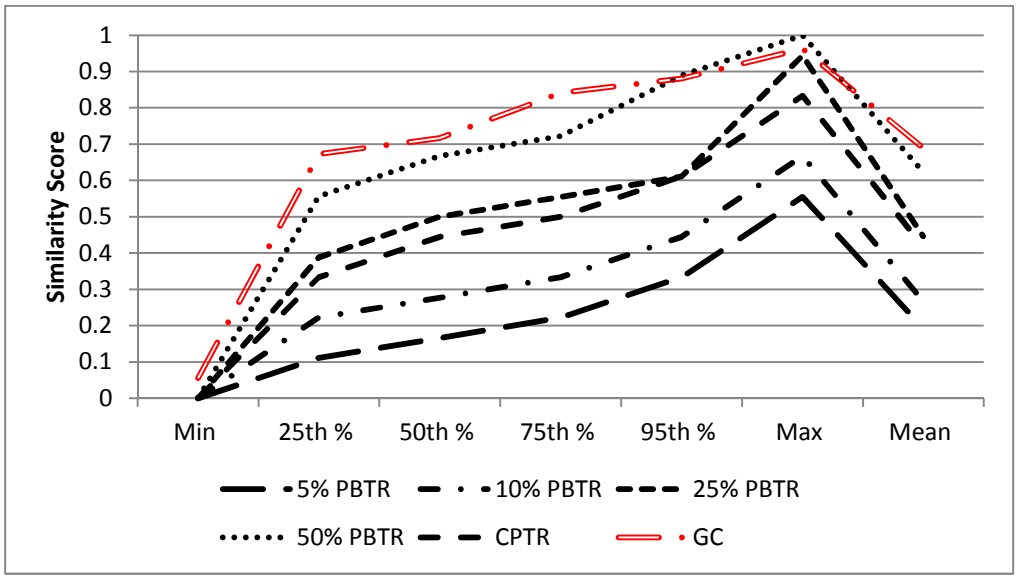

Figure 14.Type 2 diabetes JC and GC descriptive similarity statistics.

Table 13.Chronic kidney disease descriptive statistics.

Tolerance Level		Percentile						
		Min	25th	50th	75th	95th	Max	Mean
5%	TFCMC	-20	-10	-10	-8	-6	4.0	-9.458
	JC	0.045	0.272	0.272	0.3182	0.3636	0.590	0.285
10%	TFCMC	-20	-8	-6	-4	0	10	-6.129
	JC	0.045	0.318	0.363	0.409	0.5	0.727	0.360
25%	TFCMC	-20	-2	2	4	8	22	0.142
	JC	0.045	0.454	0.545	0.590	0.681	1	0.503
50%	TFCMC	-20	6	8	12	16	22	7.045
	JC	0.045	0.636	0.681	0.772	0.863	1	0.660
Cut Points	TFCMC	-20	-2	2	6	10	22	1.170
	JC	0.045	0.454	0.545	0.636	0.727	1	0.526
	GC	0.153	0.771	0.813	0.866	0.922	0.96273	0.760

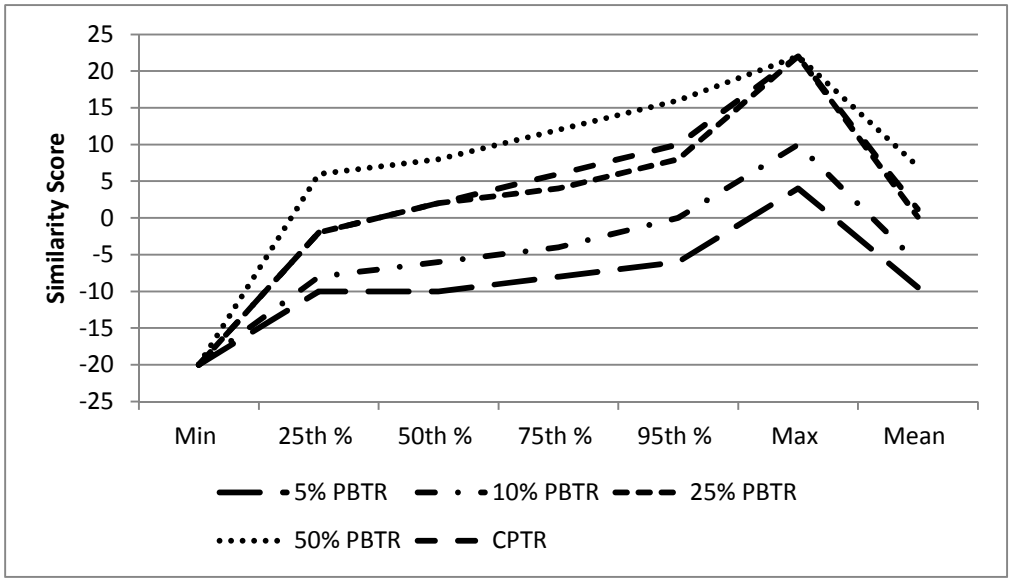

Figure 15.Chronic kidney disease TFCMC descriptive similarity statistics.

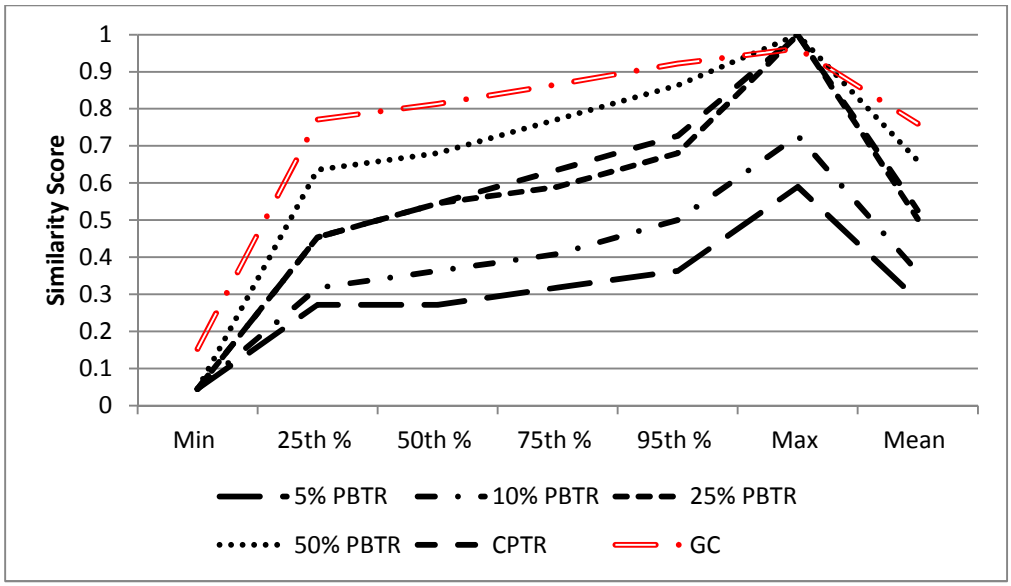

Figure 16.Chronic kidney disease JC and GC descriptive similarity statistics.

Figures 13 and 14 illustrate the correlation coefficients between each combination of similarity coefficients at PBTRs of 5%, 10%, 25%, and 50%. It also illustrates the correlation coefficient for CPTRs. Both figures show that the correlation strength between (TFCMC, GC) and (JC, GC) increases as the tolerance level increases. Note however that (TFCMC, JC) are strongly correlated at all PBTRs and the CPTR. For (TFCMC, GC), and (JC, GC) the correlation coefficient for CPTR is between the correlation coefficients at the 25% and 50% tolerance levels in the T2D example. In the CKD example, the correlation coefficient falls between the 10% and 25% PBTR tolerance levels. This may result from the implementation of TFCMC since weighting, as described in Sections 2.5.3.1 and 2.5.5, is not used in computing the similarity score.

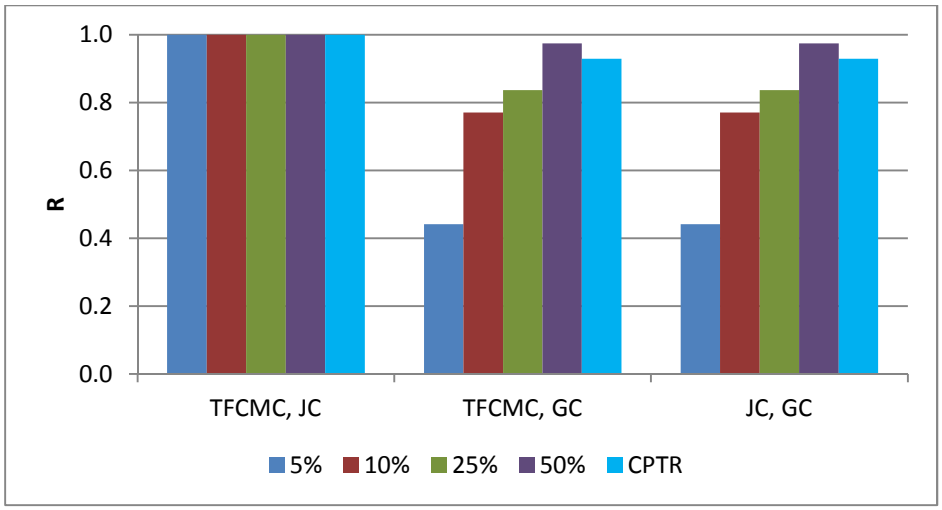

Figure 17.Correlation coefficients associated with type 2 diabetes similarity coefficients.

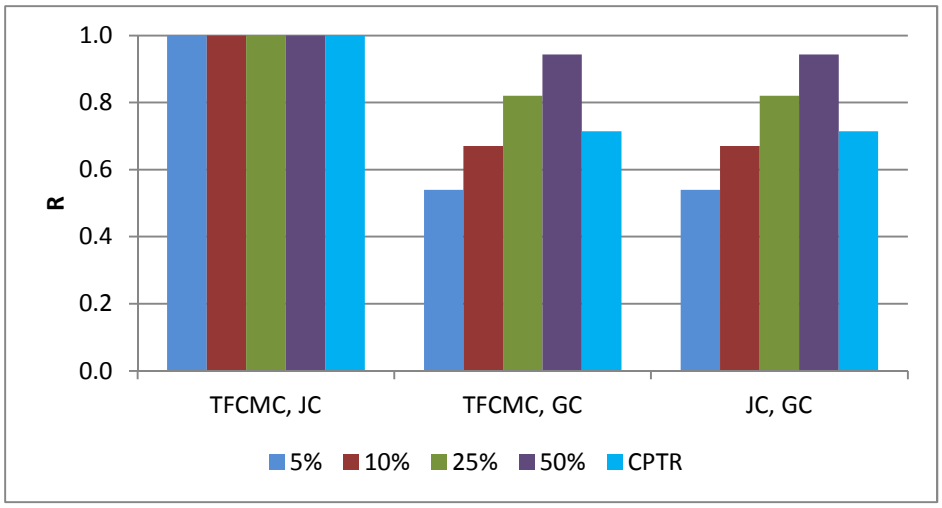

Figure 18.Correlation coefficients associated with chronic kidney disease similarity coefficients.

Chapter 5 Discussion

The purpose of this research was to build a SHN computational model for grouping individuals similar to a *candidate* individual, based on their medical measurements. The computational model consisted of filtering, data reduction, and similarity computation. Different techniques for computing similarity were compared. This research is significant because, to date, a computational model for the purpose of automatically creating SHNs has not been defined.

The simulation used a publicly available dataset and successfully demonstrated that the scale of the problem, in terms of the number of observations and feature space, can be significantly reduced using filtering and principal component analysis (PCA). In the examples chosen, filtering for a specific age range reduced the number of observations by about one-half. This will vary based on the filtering critera and the population of individuals in the dataset. In the T2D example, the feature space was reduced from 28 to 18 medical measurements using PCA. Similarly, in the CKD example the feature space was reduced from 41 to 22 medical measurements. This is a reduction of 35.7% and 46.3% for T2D and CKD.

The mean similarity scores for TFCMC, JC, and GC all increased as the PBTR increased. The increased scores imply a higher degree of likeness between the candidate and each of the other observations (i.e., individuals) in the dataset. The mean similarity score for TFCMC is low at all PBTRs and with the CPTR. It should be noted however, that a higher similarity score must be balanced against the tolerance level used with

PBTRs. Using a high tolerance level may in practice bring dissimilar individuals together into an SHN. Therefore, caution must be used in setting the tolerance level.

The strong correlation between TFCMC and JC is an unexpected finding as TFCMC lowers the similarity score due to disimilar measurements. However, JC in some sense takes dissimilar features into account as well in the denominator (refer to equation 1). Thus, the two similarity coefficients track together and are thus strongly correlated. This may not be the case however if TFCMC is weighted.

Similarity computation showed strong positive correlations between JC and GC for PBTRs of 10%, 25%, 50%, and CPTRs in both examples. For PBTRs at 5% in the CKD example, (TFCMC, GC) and (JC, GC) showed a moderate correlation. This did not hold for the T2D example where the correlation was a little bit below a moderate correlation. Using a 50% PBTR is not likely to be a good approach as it may result in ranges that cross over many cut points of healthy values for a specific medical measurement. For example, one person with an unhealthy blood pressure level might pool with people who have healthy blood pressure levels.

The similarity score results highlight two points. First, in the case of TFCMC and JC it is important to establish a threshold similirty score for grouping individuals into an SHN. This may be based on a minimum number of measurements that are considered the same. Arbitrary assignment of a threshold value should be avoided. Intuitively, one might consider that at least half the measurements should be equivalent. This would establish a TFCMC floor of zero and a JC floor of 0.50. An alternative approach is to

consider the statistical distribution of the similarity scores and choose those scores at the 95th percentile or higher. In practice, the assignment of a threshold may be based on empirical evidence. Secondly, GC scales each measurement by the range and is conceptually appealing as it is desgiend to work with mixed data types. It is true that as the GC score increases two individuals are considered more similar. However, it is not clear how the scores are to be interpreted and thus GC presents a problem. Moreover, the interpretation of the GC similarity score is not as intuituive as JC and TFCMC.

Based on the results of the simulation there are several recommendations. First, because there can be a large number of medical measurements in a medical record and the potential number of individuals to be linked in an SHN is increasing due to the use of the Internet, it is plausible that the problem space for grouping individuals in SHNs could be quite large. While it is recommended that filtering and data reduction steps be used as described in this dissertation, this step should be explored further to determine the sensitivity to specific measures or types of measures.

Secondly, different similarity coefficients were explored. All are relatively straight forward to calculate. TFCMC and JC require the additional step of producing binary datasets. Two approachesfor computing binary datasets, PBTR and CPTR, were explored. Using PBTRs alone is probably not an ideal solution as it is not clear on which tolerance level to choose. Conversely, CPTRs are logical and should be employed with one exception. For example, when a measurement is close to a cut point between different health states, it may be more appropriate to use a PBTR. Thus, when a

measurement is on the borderline between one category and another, rather than setting a hard threshold using a CPTR, a PBTR is used.

A third recommendation is that the model should be augmented to consider how new individuals enter an existing SHN. This is an important factor since each new individual attempting to enter the SHN should not trigger a complete recomputation of the SHN. One approach is to represent the SHN by an n-dimensional vector, which includes a scalar value, r, and the measurements associated with the *candidate* individual. The scalar value is the distance from the *candidate* to the individual furthest away from the *candidate* in the SHN. In essence, the scalar value is the radius of a hypersphere, which is the general n-dimensional representation of a sphere. All individuals that are part of the SHN are within or on the hypersphere. Any new individuals wanting to join the SHN must be within or on the hypersphere. Membership in the SHN requires that the distance from the candidate to a potential member to be less than the radius of the hypersphere. However, this would mean that the distance computation after the initial SHN formation is a geometric distance and would thus suffer from the limitations previously described such as the limitation of measurement types.

Alternatively, the SHN is represented by two vectors $\boldsymbol{m}$ and $\boldsymbol{t}$, as illustrated in equations 6 and 7. The vector $\boldsymbol{m}$ consists of the similarity threshold used for SHN inclusion and the measurement type and values for the *candidate's* measurements. Measurement types can be nominal, ordinal, interval, or ratio. The vector $\boldsymbol{t}$ is used for defining the tolerances around each of the *candidate's* measurements for the production

of the binary array used by JC or TFCMC. Each element of the vector is a four-element tuple. The first element of the tuple is the type of tolerance used for the specific measure. This can be PBTR, CPTR, or a hybrid between PBTR and CPTR. The second and third elements are the minimum and maximum tolerance value for the specific measure. THe fourth element is the percentage used when a PBTR is used for a specific measure.

$$\boldsymbol{m} = \left(threshold, (m_{c1}, type_1), (m_{c2}, type_2), \dots, (m_{cn}, type_n)\right) \tag{6}$$

$$t = \begin{pmatrix} \left(t_{1_{type}}, t_{1_{min}}, t_{1_{max}}, t_{1_{percent}}\right), \left(t_{2_{type}}, t_{2_{min}}, t_{2_{max}}, t_{2_{percent}}\right), \dots, \\ \left(t_{n_{type}}, t_{n_{min}}, t_{n_{max}}, t_{n_{percent}}\right) \end{pmatrix} \tag{7}$$

Although both of these approaches are plausible, the first approach may require re-computing the hypersphere each time a new individual would like to join the SHN. The second approach precludes this issue by maintaining the values from the *candidate* and only requiring that the similarity score be computed between the *candidate* and the new individual. Hence, if the similarity score is greater than the threshold stored in $\boldsymbol{m}$, than that individual can join the SHN.

This study has several limitations. First, due to confidentiality and disclosure issues the dataset does not contain temporal information. Therefore, while filtering the dataset due to timestamps would likely result in a smaller dataset the impact of this filtering is unknown. This is particularly important as data reduction in feature space using PCA may also be impacted. Furthermore, Allen's Algebra is proposed for the temporal filter. This is a comprehensive temporal logic covering all the permutations of

temporal ordering. However, future evaluation of this technique is needed to determine if it has sufficient precision for temporal information in this computational model.

A second limitation of this model is that it does not account for any procedures that would limit the possibility of discovering the indentity of an individual in a group. This is an important consideration since a 2005 poll of the U.S. population showed 67% of U.S. adults are somewhat or very concerned about the privacy of their health data and only 27% believe that they have increased rights due to Health Insurance Portability and Accountability Act (HIPAA) [124]. These concerns have historical importance beginning with the atrocities of World War II and the resulting laws enacted to protect individuals from experimentation, research, privacy breaches, and release of health data [125, 126, 127, 128, 129].

Lastly, similarity weighting has not been incorporated into the computational model as it is difficult to assign the relative importance of each feature. However, a future consideration is to adjust the similarity scores with a posterior adjustment such that the similarity score is weighted by time differentials in overall health state between individuals. Thus, the similarity score might be proportionally weighted by the inverse of the temporal distance between the health states of the candidate and other individuals.

Chapter 6 Conclusion

Social support is an important contributor to helping individuals through health issues. The ubiquitous nature of the Internet has fostered the development of many websites for individuals to form health communities. These communities are ad hoc in nature. In contrast, this paper proposes a solution that automatically forms SHNs based on a computational model that filters the data space, reduces the feature space with PCA, and computes the likeness between individuals with similarity coefficients. Conceptually, this approach can more precisely group similar individuals. The precision of the grouping depends upon the number of features used, the tolerances around the features, and the timeframe of the match.

The two examples used to simulate the computational model illustrate the effectiveness of filtering and PCA in reducing the scale of the problem. PBTRs and CPTRs are an effective way of preparing the data for binary similarity coefficients such as JC and TFCMC. An approach that uses a hybrid between PBTR and CPTR should be explored along with posterior adjustments.

References

[1] Blumenthal D. Launching HITECH. New England Journal of Medicine, vol. 362, no. 5, February 2010, pp. 382-385.

[2] Cohen S and Syme SL. (1985). Issues in the study and application of social support. In S. Cohen & S.L. Syme (Eds), Social support and health. San Diego, CA.: Academic Press.

[3] Halamka JD, Mandl KD, Tang PC. Early experiences with personal health records. Journal of the American Medical Informatics Association, vol. 15, no. 1, Jan / Feb 2008, pp. 1-7.

[4] Oster N. A Cancer Survivor's Lifeline. Wired Ventures Ltd. January 11, 1995. http://www.silcom.com/~noster/netsurftxt.html. Accessed on February 10, 2011.

[5] Pew Internet &American Life Project. Pew Research Center. http://www.pewinternet.org/About-Us/Project-History.aspx. Accessed on September 29, 2009.

[6] United States Census Bureau. American FactFinder. http://factfinder.census.gov. Accessed on April 6, 2010.

[7] Harris Interactive. Healthcare news. Number of "Cyberchondriacs" – adults going online for health information – has plateaued or declined. http://www.harrisinteractive.com/news/newsletters/healthnews/HI_HealthCareNews2008Vol8_Iss8.pdf. Vol. 8, no. 8, August 20, 2008. Accessed on April 6, 2010.

[8] Pew Internet &American Life Project. Pew Research Center. http://www.pewinternet.org/Trend-Data/Online-Activities-Total.aspx. Accessed September 29, 2009.

[9] Pew Internet &American Life Project. Pew Research Center. http://www.pewinternet.org/Reports/2009/Adults-and-Social-Network-Websites.aspx. Accessed on April 6, 2010.

[10] Cohen S and Syme SL. (1985). Issues in the study and application of social support. In S. Cohen & S.L. Syme (Eds), Social support and health. San Diego, CA.: Academic Press.

[11] House JS. Work stress and social support. Reading, MA. Addison-Wesley Company, 1981.

[12] Gottlieb BH. Using social support to protect and promote health. Journal of Primary Prevention, Fall Winter 1987, Vol. 8 (1&2), pp.49-70.

[13] Coromilas J. Physician Credentials and ICD Implantation: Certified "Electricians" Best Deal with Electrical Problems. Journal of the American Medical Association. http://jama.ama-assn.org/cgi/content/full/301/16/1713. Accessed on April 6, 2010.

[14] Dickerson SS, Flaig DM, Kennedy MC. Therapeutic connection: Help seeking on the Internet for persons with implantable cardioverter defibrillators. Heart and Lung, 2000, vol. 29, part 4; pp. 248–255.

[15] Sharf BF. Communicating Breast Cancer On-Line: Support and Empowerment on the Internet. Women & Health, 1997, vol. 26 (1), pp. 65-84.

[16] Klemm P, Hurst M, Dearholt SL, Trone SR. Cyber Solace: gender differences on internet cancer support groups. Computers in Nursing, March/April 1999, vol. 17, no. 2, pp. 65-72.

[17] Hopecube, http://www.hopecube.com/. Accessed on April 6, 2010.

[18] DailyStrength, http://www.dailystrength.org/. Accessed on April 6, 2010.

[19] MDJunction.com, http://www.mdjunction.com/. Accessed on April 6, 2010.

[20] iMedix, http://www.imedix.com/. Accessed on April 6, 2010.

[21] Krantz DH, Luce RD, Suppes P, and Tversky A. Foundations of Measurement: Volume 1, Additive and Polynomial Representations. Dover Publications, Mineola, NY, 1999.

[22] McCall RB. Fundamental Statistics for Psychology. Second Edition, Harcout Brace Jovanovich, Inc. New York. 2nd Edition, 1975. pp. 6-9.

[23] Friedman CP, Wyatt JC. Evaluation Methods in Medical Informatics. Springer-Verlag, New York, 1997, pp. 107-108.

[24] Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med, 1993, vol. 32, no. 4, pp. 281-91.

[25] National Institutes of Health, National Library of Medicine. Fact sheet: Unified Medical Language System®. http://www.nlm.nih.gov/pubs/factsheets/umls.html. Accessed on April 6, 2010.

[26] American Medical Association. Current Procedural Terminology, CPT Code / Relative Value Search. https://catalog.ama-assn.org/Catalog/cpt/cpt_search.jsp?locality=MD. Accessed on April 6, 2010.

[27] ICD-CM-9 Online. Classification of Diseases and Injuries – Tabular Index. http://icd9cm.chrisendres.com/. Accessed on April 6, 2010.

[28] American Medical Association. About CPT®. http://www.ama-assn.org/ama/pub/physician-resources/solutions-managing-your-practice/coding-billing-insurance/cpt/about-cpt.shtml. Accessed on April 6, 2010.

[29] Centers for Disease Control and Prevention, National Center for Health Statistics. International Classification of Disease, Ninth Revision, Clinical Modification (ICD-9-CM). http://www.cdc.gov/nchs/icd/icd9cm.htm. Accessed on February 6, 2011.

[30] World Health Organization. History of the development of ICD. http://www.who.int/classifications/icd/en/HistoryOfICD.pdf. Accessed on April 6, 2010.

[31] Forrey AW, McDonald CJ, DeMoor G, Huff SM, Leavelle D, Leland D, Fiers T, Charles L, Griffin B, Stalling F, Tullis A, Hutchins K, Baenziger J. Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory results. Clinical Chemistry, 1996, vol. 42, no. 1, pp. 81-90.

[32] Regenstreif Institute. Logical Observation Identifiers Names and Codes. http://loinc.org/. Accessed on April 6, 2010.

[33] International Health Terminology Standards Development Organization. SNOMED-CT Components. http://www.ihtsdo.org/snomed-ct/snomed-ct0/snomed-ct-components/. Accessed on April 6, 2010.

[34] Chute C, Cohn SP, Campbell KE, Oliver DE, Campbell JR. The content coverage of clinical classifications. Journal of the American Medical Informatics Association, 1996, vol. 3, no. 3, pp. 224-233.

[35] Richesson RL, Andrews JE, Krischer JP. Use of SNOMED CT to represent clinical research data: a semantic characterization of data items on case report forms in vasculitis research. Journal of the American Medical Informatics Association, 2006, vol. 13, pp. 536-546.

[36] Campbell JR, Carpenter P, Sneiderman C, Cohn S, Chute CG, Warren J. Phase II Evaluation of clinical coding schemes: completeness, taxonomy, mapping, definitions, and clarity. J Am Medical Info Assoc, 1997, vol. 4, pp. 238-250.

[37] Marchiando RJ, Elston MP. Automated ambulatory blood pressure monitoring: clinical utility in the family practice setting. American Family Physician, June 2003, vol. 67, no. 11, pp. 2343-2350.

[38] Bolour A, Anderson TL, Dekeyser LJ, and Wong HKT. The role of time in information processing: a survey. ACM SIGART Bulletin, 1982, vol. 80, pp. 27-50.

[39] Combi C, Shahar Y. Temporal reasoning and temporal data maintenance in medicine: issues and challenges. Computational biology in Medicine, 1997, vol.27, no. 5, pp. 353-368.

[40] Ozsoyoglu G, Snodgrass RT. Temporal and real-time databases: a survey. IEEE Transactions on Knowledge and Data Engineering, August 1995, vol. 7, no.4, pp. 513-532.

[41] Ariav G. A temporally oriented data model. ACM Transactions on Database Systems, December 1986, vol. 11, no. 4, pp. 499-527.

[42] Dyreson CE, Cook J, and Snodgrass RT. Supporting valid-time indeterminacy. ACM Transactions on database systems (TODS), March 1998, vol. 23, issue 1, pp. 1-57.

[43] Clifford J, Rao A. A simple, general structure for temporal domains. Proceedings of the conference on temporal aspects in information systems. AFCET. France: May 1987, pp. 23-30.

[44] Wiederhold G, Fries JF, Weyl S. Structured organization of clinical databases. In Proceedings of the National Computer Conference, May 19-22, 1975, vol. 44, Anaheim, CA., AFIPS Press, Reston, VA, pp. 479-485.

[45] Snodgrass R. The temporal query language TQuel. ACM Trans Database Systems, 1987, vol. 12, no. 2, pp. 247-298.

[46] Dyreson C, Grandi F, Käfer W, Kline N, Lorentzos N, Mitsopoulos Y, Montanari A, Nonen D, Peressi E, Pernici B, Roddick JF, Sarda NL, Scalas MR, Segev A, Snodgrass RT, Soo MD, Tansel A, Tiberio P, Wiederhold G. A consensus glossary of temporal database concepts. Jensen CS, Clifford J, Elmasri, Gadia SK, Hayes P, Jajodia S (editors). ACM SIGMOD Record, , March 1994, vol. 23, no. 1.

[47] Stacey M, McGregor C. Temporal abstraction in intelligent clinical data analysis: a survey. Artificial Intelligence in Medicine, 2007, vol. 39, pp. 1-24.

[48] Shahar Y, Miksch S, Johnson P. The ASGAARD Project: A Task-Specific Framework for the application and critiquing of time-oriented clinical guidelines. Artificial Intelligence in medicine, 1998, vol. 14, pp. 29-51.

[49] Aliferis CF, Cooper GF, Miller RA, Buchanan BG, Bankowitz R, Giuse N. A temporal analysis of QMR. Journal of the American Medical Informatics Association, 1996, vol. 3, pp. 79-91.

[50] Ho TB, Nguyen TD, Kawasaki S, Le SQ, Nguyen DD, Yokoi H, Takabayashi K. Mining Hepatitis Data with temporal abstraction. ACM SIGKDD '03, August 24-27, 2003, Washington, DC, pp. 369-377.

[51] Sripada, S. G., Reiter, E., Hunter, J., and Yu, J. 2003. Summarizing neonatal time series data. In Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 2 (Budapest, Hungary, April 12 - 17, 2003). European Chapter Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, pp. 167-170.

[52] Fodor IK. A Survey of Dimension Reduction Techniques. U.S. Department of Energy, Lawrence Livermore National Laboratory, UCRL-ID-148494. May, 9, 2002.

[53] Pett MA, Lackey NR, Sullivan JJ. Making Sense of Factor Analysis: The Use of Factor Analysis for Instrument Development in Health Care Research. Sage Publications, Thousand Oaks, California, 2003.

[54] Dunteman GH. Principal Component Analysis, Series: Quantitative Applications in the Social Sciences, Sage Publications, 1989, Newbury Park, CA.

[55] Tversky A and Gati I. Studies of similarity. In. E. Rosch and B. Lloyd (eds.), Cognition and Categorization (pp. 79-98). Hillsdale, NJ: Erlbaum.

[56] Corter JE and Tversky A. Extended Similarity Trees. Psychometrika, September 1986, vol. 51, no. 3, pp. 429-451.

[57] Gower JC. A general coefficient of similarity and some of its properties. Biometrics, December 1971, vol. 27, pp. 857-874.

[58] Yin Y and Yasuda K. Similarity coefficient methods applied to cell formation problem: a comparative investigation. Computers & Industrial Engineering, 2005, vol. 48,pp. 471-489.

[59] Reif, JC, Melchinger, AE, Frisch, M. Genetical and Mathematical Properties of Similarity and Dissimilarity Coefficients Applied in Plant Breeding and Seed Bank Management Crop Sci, 2005, vol. 45, pp. 1-7.

[60] Willett P. Similarity-based virtual screening using 2D fingerprints. Drug Discovery Today, December 2006, vol 11, no. 23/24, pp. 1046-1053.

[61] Kosman E., Leonard KJ. Similarity coefficients for molecular markers in studies of genetic relationships between individuals for haploid, diploid, and ploypoid species. Molecular Ecology, 2005, vol. 14, pp. 415-424.

[62] Flake RH, Satterfield JH. ACM Southeast Regional Conference: Proceedings of the 6th annual Southeastern regional meeting of the Association of Computing Machinery and national meeting of Biomedical Computing – Volume 1. Chapel Hill, North Carolina, Session: Technical session IIB, 1967, pp. 1-18.

[63] Van den Akker, Buntinx F, Metsemakers JFM, Roos S, Knottnerus JA. Multimorbidity in General Practice: Prevalence, Incidence, and Determinants of Co-Occuring Chronic and Recurrent Diseases. J. Clin Epidemiol, May 1998, vol. 51, no. 5, pp. 367-375.

[64] Cornell JE, Pugh JA, Williams JW, et al. Multimorbidity clusters: clustering binary data from multimorbidity clusters: clustering binary data from a large administrative medical database. Applied multivariate research, 2007, vol. 12, no. 3, pp. 163-182.

[65] Goodall DQ. A new similarity index based on probability. Biometrics, December 1966, pp. 882-907.

[66] Jaccard P. The distribution of the flora in the alpine zone. The New Phytologist, vol. XI, no. 2, pp. 37-50, Feb. 1912.

[67] Santini S and Jain R. Similarity Measures. IEEE Trans. Pattern Analysis and Machine Intelligence, Sept. 1999, vol. 21, no. 9, pp. 871-883.

[68] Caviedes JE and Cimino JJ. Towards the Development of a Conceptual Distance Metric for the UMLS. Journal of Biomedical Informatics, April 2004, vol. 37, no. 2, pp. 77-85.

[69] Pedersen T, Pakhomov SVS, Patwardhan S, Chute CG. Measures of Semantic Similarity and Relatedness in the Biomedical Domain. Journal of Biomedical Informatics, June 2007, vol. 40, no. 3, pp. 288-299.

[70] Leacock C, Chodrow M. Combining Local Context and WordNet Similarity for Word Sense Identification. In: Fellbaum C, Editor. WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press, 1998, pp. 265-283.

[71] Resnik P. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence. Montreal, Canada; 1995.

[72] Resnik P. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. Journal of Artificial Intelligence Research. July 1999, pp. 95-130.

[73] Alderderfer MS and Blashfield RK. Cluster Analysis, Series: Quantitative Applications in the Social Sciences. Series/Number 07-044. Newbury Park: Sage Publications, 1984.

[74] StataCorp LP. Stata 11 Help for Measure option. http://www.stata.com/help.cgi?measure_option. Accessed on April 6, 2010.

[75] Sokal R, Sneath P. Principles of Numerical Taxonomy. San Francisco: W.H.Freeman, 1963.

[76] Tversky A. Features of Similarity. Psychological Review, July 1977, vol. 84, no. 4, pp. 327 – 352.

[77] Sneath P, Sokal R. Numerical Taxonomy. 1973, San Francisco: W.H. Freeman.

[78] Romesburg HC. Cluster analysis for researchers. Lulu Press, North Carolina, 2004.

[79] National Cholesterol Education Program. Detection, Evaluation, and Treatment of High Cholesterol in Adults (Adult Treatment Panel III): Executive Summary. U.S. Department of Health and Human Services, NIH Publication No. 01-3670, May 2001, pp. 3. http://www.nhlbi.nih.gov/guidelines/cholesterol/atp3xsum.pdf. Accessed on April 6, 2010.

[80] Grundy SM, Brewer B, Cleeman JI, Smith SC, Lenfant C. Definition of Metabolic Syndrome: Report of the National Heart, Lung, and Blood Institute/American Heart Association Conference on Scientific Issues Related to Definition. Circulation, 2004: 109;433-438

[81] Allen JF. Maintaining knowledge about temporal intervals. Communications of the ACM, 1983, vol. 26, no. 1, pp. 832-843.

[82] Goddard J and Kirby A. An introduction to factor analysis. Norwich, UK: Geo Abstracts, 1976.

[83] The Practical Guide Identification, Evaluation, and Treatment of Overweight and Obesity in Adults. U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health, National Heart, Lung, and Blood Institute. NIH Publication No. 00-4084. October 2000. Available at http://www.nhlbi.nih.gov/guidelines/obesity/prctgd_c.pdf. Accessed on January 4, 2011.

[84] About the Framingham Heart Study. National Institutes of Health, National Heart Lung and Blood Institutes, and the Boston University. http://www.framinghamheartstudy.org/about/index.html. Accessed on April 6, 2010.

[85] About the National Health and Nutrition Examination Survey (NHANES). United States Centers for Disease Control and Prevention, National Center for Health Statistics. http://www.cdc.gov/nchs/nhanes/about_nhanes.htm. Accessed on April 6, 2010.

[86] National Health and Nutrition Examination Survey: 1999-2010 Survey Content. United States Centers for Disease Control and Prevention, National Center for Health. Statistics. http://www.cdc.gov/nchs/data/nhanes/survey_content_99_10.pdf. Accessed April 6, 2010.

[87] About the National Health Interview Survey. United States Centers for Disease Control and Prevention, National Center for Health Statistics. http://www.cdc.gov/nchs/nhis/about_nhis.htm. Accessed on April 6, 2010.

[88] NHIS Questionnaires, Datasets, and Related Documentation 1997 to the Present. United States Centers for Disease Control and Prevention, National Center for Health Statistics. http://www.cdc.gov/nchs/nhis/quest_data_related_1997_forward.htm. Accessed on April 6, 2010.

[89] Brick JM and Kalton G. Handling missing data in survey research. Stat Methods Med Res. September 1996, vol. 5, pp. 215-238.

[90] National Health and Nutrition Examination Survey: NHANES 1999-2000. Centers for Disease Control and Prevention. http://www.cdc.gov/nchs/nhanes/nhanes1999-2000/nhanes99_00.htm. Accessed on January 4, 2011.

[91] National Health and Nutrition Examination Survey: NHANES 2001-2002. Centers for Disease Control and Prevention. http://www.cdc.gov/nchs/nhanes/nhanes2001-2002/nhanes01_02.htm. Accessed on January 4, 2011.

[92] National Health and Nutrition Examination Survey: NHANES 2003-2004. Centers for Disease Control and Prevention. http://www.cdc.gov/nchs/nhanes/nhanes2003-2004/nhanes03_04.htm. Accessed on January 4, 2011.

[93] Diagnosis of Diabetes. National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases. http://diabetes.niddk.nih.gov/dm/pubs/diagnosis/index.htm. Accessed on January 4, 2011.

[94] National Diabetes Fact Sheet, 2007. Centers for Disease Control and Prevention. http://www.cdc.gov/diabetes/pubs/pdf/ndfs_2007.pdf. Accessed on January 4, 2011.

[95] Medline Plus: Type 2 Diabetes - Risk Factors. National Institutes of Health, National Library of Medicine. http://www.nlm.nih.gov/medlineplus/ency/article/002072.htm. Accessed on January 4, 2011.

[96] Diabetes Health Center: Risk Factors for Diabetes. WedMD. http://diabetes.webmd.com/risk-factors-for-diabetes. Accessed on January 4, 2011.

[97] Diabetes Basics: Symptoms. American Diabetes Association. http://www.diabetes.org/diabetes-basics/symptoms/. Accessed on January 4, 2011.

[98] Living with Diabetes: Complications. American Diabetes Association. http://www.diabetes.org/living-with-diabetes/complications/. Accessed on January 4, 2011.

[99] Complications of Diabetes. National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases. http://diabetes.niddk.nih.gov/complications/. Accessed on January 4, 2011.

[100] Diabetes Guide: Diabetes Testing. *Web*MD. http://diabetes.webmd.com/guide/diagnosing-type-2-diabetes. Accessed on January 4, 2011.

[101] Mayfield, J. Diagnosis and Classification of Diabetes Mellitus: New Criteria. American Family Physician. http://www.aafp.org/afp/981015ap/mayfield.html. Accessed on January 4, 2011.

[102] American Diabetes Association. Position Statement: Diagnosis and Classification of Diabetes Mellitus. Diabetes Care. Volume 27, Supplement 1, January 2004, pp. s5-s10. http://care.diabetesjournals.org/content/27/suppl_1/s5.full.pdf+html. Accessed on January 4, 2011.

[103] Diabetes. Lab Tests Online. http://www.labtestsonline.org/understanding/conditions/diabetes-6.html. Accessed on January 4, 2011.

[104] Kidney Disease Outcomes Quality Initiative Clinical Practice Guidelines for Chronic Kidney Disease: Evaluation, Classification, and Stratification. Part 1. Executive Summary. National Kidney Foundation®, 2002.

[105] U.S. Renal Data System, USRDS 2009 Annual Data Report: Atlas of Chronic Kidney Disease and End-Stage Renal Disease in the United States, National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD, 2009.

[106] Johnson CA, Levey AS, Coresh J, Levin A, Lau J, Eknoyan G. Clinical Practice Guidelines for Chronic Kidney Disease in Adults: Part I. Definition, Disease Stages, Evaluation, Treatment, Risk Factors. American Family Physician, September 1, 2004, vol. 70, no. 5, pp.869-876.

[107] Chronic Kidney Disease. National Kidney Foundation. http://www.kidney.org/kidneydisease/ckd/index.cfm#whatare. Accessed on January 4, 2011.

[108] Levey, AS, Coresh, J, Balk E, Kausz AT, Levin A, Steffes MW, Hogg, RJ, Perrone RD, Lau J, Eknoyan G. National Kidney Foundation Practice Guidelines for Chronic Kidney Disease: Evaluation, Classification, and Stratification. Annals of Internal Medicine, July 2003, vol. 139, no. 2, pg 137-147.

[109] Johnson CA, Levey AS, Coresh J, Levin A, Lau J, Eknoyan G. Clinical Practice Guidelines for Chronic Kidney Disease in Adults: Part I. Definition, Disease Stages, Evaluation, Treatment, Risk Factors. American Family Physician, September 2004, vol. 70, no. 5, pp. 869-876.

[110] Glomerular Filtration Rate (GFR). National Kidney Foundation. http://www.kidney.org/kidneydisease/ckd/knowgfr.cfm. Accessed February 6, 2011.

[111] Saydah S, Eberhardt M, Williams D, Geiss L, Rosey R. Prevalence of Chronic Kidney Disease and Associated Risk Factors --- United States, 1999--2004. Mortality and Morbidity Weekly Report, March 2, 2007, vol. 56, no. 8, 161-165.

[112] Healthy Weight - it's not a diet, it's a lifestyle!: About BMI for Adults. http://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html. Accessed on January 4, 2011.

[113] Weight-control Information Network: Weight and Waist Measurement: Tools for Adults. National Institutes of Health, National Institute of DIabetes and Digestive and Kidney Disease. http://www.win.niddk.nih.gov/publications/tools.htm#circumf. Accessed on January 4, 2011.

[114] Medline Plus: High Blood Pressure. National Institutes of Health, National Library of Medicine. http://www.nlm.nih.gov/medlineplus/highbloodpressure.html. Accessed on January 4, 2011.

[115] Tietz NW. Clinical Guide to Laboratory Tests. 3rd Edition. Edited by Norbert W. Tietz. W. B. Saunders Company, Philadelphia, 1995.

[116] Diabetes Health Center: Blood Glucose. *Web*MD. http://diabetes.webmd.com/blood-glucose?page=3. Accessed on January 4, 2011.

[117] Diabetes Health Center: Microalbumin Urine Test. *WebMD.* http://diabetes.webmd.com/microalbumin-urine-test?page=2. Accessed on January 4, 2011.

[118] Diabetes Health Center: Hyperglycemia and Diabetes. *Web*MD. http://diabetes.webmd.com/diabetes-hyperglycemia. Accessed on January 4, 2011.

[119] Medline Plus: Hematocrit. National Institutes of Health, National Library of Medicine. http://www.nlm.nih.gov/medlineplus/ency/article/003646.htm. Accessed February 6, 2011.

[120] Medline Plus: RBC Count. National Institutes of Health, National Library of Medicine. http://www.nlm.nih.gov/medlineplus/ency/article/003644.htm. Accessed February 6, 2011.

[121] Medline Plus: BUN. National Institutes of Health, National Library of Medicine. http://www.nlm.nih.gov/medlineplus/ency/article/003474.htm. Accessed February 6, 2011.

[122] Medline Plus: Creatinine - blood. National Institutes of Health, National Library of Medicine. http://www.nlm.nih.gov/medlineplus/ency/article/003475.htm. Accessed February 6, 2011.

[123] WebMD: Microalbumin Urine Test. WebMD. http://diabetes.webmd.com/microalbumin-urine-test?page=2. Accessed February 6, 2011.

[124] Forrester Research. 2005. National Consumer Health Privacy Survey 2005. http://www.chcf.org/topics/view.cfm?itemID=115694. Accessed on April 6, 2010.

[125] Nuremberg Code. National Institutes of Health, Office of Human Subjects Research. http://ohsr.od.nih.gov/guidelines/nuremberg.html. Accessed on April 6, 2010.

[126] Regulations and Ethical Guidelines: The Belmont Report, Ethical Principles and Guidelines for the protection of human subjects of research. National Institutes of Health, Office of Human Subjects Research. http://ohsr.od.nih.gov/guidelines/belmont.html. Accessed on April 6, 2010.

[127] The Privacy Act. U.S. Department of Health and Human Service. http://www.hhs.gov/foia/privacy/index.html. Accessed on April 6, 2010.

[128] United States Department of Justice. The Privacy Act of 1974. http://www.usdoj.gov/opcl/1974privacyact-overview.htm. Accessed on April 6, 2010.

[129] Department of Health and Human Services, Office for Civil Rights. HIPAA: Frequently Asked Questions, Health Information Privacy: What does the HIPPA Privacy Rule do? http://www.hhs.gov/ocr/privacy/hipaa/faq/about/187.html. Accessed on April 6, 2010.

APPENDIX A

Table 14. NHANES 1999-2003 data items that correspond to diabetes.

Measurement Area	Item #	Description of Measure	NHANES Measure Variable Name [90, 91, 92]			Notes
			1999-2000	2001-2002	2003-2004	
Demographic Data	**1**	Gender	RIAGENDR	RIAGENDR	RIAGENDR	
	2	Age	RIDAGEYR	RIDAGEYR	RIDAGEYR	
Examination Data	**3**	Systolic Blood Pressure	BPXSY1	BPXSY1	BPXSY1	
	4	Diastolic Blood Pressure	BPXDI1	BPXDI1	BPXDI1	
	5	Body Mass Index	BMXBMI	BMXBMI	BMXBMI	
	6	Waist circumference	BMXWAIST	BMXWAIST	BMXWAIST	
Questionnaire Data	**7**	How long taking insulin	DIQ060U/Q	DIQ060U/Q	DIQ060U/Q	Recoded to months on insulin which is measure variable name DID060MN
	8	Take diabetic pills to lower blood sugar	DIQ070	DIQ070	DIQ070	
	9	Diabetes affected eyes / had retinopathy	DIQ080	DIQ080	DIQ080	
	10	Ulcers / sores not healed within 4 weeks	DIA090	DIA090	DIA090	Merged into 1 data item reflecting pain / numbness / tingling
		Numbness in hands / feet past 3 months	DIQ100	DIQ100	DIQ100	

		Numbness in hands / feet or both	DIQ110	DIQ110	DIQ110	
		Pain / tingling in hands / feet past 3 months	DIQ120	DIQ120	DIQ120	
		Where was pain or tingling	DIQ130	DIQ130	DIQ130	
		Pain in either leg while walking	DIQ140	DIQ140	DIQ140	
		Pain in calf or calves	DIQ150	DIQ150	DIQ150	
	11	Mother with diabetes	MCQ260AA	MCQ260AA	MCQ260AA	Merged into 1 data item reflecting family history of diabetes
		Father with diabetes	MCQ260AB	MCQ260AB	MCQ260AB	
		maternal grandmother with diabetes	MCQ260AC	MCQ260AC	MCQ260AC	
		Paternal grandmother with diabetes	MCQ260AE	MCQ260AE	MCQ260AE	
		Maternal grandfather with diabetes	MCQ260AD	MCQ260AD	MCQ260AD	
		Paternal grandfather with diabetes	MCQ260AF	MCQ260AF	MCQ260AF	
		Brother with diabetes	MCQ260AG	MCQ260AG	MCQ260AG	
		Sister with diabetes	MCQ260AH	MCQ260AH	MCQ260AH	
		Other relative with diabetes	MCQ260AI	MCQ260AI	MCQ260AI	
	12	Mother with hypertension	MCQ260FA	MCQ260FA	MCQ260FA	Merged into 1 data item reflecting family history of hypertension
		Father with hypertension	MCQ260FB	MCQ260FB	MCQ260FB	
		maternal grandmother with hypertension	MCQ260FC	MCQ260FC	MCQ260FC	

		Paternal grandmother with hypertension	MCQ260FE	MCQ260FE	MCQ260FE	
		Maternal grandfather with hypertension	MCQ260FD	MCQ260FD	MCQ260FD	
		Paternal grandfather with hypertension	MCQ260FF	MCQ260FF	MCQ260FF	
		Brother with hypertension	MCQ260FG	MCQ260FG	MCQ260FG	
		Sister with hypertension	MCQ260FH	MCQ260FH	MCQ260FH	
		Other relative with hypertension	MCQ260FI	MCQ260FI	MCQ260FI	
	13	Told to take prescribed medicine for BP	BPQ040A	BPQ040A	BPQ040A	
Laboratory Data	**14**	Glycohemoglobin	LBXGH	LBXGH	LBXGH	
	15	High Density Lipoprotein	LBDHDL	LBDHDL	LBXHDD	
	16	Hematocrit	LBXHCT	LBXHCT	LBXHCT	
	17	Hemoglobin	LBXHGB	LBXHGB	LBXHGB	
	18	Hepatitis C	LBDHCV	LBDHCV	LBDHCV	
	19	Insulin	LBXIN	LBXIN	LBXIN	
	20	Low Density Lipoprotein	LBDLDL	LBDLDL	LBDLDL	
	21	Phosphorus	LBXSPH	LBDSPH	LBXSPH	
	22	Plasma Glucose	LBXGLU	LBXGLU	LBXGLU	
	23	Potassium	LBXSKSI	LBXSKSI	LBXSKSI	
	24	Serum Glucose	LBXSGL	LBXSGL	LBXSGL	
	25	Total Cholesterol	LBXTC	LBXTC	LBXTC	
	26	Triglyceride	LBXTR	LBXTR	LBXTR	

	27	Urine Albumin	URXUMA	URXUMA	URXUMA	
	28	White Blood Cell Count	LBXWBCSI	LBXWBCSI	LBXWBCSI	

Table 15. NHANES 1999-2003 data items that correspond to chronic kidney disease.

Measurement Area	Item #	Description of Measure	NHANES Measure Variable Name [90, 91, 92]			Notes
			1999-2000	2001-2002	2003-2004	
Demographic Data	**1**	Gender	RIAGENDR	RIAGENDR	RIAGENDR	
	2	Race / ethnicity	RIDRETH1	RIDRETH1	RIDRETH1	
	3	Age	RIDAGEYR	RIDAGEYR	RIDAGEYR	
Examination Data	**4**	Systolic Blood Pressure	BPXSY1	BPXSY1	BPXSY1	
	5	Diastolic Blood Pressure	BPXDI1	BPXDI1	BPXDI1	
	6	Body Mass Index	BMXBMI	BMXBMI	BMXBMI	
	7	Waist circumference	BMXWAIST	BMXWAIST	BMXWAIST	
Questionnaire Data	**8**	Doctor told you have diabetes	DIQ010	DIQ010	DIQ010	
	9	Taking insulin now	DIQ050	DIQ050	DIQ050	
	10	How long taking insulin	DIQ060U/Q	DIQ060U/Q	DIQ060U/Q	Recoded to months on insulin which is measure variable name DID060MN
	11	Take diabetic pills to lower blood sugar	DIQ070	DIQ070	DIQ070	
	12	Ulcers / sores not healed within 4 weeks	DIA090	DIA090	DIA090	Merged into 1 data item reflecting pain / numbness / tingling which is measurement
		Numbness in hands / feet past 3 months	DIQ100	DIQ100	DIQ100	

		Numbness in hands / feet or both	DIQ110	DIQ110	DIQ110	variable name PAINNUMB
		Pain / tingling in hands / feet past 3 months	DIQ120	DIQ120	DIQ120	
		Where was pain or tingling	DIQ130	DIQ130	DIQ130	
		Pain in either leg while walking	DIQ140	DIQ140	DIQ140	
		Pain in calf or calves	DIQ150	DIQ150	DIQ150	
	13	Ever told you had weak/ failing kidneys	KIQ020	KIQ022	KIQ022	
	14	Ever told had congestive heart failure	MCQ160B	MCQ160B	MCQ160B	Merged into 1 data item reflecting personal history of congestive heart failure, angina, heart attack, heart failure, and coronary heart disease with a measurement variable name of HRTHIST.
		Ever told had angina / angina pectoris	MCQ160D	MCQ160D	MCQ160D	
		Ever told you had heart attack	MCQ160E	MCQ160E	MCQ160E	
		Ever told you had heart failure	MCQ160B	MCQ160B	MCQ160B	
		Ever told you had coronary heart disease	MCQ160C	MCQ160C	MCQ160C	
	15	Mother had heart attack	MCQ260GA	MCQ260GA	MCQ260GA	Merged into 1 data item reflecting family history of heart attack with a measurement variable name of FAMHRTAT
		Father had heart attack	MCQ260GB	MCQ260GB	MCQ260GB	
		Maternal grandmother had heart attack	MCQ260GC	MCQ260GC	MCQ260GC	
		Paternal grandmother had heart attack	MCQ260GE	MCQ260GE	MCQ260GE	

		Maternal grandfather had heart attack	MCQ260GD	MCQ260GD	MCQ260GD	
		Paternal grandfather had heart attack	MCQ260GF	MCQ260GF	MCQ260GF	
		Brother with heart attack	MCQ260GG	MCQ260GG	MCQ260GG	
		Sister with heart attack	MCQ260GH	MCQ260GH	MCQ260GH	
		Other relative with heart attack	MCQ260GI	MCQ260GI	MCQ260GI	
	16	Taking prescription for hypertension	BPQ040A	BPQ040A	BPQ040A	
	17	Mother with diabetes	MCQ260AA	MCQ260AA	MCQ260AA	Merged into 1 data item reflecting family history of diabetes which is variable name FAMDIA.
		Father with diabetes	MCQ260AB	MCQ260AB	MCQ260AB	
		maternal grandmother with diabetes	MCQ260AC	MCQ260AC	MCQ260AC	
		Paternal grandmother with diabetes	MCQ260AE	MCQ260AE	MCQ260AE	
		Maternal grandfather with diabetes	MCQ260AD	MCQ260AD	MCQ260AD	
		Paternal grandfather with diabetes	MCQ260AF	MCQ260AF	MCQ260AF	
		Brother with diabetes	MCQ260AG	MCQ260AG	MCQ260AG	
		Sister with diabetes	MCQ260AH	MCQ260AH	MCQ260AH	
		Other relative with diabetes	MCQ260AI	MCQ260AI	MCQ260AI	
	18	Mother with hypertension	MCQ260FA	MCQ260FA	MCQ260FA	Merged into 1 data item reflecting family
		Father with hypertension	MCQ260FB	MCQ260FB	MCQ260FB	

		maternal grandmother with hypertension	MCQ260FC	MCQ260FC	MCQ260FC	history of hypertension which is measurement variable name FAMHYP.
		Paternal grandmother with hypertension	MCQ260FE	MCQ260FE	MCQ260FE	
		Maternal grandfather with hypertension	MCQ260FD	MCQ260FD	MCQ260FD	
		Paternal grandfather with hypertension	MCQ260FF	MCQ260FF	MCQ260FF	
		Brother with hypertension	MCQ260FG	MCQ260FG	MCQ260FG	
		Sister with hypertension	MCQ260FH	MCQ260FH	MCQ260FH	
		Other relative with hypertension	MCQ260FI	MCQ260FI	MCQ260FI	
Laboratory Data	**19**	Blood Urea Nitrogen (BUN)	LBXSBU	LBXSBU	LBXSBU	
	20	Glomerular Filtration Rate (GFR)	-	-	-	This is a derived measurement.
	21	Glycohemoglobin	LBXGH	LBXGH	LBXGH	
	22	High Density Lipoprotein	LBDHDL	LBDHDL	LBXHDD	
	23	Hematocrit	LBXHCT	LBXHCT	LBXHCT	
	24	Hemoglobin	LBXHGB	LBXHGB	LBXHGB	
	25	Hepatitis C	LBDHCV	LBDHCV	LBDHCV	
	26	Insulin	LBXIN	LBXIN	LBXIN	
	27	Low Density Lipoprotein	LBDLDL	LBDLDL	LBDLDL	
	28	Phosphorus	LBXSPH	LBDSPH	LBXSPH	
	29	Plasma Glucose	LBXGLU	LBXGLU	LBXGLU	
	30	Potassium	LBXSKSI	LBXSKSI	LBXSKSI	

	31	Red Blood Cell Count	LBXRBCSI	LBXRBCSI	LBXRBCSI	
	32	Serum Albumin	LBXSAL	LBXSAL	LBXSAL	
	33	Serum Creatinine	LBXSCR	LBDSCR	LBXSCR	
	34	Serum Glucose	LBXSGL	LBXSGL	LBXSGL	
	35	Total Cholesterol	LBXTC	LBXTC	LBXTC	
	36	Total Protein	LBXSTP	LBXSTP	LBXSTP	
	37	Triglyceride	LBXTR	LBXTR	LBXTR	
	38	Uric Acid	LBXSUA	LBXSUA	LBXSUA	
	39	Urine Albumin	URXUMA	URXUMA	URXUMA	
	40	Urine Creatinine	URXUCR	URXUCR	URXUCR	
	41	White Blood Cell Count	LBXWBCSI	LBXWBCSI	LBXWBCSI	

Table 16.NHANES 1999-2000, 2001-2002, 2003-2004 Data file URLs.

Measurement Area	Survey Cycle	Data File Link
Demographics Files	1999-2000	http://www.cdc.gov/nchs/nhanes/nhanes1999-2000/demo99_00.htm
	2001-2002	http://www.cdc.gov/nchs/nhanes/nhanes2001-2002/demo01_02.htm
	2003-2004	http://www.cdc.gov/nchs/nhanes/nhanes2003-2004/demo03_04.htm
Examination Files	1999-2000	http://www.cdc.gov/nchs/nhanes/nhanes1999-2000/exam99_00.htm
	2001-2002	http://www.cdc.gov/nchs/nhanes/nhanes2001-2002/exam01_02.htm
	2003-2004	http://www.cdc.gov/nchs/nhanes/nhanes2003-2004/exam03_04.htm
Laboratory Files	1999-2000	http://www.cdc.gov/nchs/nhanes/nhanes1999-2000/lab99_00.htm
	2001-2002	http://www.cdc.gov/nchs/nhanes/nhanes2001-2002/lab01_02.htm
	2003-2004	http://www.cdc.gov/nchs/nhanes/nhanes2003-2004/lab03_04.htm
Questionnaire Files	1999-2000	http://www.cdc.gov/nchs/nhanes/nhanes1999-2000/quex99_00.htm
	2001-2002	http://www.cdc.gov/nchs/nhanes/nhanes2001-2002/quex01_02.htm
	2003-2004	http://www.cdc.gov/nchs/nhanes/nhanes2003-2004/quex03_04.htm