2007-12-13

A Framework for the Creation of a Unified Electronic Medical Record Using Biometrics, Data Fusion and Belief Theory

Dwayne Christopher Leonard
University of Miami, d.leonard@umiami.edu

UNIVERSITY OF MIAMI

A FRAMEWORK FOR THE CREATION OF A UNIFIED ELECTRONIC MEDICAL RECORD USING BIOMETRICS, DATA FUSION AND BELIEF THEORY

By

Dwayne C. Leonard

A DISSERTATION

Submitted to the Faculty
of the University of Miami
in partial fulfillment of the requirements for
the degree of Doctor of Philosophy

Coral Gables, Florida

December 2007

UNIVERSITY OF MIAMI

A dissertation submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

A FRAMEWORK FOR THE CREATION OF A UNIFIED ELECTRONIC MEDICAL RECORD USING BIOMETRICS, DATA FUSION AND BELIEF THEORY

Dwayne C. Leonard

Approved:

_____	_____
Dr. Shihab S. Asfour Committee Chairman Professor and Associate Dean College of Engineering	Dr. Terri A. Scandura Dean of the Graduate School
_____	_____
Dr. Alexander P. Pons Committee Co-Chairman Associate Professor of Computer Information Systems	Dr. Murat Erkoc Assistant Professor of Industrial Engineering
_____	_____
Dr. Sohyung Cho Assistant Professor of Industrial Engineering	Dr. Moiez A. Tapia Professor of Electrical and Computer Engineering

LEONARD, DWAYNE C. (Ph.D., Industrial Engineering)

A Framework for the Creation of a Unified Electronic Medical Record Using Biometrics, Data Fusion and Belief Theory (December 2007)

Abstract of a dissertation at the University of Miami.

Dissertation supervised by Professor Shihab S. Asfour.
No. of pages in text. (152)

The technology exists for the migration of healthcare data from its archaic paper-based system to an electronic one and once in digital form, to be transported anywhere in the world in a matter of seconds. The advent of universally accessible healthcare data benefits all participants, but one of the outstanding problems that must be addressed is how to uniquely identify and link a patient to his or her specific medical data. To date, a few solutions to this problem have been proposed that are limited in their effectiveness. We propose the use of biometric technology within our FIRD framework in solving the unique association of a patient to his or her medical data distinctively. This would allow a patient to have real time access to all of his or her recorded healthcare information electronically whenever it is necessary, securely with minimal effort, greater effectiveness, and ease.

ACKNOWLEDGMENTS

I wish to express my deepest gratitude first to God, through all things are possible. My Mother who instilled in me that I could do anything I put my mind to. I would like to thank my family, friends and everyone that I have met along this journey who encouraged me both positively and negatively in my pursuit of the things in life that interested me.

I would like to express my thanks, appreciation and grateful acknowledgement to my advisor Professor Shihab S. Asfour for his supervision, guidance and support both intellectual and financial throughout my journey towards this accomplishment. His help and advice are major components of my successful completion of this work.

I would like to also express my thanks to my doctoral committee co-chair Professor Alexander Perez Pons for mentoring me throughout this process as well as his support and counsel. Thank you for starting this journey with me when it was just an idea and seeing it through to completion.

My deep appreciation and acknowledgement are also due to the members of my doctoral examination committee, Professor Sohyung Cho, Professor Murat Erkoc, Professor Moiez A. Tapia and Professor Norman G. Einspruch for their enrichment, suggestions and evaluation of my research work. Their support and assurance will always be remembered as valuable and motivating for achievement of this endeavor.

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION

Gone are the days when a patient's healthcare needs were provided by a single physician that knew all of his or her patients by first and last name. In today's healthcare a team of healthcare professionals from different disciplines and institutions are responsible for providing services for a patient's health and well being. A high degree of data integration, data interoperability and data sharing among healthcare professionals and healthcare institutions is required in ordered to deliver high quality healthcare to the patient it serves. Proper delivery of patient care is critical and is dependent on the ability to retrieve relevant information about the patient as quickly as possible. The accessibility of complete healthcare information needs to be available to everyone involved in the delivery of patient healthcare [NCVHS, 1997]. From researchers attempting to find causes, treatments and cures for diseases to the patient themselves. The ability to universally access all patient healthcare information in a timely fashion is of utmost importance.

In this dissertation a framework called the FIRD framework is proposed that utilizes a patient's biometric characteristics to uniquely associate them to his or her medical data. The framework establishes an infrastructure that will distinctively identify a patient to his or her complete electronic healthcare record (EHCR) with exact precision and accuracy. The framework's inner workings will collect records that are not properly assigned to the universal patient identifier (UPI), remove records that do not belong to the patient and correct errors and omissions within the patient's EHCR. Resulting in a final

information compilation that provides a complete healthcare history to the healthcare provider, while reducing medical errors and lower healthcare cost.

1.1 Shift to Electronic Medical Records

Medical information is now collected from the moment of conception until autopsy. Up until present day (and still very prevalent) patients received his or her healthcare information in hand written paper based notation and record keeping. Traditionally, this data is passed between healthcare locations and patients themselves through paper request of his or her medical records. The healthcare industry has operated with this antiquated method of paper-based record keeping for a very long time. However 21st century technology has created the ability to electronically store, maintain and move data across the world in a matter of seconds. This technology provides healthcare with tremendous potential to increase productivity and quality of service. Although the application of Information Technology strategies and best practice to healthcare medical records may bring vast benefits, they also provide significant difficulties [Dick, 1991; IOM, 2001]. Healthcare information systems must deal with incredible amounts of data pertaining to all areas of a person's mental and physical health. This task alone can impede efficient operations within healthcare leading to rising cost in rendering services.

In recent years there has been a groundswell for the advent of a national database system for EHCR in the United States [Safran, 2001]. According to testimony given to the United States government accountability office (GAO) the use of information technology

(IT) has enormous potential to improve the quality of healthcare and is critical in improving the performance of the U.S. healthcare system as a whole [GAO, 2006]. The president of the United States recently announced the goal is for most Americans to have an electronic healthcare record by the year 2010 [The White House, 2004] and appointed a National Health Information Technology Coordinator to oversee the nationwide implementation of interoperable health information technology. The rising cost and other issues such as declining quality of service in healthcare have led to the development of the EHCR information systems in an effort to standardize non value added essential healthcare operations. An EHCR system will facilitate access to medical data and the migration of medical data stored in antiquated (paper-based) methods to a more information exchange friendly environment.

1.2 A Universal Patient Record in Context

The implementation of a universal electronic healthcare record would affect patient care from several points of view:

Effect on the patient

- A complete, rather than partial, history of a patient's past medical care regardless of where the patient received medical treatment.
- Access to critical information during medical emergencies when a patient is unable to provide it.
- Eliminate common delays in medical treatment due to the need of previous medical history, allergies, patient family history, and current medical treatment.

- Improve dialogue between patient and physician allows for a more informative visit.
- Improves the quality of service between physicians during consultations and collaborations of medical cases. Since both physicians would have a copy of the patient's medical history in real time, regardless of demographic location.

Effect on Society as a whole

- Create a digital pipeline between healthcare delivery and healthcare research environments, leading to significant improvements in the statistical validity, efficiency and effectiveness of research.
- Provides a larger more representative data set in which to discover correlation to root causes of disease for patients as well as their offspring.
- Allows for the research community to tailor drug treatment to specific genetic variation of patient populations.

Effect on Healthcare Organizations

- Administrative tasks such as billing can be integrated and streamlined which will lower healthcare cost.
- Instrumental in the standardization of cost and reimbursement structure for healthcare services covered by local and federal government agencies. (i.e. worker's compensation)

- Would consolidate all types of records for a single patient (admitting, billing, pharmacy, inpatient, outpatient, etc.) within the internal healthcare organization's system
- Actual costs of medical care could be studied and evaluated. To make healthcare more cost-effective.
- Medical fraud could be more easily detected.
- Best practice models could be developed and enforced across healthcare organizations, ensuring a continuum of care for patients no matter where they are seen.

In 1997, the National Committee on Vital and Health Statistics (NCVHS) began to explore the concept of a national health information infrastructure (NHII), the framework that would support the appropriate and secure exchange of health information [NCVHS, 1997]. The NCVHS, which advises the Department of Health and Human Services (HHS) on national health information policy wrote, that the NHII included, "values, practices, relationships, laws, standards, systems, and applications that support all facets of individual health, health care, and public health" [NCVHS, 1997].

The following table outlines some of the necessary attributes of the functionality of standard universal electronic healthcare record data retrieval and exchange system:

To allow communication between differing healthcare institutions, the universal electronic patient record must have a commonality of information.
To support the needs of different types of healthcare organizations, the universal electronic healthcare record must also allow for diversity in the access of the record for documentation.
The universal electronic healthcare record must use agreed-upon healthcare industry data standards (Which is the current HL7 standard).
The universal electronic healthcare record may be stored anywhere and retrievable from anywhere and thus requires a common secure network shared by healthcare organizations.
The universal electronic healthcare record must be able to employ security measures to control the visibility and availability of information. Only authorized parties should be able to access a patient record in the presence of the owner (the patient) of that record.
The universal electronic healthcare record must be able to accommodate information in any language that care is given in.
The universal electronic healthcare record must be developed judiciously so upon its introduction the physician can have confidence that, for the particular patient in front of him or her, the electronic healthcare record is a complete medical history of that patient.

Table 1: Necessary Attributes of EHCR system

The absence of this information, may lead to unnecessary medical procedures or misdiagnosis resulting in medical errors and in the case of emergencies, lifesaving information may be unavailable during a critical decision making moment. The accessibility of healthcare information needs to be securely available to everyone involved in the delivery of patient healthcare. Because of this and other operational needs to streamline the processes involved in the high quality delivery of healthcare a tremendous focus on the development and implementation of the electronic healthcare record systems is at the forefront of the healthcare community. The creation of an electronic unified record of a patient's health and previous medical treatment from the

moment of conception until autopsy will allow national and eventually global secure access to pertained patient information. Once a reality, the advent of a national system for electronic healthcare records will create an infrastructure required for a central depository or a network of federated electronic medical records database for medical practitioners to securely identify the patient to that information infrastructure for proper access and or retrieval of that particular patient's medical information. Electronic healthcare records deal with the issue of moving data stored in antiquated methods (paper-based) to a more information exchange friendly environment. EHCR systems format, transport, and properly describe medical data which is essential to healthcare information exchange and storage. However a bigger problem lies in the fact that if all of the recorded healthcare data is electronic and interchangeable, who does what data belong to? How is that data verified to belong to that patient? How is that patient's identity verified? Is the person accessing the data supposed to have access? Regrettably, some of these questions regarding the security of the healthcare information system have been neglected by both the government and society. The questions that were addressed within healthcare used various security verification methods which include using biometrics. Biometrics measures were developed to ensure infant safety in the maternity ward in hospitals and other restricted areas in hospitals. These examples are encouraging in the fact that they provide a template as to the functionality of biometrics as a form of both verification and identification of authorized personnel with access to medical or healthcare data.

With the increase of legislative requirements, a heightened awareness to patient privacy, and this newly created venue of electronically access to medical information;

secure access to relevant patient information can be seen as vital to the evolution of healthcare and lowering the cost of providing that care to hospitals and doctor's offices alike. Therefore it is essential to develop a framework for a universal patient identifier that standardizes the secure access and proper identification of both the input and output of a patient's healthcare data (for diagnosis, update, or research).

A universal patient identifier framework specifically designed for the integration of healthcare information would facilitate the evolution of healthcare into the 21st century by allowing patients to authorize access to all of the his or her medical data files without barriers (paper request, transportation, etc). A universal patient identifier will also ensure that the correct and complete medical history for that patient is the medical history that the healthcare practitioner has in front of him or her. The primary objective of this research is to develop a universal patient identifier for the integration of patient's medical information that is unique, nontransferable and flexible yet secure enough that it cannot be forged or duplicated in any way.

Therefore the focal point of this dissertation is to develop and demonstrate how biometric based system can be developed as the foundation of universal patient identifier to increase the quality of healthcare and patient safety; while maintaining confidentiality, privacy and security. The result of this framework will create a system of checks and balances through a newly created FIRD™ biometric checkpoint system to describe the rules that allow the verification of a patient's identity in order to access all of the correct medical information for that patient. With access to proper and complete integrated healthcare information, researchers and physicians can analyze what treatments work in

what population because a substantial if not all of the healthcare data for that population is available for statistical analysis. This allows for the research outcome to tailor drug treatment to the specific genetic variation of patient populations.

By placing the focus on the proper acquisition of integrated healthcare information, healthcare can concentrate on the treatment and cure of patients and diseases not the cumbersome task of incorrect access and storage of healthcare data which will lead to medical errors.

CHAPTER 2 CONCEPT OF UNIVERSAL PATIENT IDENTIFICATION AND PROPOSED SOLUTIONS

2.1 Uniquely Associating a Patient to his or her Healthcare History

The EHCR system primary function is to format, transport, and properly describe medical data from its antiquated (paper-based) to a more information exchange friendly environment which is essential to improving the quality of healthcare by streamlining its information exchange and storage. According to the Centers for Disease Control's (CDC) National Center for Health Statistics as of 2005:

- Nearly one in four (23.9 percent) of physicians reported using full or partial electronic medical records (EMRs) in their office-based practice in 2005, a 31 percent increase from the 18.2 percent reported in 2001.
- Physicians in the Midwest (26.9 percent) and West (33.4 percent) were more likely to use EMRs than those in the Northeast (14.4 percent).
- Physicians in metropolitan statistical areas (nearly 24.8 percent) were more likely to use EMRs than were those in non-metropolitan areas (16.9).
- Only one in 10 (9.3 percent) physicians, however, used EMRs with all four of the basic functions (computerized orders for prescriptions, computerized orders for tests, reporting of test results, and physician notes) considered necessary for a complete EMR system.

However a bigger problem exists in the fact that if all of the recorded healthcare data is electronic and interchangeable:

- Who does what data belong to?
- How is that data verified to belong to that patient?
- How is that patient's identity verified?
- How is the data that is identified to belong to that particular patient confirmed to be all of the data for that patient from all repositories in the network?
- How is that data checked for errors?

Current systems in place now are demographic-based, where individuals are known by variables that could change over time and could lead to medical errors. A small change in data entry, and a patient could be known as John Doe or J. W. Doe or John W. Doe, and when you look at all the variables and put this information together within a multi-facility healthcare organization be it electronic or not, there still exist the possibility for patient misidentification. Upon investigation into avoidable medical errors that occur on a daily basis, many are associated with patient misidentification or lack of complete medical information.

Patient Identifiers are essential to the day to day operations of any healthcare facility. Tasks such as the proper delivery of care, administrative processes, support services, record keeping, information management, and follow-up and preventive care all rely on the proper identification of patients. The current method of patient identification involves the use of a medical record number, issued and maintained locally by the healthcare practitioner or a healthcare insurance provider. This identification number is based on an internal numbering system is specific to the issuing organization. Different

healthcare organizations use different numbering systems. This can leave patients with several different healthcare identification numbers, each issued by the healthcare organization that provided healthcare service. These proprietary numbers provide unique identification only within the healthcare organization that issued them. This type of patient identifier is in fact counterproductive since, it furthers incompatibility and is completely inadequate to support the push for a national healthcare system and the EHCR. Although the electronic healthcare record, medical data repositories and the network for data interchange would exist, the system still requires a way to uniquely identify and integrate data of the same patient from various data sources. This universal patient identifier will allow a mechanism for the consolidation of all of a patient's electronic healthcare record entries.

A universal healthcare patient Identifier framework specifically designed for the integration of healthcare information would allow a patient to authorize access to all of his or her medical healthcare data files without barriers (paper request, transportation, etc). The universal patient identifier addresses the critical aspect of combining and relating patient data to a particular patient. Therefore, we will discuss various criteria and proposals of universal patient identifiers to establish a context to evaluate our proposed biometric-centric patient identifier. In a report submitted to the Secretary of Health and Human Services [Appavu, 1997] there are four basic functions that a Unique Patient Identifier must support:

Positive identification of the individual: for delivery of care (e.g. diagnosis, treatment, blood transfusion and medication) for administrative functions (e.g. eligibility, reimbursement, billing and payment)
Identification of information: Identification to access patient information for prompt delivery of care, coordination of multi-disciplinary patient care services during current encounters and communication of orders, results, supplies, etc. Organization of patient care information into a manual medical record chart or an automated electronic medical record for both current and future use Manual and automated linkage of various clinical records pertaining to a patient from different practitioners, sites of care and times to form a lifelong view of the patient's record and facilitate continuity of care in future Aggregation of information across institutional boundaries for population- based research and planning
Support the protection of privacy and confidentiality through, accurate identification (explicit identification of patient information) and dis-identification (mask/encrypt/hide patient information).
Reduce healthcare operational cost and enhance the health status of the nation by supporting both automated and manual patient record management, access to care and information sharing.

Table 2- Universal Patient Identifier Basic Functions.

2.2 Required Components of Unique Patient Identifier

In order for a universal patient identifier to be considered for use it must be supported by adequate identification information of the individual it identifies. Such information must be current; indexed and stored properly. The identification process must include searching a master patient indexing (MPI) scheme, matching identifiers and verifying information. The identifier's scope and level of use should include the scalability to access information from any healthcare data source raging from a single healthcare provider organization to national or even global healthcare system. In order to meet these requirements a universal patient identifier would require a robust technical and administrative infrastructure. The following six (6) components are integral parts of the Universal Patient Identifier. They must work together in order for it to perform its functions and fulfill its objectives:

1. An Identifier (numeric, alphanumeric, etc.) Scheme
2. Identification Information
3. Index
4. Mechanism to hide or encrypt the Identifier
5. Technology infrastructure with the ability to; search, identify, match, encrypt, etc.
6. Administrative infrastructure including the Central Governing Authority.

Table 3- Universal Patient Identifier Components

Concerns over exactly how patients will be identified surround the adaptation of a universal patient identifier. Topics like privacy, usability, adaptation of a new way to identify a patient are being constantly debated. Is the new universal patient identifier going to be an alpha-numeric medical necessity or another nuisance number to be stored in our memory banks?

Although the issue of a universal patient identifier is deemed necessary by many, it is also considered controversial by those with concerns about privacy and security. According to healthcare and legal experts, the need for individuals to be issued a unique healthcare identifier, preferably one not tied to a Social Security number is significant to the evolution of healthcare. Some fear that a single unique identifier would make it easier for someone to access their healthcare information and use it for unintended purposes. In light of the increase of identity theft in the financial sector, some healthcare insurance carriers have implemented procedures to remove the Social Security numbers from subscribers' identifying information and instituted assigning unique healthcare identifiers that do not coincide with individual Social Security numbers.

The American Medical Informatics Association (AMIA) has always been a leading proponent of the use of the health information technology throughout the healthcare system to improve patient safety, reduce medical errors, and achieve streamlined systems that assure 'just-in-time' knowledge and service and decision support [AMIA, 1993]. They suggest that the Voluntary Health Care Identifier is just one way to enhance dissemination of Electronic Health Records and support interoperability

of those systems. The AMIA noted that the issue of creating a universal method for identifying patients must be a top priority if the nation is to reduce medical errors.

Threats to privacy are inherent in any unique identifier for individuals. Having different identifiers for the same individual across organizations is sometimes perceived to be a protective solution for individual privacy because potential linkages across data systems are impeded. The rationale that an electronic healthcare environment poses greater risks than one that relies on paper records is based on the belief the use of a single identifier across all healthcare organizations will increase the threat to an individual's privacy by facilitating unauthorized linkages of information about an individual within and across organizations.

The current healthcare environment of record keeping and usage exposes more personal identifying information to the threat of improper access then the EHCR system would if a single unique universal patient identifier were used. Typical healthcare records contain a patient's demographic information name, gender, address, phone number, birth date, SSN, healthcare provider insurance information, employer information, and other personal information. Healthcare organizations use a multitude of these data items to ensure a correct match between the records in hand and a specific patient. In reality, a medical record or transaction bearing merely a person's name and address may make the information more vulnerable to exposure to anyone who deliberately or accidentally comes in contact with it. Ironically, this use of personal information for matching people and records generates little controversy compared to controversy over the use of a

universal patient identifier for EHCR, despite the lack of security standards and privacy protections in place today.

Protection of healthcare information from inadvertent or unauthorized disclosure would become easier with a unique universal patient individual identifier that is used for healthcare, but not for other purposes. Such an identifier would be used in a similar manner to the way that HIV testing is often conducted anonymously, by assigning an individual a number that is not otherwise known or used. This number, which is used to track and retrieve the test result, cannot easily be used to identify the individual, whereas name and other identifiers could be. A test result bearing only a protected number cannot be associated easily with an individual [Appavu, 1997].

In a letter to the Secretary of Health and Human Services, five major standards development organizations and associations that are described as clinical domain experts recommended the prompt adoption of a unique individual identifier. These organizations are: American Nurses Association, Digital Imaging and Communication in Medicine, Health Level Seven (HL7), National Association of Chain Drug Stores, and National Council for Prescription Drug Programs. The reasons they cited were to reduce administrative workloads and costs, enable faster access to critical healthcare information, and increase efficiency in the exchange of electronic healthcare data. Because any national EHCR system created in the United States is going to be operating on a peer-to-peer platform from multiple data repositories as opposed to a single centralized system, such as the ones found in the U.K. or Australia. For the critical

function of proper delivery healthcare data the right data must be routed and deposit it into the right record. Which means data must be associated with a unique identifier.

2.3 Criteria for Evaluation of Candidate Identifiers

The American Society for Testing and Materials (ASTM), a standards development organization accredited by the American National Standards Institute, published the Standard Guide for Properties of a Universal Healthcare Identifier (UHID) [ASTM, 1995]. The Standard Guide provides 30 criteria which any proposed universal patient identifier must be evaluated against. The 30 criteria from the standard guide are:

Accessible (available when required).
Assignable (assign when needed by trusted authority after properly authenticated request).
Atomic (single data item--no sub-elements having meaning).
Concise (as short as possible).
Content-free (no dependence on possibly changing or unknown information).
Controllable (only trusted authorities have access to linkages between encrypted and non-encrypted identifiers).
Cost-effective (maximum functionality with minimum investment to create and maintain).
Deployable (implementable using a variety of technologies).
Disidentifiable (possible to create a number of encrypted identifiers with same properties).
Focused (created and maintained solely for supporting health care--form, usage, and policies not influenced by other activities).
Governed (has entity responsible for overseeing system--determines policies, manages trusted authorities, and ensures proper and effective support for health care).
Identifiable (possible to identify the person with such properties as name, birth date, sex, etc, by associating these with the identifier).
Incremental (capable of being phased in).
Linkable (can link health records together in both automated and manual systems).
Longevity (designed to function for foreseeable future with no known limitations).
Mappable (able to create bidirectional linkages between new and existing identifiers during incremental implementation of a new identifier).
Mergeable (can merge duplicate identifiers to apply to the same individual).
Networked (supported by a network that makes services available universally).
Permanent (never to be reassigned, even after a holder's death).
Public (meant to be an open data item--person can reveal it).
Repository-based (secure, permanent repository exists to support functions).
Retroactive (can assign identifiers to all existing individuals when system is implemented).
Secure (can encrypt and decrypt securely).
Splittable (able to assign new identifier to one or both people if the same identifier is assigned to two people).
Standard (compatible if possible with existing or emerging standards).
Unambiguous (minimizes risk of misinterpretation such as confusing number zero with letter O).
Unique (identifies one and only one individual).
Universal (able to support every living person for the foreseeable future).
Usable (processable by both manual and automated means).
Verifiable (can determine validity without additional information).

Table 4- The 30 criteria from the Standard Guide for Properties of a Universal Healthcare Identifier (UHID) [ASTM, 1995]

2.4 Potential Solutions for a Universal Patient Identifier

Based on a government study conducted to analyze the available Healthcare Universal Patient Identifier proposals that currently exist; there are six options for a Unique Universal Patient Identifier, three for Non Unique Patient Identifiers and five as alternatives to Unique Patient Identifier. The analysis used the 30 ASTM criteria for a Universal Healthcare Patient Identifier in a two step process to determine suitable Universal Patient Identifier Options [ASTM, 1995]. The first step analyzed the various issues surrounding each of the Universal Patient Identifiers proposals including its required characteristics, capabilities, components, and functions. The second step analyzed each of the available Universal Patient Identifier proposals individually. The table below lists the various options of proposed Universal Patient Identifiers:

Unique Patient Identifier Options Social Security Number: ASTM Sample UHID Implementation Patient Identification Number based on Bank Card Method Model UPI based on Personal Immutable Properties Lifetime Human Service and Treatment Record (LHSTR) Number based on the Birth Certificate Biometric Identification
Non Unique Patient Identifiers Options 1) Medical Record Number 2) Medical Record Number with a Provider Prefix 3) Cryptography-based Identifier
Alternatives to Unique Patient Identifier 1) Manual Process 2) CORBAMed Person Identification Service 3) HL7 MPI Mediation 4) FHOP's Standard Data Set as Common Patient Identifier 5) Directory Service

Table 5- Universal Patient Identifiers Proposed

CHAPTER 3 BIOMETRICS AS A POTENTIAL SOLUTION FOR THE UNIVERSAL PATIENT IDENTIFIER IN ELECTRONIC MEDICAL RECORDS

3.1 Benefits/Cost of Biometric as a Universal Identifier

The newly created venue of the electronic storage of medical information prompts a vital need for a secure method to access relevant patient information that is unique to an individual. Patient Identifiers are at the center of healthcare organization's day to day operations. National initiatives are taking place for the development of a national healthcare delivery system through the use information technology in the form of electronic healthcare records, which the patient identifier is at the center. Therefore, to improve the quality of healthcare service and patient safety, it is essential that the system have the ability to uniquely identify patients across multiple providers and access his or her information from multiple locations. We develop a framework for a universal patient identifier that standardizes the secure access and proper identification of both the input and output of a patient's healthcare data (for diagnosis, update, or research).

Of all of the proposed options for universal patient identifiers stated in the previous chapter, the use of a biometrics system can be developed as the core element of a framework for a universal patient identifier that would increase the quality of healthcare and patient safety while maintaining confidentiality, privacy and security. The framework will create a system of checks and balances through a newly created FIRD biometric checkpoint system to describe the rules that allow the verification of a

patient's identity in order to access all of the correct medical information for that patient. The medical data files of that patient will be merged into a single integrated data set using the FIRD™ as it universal patient identifier or what is commonly know in database structures as the composite primary key.

The "F.I.R.D." is an acronym for the first area of the framework. The first area of the framework is a multi layered biometric system consisting of four types of biometric identifiers. The definition of a biometric is a "measurable physiological and/or behavioral trait that can be captured and subsequently compared with another instance at the time of verification" [Ashbourn, 2000]. Basically, biometrics is the process of automatically recognizing a person using distinguishing traits [The Biometric Consortium, 2001a]. "Matching of finger prints, voice patterns, hand geometry, iris and retina scans, vein patterns and other such methodologies" are more physiological and "signature verification, keystroke patterns and other methodologies are weighted towards individual behavior" [Ashbourn, 2000]. The system verifies the identity of the person by processing biometric data, which refers to the person who asks and takes a yes/no decision (1:1 comparison).

The first process is Enrollment which is "where each new user is required to register onto the biometric system" [Hefferman, 1999]. Samples of the particular biometric will be taken to measure the characteristics of the sample, whether it's a fingerprint, iris or retinal scan, or other biometric. An "average" is taken from these readings, which will then be used to produce a template [Harris and Yen 2002)]. A template is "a very small amount of information when compared to the original

measurement of the biometric and is nothing more than a collection of numbers, which have no meaning except to the biometric system that produced them" [Hefferman, 1999]. A template, in other words, is a "run down" version of an actual reading [Harris and Yen 2002].

3.2 Multi Modal Biometrics

A multi modal biometric system is a biometric system that uses information extracted from more than one biometric identifier from a person's physical attributes for identification and/or authentication. The information extracted may be consolidated at various levels [Jain 2004].

- At the feature extraction level, the feature sets of multiple modalities are integrated and a new feature set is generated; the new feature set is then used in the matching and decision making modules of the biometric system.
- At the matching score level, the matching scores output by multiple matchers are integrated.
- At the decision level, the final decisions made by the individual systems are consolidated by employing techniques such as majority voting.

The key to Biometrics-based systems are their ability to provide automatic identification and/or authentication of individuals. Biometric authentication answers the question: "Am I who I claim to be?". On the other hand, identification answers the question: "Who am I?". The system recognizes the individual who asks by distinguishing

him from other persons whose biometric data is also stored in the database. With this knowledge, any human physiological or behavioral characteristics can serve as biometrics whether for authentication or identification, if they fulfill the following properties in the table below [Vaclav and Zdenek, 2000; Prabhakar, 2003]:

• Universal: the biometric element exists in all people. In this respect, not all biometric elements are equivalent and the rate of distinguishing one person from another is very different, according to the type of biometrics used.
• Distinctiveness: the biometric element must be distinctive to each person, i.e. no two persons should be the same in terms of the biometrics. Fingerprints have a high diversification and the probability of two persons to have the same iris is estimated as negligible. The most distinctive elements seem to be DNA, iris, retina and fingerprint.
• Permanence: the property of the biometric element remains invariant over time for each person. While some biometrics such as iris remain stable over decades, other biometrics such as a person's face or his signature's dynamics change over time. Also, fingers are frequently injured.
• Collectibility: the biometric characteristic should be quantitatively measurable and easy to collect. Although Retina scan and DNA analysis are quite intrusive, they are also the most accurate.
• Performance: accuracy, speed, and resource requirements should be satisfied, in order for a biometrics-based system to be practical.
• Acceptability: indicates the extent to which a system is harmless and accepted by the intended users, in order to be of practical value.
• Circumvention: refers to the robustness of a system against various fraudulent methods and attacks, for instance against fake fingerprints.

Table 6- Properties of Biometrics

Each of the above properties refers to key requirements stated in the report presented to the United States Department of Health on Universal Patient Identifiers. It this reasoning that leads this research to believe the use of biometrics can be considered an ideal foundation for uniquely identifying a particular patient's record. The following table lists the types of biometrics that have been created based by body parts:

Hand area

Fingerprint

- Three characteristics of fingerprints that make each unique:
  - The loop
  - The whorl
  - The arc
- Ten to 16 of these patterns to match a print, a positive ID will generally be located [Ashbourn, 2000].

Hand geometry. Scanners used for this form take a three-dimensional snapshot of the hand.

- Scanners used have a large surface area for scanning the hand and have explicit guidance and instruction
- If part of a hand is not scanned or is off the scanner's platform, it will not be a significant hindrance to the system verifying the user [Ashbourn, 2000].

Eye Area

Iris scan

- Uses the iris patterns around the pupil (called "trabecular meshwork"). This is the elastic structure of fibers, which will change positions as the pupil dilates. As a result, different rings and zones of this area are scanned and collected to form a sample.

Retinal scan

- Looks at the blood vessels of the retina ("retinal vascular patterns"), using a low intensity light source. A 360-degree scan is performed during which several different readings are collected and converted into a reference point template [Ashbourn, 2000].

DNA

- Most controversial and ironically the most accurate of all biometrics.
- DNA is isolated from a sample such as blood, saliva, semen, tissue, or hair. The genome is divided into smaller, manageable DNA fragments with restriction enzymes. The bacterial enzymes recognize four to six specific base sequences and reliably separate DNA at a specific base pair. By separating human DNA with one of these enzymes breaks the chromosomes down into millions of differently sized DNA fragments ranging from 100 to more than 10,000 base pairs long.[from how does stuff work .com]
- DNA has the smallest margin of error in correctly identifying an individual with the exception of twins in which case both individuals have the same DNA structure.

Behavioral

Voice recognition.

- Based on the physical construction of an individual's vocal chords, vocal tract, palate, teeth, sinuses and tissue within the mouth will affect the dynamics of speech" [Ashbourn, 2000].

Signature verification.

- Measure of how a person signs their name and the many dynamics of writing:
- How hard did the writer press down
- How quickly was the name signed
- How fast was each stroke made
- When the "i"'s were dotted and "t"'s crossed, and of course the overall appearance of the signature [Ashbourn, 2000].

Table 7-Types of biometric authentication (By Body part)

Biometric identification application should be regarded distinctly from a biometric authentication application. Most biometric systems integrate the two functions, identification and authentication; since the identification function is just repetitive loop of the authentication function. So a highly effective biometric identifier must do two things:

- Must correctly identify the person requesting access
- Must provide proper authentication as to the rights or permissions that the person the system identified has access to

Table 8 below shows what functional attributes each of the biometric identifiers can provide:

Biometric	Verify	ID	Accuracy	Reliability	Error Rate	Errors	False Pos.	False Neg.
Fingerprint	✓	✓	◎◎◎◎	▶▶▶	1 in 500+	dryness, dirt, age	Ext. Diff.	Ext. Diff.
Facial Recognition	✓	✗	◎◎◎	▶▶	no data	lighting, age, glasses, hair	Difficult	Easy
Hand Geometry	✓	✗	◎◎◎	▶▶	1 in 500	hand injury, age	Very Diff.	Medium
Voice Recognition	✓	✗	◎◎	▶	1 in 50	noise, weather, colds	Medium	Easy
Iris Scan	✓	✓	◎◎◎◎	▶▶▶	1 in 131,000	poor lighting	Very Diff.	Very Diff.
Retinal Scan	✓	✓	◎◎◎◎	▶▶▶	1 in 10,000,000	glasses	Ext. Diff.	Ext. Diff.
Signature Recognition	✓	✗	◎◎	▶	1 in 50	changing signatures	Medium	Easy
Keystroke Recognition	✓	✗	◎	▶	no data	hand injury, tiredness	Difficult	Easy
DNA	✓	✓	◎◎◎◎	▶▶▶	no data	none	Ext. Diff.	Ext. Diff.

Table 8- Functional Attribute of Biometric Identifiers

- The verify column indicates the biometric performs the function of verification or authentication.
- The ID column indicates the biometric performs the function of identification.
- The Accuracy column indicates how accurate the biometric is in the functions of verification/authentication and or identification (4 indicates highest accuracy).
- The reliability column indicates how reliable the biometric is in the functions of verification/authentication and or identification (4 green indicates highest reliability).
- The error rate column indicates the probability that the biometric identifier returns an error (Type 1 or Type 2).
- The error column indicates the likely causes of a returned error during either the biometric identification or biometric authentication application.
- The reliability column indicates how reliable the biometric is in the functions of verification/authentication and or identification (4 green indicates highest reliability).

Multi modal biometric systems address the problem of non-universality, since multiple traits can ensure sufficient population coverage. Furthermore, multi-biometric systems provide anti-spoofing measures by making it difficult for an intruder to simultaneously spoof the multiple biometric traits of a legitimate user. By asking the user to present a random subset of biometric traits, the system ensures a live user in this case the patient is indeed present at the point of data acquisition which is the request for his or her HIPAA protected medical records.

The choice and number of biometric traits is largely driven by the nature of the application, the overhead introduced by multiple traits (computational demands and cost, for example), and the correlation between the traits considered. In a cell phone equipped with a camera it might be easier to combine the face and voice traits of a user, while in an ATM application it might be easier to combine the fingerprint and face traits of the user. A commercial multi-biometric system called BioID (www.bioid.com) integrates the face, voice, and lip movement of an individual [Jain 2004].

3.3 Data Analysis

The motivation of this dissertation is to design a system that extracts information from what can be termed as an infinite database one must first understand data itself. The term “data", is a complex concept and not easy to universally define. Often data is confused and interchanged with information. Data is any and everything that can be processed. Where information is a pattern recognize within data. That pattern describes a set of objects or patterns in data that can be processed by a computer. The objects are

assumed to have some commonalities, so that the same systematic procedure can be applied to all the objects to generate the description.

3.3.1 Types of Data

Data can be classified into different types. Most often, an object is represented by the results of measurement of its various properties. A measurement result is called “a feature" in pattern recognition or “a variable" in statistics.

A third possibility to represent an object is by discrete structures, such as parse trees, ranked lists, or general graphs. Objects such as chemical structures, web pages with hyperlinks, DNA sequences, computer programs, or customer preference for certain products have a natural discrete structure representation. Graph-related representations have also been used in various computer vision tasks. For example both object recognition [W.-Y. Kim and A.C. Kak, 1991] and shape-from-shading [A. Robles-Kelly and E.R. Hancock, 2004] used graphical representations of objects. Representing structural objects using a vector of attributes can discard important information on the relationship between different parts of the objects. On the other hand, coming up with the appropriate dissimilarity or similarity measure for such objects is often difficult. New algorithms that can handle discrete structure directly have been developed. An example of this uses a kernel function (diffusion kernel) defined on different vertices in a graph [R. Kondor and J. Lafferty, 2002]. This leads to improved classification performance for categorical data. Learning with structural data is sometimes called “learning with relational data".

3.3.2 Types of Data Features

Even within the feature vector representation, descriptions of an object can be classified into different types. A feature is essentially a measurement, and the “scale of measurement" [Stevens, 1946] can be used to classify features into different categories. They are:

Type	Feature	Example
Nominal	Discrete Unordered Measurement	Number of Apples or Oranges
Ordinal	Discrete Unordered categorical measurement	Conservative, Liberal, Moderate Political views
Interval	continuous no absolute zero can be negative	Temperature in Fahrenheit.
Ratio	Continuous with absolute zero positive	Length/Width, Weight

Table 9- Data Types and Features

The classification scheme does have drawbacks some measurements may not fit well into any of the categories listed or be misclassified. Some examples are the following types of measurements:

- **College Grade Level:** ordered labels such as: Freshmen, Sophomore, Junior, and Senior.
- **Rankings:** starting from 1, which depending on the case may be the largest or the smallest.
- **Counted fractions:** that is bounded by zero and one, which includes percentage, for example.
- **Counts:** non-negative integers.

- **Amounts:** non-negative real numbers.
- **Balances:** unbounded, positive, or negative values.

Most would agree that these six types of data are different, yet all but the third and the last would be considered “ordinal" in the scheme by Stevens. This consideration of different types of features can aid in the design of appropriate algorithms for handling different types of data arising from different domains [Stevens, 1946].

3.4 Data Clustering

The process of grouping together things described by different kinds of data is very important. Records in general can consist of information recorded in many different data types. We have a great need to differentially group records containing this diverse data. In some sense, this can be considered as “data fusion.” Data fusion is a critical necessity for medicine, military sensor integration, and studies of the human population. Clustering data records offers information. This information includes:

- Discovery of groups of records that contain data that is similar to each other
- Discovery of one or more records that are outliers, records which of data that is largely dissimilar to the data in other records. (85 year old woman who purchases a motorcycle is a great indicator of fraudulent use of the credit card.)
- A description of similar records, (Online store suggestions of what other products were purchases by customers who buy a particular product)

Note that the two above examples rely on the fact that clustering discovers the distribution of the data. This allows for the purpose of clustering to be defined by the user and their information needs.

Consider a set of data containing medical information in regards to treatment testing of patients having a particular disease and who were given different treatments. The patient records may consist of a combination of different kinds of data with different data types: qualitative data values such as blood type, and quantitative data values such as age and weight. We use classification on the data set to discover information such as which type of patient responds to a particular treatment. Clustering is classification's unsupervised counterpart offering more potential information. Clustering does not group according to one attribute as classification does. This means that there is potential information offered by clustering that is not offered by classification.

The notion of clustering is closely related to classification and is also used in our own learning of concepts in the real world. Clustering is the grouping of objects into clusters such that the similarity among objects within the same cluster is maximized (intra-cluster similarity) and the similarity between objects in different clusters (intercluster similarity) is minimized [Everitt, 1993] [Jain, 1988].

3.4.1 Types of Data Analysis

The analysis to be performed on the data can also be classified into different types. The analysis can be exploratory or descriptive, meaning that the investigator does not have a specific goal and only wants to understand the general characteristics or

structure of the data. Or the analysis can be confirmatory or inferential, meaning that the investigator wants to confirm the validity of a hypothesis/model or a set of assumptions using the available data. Many statistical techniques have been proposed to analyze data, such as analysis of variance (ANOVA), linear regression, canonical correlation analysis (CCA), multidimensional scaling (MDS), factor analysis (FA), or principal component analysis (PCA), to name a few. A useful overview is given in [Sungur].

3.5 Data Mining and Learning Classifier System

3.5.1 Data Mining

Data mining's main objective is to discover novel, interesting and useful knowledge from databases [Fayyad, 1996]. Conventional data analysis techniques, analyze data manually. This leads to many hidden and potentially useful relationships overlooked by the analyst. As stated earlier, hospitals are capable of generating and collecting a huge amount of data. This vast storage of data requires an automated way to extract critical information and knowledge. These attributes, make healthcare and its large data stores an ideal environment for applying data mining techniques.

Utilization of clustering as a data mining and data analysis technique can lead to the discovery of the general distribution of the data. It also allows for discovery of similar objects described in the data set. Usually, a good characterization of the resulting clusters is also an objective. Another objective is scalability. An algorithm is considered scalable if its cost increases linearly with the number of records.

In pattern recognition, most of the data analysis is concerned with predictive modeling: given some existing data (“training data"), we want to predict the behavior of the unseen data (“testing data"). This is often called “machine learning" or simply “learning." Depending on the type of feedback one can get in the learning process, three types of learning techniques have been suggested. In supervised learning, labels on data points are available to indicate if the prediction is correct or not. In unsupervised learning, such label information is missing. In reinforcement learning, only the feedback after a sequence of actions that can change the possibly unknown state of the system is given. In the past few years, a hybrid learning scenario between supervised and unsupervised learning, known as semi-supervised learning, transductive learning [Joachims, 1999], or learning with unlabeled data [K. Nigam, A. McCallum, S. Thrun, and T. Mitchell, 2000], has emerged, where only some of the data points have labels. This scenario happens frequently in applications, since data collection and feature extraction can often be automated, whereas the labeling of patterns or objects has to be done manually and this is expensive both in time and cost.

3.5.2 Evolutionary Computation

The term evolutionary computation is used to describe algorithms that simulate the natural evolution to perform function optimization and machine learning. They are based on the Darwinian principle of evolution through natural selection. The algorithms maintain a group of individual attributes to explore the search space. Examples of evolutionary computation include genetic algorithms (GA), genetic programming (GP),

generic genetic programming (GGP), evolutionary programming (EP) and evolution strategy (ES). Genetic Algorithms use a fixed-length binary bit string as an individual. Three Genetic Operators are used to search for better attributes. Reproduction operator copies the unchanged individual. Crossover operator exchanges bits between two parents. Mutation operator randomly changes individual bits. Genetic Programming extends Genetic Algorithms by using a tree structure as the individual. Evolutionary Programming emphasizes on the behavioral linkage between parents and their offspring. Mutation is the only genetic operator within EP. There is no constraint on the representation in EP. Evolutionary Strategy emphasizes on the individual, i.e. the phenotype, to be the object to be optimized. A genetic change in the individual is within a narrow band of the mutation step size and the step size has self-adaptations.

Data mining can be considered as a search problem, which tries to find the most accurate knowledge from all possible hypotheses. Since evolutionary computation is a subfield of artificial intelligence (more particularly computational intelligence) involving combinatorial optimization problems it can be considered a valuable asset as a robust and parallel search algorithm, when used in data mining it may find interesting knowledge inside a noisy environment.

A learning classifier system, or LCS, is a machine learning system with close links to reinforcement learning and genetic algorithms. First described by John Holland, an LCS consists of a population of binary rules on which a genetic algorithm altered and selected the best rules. Instead of a using goodness of fits function, rule utility is decided by a reinforcement learning technique.

Learning classifier systems can be split into two types depending upon where the genetic algorithm acts. A Pittsburgh-type LCS has a population of separate rule sets, where the genetic algorithm recombines and reproduces the best of these rule sets. In a Michigan-style LCS there is only a single population and the algorithm's action focuses on selecting the best classifiers within that rule set. Michigan-style LCSs have two main types of reinforcement learning, fitness sharing (**ZCS**) and accuracy-based (**XCS**).

Initially the classifiers or rules were binary, but recent research has focused on improving this representation. This has been achieved by using populations of neural networks and other methods. Learning classifier systems are not well-defined mathematically and doing so remains an area of active research. Despite this, they have been successfully applied in many problem domains.

3.5.3 Rule Learning

A rule is a sentence of the form `if antecedents, then consequent'. Rules are commonly used in expressing knowledge and are easily understood by human. Rule learning is the process of inducing rules from a set of training examples. Classical algorithms in this field include AQ15 [Michalski, 1986] and CN2 [Clark, 1986]. Previous works in rule learning using evolutionary computation mainly use GA [Holland, 1992; Goldberg, 1989]. There are two different approaches. In the Michigan approach [Holland, 1978; Booker, 1998], each individual in the GA corresponds to a rule, while in the Pittsburgh approach [Smith, 1980; Smith, 1983] it corresponds to a set of rules. The system REGAL [Giordana, 1995] uses the Michigan approach and a distributed genetic

algorithm to learn first-order logic concept descriptions. It uses a selection operator, called Universal Suffrage operator, to achieve the learning of multi modal concepts. Another system GABIL [Jong 1993] uses the Pittsburgh approach. It can adaptively allow or prohibit certain genetic operations for certain individuals. GIL [Janikow, 1993] also uses the Pittsburgh's approach and utilizes 14 genetic operators. These operators perform generalization, specialization or other modifications to the individuals at the rule set level, the rule level and the condition level.

3.5.4 Grammar of Rule Based Algorithms

Specific grammar is necessary for rules to be structured for LCS algorithms. The format for rules in each problem can be different. Thus, for each problem a specific grammar is written so that the format of the rules can best fit the domain. In general, the grammar specifies a rule is of the form `if antecedents then consequent'. The antecedent part is a conjunction of attribute descriptors. The consequent part is an attribute descriptor as well. An attribute descriptor assigns a value to a nominal attribute, a range of values to a continuous attribute, or can be used to compare attribute values. The tables below are examples of LCS rules grammar and derivations.

Rule→if Antes, then Consq.
Antes→Attr1 and Attr2 and Attr3
Attr1→any Attr1_descriptor
Attr2→any Attr2_descriptor
Attr3→any Attr3_descriptor
Attr1_descriptor→attr1=erc1
Attr2_descriptor→attr2 between erc2 erc2
Attr3_descriptor→attr3 Comparator Attr3_term
Comparator→ ≠ <= >=< >
Consq→Attr4_descriptor
Attr4_descriptor→attr4=boolean_erc

Table 10.Example of grammar used for creating rules in rule learning

	Rule
⇒	if Antes, then Consq.
⇒	if Attr1 and Attr2 and Attr3, then Consq.
⇒	if Attr1_descriptor and Attr2_descriptor and Attr3_descriptor, then Attr4_descriptor.
⇒	if attr1=erc1 and attr2 between erc2 erc2 and attr3 Comparator Attr3_term, then attr4=boolean_erc.
⇒	if attr1=erc1 and attr2 between erc2 erc2 and attr3≠erc3, then attr4=boolean_erc.
⇒	if attr1=0 and attr2 between 100 150 and attr3≠50, then attr4=T.

Table 11.Example derivation of the rules created for rule learning

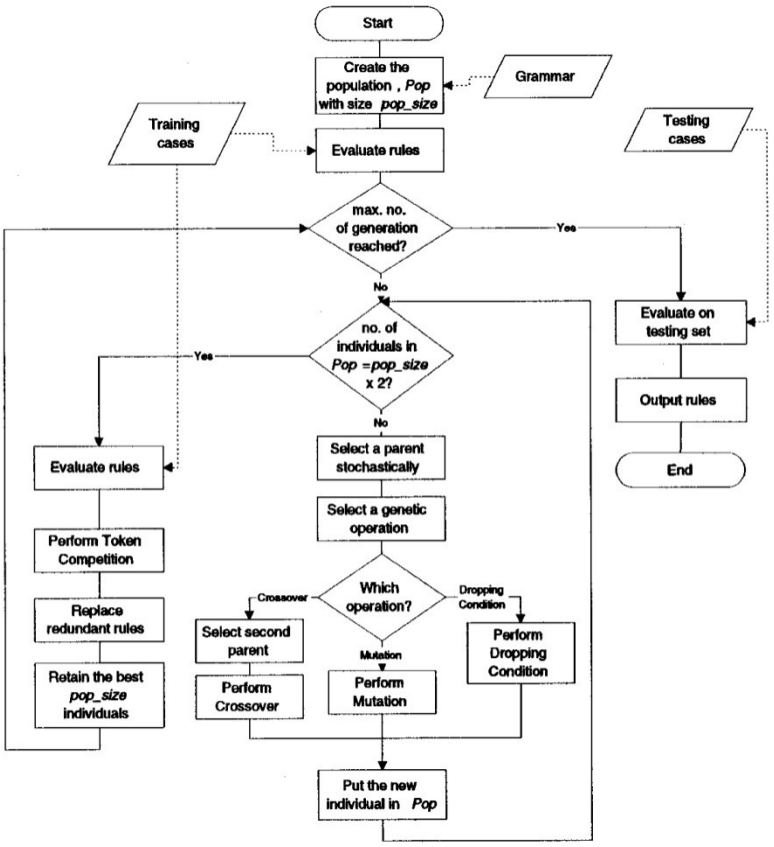


Fig 1. Example flowchart of the rule learning process.

The above figure is an example of a flowchart to develop a set of rules to create an initial population using generic genetic programming (GGP). The grammar is used to derive rules to make up the initial population. The start symbolic the first symbol of the first line of the grammar. From the start symbol, a complete derivation is performed. An example of how a rule is derived from the grammar is as follows:

“If attr1 = 0 and attr2 between 100 and 150 and attr3 " 50, then attr4 =T.

“If attr1 = 1 and any attr3 or attr2, then attr4 =F.

The representation of rules is not fixed but depends on the grammar. The descriptor is not restricted to compare attributes with values. Rather, the descriptors can be comparisons between attributes. Rules with other formats can be learned, provided that the suitable grammar is supplied. Moreover, rules with the user-desired structure can be learned because the user can specify the required rule format in the grammar.

3.6 Belief Theory

3.6.1 Frequentist approach to probability

Probability theory is the body of knowledge that enables us to reason formally about uncertain events. The populist view of probability is the so-called frequentist approach whereby the probability P of an uncertain event A, written P(A), is defined by the frequency of that event based on previous observations. For example, suppose that in the United States 49.1% of all babies born are girls; suppose then that we are interested in the event A: 'a randomly selected baby is a girl'. According to the frequentist approach P(A)=0.491.

3.6.2 Bayesian approach to probability

The Frequentist approach for defining the probability of an uncertain event is all well and good providing that we have been able to record accurate information about many past instances of the event. However, if no such historical database exists, then we have to consider a different approach. Suppose, for example, we want to know the

probability that a newly developed flight control system contains a critical fault. Since there are no previous instances of such systems we cannot use the frequentist approach to define our degree of belief in this uncertain event.

Bayesian probability is a formalism that allows us to reason about beliefs under conditions of uncertainty. If we have observed that a particular event has happened, such as the Miami Heat winning the NBA Championship in 2006, then there is no uncertainty about it. However, suppose ‘a’; is the statement

"The Miami Heat will win the NBA Championship in the year 2011"

Since this is a statement about a future event, nobody can state with any certainty whether or not it is true. Different people may have different beliefs in the statement depending on their specific knowledge of factors that might affect its likelihood. For example, John may have a strong belief in the statement a based on his knowledge of the current team and past achievements. Joe, on the other hand, may have a much weaker belief in the statement based on some inside knowledge about the status of the Heat; for example, he might know that the club is going to have to trade some of its best players in the year 2009 for salary cap reasons.

Thus, in general, a person's subjective belief in a statement ‘a’ will depend on some body of knowledge K. We write this as P(a|K). John's belief is different from Joe's because they are using different K's. However, even if they were using the same K they might still have different beliefs in ‘a’. The expression P(a|K) thus represents a belief measure. Sometimes, for simplicity, when K remains constant we just write P(a), but you must be aware that this is a simplification.

3.6.3 Probability axioms

While different people may give P(a) a different value there are nevertheless certain axioms which should always hold for internal consistency. These are the axioms of probability theory (which can be proved to be valid when P(a) represents the frequentist approach):

1. P(a) should be a number between 0 and 1
2. If a represents a certain event then P(a)=1.
3. If a and b are mutually exclusive events then P(a or b) = P(a)+P(b)

(Mutually exclusive means that they cannot both be true at the same time; for example, if a represents the proposition that our control system contains 0 faults, while b represents the proposition that our control system contains 1 fault)

3.6.4 Variables and probability distributions

Take the example of the uncertain event a = "The Miami Heat will win the NBA Championship in the year 2011". We can think of this event as just one state of the variable A which represents "NBA Championship winners in 2011". In this case A has many states, one for each team entering the FA Cup. We write this as A = {a1, a2, ..., an}

Where a1 = "Heat", a2 = "Pistons", a3 = "Lakers", etc.

Since in this case the set A is finite we say that A is a finite discrete variable.

As another example, suppose we are interested in the number of critical faults in our control system. The uncertain event is A = "Number of critical faults". Again it is best to think of A as a variable which can take on any of the discrete values 0,1,2,3,... thus

A={0,1,2,3,....;}.

In this case we

Let us define a1 as the event "A=0", and a2 as the event "A=1".

Clearly the events a1 and a2 are mutually exclusive and so P(a1 or a2)=P(a1)+P(a2). However, we cannot say that

P (a1 or a2) = 1

Because a1 and a2 are not exhaustive, that is, they do not form a complete partition of A. However, if we define a3 as the event "A>1" then a1, a2, and a3 are complete and mutually exhaustive and in this case

P(a1)+P(a2)+P(a3) = 1

In general if A is a variable with states a1, a2, ..., an:

$$\sum_{i=1}^{n} P(a_i) = 1$$

The probability distribution of A, written P(A), is simply the set of values {P(a1), P(a2), ..., P (an)}

3.6.5 Joint events and marginalization

Suppose that our control system X is made up of two subsystems. Let A be the number of critical faults in the first subsystem and let B be the number of critical faults in the second subsystem.

Suppose that

A = {a1, a2, a3} where a1=0, a2=1, a3=">1"

B = {b1, b2, b3} where b1=0, b2=1, b3=">1"

If we are interested in the overall number of critical faults in the system, then we speak about the joint event A and B. We write the probability of this event as

P (A,B)

P (A,B) is called the joint probability distribution of A and B. Specifically, P(A,B) is the set of probabilities:

{P (a1, b1), P(a1, b2), P(a1, b3), P(a2, b1), P(a2, b2), P(a2, b3), P(a3, b1), P(a3, b2), P(a3, b3)}

Where for any i and j; P (a_i, b_j) is the probability of the event a_i and b_j.

In general, if A and B are variables with possible states {a1, a2... an} and {b1, b2... bm} respectively then the joint probability distribution P (A,B) is the set of probabilities

{P (a_i, b_j) | i=1,…,n and j=1,..., m}

If we know the joint probability distribution P (A,B) then we can calculate P(A) by a formula (called marginalization) which comes straight from the third axiom, namely:

$$P(a) = \sum_i P(a, b_i)$$

This is because the events (a, b1), (a, b2), ..., (a, bm) are mutually exclusive. When we calculate P(A) in this way from the joint probability distribution we say that the variable

B is marginalized out of P(A,B). It is a very useful technique because in many situations it may be easier to calculate P (A) from P (A,B).

3.6.6 Conditional probability

In the introduction to Bayesian probability we explained that the notion of degree of belief in an uncertain event A was conditional on a body of knowledge K. Thus, the basic expressions about uncertainty in the Bayesian approach are statements about conditional probabilities. This is why we used the notation P (A|K) which should only be simplified to P (A) if K is constant. Any statement about P (A) is always conditioned on a context K

In general we write P (A|B) to represent a belief in A under the assumption that B is known. Even this is, strictly speaking, shorthand for the expression P (A|B, K) where K represents all other relevant information. Only when all such other information is irrelevant can we really write P (A|B).

The traditional approach to defining conditional probabilities is via joint probabilities. Specifically we have the well known 'formula':

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

This should really be thought of as an axiom of probability. Just as we saw the three probability axioms were 'true' for frequentist probabilities, so this axiom can be similarly justified in terms of frequencies:

Example: Let A denote the event 'student is female' and let B denote the event 'student is Chinese'. In a class of 100 students suppose 40 are Chinese, and suppose that 10 of the Chinese students are females. Then clearly, if P stands for the frequency interpretation of probability we have:

P (A, B) = 10/100 (10 out of 100 students are both Chinese and female)

P (B) = 40/100 (40 out of the 100 students are Chinese)

P (A|B) = 10/40 (10 out of the 40 Chinese students are female)

It follows that the formula for conditional probability 'holds'.

In those cases where P (A|B) = P (A) we say that A and B are independent. If P (A|B, C) = P (A|C) we say that A and B are conditionally independent given C.

3.6.7 Bayes Rule

True Bayesians actually consider conditional probabilities as more basic than joint probabilities. It is easy to define P (A|B) without reference to the joint probability P (A, B). To see this note that we can rearrange the conditional probability formula to get:

P (A|B) P (B) = P (A, B) but by symmetry we can also get:

P (B|A) P (A) = P (A, B)

It follows that:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

This is referred to as Bayes Rule.

It is common to think of Bayes rule in terms of updating our belief about a hypothesis A in the light of new evidence B. Specifically, our posterior belief P(A|B) is calculated by multiplying our prior belief P(A) by the likelihood P(B|A) that B will occur if A is true. The power of Bayes' rule is that in many situations where we want to compute P (A|B) it turns out that it is difficult to do so directly, yet we might have direct information about P (B|A). Bayes' rule enables us to compute P (A|B) in terms of P (B|A). For example, suppose that we are interested in diagnosing cancer in patients who visit a chest clinic.

Let A represent the event "Person has cancer"

Let B represent the event "Person is a smoker"

We know the probability of the prior event P (A) =0.1 on the basis of past data (10% of patients entering the clinic turn out to have cancer). We want to compute the probability of the posterior event P (A|B). It is difficult to find this out directly. However, we are likely to know P (B) by considering the percentage of patients who smoke – suppose P (B) =0.5. We are also likely to know P (B|A) by checking from our record the proportion of smokers among those diagnosed. Suppose P (B|A) =0.8.

We can now use Bayes' rule to compute:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} = \frac{(0.8)(0.1)}{(0.5)} = 0.16$$

Thus, in the light of evidence that the person is a smoker we revise our prior probability from 0.1 to a posterior probability of 0.16. This is a significance increase, but it is still unlikely that the person has cancer.

The denominator P (B) in the equation is a normalizing constant which can be computed, for example, by marginalization whereby

$$P(B) = \sum_i P(B, A_i) = \sum_i P(B|A_i) \cdot P(A_i)$$

Hence we can state Bayes rule in another way as:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{\sum_i P(B|A_i) \cdot P(A_i)}$$

3.6.7.1 Bayes Rule Example

Suppose that we have two bags each containing black and white balls. One bag contains three times as many white balls as blacks. The other bag contains three times as many black balls as white. Suppose we choose one of these bags at random. For this bag we select five balls at random, replacing each ball after it has been selected. The result is that we find 4 white balls and one black. What is the probability that we were using the bag with mainly white balls?

Solution: Let A be the random variable "bag chosen" then A= {a1, a2} where a1 represents "bag with mostly white balls" and a2 represents "bag with mostly black balls". We know that P (a1) =P (a2) =1/2 since we choose the bag at random.

Let B be the event "4 white balls and one black ball chosen from 5 selections".

Then we have to calculate P (a1|B). From Bayes' rule this is:

$$P(a_1|B) = \frac{P(B|a_1) \cdot P(a_1)}{P(B|a_1) \cdot P(a_1) + P(B|a_2) \cdot P(a2)}$$

Now, for the bag with mostly white balls the probability of a ball being white is ¾ and the probability of a ball being black is ¼. Thus, we can use the Binomial Theorem, to compute P (B|a1) as:

$$P(B|a_1) = \binom{5}{1}\left(\frac{3}{4}\right)^4\left(\frac{1}{4}\right)^1 = \frac{405}{1024}$$

Similarly

$$P(B|a_2) = \binom{5}{1}\left(\frac{1}{4}\right)^4\left(\frac{3}{4}\right)^1 = \frac{15}{1024}$$

hence

$$P(a_1|B) = \frac{405/1024}{405/1024 + 15/1024} = \frac{405}{420} = 0.964$$

3.6.8 Likelihood Ratio

We have seen that Bayes' rule computes P (A|B) in terms of P (B|A). The expression P (B|A) is called the likelihood of A. In the example above A had two values a1 and a2. The ratio:

$$\frac{P(B|a_1)}{P(B|a_2)}$$

is called the likelihood ratio. In the example the likelihood ratio computes to 405/15 = 27. This tells us that the 'odds' on the bag being the one mainly with white balls is 27 to 1.

3.6.9 Chain rule

We can rearrange the formula for conditional probability to get the so-called product rule:

$P(A, B) = P(A|B)\, P(B)$

We can extend this for three variables:

$P(A,B,C) = P(A| B,C)\, P(B,C) = P(A|B,C)\, P(B|C)\, P(C)$

and in general to n variables:

$P(A1, A2, ..., An) = P(A1| A2, ..., An)\, P(A2| A3, ..., An)\, P(An-1|An)\, P(An)$

In general we refer to this as the chain rule.

3.6.10 Independence and conditional independence

The conditional probability of A given B is represented by P (A|B). The variables A and B are said to be independent if P (A) = P (A|B) (or alternatively if P(A,B)=P(A) P(B) because of the formula for conditional probability).

3.6.10.1 Example 1

Suppose Rayde and Henry each toss separate coins. Let A represent the variable "Rayde's toss outcome", and B represent the variable "Henry's toss outcome". Both A and B have two possible values (Heads and Tails). It would be uncontroversial to assume that A and B are independent. Evidence about B will not change our belief in A.

3.6.10.2 Example2

Now suppose both Rayde and Henry toss the same coin. Again let A represent the variable "Rayde's toss outcome", and B represent the variable "Henry's toss outcome". Assume also that there is a possibility that the coin in biased towards heads but we do not know this for certain. In this case A and B are not independent. For example, observing that B is Heads causes us to increase our belief in A being Heads (in other words P(a|b)>P(b) in the case when a=Heads and b=Heads).

In Example 2 the variables A and B are both dependent on a separate variable C, "the coin is biased towards Heads" (which has the values True or False). Although A and B are not independent, it turns out that once we know for certain the value of C then any evidence about B cannot change our belief about A. Specifically:

P(A|C) = P(A|B,C)

In such case we say that A and B are conditionally independent given C.

In many real life situations variables which are believed to be independent are actually only independent conditional on some other variable.

3.6.10.3 Example 3

Suppose that Rayde and Henry live on opposite sides of the City and come to work by completely different means, say Rayde comes by train while Henry drives. Let A represent the variable "Rayde late" (which has values true or false) and similarly let B represent the variable "Henry late". It would be tempting in these circumstances to assume that A and B must be independent. However, even if Rayde and Henry lived and

worked in different countries there may be factors (such as an international fuel shortage) which could mean that A and B are not independent. In practice any model of uncertainty should take account of all reasonable factors. So, the probability of a meteorite hitting the Earth might be reasonably excluded it does not seem reasonable to exclude the fact that both Rayde (A) and Henry (B) being late may be affected by a Train strike (C). Clearly P(A) will increase if C is true; but P(B) will also increase because of extra traffic on the roads.

3.7 Dempster Shafer Belief Theory

The aim of belief theory is to give a corrected formal representation of the inaccurate and uncertain aspect of information. Belief theory is derived from the concepts of Probability theory and possibility theory. Both Probability theory and possibility theory are derivates of evidence theory. Probability theory and possibility theory both have similar properties and thus they cannot be mixed up, as they are based on incompatible assertions. [Dubois and Prade, 1987a] "Inaccuracy" and "uncertainty" in belief theory can be considered as two antagonistic points of view concerning the same actual imperfection of the information in that:

- Inaccuracy is concerned with the correctness in the content of information.
- The uncertainty of information deals with the insufficient knowledge of its (information) truth.

Uncertainty of information is considered with the help of qualifiers such as “probable ", " possible ",” necessary ", " plausible ", " credible " [Dubois and Prade, 1987a].

For any real world situation, there are difficulties that researchers face when using information derived from various sources (i.e. human recall or opinion, statistics, etc.).

The Dempster-Shafer (DS) belief theoretic model is designed where each attribute is modeled to have its own belief function. The functionality of DS belief theory model is a superlative application to address several common types of data imperfections: missing data, incomplete data and ambiguities. The normal probabilistic approaches required to design the initial assumptions in the model (e.g., independence of events, equiprobabilities, etc.) depend on the dataset. The performance of the model is directly correlated to how closely are the assumptions based to reality; the more realistic the assumptions, the better the performance.

Let Θ be a finite set of mutually exclusive and exhaustive proposition or commonly known as frame of discernment. The power set 2^{Θ} is the set of all subsets of Θ including itself and the null set φ. Each subset in the power set is called focal element. Based on the evidence, a value between [0, 1] is assigned to each focal element with 0 representing no belief and 1 representing total belief. Basic belief assignment (bba) is assigned to the individual propositions and is also known as the mass of the individual proposition. It is assigned to every subset of the power set. If bba of an individual proposition A is m(A) then,

$$\sum_{A \subset \Theta} m(A) = 1 \qquad (1)$$

Also, bba of a null set is zero, i.e.

$$m(\varphi) = 0 \tag{2}$$

Ignorance is represented by assigning the complementary probability to m(Θ). Measure of total belief committed to A, Bel (A), is computed using Eq. (3).

$$\text{Bel (A)} = \sum_{B \subset A} m(B) = 1 \tag{3}$$

According to Smets [7], formal notation of Bel is given as,

$$Bel \ {}^{\Theta,\Re}_{Y,,t}[E_{Y,t}](\omega_o \in A) = x \tag{4}$$

This equation denotes the degree of belief x of the classifier Y at time t when ω_o belongs to set A, where A is the subset of Θ and $A \in \Re; \Re$ is a Boolean algebra of Θ. Belief is based on the evidential corpus $E_{Y,t}$ held by Y at time t where $E_{Y,t}$ represents all what Y knows at time t. For simply $Bel \ {}^{\Theta,\Re}_{Y,,t}[E_{Y,t}](\omega_o \in A) = x$ can be written as Bel [E](A) or Bel (A). Further, plausibility function of A is defined as,

$$Pl(A) = 1 - Bel(-A) = \sum_{B \cap A \neq \phi} m(B) \tag{5}$$

Bel (A) represents the lower limit of probability and P l(A) represents the upper limit.

3.7.1 Example of Dempster Shafer Theory based Classifier Fusion

Let's look at an example of classifier fusion algorithm based on the DS theory as it is applied to combine the output of individual fingerprint recognition algorithms to improve the verification performance. Using the underlying concept to DS theory and basic belief assignment, classifier fusion is performed using minutiae based fingerprint recognition algorithm [Jain, 1997], ridge based recognition algorithm [Marana and Jain, 2005] and finger code based recognition algorithm [Jain, 1999]. For every input fingerprint image, each classifier assigns a label true or 1 to proposition i, $i \in \Theta$ and the remaining classes are labeled as false or 0. Thus there are two focal elements for each fingerprint recognition algorithm i and $\neg i = \Theta - i$. i is for confirming and $\neg i$ is for denying the proposition for mass assignment in the DS theory. For each fingerprint recognition algorithm, we compute the respective predictive rates used to assign their basic belief assignment. For a c class problem, let us assume that an input pattern belonging to class j ($j \in c$) is classified as one of the k ($k \in c + 1$) classes including the rejection class, i.e. $(c + 1)^{th}$ class. So, the predictive rate of a classifier P_k for an output class k is the ratio of the number of input patterns classified correctly to the total number of patterns classified as class k where input patterns belonging to all classes is presented to the classifier.

In this example, when a fingerprint recognition algorithm classifies the result $k \in c + 1$, it is considered that for all instances the likelihood of k being the actual class is Pk and the likelihood of k not being the correct class is $(1 - Pk)$. The predictive rate is used as basic belief assignment or mass m(k) and disbelief is assigned to

$m(\neg k)$; with $m(\Theta) = 1$.

Further, multiple evidences are combined using the Dempster's rule of combination. Let A and B be used for computing new belief function for the focal element C, Dempster's rule of combination is written as:

$$m(C) = \frac{\sum_{AnB=C} m(A)m(B)}{1 - \sum_{AnB=\phi} m(A)m(B)} \quad (6)$$

Let m1, m2 and m3 be the mass computed from the three fingerprint recognition algorithms or classifiers which are combined recursively as shown in Eq. (7),

$$m_{final} = m1 \oplus m2 \oplus m3 \quad (7)$$

where $\oplus$ shows the Dempster's rule of combination. Final result is obtained by applying threshold t to m_{final} ,

$$result = \begin{cases} accept, & \text{if } m_{final} \geq t \\ reject, & \text{otherwise} \end{cases} \quad (8)$$

3.7.2 Update Rule for Calculating Belief Assignment

In most cases, it is required to update the belief based on new evidences or data. Let $E \subset \Theta$ and E_v be the evidence which states that the actual world is not in $\neg E$. Now suppose that the new data or evidence provides the exact value of E_v. Belief function is revised using the Dempster's update rule,

$$Bel[Ev](A) = Bel(A \cup \neg E) - Bel(\neg E) \quad (9)$$

This rule would be used to update the basic belief assignment associated with each fingerprint algorithm when a new training data is added. With this rule, only new basic belief assignments would be used to update the classifier. The time required for updating is significantly less as it is not required to train the complete classification algorithm when new training data is added.

CHAPTER 4 THE FIRD FRAMEWORK

4.1 Biometric Phase of framework

The primary phase of the FIRD ™ framework is a composite of these four types of biometric identifiers:

1. **F**ingerprint
2. **I**ris
3. **R**etina Scan
4. **D**NA

The fusion of these four biometric identifiers promote the best 1:1 match which is essential for the EHCR system to be effective at the rudimentary level. The four biometric identifiers chosen also have the greatest accuracy and the lowest error rate. The four biometric identifiers with the exception of fingerprinting are all native elements within the healthcare environment. The enrollment of the biometric identifiers for the FIRD™ can be collected from a patient during a routine physical or office visit. Because of the current state of society and national security concerns the issue of fingerprint scanning is fast becoming a widely accept practice so its collection and enrollment during a healthcare visit would not cause any alarms. Collection of the samples for iris and retina scans as a patient has his or her eyes checked by an optometrist. The DNA sample can be collected from saliva when a patient has his or her glands examined (saying 'ahh' to check your throat).DNA can also be collected from a blood sample.

Based on the ASTM criteria a universal patient identifier has to be scalable. We choose multi modal biometrics systems as the core of our universal patient identifier for this very reason. In general multi modal biometric systems are scalable by design in the choice and number of biometric traits; the level in the biometric system at which information provided by multiple traits should be integrated; the methodology adopted to integrate the information; and the cost versus matching performance trade-off.

The scalability of FIRD allows it implementation to meet the vital criteria of cost-effectiveness. In a perfect world having all four of the identifiers within your biometric system will provide a 1:1 match each and every time; the absence of one, two or three of the four identifiers does not significantly affect the frameworks ability to produce a 1:1 match of a patient records due to the other functions within the framework. The presence of all four biometric identifiers allows for system scalability with the fingerprint identifier as the minimal requirement. But in order to capture the entire patient population of the planet all four identifiers would be necessary.

All four biometric identifiers allow for the inclusion of special populations. In the case of amputees fingerprints may cause a problem. In the case of twins, they both share the same DNA but different fingerprints. This also holds true for bone marrow transplants patients where the procedure causes a change in his or her DNA, but not his or her fingerprint, Iris or retina scans. So the FIRD™ biometric checkpoint system would allow each healthcare organization to describe the rules that allow the verification of a patient's

identity in order to access all of the proper medical information for that patient in every situation.

Once the biometric query returns data files that match the criteria of the biometric identifier used. A single integrated data set of all returned records is created. Although the biometric identifier is very reliable there may be records that are converted to electronic format but are not associated with a biometric identification sample.

There are consequences of combining two or more biometric tests of identity into an enhanced multi modal test. The common and intuitive assumption is that the combination of different tests must improve performance, because "more information is better than less information." However, a different intuition suggests that if a strong test is combined with a weaker test, the resulting decision environment is in a sense averaged, and the combined performance will lie somewhere between that of the two tests conducted individually (and hence will be degraded from the performance that would be obtained by relying solely on the stronger test).

Although there is truth in both intuitions, the key to resolving the apparent paradox is when biometric tests are combined. Let’s use combining two biometrics tests as an example. One of the resulting error rates (False Accept or False Reject rate) becomes better than that of the stronger of the two tests, while the other error rate becomes worse even than that of the weaker of the tests. If the two biometric tests differ significantly in their power, and each operates at its own cross-over point, then combining them gives significantly worse performance than relying solely on the stronger biometric.

4.1.1 Example Notation

The following is an example of how the traditional multi modal biometric error probability is calculated. Assume there are two hypothetical and independent biometric tests referred to respectively as 1 and 2[Tang, 1997]. For example, 1 might be voice-based verification, and 2 could be fingerprint verification. Each biometric test is characterized by its own pair of error rates at a given operating point; denote as the error probabilities, where y is the biometric tested (1 for voice recognition or 2 for fingerprint):

- $P_y(FA)$ = probability of a FA with test y
- $P_y(FR)$ = probability of a FR with test y

There are two possible ways to combine the outcomes of these two biometric tests when forming the conjoint ("enhanced") decision: the Subject may be required to pass both of the biometric tests, or they may be accepted if they can pass at least one of the two tests. These two cases define the disjunctive and conjunctive rules:

- Rule A: Conjunction ("AND" Rule)

Accept only if both tests 1 and 2 are passed.

- Rule B: Disjunction ("OR" Rule)

Accept if either test 1 or test 2 is passed.

We can now calculate False Accept and False Reject error rates of the combined biometric, both for conjunctive (Rule A) and disjunctive (Rule B) combinations of the two tests. These new error probabilities for x = {A, B} will be denoted:

- $P_x(FA) = \prod P_y(FA)$
- $P_x(FR) = \prod P_y(FR)$

Rule A: Conjunction (The "AND") Criteria Rule.

If Rule A (the "AND" Rule) is used to combine the two tests 1 and 2, a False Accept can only occur if both tests 1 and 2 produce a False Accept. Thus the combined probability of a False Accept, $P_y(FA)$, is the product of its two probabilities for the individual tests: $P_A(FA) = P_1(FA) \cdot P_2(FA)$ (clearly a lower probability than for either test alone). But the probability of a False Reject when using this Rule, which can be expressed as the complement of the probability that neither test 1 nor 2 produces a False Reject, is higher than it is for either test alone:

- $P_A(FR) = 1- [1- P_1(FR)] \cdot [1- P_2(FR)] = P_1(FR) + P_2(FR) - P_1(FR) \cdot P_2(FR)$

Rule B: Conjunction (The "OR") Criteria Rule.

If Rule B (the "OR" Rule) is used to combine the two tests 1 and 2, a False Reject can only occur if both tests 1 and 2 produce a False Reject. Thus the combined probability of a False Reject, $P_y(FR)$, is the product of its two probabilities for the individual tests: $P_B(FR) = P_1(FR) \cdot P_2(FR)$ (clearly a lower probability than for either test

alone). But the probability of a False Accept when using this Rule, which can be expressed as the complement of the probability that neither test 1 nor 2 produces a False Accept, is higher than it is for either test alone:

- $P_B(FA) = 1-[1- P_1(FA)] \cdot [1- P_2(FA)] = P_1(FA) + P_2(FA) - P_1(FA) \cdot P_2(FA)$

This dissertation uses both a multi layered and multi modal approach of the biometrics identifiers but incorporates in its foundation the ability to use multi modal biometrics when the technology is becomes widely available and cost effective. While the biometrics are used to verify a patient's existence but more so to identify and verify the correct electronic medical records are collected and returned for that patient. The FIRD framework will scan through all medical records stored within the virtual data depository and return records that match the query in the following manner as shown in Figure 2. First, the fingerprint biometric which has an error rate of 1/500 will return a set of matching records. Then, the Iris biometric which has an error rate of 1/131,000 is used to return a more precise data set. Next, the retina scan which has an error rate of 1/10,000,000 will return a greater precise data set, and finally it uses DNA in the final stage which has an error rate of 1/30,000,000,000[IOM, 1994].

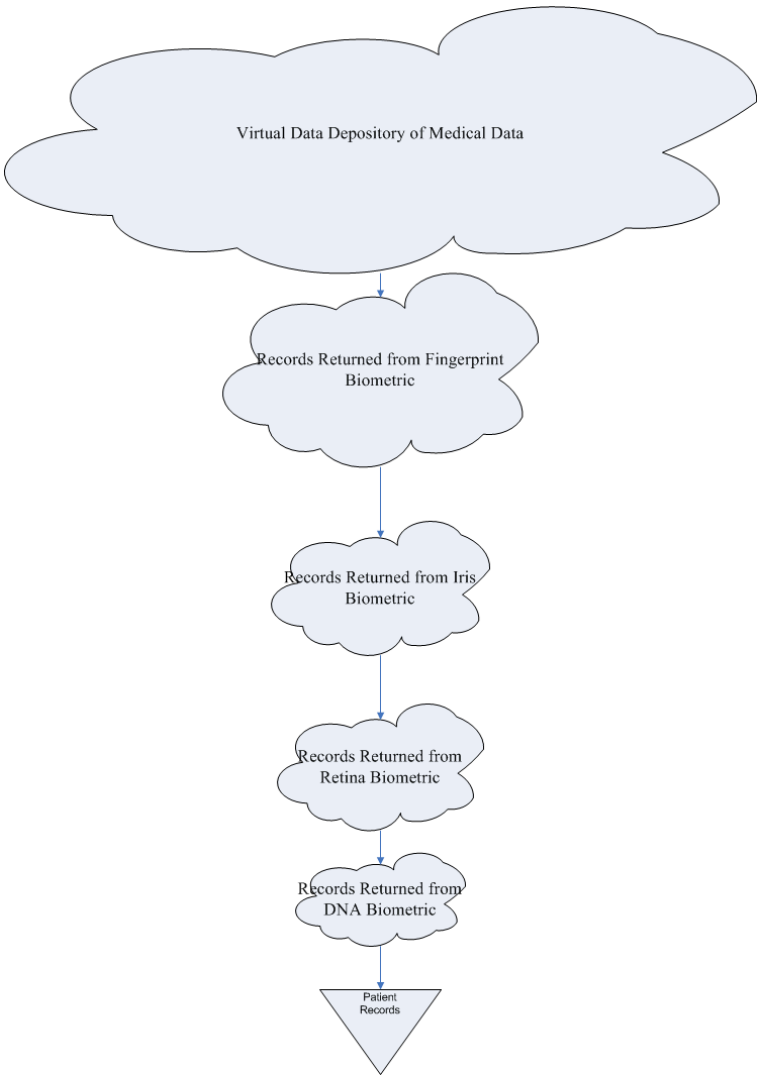


Figure 2: Multi Layer Biometric system

Depending on what criteria a returned record meets it will fall into two types:

- Type 1-1: Record conforms to conjunction rule meaning it has all four biometric identifiers
- Type 1-2: Record conforms to disjunction rule meaning it has one, two or three of the four biometric identifiers

Therefore, the acceptance utilized in the framework is a hybrid of both conjunction and disjunction rules, as denoted in the following expressions.

4.1.2 FIRD FRAMEWORK RULE

Based on the patient's query return, all medical records that pass one and or the entire biometric test will be considered. Where y is in the set of {1...n} biometric criteria for $P_y(FA)$ and $P_y(FR)$, such that acceptance is the inclusion of the electronic record. The query is the patient requesting his or her record. There is a record that is created that contains a patients' extraction sample and registration record with all of his or her personal and identifiable information this record is called Record Zero. The framework returns records that indicate they contain the same biometric identifier as the patients' extraction sample and registration record. Records that are returned containing all four biometric ID's contained in the patients' extraction sample and registration record are considered Type 1-1 records.

- $P_1(FA) + P_2(FA) + P_3(FA) + P_4(FA) – P_1(FR) \cdot P_2(FR) \cdot P_3(FA) \cdot P_4(FA)$

If a record is returned and it contains one, two, or three of the four biometric ID's contained in the patients' extraction sample and registration record. It is considered Type 1-2 records:

- $P_1(FA) + P_2(FA) + P_3(FA) + P_4(FA) – P_1(FA) \cdot P_2(FA) \cdot P_3(FA) \cdot P_4(FA)$

The developed Biometric phase is layered system designed to use one or all of the identifiers outlined. If all are identifiers are included the system will have a 99.9999% accuracy rate and will allow the framework to encompass all possible populations. Since the technology for instantaneous identification of the Iris, Retina, and DNA are not wide spread or do not exist and fingerprints are accepted by society then the framework needs a way to compensate for the error rate thus the second phase. The second phase also will return records that are not associated with a biometric tag.

4.2 Data Fusion Phase

The primary function of the second phase of the framework is to return records belonging to the patient in question that were not returned with the biometric sample. This is achieved through the creation of a set of data fusion algorithms for data mining the EHCR database. The framework uses data mining in a non-conventional manner to extract specific knowledge in this case all of a specific patient's electronic healthcare information from the de facto healthcare data system. The information collected in the first phase is used to create a baseline record for the patient. This record zero contains all

of a patient's up to date life data. With this record zero data analysis can be started using steps to induce knowledge in our case records with matching information from the preprocessed data within record zero. We then use causality and structure analysis to collect the overall relationships between the four protected health information variables that provide unique patient identification.

This phase allows for the EHCR system to use only one of the biometric identifiers in the first phase and still achieve a high level of accuracy. This research presents the two knowledge learning steps which are the core of the second phase of our framework. They both employ a type of evolutionary computation as the search algorithms. This becomes a form of secondary filtering and unification of the returned data set by using data fusion algorithms consisting of four additional unique identifiers within the HL7[HIPAA, 1996] standard of the EHCR. These four identifiers cannot be used as a universal patient identifier because they fall under protected healthcare information (PHI) [IOM, 1994, HIPAA, 1996]. The four identifiers are contained in the returned biometric sample record which is denoted a record 0 (zero). So the purpose of the Machine Learning System phase will be to determine if there are records that do not have the biometric identification marker present or less accurate biometric for example several records can be considered correct based on the fingerprints. But once these records are filtered with the 4 PHI identifiers belonging to the patient named on record zero the records that are 'false positive' will be removed because the four protected health information (PHI)[Terry, 2003, NCVHS, 1997] attributes are present are the same. Also these 4 PHI attributes will return the records that were excluded as 'true negatives'.

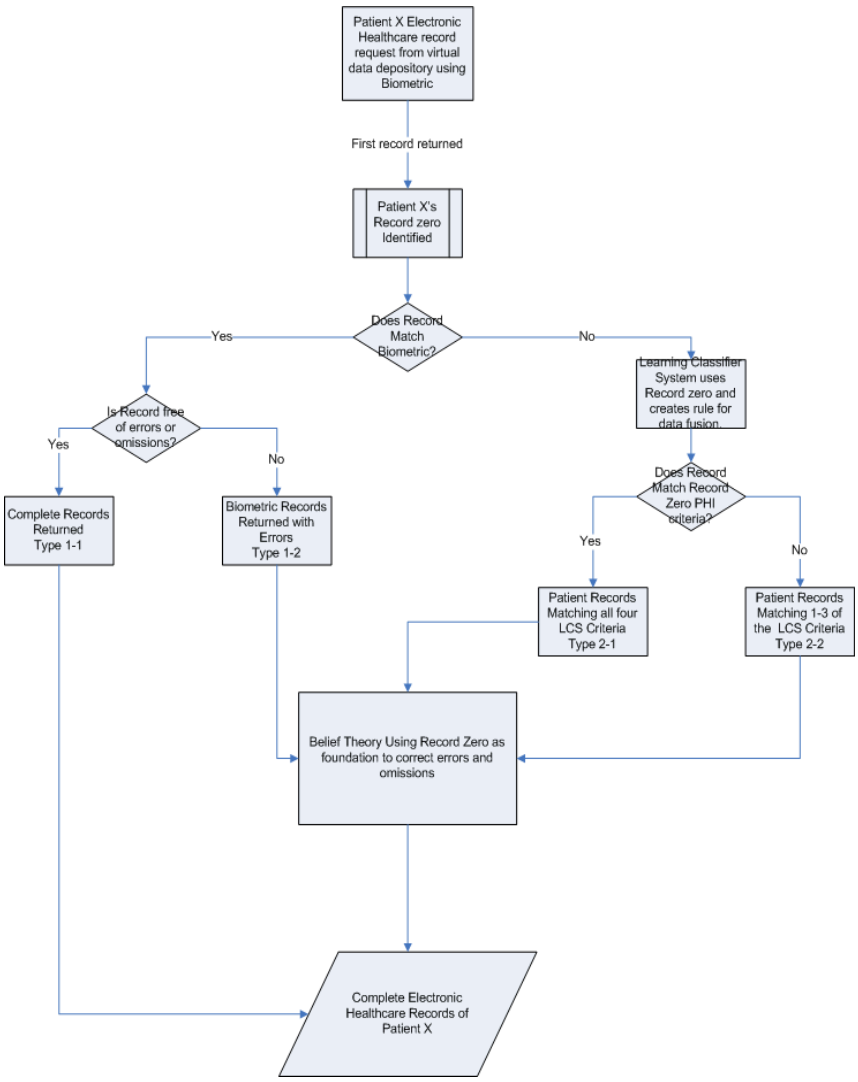
The framework's data fusion system uses a patient's Date of Birth, Race and Ethnic Background, Blood type and Gender to further consolidate and filter the dataset that was returned by the patient's biometric identifier. The use of these four protected health information identifiers allow the framework to analyze the records returned based on the biometrics which is the foundation of the framework. The second phase of the framework cleanses the first phase by removing the false positives records (Patient records returned that don't belong to the patient) returned that that were included that should not have been included and retrieve records that were passed over which are true negatives (Patient records that were not included that truly belong to the patient) that were missed by the scaled down use of the FIRD or electronic healthcare records which were not associated with a biometric identification sample or are incomplete.

Therefore, the search algorithms return records that were missed by the biometric phase, these records will be classified in two categories:

- Type 2-1 Records that belong to the patient and based on the PHI of the patient however do not have any biometric identifiers associated with it
- Type 2-2 Records that belong to the patient and based on the biometric identifiers of the patient however there are some missing values within the PHI of the patient's records.

4.3 Belief Theory Phase

The underline assumption in data mining and data fusion is that the data to be mined is complete. Healthcare data is not complete do to human error or oversight some records for a particular patient may contain errors or omissions. This area is often overlooked in data management. Electronic format or not, data integrity within medical records is essential for error free high quality healthcare. This framework addresses this problem in the data to ensure the maximum amount of patient information is included in his or her complete healthcare history. The FIRD incorporates the concept of belief theory to allocate a value to inconsistent data within the collected dataset which is in this case a patient's electronic medical record. The belief theory phase of the framework takes the returned records that are incomplete or do not meet the criteria of Type 1-1 records (perfectly matched records) and analyzes the discrepancies within to see if a probability theory model can with high confidence correct those same discrepancies to determine if any of these records can be transformed into Type 1-1 records. This is achieved by either putting in place the biometric identifier for record were no biometrics identifier exist and or by correcting or inserting the 4 PHI attributes in those patient records that do not have any. Figure 3 below shows a visual display of the functionality of the framework in managing the formulation of a complete medical record history of a patient.



*- Biometric Phase in progress
**- Data Fusion Phase in progress
***- Belief Theory Phase in progress

Figure 3. FIRD Framework Flow Diagram

4.4 The FIRD Solution

Under these conditions, implementation of a universal patient identifier would be quite feasible with the use of a multi modal biometric system as its core. The idea of identifying patients through a fingerprint scan eliminates the need for issuing more numeric codes. The FIRD framework technology is not just a patient identifier but rather a patient data identity and integrity management infrastructure that is scalable and has the ability to adapt with the advancement of technology. The framework can be retrofitted to any developed backend Master Patient Indexing System that links vital patient identification and medical record information from disparate databases. The system ensures accurate patient identification and facilitates rapid dissemination of clinical information across various departments throughout a healthcare facility.

The FIRD framework will identify a patient not by name, address, or telephone number but by a physical characteristic, i.e. their fingerprint for the primary deployment of the system. If any of the other unique identifiers are issued how will they link to the patient exclusively? If the system is not combined with biometrics, then that unique number is as suspect to theft, fraud, or error just like Access codes and PIN [Jain, 2003; NCVHS, 1997]. The goals that healthcare are expecting from the implementation of a universal patient identifier are accomplished by the FIRD framework. For there is a no more unique identifier of a person than the person themselves, and the use of data fusion and belief theory on specific collected information within the record will ensure data integrity and completeness of the data set.

4.5 Purpose of Framework

The creation of a standardize nationwide electronic healthcare record system in the United States would require a way to match a composite of an individual's recorded healthcare information to an identified individual patient out of approximately 300 million individuals to a 1:1 match. The technology exists for the migration of healthcare data from its archaic paper-based system to an electronic one. The technology also exist for the new digitized healthcare data to be transported anywhere in the world in a matter of seconds. This would allow all of the healthcare industry to store and exchange all of it healthcare data with one another whenever it is necessary, leading to increase in quality of service/treatment and lower cost. A critical element of the functionality of this system is the ability to uniquely identify a patient and match them to all of his or her medical records regardless of location. However a considerable problem lies in the fact that if all of the recorded healthcare data is electronic and interchangeable, who does what data belong to? How is that data verified to belong to that patient? How is that patient's identity verified? The research proposed in this paper presents the use of a multi layered biometric system as a foundation for an electronic healthcare record unification framework. The presented framework includes a secondary layer to capture records that belong to the patient query but do not have a biometric tag associated with it. This secondary layer also verifies that the records captured in the biometric collection truly belong to the patient query, thus eliminating the falsely accepted records from the query. The third layer of the framework is vital to the unification process in that it corrects omissions and discrepancies within the

patients' records that are the results of errors. The result of this framework should create a system of checks and balances through a newly created FIRD biometric checkpoint system to describe the rules that allow the verification of a patient's identity in order to access all of the correct medical information for that patient and merged it into a single integrated data set for medical use.

4.6 Questions Research Seek To Address

The objective of this dissertation is to create a framework that will be able to unify a patient's electronic medical records at the time of a biometric request query. This will involve clustering of mixed data types (quantitative and qualitative) efficiently and usefully. It is likely that not all questions stated will be dealt with in this research.

Focusing on the concept of universal identifying a patient to his or her records, to date this research has accomplished the following:

- Development and creation of universal patient identifier using multi layer and multi modal biometrics.
- The development and creation of the enrollment record named Record Zero.
- Identification and primary development of the PHI attributes and their data mining algorithms.
- Identification of the attributes needed for belief theory equations.

The primary focus of this dissertation is to retrieve query activated patient data by developing an object oriented framework for data mining electronic healthcare data. This framework will operate upon a selected data source and produce a result file. Certain core functions are performed by the framework, which interact with the extensible function. This separation of core and extensible functions allows scalability within the framework by the separation of the specific processing sequence and requirement of a specific data mining operation from the common attribute of all data mining operations. This separation will allow the end user to define extensible functions that allow the framework to perform new data mining operations without the framework having the knowledge of the specific processing required by those operations. This work may extend or combine existing methods or develop an entirely new approach. Although the concept of universal identifying a patient to his or her records is at the center of this research; the question of unifying electronic medical records will be explored. The end result will be the development of a foundation framework methodology to data mine records within the electronic medical record data depository.

Testing will be performed on synthetic data sets that will emulate patients electronic medical record data with a known classification and distribution for each class attribute pair. For example, there are x number of classes, y number of qualitative attributes and z number of quantitative attributes. For each class, a distribution is specified for each attribute. The data is then created according to this distribution. More data sets are then created through incorporating different levels of error into each of the attribute distributions during the creation process. Testing will be for: evaluating whether

there is variation in effectiveness of methods developed in phase 2 and 3 in an attempt to determine the validity of the clustering results. So in order for this research to explore the question of unifying a patient electronic medical record; the proposed research will develop:

- Machine Learning Data Mining rules or algorithms based on PHI contained in the returned records from domain of the data repository.
- This work will develop a novel framework for the identification and unification of patient data. This work may extend or combine existing methods or develop an entirely new approach.
- The end result will be the development of a foundation framework methodology to data mine records within the electronic medical record data depository.
- The development of a framework methodology for data integrity verification and correction through the use of belief theory.
- This work will allow a novel approach to the improvement of data quality. This work may extend or combine existing methods or develop an entirely new approach.
- Testing will be performed on a sample of a mathematical based synthetic data set that will emulate patients' electronic medical record data of the United States population.
- The scenario based analysis will present the effectiveness of methods developed in an attempt to determine the validity of the framework.

The goal of this area of the dissertation is to discover a metric or method to retrieve and return records that belong to a patient but are not identified by all phases of the framework. The research will utilize scenario based experiments to investigate the overall effectiveness of the framework phase by phase. The results will be presented systematically to appropriateness of the phases of the framework. Analytical insights will be discussed as to why the results allow a foundation for the creation of an information system to address current and future needs in the realm of electronic healthcare data. Intuitively, it would seem that different classes of problems would lead to different results. It is the inherent nature of the overall purpose of this research that all results come to the same conclusion. That a 1:1 match to all or as many as possible records belonging to a patient at the time of query.

CHAPTER 5 METHODOLOGY

5.1 FIRD Framework Architecture and Design

Recently, Healthcare's capabilities of both generating and collecting data have been increasing rapidly. This explosive growth in data and databases has generated an urgent need for new techniques and tools that can intelligently and automatically transform the processed data into useful information and knowledge. As with all research areas there are requirements to be met and challenges to face. Below is an overview of both the requirements and challenges facing this research and data mining in general. Some of which have been answered and others require development to validate the proposed framework. The rest of this section will present the challenges and requirements in data mining. The chapter will then introduce the FIRD data model in Section 5.2. Section 5.3 describes the Phase 1 method and request and response in order to support atomicity of each SQL statement in the FIRD framework. Phase 2 is presented for searching non biometric tagged records in Section 5.4. This is followed by a description of the data correction (Phase 3) in FIRD framework in Section 5.5. Finally Section 5.6 provides insights to the results of the scenario based analysis.

5.1.1 Requirements and challenges of data mining

In order to conduct effective data mining, one needs to first examine what kind of features an applied knowledge discovery system is expected to have and what kind of challenges one may face at the development of data mining techniques.

1. *Handling of different types of data.*

Because there are many kinds of data and databases used in different applications, one may expect that a knowledge discovery system should be able to perform effective data mining on different kinds of data. Since most available databases are relational, it is crucial that a data mining system performs efficient and effective knowledge discovery on relational data. Moreover, many applicable databases contain complex data types, such as structured data and complex data objects, hypertext and multimedia data, spatial and temporal data, transaction data, legacy data, etc. A powerful system should be able to perform effective data mining on such complex types of data as well. However, the diversity of data types and different goals of data mining make it unrealistic to expect one data mining system to handle all kinds of data. Specific data mining systems should be constructed for knowledge mining on specific kinds of data, such as systems dedicated to knowledge mining in relational databases, transaction databases, spatial databases, multimedia databases, etc.

2. *Efficiency and scalability of data mining algorithms.*

To effectively extract information from a huge amount of data in databases, the knowledge discovery algorithms must be efficient and scalable to large databases. That is, the running time of a data mining algorithm must be predictable and acceptable in large databases. Algorithms with exponential or even medium-order polynomial complexity will not be of practical use.

3. *Usefulness, certainty and expressiveness of data mining results.*

The discovered knowledge should accurately portray the contents of the database and be useful for certain applications. The imperfectness should be expressed by measures of uncertainty, in the form of approximate rules or quantitative rules. Noise and exceptional data should be handled elegantly in data mining systems. This also motivates a systematic study of measuring the quality of the discovered knowledge, including interestingness and reliability, by construction of statistical, analytical, and simulative models and tools.

4. *Expression of various kinds of data mining results.*

Different kinds of knowledge can be discovered from a large amount of data. Also, one may like to examine discovered knowledge from different views and present them in different forms. This requires us to express both the data mining requests and the discovered knowledge in high-level languages or graphical user interfaces so that the data mining task can be specified by non-

experts and the discovered knowledge can be understandable and directly usable by users. This also requires the discovery system to adopt expressive knowledge representation techniques.

5. *Interactive mining knowledge at multiple abstraction levels.*

Since it is difficult to predict what exactly could be discovered from a database, a high-level data mining query should be treated as a probe which may disclose some interesting traces for further exploration. Interactive discovery should be encouraged, which allows a user to interactively refine a data mining request, dynamically change data focusing, progressively deepen a data mining process, and flexibly view the data and data mining results at multiple abstraction levels and from different angles.

6. *Mining information from different sources of data.*

The widely available local and wide-area computer networks, including the Internet, connect many sources of data and form a huge distributed heterogeneous database. Mining knowledge from different sources of formatted or unformatted data with diverse data semantics poses new challenges to data mining. On the other hand, data mining may help disclose the high-level data regularities in heterogeneous databases which can hardly be discovered by simple query systems. Moreover, the huge size of the database, the wide distribution of

data, and the computational complexity of some data mining methods motivate the development of parallel and distributed data mining algorithms.

7. *Protection of privacy and data security.*

When data can be viewed from many different angles and at different abstraction levels, it threatens the goal of protecting data security and guarding against the invasion of privacy. It is important to study when knowledge discovery may lead to an invasion of privacy, and what security measures can be developed for preventing the disclosure of sensitive information. Notice that some of these requirements may carry conflicting goals. For example, the goal of protection of data security may conflict with the requirement of interactive mining of multiple -level knowledge from different angles. Moreover, this research addresses only some of the above requirements, with an emphasis on the efficiency and scalability of data mining algorithms.

5.1.2 An Overview of Data Mining Techniques

Since data mining poses many challenging research issues, direct applications of methods and techniques developed in related studies in machine learning, statistics, and database systems cannot solve these problems. It is necessary to perform dedicated studies to invent new data mining methods or develop integrated techniques for efficient

and effective data mining. In this sense, data mining itself has formed an independent new field.

5.1.2.1 Classifying data mining techniques

There have been many advances on researches and developments of data mining, and many data mining techniques and systems have recently been developed. Different classification schemes can be used to categorize data mining methods and systems based on the databases to be analyzed, what knowledge to be discovered, and the techniques to utilized, as shown below.

- Types of databases analyzed.

  A data mining system can be classified according to the kinds of databases on which the data mining is performed. For example, a system is a relational data miner if it discovers knowledge from relational data or an object-oriented one if it mines knowledge from object-oriented databases. In general, a data miner can be classified according to its mining of knowledge from the following different kinds of databases: relational databases, transaction databases, object oriented databases, deductive databases, spatial databases, temporal databases, multimedia databases, heterogeneous databases, active databases, legacy databases, and the Internet information-base.

- What knowledge to be mined.

  Several typical kinds of knowledge can be discovered by data miners, including association rules, characteristic rules, classification rules, discriminant rules, clustering, evolution, and deviation analysis, which will be discussed in detail in the next subsection. Moreover, data miners can also be categorized according to the abstraction level of its discovered knowledge which may be classified into generalized knowledge, primitive-level knowledge, and multiple-level knowledge. A flexible data mining system may discover knowledge at multiple abstraction levels.

- What techniques to utilize.

  Data miners can also be categorized according to the underlying data mining techniques. For example, it can be categorized according to the driven method into autonomous knowledge miner, data-driven miner, query-driven miner, and interactive data miner. It can also be categorized according to its underlying data mining approach into generalization-based mining, Pattern based mining; mining based on statistics or mathematical theories, and integrated approaches, etc. Among many different classification schemes, this research is on the classification scheme of the precise identification of all of a particular patient's electronic medical record; which is a subset of the data depository to be mined. It is because of such a unique demand that this particular classification problem exists. A clear picture on different data mining requirements and

techniques and their possible evolution is needed. General methods for mining different kinds of knowledge, including association rules, characterization, classification, clustering, etc. are examined in depth. For mining this particular kind of knowledge, different approaches, such as machine learning approach, statistical approach, and large database-oriented approach, are compared, with an emphasis on the database issues, such as efficiency and scalability.

5.1.2.2 Mining different kinds of knowledge from databases

Data mining is an application-dependent issue and different applications may require different mining techniques to cope with. In general cases, the kinds of knowledge which can be discovered in a database are categorized as follows. Mining association rules in transactional or relational databases has always been a topic of interest in the research community. The general task of the creation of association rule is to derive a set of strong association rules in the form of:

$$A_1 \square \ldots\ldots \square A_m \Rightarrow B_1 \square \ldots\ldots \square B_n$$

Where

A_i (for $i \in \{1 \ldots\ldots m\}$

And

B_j (for $j \in \{1 \ldots\ldots n\}$

are the sets of attribute-values, from the relevant data sets in a database. For example, on a large set of transaction data, an applicable association rule may look for, if a customer

buys (one brand of) milk, he/she usually buys (another brand of) bread in the same transaction. Since mining with association rules may require the system to repeatedly scan through a large transaction database to identify different association patterns, the amount of processing could be huge, and performance improvement is an essential concern at mining such rules. Efficient algorithms for mining association rules and some methods for further performance enhancements are often utilized. Some of the most popular data mining and data analysis tools associated with database system products are data generalization and summarization tools, which carry several alternative names, such as on-line analytical processing (OLAP), multiple-dimensional databases, data cubes, data abstraction, generalization, summarization, characterization, etc. Data generalization and summarization presents the general characteristics or a summarized high-level view over a set of user-specified data in a database. For example, the general characteristics of the technical staffs in a company can be described as a set of characteristic rules or a set of generalized summary tables. Moreover, it is often desirable to present generalized views about the data at multiple abstraction levels. Another important application of data mining is the ability to perform classification in a huge amount of data. This is referred to as mining classification rules. Data classification is to classify a set of data based on their values in certain attributes. A classic classification example is the one of an automobile dealership's goal was to say classify its customers according to their automobile preferences so that their sales staff will know which potential customers to approach and catalogs of new models can be mailed directly to those customers with identified features so as to maximize the business opportunity. Basically, data clustering is to group a set of

data (without a predefined class attribute), based on the conceptual clustering principle: maximizing the intra-class similarity and minimizing the interclass similarity. In the case of this research, a group of patients can be first clustered into a set of classes and then a set of rules can be derived based on such a classification. Such clustering may facilitate taxonomy formation, which means the organization of observations into a hierarchy of classes that group similar events together. Temporal or spatial-temporal data constitutes a large portion of data stored in computers. Examples of this type of database include: financial database for stock price index, medical databases, and multimedia databases, to name a few. Searching for similar patterns in a temporal or spatial-temporal database is essential in many data mining operations in order to discover and predict the risk, causality, and trend associated with a specific pattern. Typical queries for this type of database include identifying companies with similar growth patterns, products with similar selling patterns, stocks with similar price movement, images with similar weather patterns, geological features, environmental pollutions, or astrophysical patterns. These types of queries invariably require similarity matches as opposed to exact matches required for this research. In a distributed information providing environment, documents or objects are usually linked together to facilitate interactive access. Understanding user access patterns in such environments will not only help improving the system design but also be able to lead to better decisions. Capturing user access patterns in such environments is referred to as mining path traversal patterns.

5.1.3 Object-oriented Technology vs. Procedural Technology

Though the present invention relates to a particular OO technology (i.e., OO framework technology), the reader must first understand that, in general, OO technology is significantly different than conventional, process-based technology (often called procedural technology). While both technologies can be used to solve the same problem, the ultimate solutions to the problem are always quite different. This difference stems from the fact that the design focus of procedural technology is wholly different than that of OO technology. The focus of process-based design is on the overall process that solves the problem; whereas, the focus of OO design is on how the problem can be broken down into a set of autonomous entities that can work together to provide a solution. The autonomous entities of OO technology are called objects. Said another way, OO technology is significantly different from procedural technology because problems are broken down into sets of cooperating objects instead of into hierarchies of nested computer programs or procedures.

5.1.4 The Concept of a Framework

There has been an evolution of terms and phrases which have particular meaning to experts skilled in designing Object Oriented Databases. However, one of loosest definitions in the OO design is the definition of the word framework. The word framework means different things to different people. Therefore, when comparing the characteristics of two supposed framework mechanisms, the reader should take care to

ensure that the comparison is indeed "apples to apples." As will become evidently clear in the forthcoming paragraphs, the term framework is used in this research to describe a mechanism rooted in OO that has been designed to have core function and extensible function. The core function is that part of the framework mechanism that is not subject to modification by the framework user. The extensible function, on the other hand, is that part of the framework mechanism that has been explicitly designed to be customized and extended by the framework user.

5.2 FIRD Synthesize Data Model

Due to the mandate of HIPPA actual medical data for these experiments is unavailable for analysis. Since this framework is focused on data that resides in the virtual data depository of electronic medical data. It is necessary to create a data set to emulate that data. This medical data must comply with the HL7 (Health System Seven) standards.

The basic elements of this research are the presentation of the FIRD framework:

1. Outlining the framework for the correct identification of a patient's electronic medical data based on the biometric identifier query. (Phase 1)
2. The subsequent identification of a patient's correct electronic medical data in the absence of Phase 1 data and or Type 1 or Type 2 biometric error. (Phase 2)

3. The correction of missing or incorrect identification data within an electronic medical record. (Phase 3)

5.2.1 Medical Data Set Description

The synthesized medical data set contains the profiles of n = 1000 patients and has p = 17 attributes corresponding to the numeric and categorical attributes listed in Table 1. The data set contains the biometric identifiers for each areas of the FIRD framework phase 1 (**F**ingerprint, **I**ris, **R**etina and **D**NA) for each of the patients. The data set also contains the second phase of personal identification information protected by HIPPA which includes Date of Birth (DOB), race, gender ethnic background and blood type. The synthesized data set contains other information including a patients' place of birth region and state, Country of Birth, Date of enrollment for the creation of record Zero or date of visit to coincide with the creation of a medical record. The data set also contains naturalization status. Finally, the data set contains the patients' first and last name which is vital for the data fusion algorithms' to function properly.

Fingerprint Biometric	Iris Biometric	Retina Biometric	DNA Biometric	Patient First	Patient Last	DOB	Age	Race	Ethnic Background	Gender	Date of Enrollment/Date of Visit	Blood Type			Country of Birth	Naturalization
Fingerprint 0011	Iris 0011	Retina 0011	DNA 0011	First Name 0011	Last Name 0011	8/21/1967	40	Amer Indian Aleut or Eskimo	Central or South American	Female	8/11/2005	2	Northeast	Michigan	United-States	Native- Born in the United States
Fingerprint 0012	Iris 0012	Retina 0012	DNA 0012	First Name 0012	Last Name 0012	8/31/1924	83	Amer Indian Aleut or Eskimo	All other	Female	8/11/2004	3	Not in universe	Not in universe	United-States	Native- Born in the United States
Fingerprint 0025	Iris 0025	Retina 0025	DNA 0025	First Name 0025	Last Name 0025	9/1/1930	87	Amer Indian Aleut or Eskimo	All other	Female	8/11/2005	2	Not in universe	Not in universe	United-States	Native- Born in the United States
Fingerprint 0031	Iris 0031	Retina 0031	DNA 0031	First Name 0031	Last Name 0031	8/30/1931	76	Amer Indian Aleut or Eskimo	Mexican-American	Female	8/11/2005	2	Not in universe	Not in universe	United-States	Native- Born in the United States
Fingerprint 0039	Iris 0039	Retina 0039	DNA 0039	First Name 0039	Last Name 0039	8/16/1987	20	Amer Indian Aleut or Eskimo	Mexican-American	Male	8/11/2005	2	South	Utah	United-States	Native- Born in the United States
Fingerprint 0046	Iris 0046	Retina 0046	DNA 0046	First Name 0046	Last Name 0046	8/13/1996	11	Amer Indian Aleut or Eskimo	All other	Male	8/13/1999	8	Not in universe	Not in universe	United-States	Native- Born in the United States
Fingerprint 0065	Iris 0065	Retina 0065	DNA 0065	First Name 0065	Last Name 0065	9/1/1930	87	Amer Indian Aleut or Eskimo	Mexican-American	Male	8/12/2001	6	Not in universe	Not in universe	United-States	Native- Born in the United States
Fingerprint 0078	Iris 0078	Retina 0078	DNA 0078	First Name 0078	Last Name 0078	8/28/1939	68	Amer Indian Aleut or Eskimo	All other	Male	8/12/2002	5	Not in universe	Not in universe	Vietnam	Foreign born- Not a citizen of U S
Fingerprint 0089	Iris 0089	Retina 0089	DNA 0089	First Name 0089	Last Name 0089	8/26/1945	62	Amer Indian Aleut or Eskimo	All other	Female	8/11/2004	3	Not in universe	Not in universe	United-States	Native- Born in the United States
Fingerprint 0090	Iris 0090	Retina 0090	DNA 0090	First Name 0090	Last Name 0090	8/23/1956	51	Amer Indian Aleut or Eskimo	All other	Female	8/11/2004	3	Not in universe	Not in universe	United-States	Native- Born in the United States
Fingerprint 0138	Iris 0138	Retina 0138	DNA 0138	First Name 0138	Last Name 0138	9/4/1910	97	Asian	Mexican-American	Male	8/11/2006	1	Not in universe	Not in universe	United-States	Native- Born in the United States
Fingerprint 0146	Iris 0146	Retina 0146	DNA 0146	First Name 0146	Last Name 0146	9/3/1915	92	Asian	All other	Male	8/11/2005	2	Not in universe	Not in universe	United-States	Native- Born in the United States

Figure 4: Snapshot of Medical Data Set Attributes.

The attributes are designed as follows:

- The Biometric attribute is representative of the digitized standard of what each biometric attribute needs to positively identify an individual.
    - Fingerprint ID
    - Iris ID
    - Retina ID
    - DNA ID

- The PHI attributes are as follows:
    - The date of birth is designed of the two digit month of birth, two digit day of birth and the four digit year of birth
    - The racial backgrounds are:
        - American Indian Aleut or Eskimo
        - Asian
        - Black (Including African American)
        - Other
        - White
    - The Ethnic Background as deem by the United States 2000 census all races can contain a Hispanic ethnic background this synthesize data set contains this information by including the region of origin of the ethnic background (i.e. Mexican-American, Cuban, etc.).
    - Patients' gender is either classified as either Male or Female.
    - Date of enrollment/visit is designed of the two digit month of birth, two digit day of birth and the four digit year of enrollment or visit.
    - Blood type is coded as follows:

Blood Type	Positive Antigen	Negative Antigen
A	A+	A-
B	B+	B-
AB	AB	AB-
O	O+	O-

Table 12: Table of Blood Type Data Set Attributes.

5.2.2 Default Data Set Settings for the creation of Record Zero

The initial one thousand patients created for testing the framework is a subset of the correct sample size of the number of patients needed representative of the over 300 million people residing in the United States based on the Sample size equation:

$$ss = \frac{Z^2 * (p) * (1-p)}{c^2}$$

Where:

Z = Z value (for 99% confidence level)

p = percentage picking a choice, expressed as decimal

c = confidence interval, expressed as decimal

(e.g., .01 = ±1)

The equations' result is a total of 16640 patients' enrollment records. The functionality of the framework is the key to the research so a subset of the sample size of patient complete electronic medical records is adequate for experimentation purposes. The 1000 patients were divided by the demographic information from the United States year 2000 census. Once the patients were placed in the appropriate age, gender and racial/ethnic group their enrollment record was created with the information in table 1. These records

are the primary records for the virtual medical data depository since they contain all of a patient's critical identification information that is vital for phase 2 and 3 of the FIRD framework. The default data settings for each of the patient's record are based on the following:

- Fingerprint ID is denoted by Fingerprint XXXX
  Where X is the patient number
- Iris ID is denoted by Iris XXXX
  Where X is the patient number
- Retina ID is denoted by Retina XXXX
  Where X is the patient number
- DNA is denoted by DNA XXXX
  Where X is the patient number

- Patient's First Name is denoted by First Name XXXX
  Where X is the patient number
- Patient's Last Name is denoted by Last Name XXXX
  Where X is the patient number
- Date of Birth (DOB) is denoted by XX/XX/XXXX
- Age is denoted by XXX
  Where X is the patient age in years

5.2.3 Expansion of 'Record Zero' data set for the creation of Synthesize virtual data depository

Once the initial patient database is develop, now the data set needs to be transformed or expanded to increase the number of records to analyze the search and correction capabilities algorithms developed for phases 2 and 3. The expansion parameter decided upon was to take the default values and create a virtual medical data depository based on the American Medical Association (AMA) advice of semiannual medical checkup for everyone. This expansion would allow for each patient to generate a new medical record twice a year for every year of their life from birth. This means that each patient will have 2X + 1 records where X is their age in years. Using the above formula;

the expansion of the initial 1000 patients' records result in a database with a total of 98564 records. The clustering problem of the first layer of FIRD framework is a multi layer one (as opposed to multi-class which usually refers to simply having more than two possible disjoint classes for the classifier to learn). Moreover, the first phase of the framework is not looking for a classifier to give a range of possible/probable classes.

5.3 Correct identification of a patient's electronic medical data based on the biometric identifier query. (Phase 1)

This dissertation uses a multi layered approach of the biometrics identifiers but incorporates the scalability within its foundation to use multi modal biometrics when the technology is becomes widely available and cost effective. While the biometrics are used to verify a patient's existence but more so to identify and verify the correct electronic medical records are collected and returned for that patient. The FIRD framework will scan through all medical records stored within the virtual data depository and return records that match the query request.

Depending on what criteria a returned record meets it will fall into two types:

- Type 1-1: Record conforms to conjunction rule meaning it has all four biometric identifiers
- Type 1-2: Record conforms to disjunction rule meaning it has one, two or three of the four biometric identifiers

The query is the patient requesting his or her record. There is a record that is created that contains a patients' extraction sample and registration record with all of his

or her personal and identifiable information this record is called Record Zero. As shown in the figure below the area the is in circle in red is phase 1 of the framework

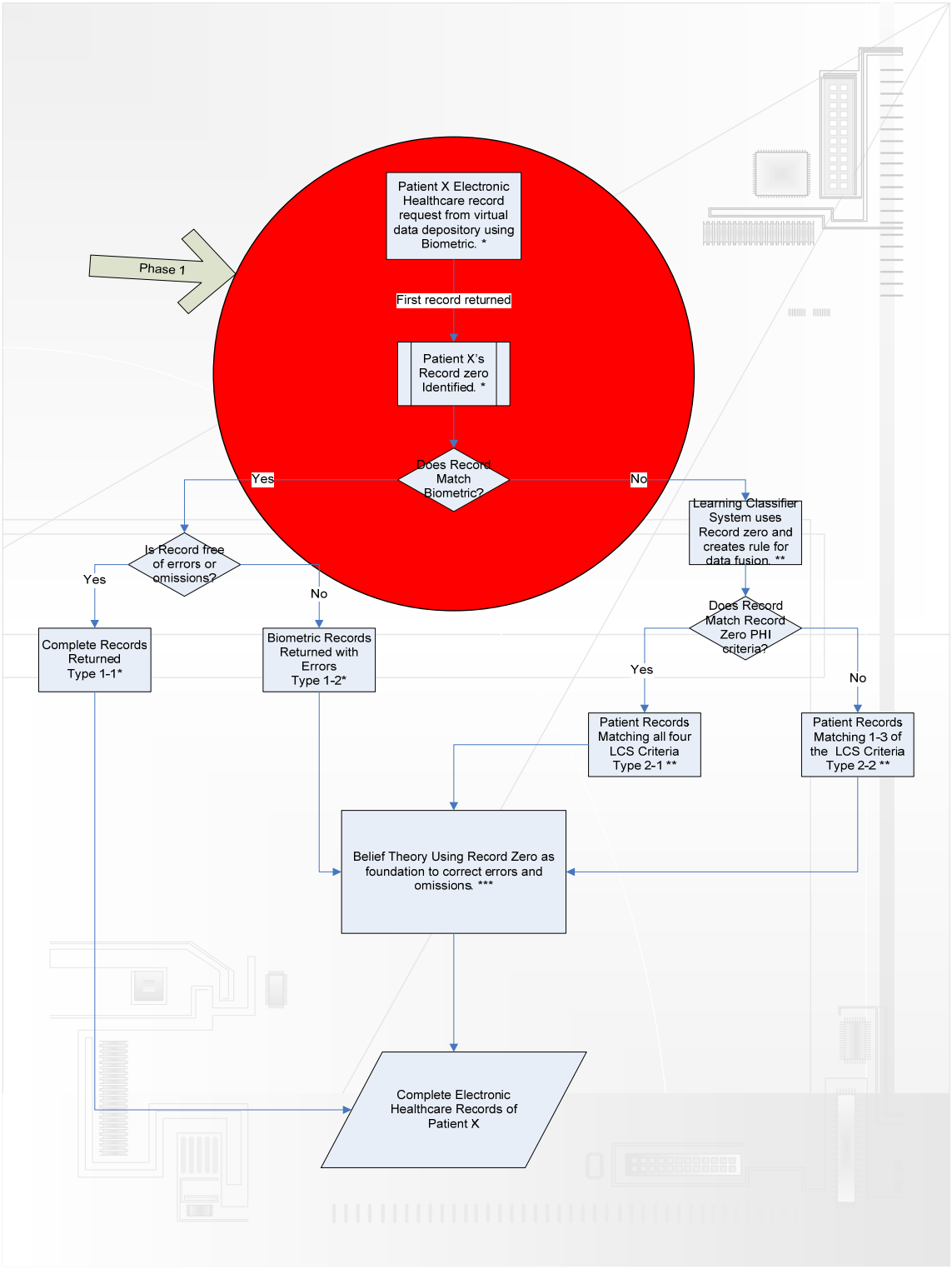


Figure 5: Highlighted Area of Phase 1 of the FIRD Framework

The framework returns records that indicate they contain the same biometric identifier as the patients' extraction sample and registration record. Records that are returned containing all four biometric ID's contained in the patients' extraction sample and registration record are considered Type 1-1 records.

- $P_1(FA) + P_2(FA) + P_3(FA) + P_4(FA) – P_1(FR) \cdot P_2(FR) \cdot P_3(FA) \cdot P_4(FA)$

If a record is returned and it contains one, two, or three of the four biometric ID's contained in the patients' extraction sample and registration record. It is considered Type 1-2 records:

- $P_1(FA) + P_2(FA) + P_3(FA) + P_4(FA) – P_1(FA) \cdot P_2(FA) \cdot P_3(FA) \cdot P_4(FA)$

The developed Biometric phase is layered system designed to use one or all of the identifiers outlined. If all are identifiers are included the system will have a 99.9999% accuracy rate and will allow the framework to encompass all possible populations. Since the technology for instantaneous identification of the Iris, Retina, and DNA are not wide spread or do not exist and fingerprints are accepted by society then the framework needs a way to compensate for the error rate thus the second phase. The second phase also will return records that are not associated with a biometric tag.

Since the data in the patient database can be stored in either a relational or object oriented database. The goal of phase 1 of the framework is to create an inclusion/exclusion process based on a biometric query, to cope with the Healthcare DBMS environments. In the FIRD framework the attempt is to architect and implement a continuously adaptive query engine suitable for global-area systems, massive parallelism,

and sensor networks of healthcare data. While in general terms the FIRD framework mechanism can be properly characterized as a specialized OO solution, there is nevertheless a fundamental difference between a framework mechanism and a basic OO solution. The difference is that FIRD framework mechanisms are designed in a way that permits and promotes customization and extension of certain aspects of the solution. In other words, framework mechanisms amount to more than just a solution to the problem. The framework mechanisms provide a living evolving solution that can be customized and extended to address individualized requirements that change over time. Of course, the customization/extension quality of framework mechanisms is extremely valuable to users (referred to herein as framework users) because the cost of customizing or extending a framework is much less than the cost of a replacing or reworking an existing solution.

Therefore, when this research set out to solve this particular problem, it was more than merely design individual objects and how those objects interrelate. The research also designed the core function of the framework (i.e., that part of the framework that is not to be subject to potential customization and extension by the framework user) and the extensible function of the framework (i.e., that part of the framework that is to be subject to potential customization and extension). In the end, the ultimate worth of a framework mechanism rests not only on the quality of the object design, but also on the design choices involving which aspects of the framework represent core function and which aspects represent extensible function.

Since it has yet to be determined which form the virtual data depository for electronic medical data will present itself. This research was developed using Object-oriented (OO) framework technology. Since the goal of this research is to see that move from paper-based records to electronic records truly lead to quality healthcare. Its database should not be considered a database in the truest sense of the word and governed by those rules DBMS. The virtual data depository or healthcare data store should be considered an Information or Knowledge base and governed by the rules of Knowledge Discovery Management System (KDMS). A KDMS is a system which would provide storage, querying and further mining on previously discovered knowledge [Imielinski and Virmani, 1999]. In such a system the discovered knowledge from possibly many mining sessions (and different data miners) can be stored and queried using the M-SQL features. Furthermore, one can build upon the knowledge, or "collective wisdom", accumulated over time. This type of system is ideal for the environment of healthcare data. This research is designed to develop a framework with KDMS in mind.

5.3.1 Basic Concepts of Phase 1

This research will now formalize some of the basic concepts used in phase one of the framework and throughout the remainder of this research. The concept of a descriptor is an expression of the form *(Ai = aij),* where *aij* belongs to the domain of *Ai.* For continuous valued attributes, a descriptor of the form (A~= *[lo, hi])* is allowed, where *[lo, hi]* represents a range of values over the domain of *Ai.* A conjunctset stands for a

conjunction of an arbitrary number of descriptors, such that no two descriptors are formed using the same attribute. The length of a conjunctset is the number of descriptors which form the conjunctset. A descriptor is thus the special case of a singleton conjunctset. A record (tuple) in R is said to satisfy a descriptor *(Ai, = vii),* if the value of *Ai* in the record equals *vii.* To satisfy a conjunctset, a record must satisfy all k descriptors forming the conjunction [Imielinski and Virmani, 1998].

Example: Let R be a relation represented by the table shown below:

EmpId	Job	Sex	Car
1	Doctor	Male	BMW
2	Lawyer	Female	Lexus
3	Consultant	Male	Toyota
4	Doctor	Male	Volvo

Table 13: Relation table of Occupation, Gender and Vehicle Choice

Then (Job = Doctor) is an example of a descriptor in the above data, satisfied by records with EmpId values 1 and 4. Along the same lines, (Sex = Female) A (Car = Lexus) is an example of a conjunctset of length 2 in the above data. By a propositional rule over R, we mean a tuple of the form (B, C, s, c), where B is a conjunctset called the Body of the rule, C is a descriptor called the Consequent of the rule, s is an integer called the support of the rule, and c is a number between 0 and 1 called the confidence of the rule. Support is defined as the number of tuples in R which satisfy the body of the rule, and confidence is defined as the ratio of the number of tuples satisfying both the body and the consequent

to the number of tuples which satisfy just the body of the rule [Imielinski and Virmani, 1999].

Intuitively, these rules are in the form of if-then statements where "if Body then Consequent", with support (denoted by s) and confidence (denoted by c) being the quality measures computed within relation R. This is represented by association rules in the following syntactic form:

Body ==~ Consequent [support, confidence]

Given the above example, the following is a rule based on that relation:

(Job = Doctor) A (Sex = Male) ====> (Car = BMW) [2, 0.5]

A rule in this case is a generalization of the association discussed earlier. Since this framework has defined procedures and functions triggered by the user in our case the patient. The expressive power of propositional rules is actually, for all practical purposes equivalent to non-recursive predicate rules. These rules can also be viewed as a query when applied to a relation. For example a relation R satisfies a rule r = (B, C, s, c) if there are at least s tuples in R which satisfy B and at least a fraction c of them satisfy the conjunction B A C. This can also be expressed by saying that r holds true in R. If R does not satisfy r, then it can be said that R violates r, or, r does not hold true in R. Generally these rules represent aggregates over a set of tuples, in that case the relationship between a rule and an individual tuple cannot be similarly defined. However if only the rule-

pattern (B, C) is considered without the associated support and confidence, the following relationships between can be defined. A tuple t satisfies a rule pattern (B, C), if it satisfies the conjunction B A C, and it violates the above pattern if it satisfies B, but not C.

Using this formulation as the foundation of phase 1, this research adopted to use MSQL. MSQL is a language developed with the SQL92 standard yet adds support for rule-manipulation operations in a familiar SQL-like syntax. Below are an overview and a how the query language was used to construct Phase 1. MSQL can be described under four main subsections, as shown below.

<MSQL Stmt> ::=	<GetRules query>	Rule-generation
|	<SelectRules Query>	Query rules from existing rulebase
|	<SatSatisfy SubQuery>	GetRules Subquery w/where clause
|	<SatViolate SubQuery>	
|	<Encode Stmt>	Provides pre/post processing

Table 14: MSQL Four Subsections and commands

An overview of the above code is as follows. The GetRules query is used for rule-generation, and the SelectRules query, which follows the same syntax and is used to query rules from an existing rulebase. In addition, a standard SQL query on a database table can have a nested GetRules sub-query in its "where" clause connected via the Satisfy or Violate keyword. Syntax for this clause is referred to as the Sat-Violate-SubQuery statement. The Encode statement provides pre- and post-processing support for continuous valued attributes. These primitives can also be supported in the object oriented API, where a relational table corresponds to a class, which in that case the terms "table" and "class" become interchangeable.

5.3.2 General query syntax

The most general formulation of the GetRules Query is as follows where C is a database table, and R1 is an alias for the generated rulebase.

[Project Body, Consequent, confidence, support]
GetRules(C) [as R1]
[into <rulebase_name>]
[where <conds>]
[sql-group-by clause]
[using-clause]

In addition, (Conds) may itself contain:

<Rule Format Conditions RC> |
<Pruning Conditions PC> |
<Mutex Conditions MC> |
<Stratified Subquery Conditions SSQ> |
<Correlated Subquery Conditions CSQ>

Figure 6: Pseudo Code of GetRules Query and commands

The GetRules operator generates rules over elements of the argument class C, satisfying the conditions described in the "where" clause. The results are placed into a rule class optionally named by the user, else named by suffixing 'RB' to the name of the source class. So for patient database created in section 5.2, the rulebase PatientRB would be generated. The projection and group-by operations can optionally be applied, and their meaning is the same as defined in SQL. Since they basically post-process the generated rules, they do not affect the semantics of rule generation.

Another important thing to point out is that the GetRules query operates on the complete class C, rather than a subset of it. There is a difference between the two classes. All rules from the subset of data with (A1 = a) in them is not the same as all rules on the whole data with (A1 = a) in the Body, since if we subset and then mine for rules, the confidence and support in the rules generated will change. Besides, if one mines for rules about a subset of the data, then technically, it is a different class and therefore, there should be a different rulebase corresponding to it.

Given the above reasoning, the GetRules operator disallows any "where" clause conditions on pure attributes of the source class C. These can always be performed by creating a view on C with the appropriate selections/projections and then using GetRules on that particular view. The only conditions allowed are the ones on rule components: Body and Consequent. Note that the evaluation of GetRules internally may involve selecting/projecting the data for efficiency, but it will preserve the query semantics.

5.3.3 Generating and retrieving rules for Phase 1 Record Zero

All examples in this section are based on the creation of the synthetic dataset from section 5.2 of the patient database in the following schema:

Patient(Fingerprint_ID, Iris_ID, Retina_ID, DNA_ID, Patient_Fname, Patient_Lname, DOB, Age, Race, Ethnic Background, Gender, Date of Enrollment/Date of visit, Blood type, State of Birth, Region, Country of Birth, Nationalization)
To generate rules from the Patients table, one uses the GetRules command
as follows:
GetRules(Patient)
where Consequent in { (DOB=DOE),
and Body has { (Fingerprint_ID=Biometric Query) }
and confidence >1.0
and support > 1.0
}

Figure 7: GetRules Schema for the Patients table

The above query would generate rules to return records' having the fingerprint of the biometric query in the antecedent where as the Consequent methods verifies that the values of date of birth (DOB) and date of enrollment (DOE) are the same. The corresponding association rule for the above query is given below:

For Fingerprint Biometric Query (FPIDQ):
IF (FPIDQ= FPID)
AND (DOB= DOE)
THEN RECORD = RECORD ZERO

Figure 8: GetRules query schema Association rules for the identification of 'Record Zero based on the fingerprint biometric

The following schemas are for the remaining three biometric identifiers (Iris, Retina, and DNA) along with their association rules:

Patient(Fingerprint_ID, Iris_ID, Retina_ID, DNA_ID, Patient_Fname, Patient_Lname, DOB, Age, Race, Ethnic Background, Gender, Date of Enrollment/Date of visit, Blood type, State of Birth, Region, Country of Birth, Nationalization)
To generate rules from the Patients table, one uses the GetRules command
as follows:
GetRules(Patient)
where Consequent in { (DOB=DOE),
and Body has { (Iris_ID=Biometric Query) }
and confidence >1.0
and support > 1.0
}

For Iris Biometric Query (IIDQ):
IF (IIDQ= IID)
AND (DOB= DOE)
THEN RECORD = RECORD ZERO

Figure 9: GetRules query schema and Association rules for the identification of 'Record Zero based on Iris biometric request.

Patient(Fingerprint_ID, Iris_ID, Retina_ID, DNA_ID, Patient_Fname, Patient_Lname, DOB, Age, Race, Ethnic Background, Gender, Date of Enrollment/Date of visit, Blood type, State of Birth, Region, Country of Birth, Nationalization)
To generate rules from the Patients table, one uses the GetRules command
as follows:
GetRules(Patient)
where Consequent in { (DOB=DOE),
and Body has { (Retina_ID=Biometric Query) }
and confidence >1.0
and support > 1.0
}

For Retina Biometric Query (RIDQ):
IF (RIDQ= RID)
AND (DOB= DOE)
THEN RECORD = RECORD ZERO

Figure 10: GetRules query schema and Association rules for the identification of 'Record Zero based on the Retina biometric request.

Patient(Fingerprint_ID, Iris_ID, Retina_ID, DNA_ID, Patient_Fname, Patient_Lname, DOB, Age, Race, Ethnic Background, Gender, Date of Enrollment/Date of visit, Blood type, State of Birth, Region, Country of Birth, Nationalization)
To generate rules from the Patients table, one uses the GetRules command
as follows:
GetRules(Patient)
where Consequent in { (DOB=DOE),
and Body has { (DNA_ID=Biometric Query) }
and confidence >1.0
and support > 1.0
}

For DNA Biometric Query (DNAIDQ):
IF (DNAIDQ = DNAID)
AND (DOB= DOE)
THEN RECORD = RECORD ZERO

Figure 11: GetQuery Schema and Association rules for the identification of 'Record Zero based on DNA request.

Upon the arrival of record zero for the patient query, a set of sub-queries based on the remaining biometric identifiers are set in motion to return results of the remaining records belonging to the patient within the patient query. As stated earlier Type 1-1 records; are records that include all four biometric identifiers and all four PHI identifiers.

Also if you recall from chapter four the biometric phase is a multi layer/multi modal biometric type of system which looks at both conjunction (“and”) and disjunction (“or”). So this research requires two distinct sets of subquery systems one for conjunction (Type 1-1) and one for disjunction (Type 1-2).

5.3.4 Creation of Conjunction and Disjunction SubQuery systems for patient record retrieval

5.3.4.1 Type 1-1 Records-Conjunction

In order to search the space of virtual medical data depository based on the biometric request of electronic patient data this framework has to be able to distinguish if the records returned are of which sub type in the Type 1 either conjunction or disjunction. Where the returned record falls in to the conjunction category if it contains both the biometric identifier used in the initial patient request as well as the other three biometric identifiers verified from record zero of the requesting patient. The schema presented in the figure below is used to explicitly evaluate the virtual data depository based to return these types of records based on the rule patterns generated. The following query, in this case will select the records from the rules from R, and evaluate them across all databases.

Patient(Fingerprint_ID, Iris_ID, Retina_ID, DNA_ID, Patient_Fname, Patient_Lname, DOB, Age, Race, Ethnic Background, Gender, Date of Enrollment/Date of visit, Blood type, State of Birth, Region, Country of Birth, Nationalization)
To generate rules from the Patients table, one uses the GetRules command
as follows:
GetRules(Patient)
where Consequent in { (Patient_Fname=Patient_Fname, Patient_Lname= Patient_Lname),
and Body has { (Fingerprint_ID=Fingerprint_IDR0), (Iris_ID=Iris_IDR0), (Retina_ID=Retina_IDR0), (DNA_ID=DNA_IDR0), (DOB=DOBR0), (Race=RaceR0),Ethnic Background=Ethnic BackgroundR0), Gender=GenderR0, Blood Type=Blood TypeR0) }
and confidence >1.0
and support > 1.0
}

For Type 1-1 Records from Based on Record Zero (1-1ID):
IF (Fingerprint_ID=Fingerprint_IDR0, Iris_ID=Iris_IDR0, Retina_ID=Retina_IDR0, DNA_ID=DNA_IDR0, Patient_Fname=Patient_Fname, Patient_Lname= Patient_Lname, (DOB=DOBR0), (Race=RaceR0),Ethnic Background=Ethnic BackgroundR0), Gender=GenderR0, Blood Type=Blood TypeR0)
THEN RETURN RECORD

Figure 12: GetQuery Schema and Association rules for the identification of patient records meeting Type 1-1 criteria.

The above returns records based on the Rulesbase generated by the Patient table across the virtual data depository containing the attributes required in the query (in this case, Fingerprint, Iris, Retina and DNA) from record zero which was returned based on the patient biometric query. In the case, the query will only return all records that contain all

four correct biometric identifiers matching the enrollment record of the patient requesting the records through a biometric query. In addition to the biometric identifiers, this query will return records containing all four PHI identifiers as well. This is due to the fact for a record to be considered a Type 1-1 it must contain and return all four correct identifiers from both biometric and PHI. If there is an error in any of the four biometric identifiers these records are not considered type 1-1. In order to capture the records that either contain all four biometrics identifiers (with errors) or records that may contain one or more of the correct biometric identifiers the disjunction rule must be utilized.

5.3.4.2 Type 1-2 Records-Disjunction

The disjunction SubQuery system is a more complicated case of the use of SelectRules query. The disjunction rule allows the system to test the records in the virtual data depository for any of the four biometric identifiers that are in the record zero returned from the patients biometric request regardless of the biometric used in the request. The system could contain a sub set of the attributes required but not all four. In order to determine if these records do belong and require some form of data correction, an approach to deals with the uncertainties in the execution of biometrics error rates by using the sub-queries, and if sub-query results are materialized below, they can also cope with records that have one or more biometric identifiers missing within the electronic medical record (which is precursor of phase 3) to some extent. Therefore if a patients'

initial request was his or her fingerprint; then that patients' records that were not returned from their fingerprint can be returned based on any of the other 3 biometrics matching the identifiers in their record zero. The records returned in this SubQuery system are considered Type 1-2 and require a level data correction provided in Phase 3. The figure below demonstrates the schema for SubQuery System:

Patient(Fingerprint_ID, Iris_ID, Retina_ID, DNA_ID, Patient_Fname, Patient_Lname, DOB, Age, Race, Ethnic Background, Gender, Date of Enrollment/Date of visit, Blood type, State of Birth, Region, Country of Birth, Nationalization)
To generate rules from the Patients table, one uses the GetRules command
as follows:
GetRules(Patient)
where Consequent in { (Patient_Fname=Patient_Fname, Patient_Lname= Patient_Lname),
and Body has { (Fingerprint_ID=Fingerprint_IDR0) }
and confidence >1.0
and support > 1.0
}

For Fingerprint Biometric from Record Zero (Fingerprint_IDR0):
IF (FPID= FPIDR0, Patient_Fname=Patient_Fname, Patient_Lname= Patient_Lname)
THEN RETURN RECORD

Patient(Fingerprint_ID, Iris_ID, Retina_ID, DNA_ID, Patient_Fname, Patient_Lname, DOB, Age, Race, Ethnic Background, Gender, Date of Enrollment/Date of visit, Blood type, State of Birth, Region, Country of Birth, Nationalization)
To generate rules from the Patients table, one uses the GetRules command
as follows:
GetRules(Patient)
where Consequent in { (Patient_Fname=Patient_Fname, Patient_Lname= Patient_Lname),
and Body has { (Iris_ID=Iris_IDR0) }
and confidence >1.0
and support > 1.0
}

For Iris Biometric from Record Zero (Iris_IDR0):
IF (Iris_ID=Iris_IDR0 Patient_Fname=Patient_Fname, Patient_Lname= Patient_Lname)
THEN RETURN RECORD

Patient(Fingerprint_ID, Iris_ID, Retina_ID, DNA_ID, Patient_Fname, Patient_Lname, DOB, Age, Race, Ethnic Background, Gender, Date of Enrollment/Date of visit, Blood type, State of Birth, Region, Country of Birth, Nationalization)
To generate rules from the Patients table, one uses the GetRules command
as follows:
GetRules(Patient)
where Consequent in { (Patient_Fname=Patient_Fname, Patient_Lname= Patient_Lname),
and Body has { (Retina_ID=Retina_IDR0) }
and confidence >1.0
and support > 1.0
}

For Retina Biometric from Record Zero (Retina_IDR0):
IF (Retina_ID= Retina_ID, Patient_Fname=Patient_Fname, Patient_Lname= Patient_Lname)
THEN RETURN RECORD

Patient(Fingerprint_ID, Iris_ID, Retina_ID, DNA_ID, Patient_Fname, Patient_Lname, DOB, Age, Race, Ethnic Background, Gender, Date of Enrollment/Date of visit, Blood type, State of Birth, Region, Country of Birth, Nationalization)
To generate rules from the Patients table, one uses the GetRules command
as follows:
GetRules(Patient)
where Consequent in { (Patient_Fname=Patient_Fname, Patient_Lname= Patient_Lname),
and Body has { (DNA_ID=DNA_IDR0) }
and confidence >1.0
and support > 1.0
}

For DNA Biometric from Record Zero (DNA_IDR0):
IF (DNA_ID = DNA_ID Patient_Fname=Patient_Fname, Patient_Lname= Patient_Lname)
THEN RETURN RECORD

Figure 13: Association rules for the identification of sub-queries based on 'Record Zero' to return records that meet the criteria of Type 1-2

Since the virtual medical data depository applies to many different types of medical databases housing electronic patient data in a standardize format(HL7) and a system to exchange that data between parties (interoperability) this framework uses above schema to explicitly evaluate the virtual data depository based on the rule patterns generated. For instance, if FL_Patients and CA_Patients are two different statewide healthcare databases defined on the Patient table/object in the states of Florida and

California respectively. If the patient that prompted the request had medical records in these states then his or her records will be retrieved based on the rulebase returned by the biometric query. The following query, in this case will select the records from the rules from R, and evaluate them across all databases.

```
Project Body, Consequent, Confidence(FL_Patients), Support(FL_Patients),
Confidence (CA_Patients), Support (CA_Patients),
SelectRules (R)
where Body has { (Age=*), (Sex=*) }
and Consequent is { (Query=*) }
```

Figure 14: GetQuery Schema and Association rules example to display how the interoperability across multiple databases for the identification patients' record based on query request.

Please note that there is an implicit line before the GetRules command:

Project Body, Confidence, Confidence(Patient), support(Patient)

The command line above can be altered to evaluate existing rules on different sets allowing the rulebase to apply across the virtual medical data depository. Typically this mechanism is commonly used in query based mining where users first do a fairly general GetRules query and store the result persistently, and then follow with a series of SelectRules queries, each of which selects a small subset of the entire rulebase. To generate rules for a given database table, the GetRules operator must be used with a table argument as follows:

GetRules (T)
into R
where confidence > 1.0
and support > 1.0

Figure 15: GetRules Schema to generate rules matching confidence and support requirements.

This will generate all rules existing in table T matching the confidence and support requirements, and put them in a persistent rulebase named R. For these sub-queries rules, the language has the SelectRules command. SelectRules will not generate any new rules, but rather rely on the contents of the argument rulebase for providing results. For instance, the following query retrieves rules with at least Age and Sex in the Body and the car driven as a Consequent.

SelectRules **(R)**
where Body has { (Age=*), (Sex=*) }
and Consequent is { (Query=*) }

Figure 16: SelectRules Schema to generate rules matching confidence and support requirements.

Note that by default, the expected confidence and support of the rules produced by the phase 1 queries is 100 percent or a 1:1 match, since those were the parameters R was mined with.

As stated earlier the Project operator is implicit. Since the virtual medical data depository applies to many different types of medical databases housing electronic patient data in a standardize format(HL7) and a system to exchange that data between

parties (interoperability) this framework can use the Projection to explicitly evaluate rule patterns over various databases.

```
Project Body, Consequent, Confidence(X_Patients), Support(X_Patients ),
Confidence (X_Patients) , Support (X_Patients ),
SelectRules (R)
where Body has { (Age=*), (Sex=*) }
and Consequent is { (Query=*) }
```

Figure 17: GetRules Schema to generate rules matching confidence and support requirements for interoperability across databases.

SelectRules, by definition, does not generate new rule patterns. The above example brings up an interesting issue: What if R is not a rulebase generated by the Patients table, but rather, by some other table? There are two possible scenarios. In the simpler case, R could be a rulebase not containing the attributes required in the query (in this case, Age, Sex and Request Query). In that case, the query will be syntactically incorrect and will return an error. In a more complicated case, the rule table and the other data tables in the above type of SelectRules query could both contain the attributes required by the "where" clause, even when they semantically meant something totally different. Understanding that the language should enforce this "typing" between rulebases and databases; The research allows the framework design to follow strong typing between rulebases and databases gives the flexibility to interact with any API. Due to the ability to treat both rulebases and datasets as untyped relational tables in MSQL.

5.4 Subsequent identification of a patient's correct electronic medical data in the absence of Phase 1 data (Phase 2)

5.4.1 Generating and retrieving rules for Phase 2 from Record Zero

As defined in the previous section the following are the database schema for identification of a patient's record zero are same as for Phase 1:

Patient(Fingerprint_ID, Iris_ID, Retina_ID, DNA_ID, Patient_Fname, Patient_Lname, DOB, Age, Race, Ethnic Background, Gender, Date of Enrollment/Date of visit, Blood type, State of Birth, Region, Country of Birth, Nationalization)
To generate rules from the Patients table, one uses the GetRules command
as follows:
GetRules(Patient)
where Consequent in { (DOB=DOE),
and Body has { (Fingerprint_ID=Biometric Query) }
and confidence >1.0
and support > 1.0
}

Figure 18: GetRules Schema for the Patients table from Phase 1

However, in the case of phase 2 the set of sub-queries are based on the PHI information contained in the patient's records. These sub-queries search for records belonging to the patient within the patient query yet do not have a biometric tag associated with it. As stated earlier Type 2 records; are records that do not contain biometric identifiers but contain the PHI attributes of the patient query. As with the case of phase 1, phase 2 mechanisms classify its subquery systems into two distinct sets one for conjunction (Type 2-1) and one for disjunction (Type 2-2).

5.4.2 Creation of Conjunction and Disjunction SubQuery systems for patient record retrieval based on PHI attributes

5.4.2.1 Type 2-1 Records-Conjunction

In order to search the space of virtual medical data depository based on the biometric request of electronic patient data this framework has to be able to distinguish if the records returned are of which sub type in the Type 2 either conjunction or disjunction. Where the returned record falls in to the conjunction category if it contains all four of the PHI attributes identifiers from the initial patient request as verified from record zero of the requesting patient. The schema presented in the figure below is used to explicitly evaluate the virtual data depository based to return these types of records based on the rule patterns generated. The following query, in this case will select the records from the rules from R, and evaluate them across all databases.

Patient(Fingerprint_ID, Iris_ID, Retina_ID, DNA_ID, Patient_Fname, Patient_Lname, DOB, Age, Race, Ethnic Background, Gender, Date of Enrollment/Date of visit, Blood type, State of Birth, Region, Country of Birth, Nationalization)
To generate rules from the Patients table, one uses the GetRules command
as follows:
GetRules(Patient)
where Consequent in { (Patient_Fname=Patient_Fname, Patient_Lname= Patient_Lname),
and Body has { (DOB=DOBR0), (Race=RaceR0),Ethnic Background=Ethnic BackgroundR0), Gender=GenderR0, Blood Type=Blood TypeR0) }
and confidence >1.0
and support > 1.0
}

For Type 2-1 Records from Based on Record Zero (2-1ID):
IF (Patient_Fname=Patient_Fname, Patient_Lname= Patient_Lname, (DOB=DOBR0, Race=RaceR0,Ethnic Background=Ethnic BackgroundR0, Gender=GenderR0, Blood Type=Blood TypeR0)
THEN RETURN RECORD

Figure 19: GetQuery Schema and Association rules for the identification of patients' record based on PHI attributes meeting the criteria for Type 2-1 records.

The above returns records based on the rulebase generated by the Patient table across the virtual data depository containing the attributes required in the query (in this case, Date of Birth (DOB), Race, Ethnic Background, Gender, and Blood type) from record zero which was returned based on the patient query. In this case, the query will return records containing all four PHI identifiers matching the enrollment record of the patient requesting the records through a biometric query. This is due to the fact for a record to be considered a Type 2-1 it must contain and return all four correct identifiers based on the PHI attributes. If there is an error in any of the PHI attributes identifiers these records cannot be considered type 2-1. In order to capture the records that either contain all four PHI attribute identifiers (with errors) or records that may contain one or more of the correct PHI attribute identifiers the disjunction rule must be utilized.

5.4.2.2 Type 2-2 Records-Disjunction

The disjunction SubQuery system is a more complicated case of the use of SelectRules query. The disjunction rule allows the system to test the records in the virtual data depository for any of the four biometric identifiers that are in the record zero returned from the patients biometric request regardless of the biometric used in the request. The system could contain a sub set of the attributes required but not all four. In order to determine if these records do belong and require some form of data correction, an approach to deals with the uncertainties in the execution of biometrics error rates by using the sub-queries, and if sub-query results are materialized below, they can also cope with records that have one or more biometric identifiers missing within the electronic medical record (which is precursor of phase 3) to some extent. Therefore if a patients' initial request was his or her fingerprint; then that patients' records that were not returned from their fingerprint can be returned based on any of the other 3 biometrics matching the identifiers in their record zero. The records returned in this SubQuery system are considered Type 1-2 and require a level data correction provided in Phase 3. The figure below demonstrates the schema for SubQuery System:

Patient(Fingerprint_ID, Iris_ID, Retina_ID, DNA_ID, Patient_Fname, Patient_Lname, DOB, Age, Race, Ethnic Background, Gender, Date of Enrollment/Date of visit, Blood type, State of Birth, Region, Country of Birth, Nationalization)
To generate rules from the Patients table, one uses the GetRules command
as follows:
GetRules(Patient)
where Consequent in { (Patient_Fname=Patient_Fname, Patient_Lname= Patient_Lname),
and Body has { (DOB=DOBR0)}
and confidence >1.0
and support > 1.0
}

For Date of Birth PHI from Record Zero (DOB=DOBR0):
IF (DOB=DOBR0, Patient_Fname=Patient_Fname, Patient_Lname= Patient_Lname)
THEN RETURN RECORD

Patient(Fingerprint_ID, Iris_ID, Retina_ID, DNA_ID, Patient_Fname, Patient_Lname, DOB, Age, Race, Ethnic Background, Gender, Date of Enrollment/Date of visit, Blood type, State of Birth, Region, Country of Birth, Nationalization)
To generate rules from the Patients table, one uses the GetRules command
as follows:
GetRules(Patient)
where Consequent in { (Patient_Fname=Patient_Fname, Patient_Lname= Patient_Lname),
and Body has { (DOB=DOBR0)}
and confidence >1.0
and support > 1.0
}

For Race PHI from Record Zero (Race=RaceR0):
IF (Patient_Fname=Patient_Fname, Patient_Lname= Patient_Lname, Race=RaceR0)
THEN RETURN RECORD

Patient(Fingerprint_ID, Iris_ID, Retina_ID, DNA_ID, Patient_Fname, Patient_Lname, DOB, Age, Race, Ethnic Background, Gender, Date of Enrollment/Date of visit, Blood type, State of Birth, Region, Country of Birth, Nationalization)
To generate rules from the Patients table, one uses the GetRules command
as follows:
GetRules(Patient)
where Consequent in { (Patient_Fname=Patient_Fname, Patient_Lname= Patient_Lname),
and Body has { (DOB=DOBR0)}
and confidence >1.0
and support > 1.0
}

For Ethnic Background PHI from Record Zero (Ethnic Background=Ethnic BackgroundR0):
IF (Patient_Fname=Patient_Fname, Patient_Lname= Patient_Lname, ,Ethnic Background=Ethnic BackgroundR0)
THEN RETURN RECORD

Patient(Fingerprint_ID, Iris_ID, Retina_ID, DNA_ID, Patient_Fname, Patient_Lname, DOB, Age, Race, Ethnic Background, Gender, Date of Enrollment/Date of visit, Blood type, State of Birth, Region, Country of Birth, Nationalization)
To generate rules from the Patients table, one uses the GetRules command
as follows:
GetRules(Patient)
where Consequent in { (Patient_Fname=Patient_Fname, Patient_Lname= Patient_Lname),
and Body has { (DOB=DOBR0)}
and confidence >1.0
and support > 1.0
}

For Gender PHI from Record Zero (DOB=DOBR0):
IF (Patient_Fname=Patient_Fname, Patient_Lname= Patient_Lname, Gender=GenderR0)
THEN RETURN RECORD

Patient(Fingerprint_ID, Iris_ID, Retina_ID, DNA_ID, Patient_Fname, Patient_Lname, DOB, Age, Race, Ethnic Background, Gender, Date of Enrollment/Date of visit, Blood type, State of Birth, Region, Country of Birth, Nationalization)
To generate rules from the Patients table, one uses the GetRules command
as follows:
GetRules(Patient)
where Consequent in { (Patient_Fname=Patient_Fname, Patient_Lname= Patient_Lname),
and Body has { (DOB=DOBR0)}
and confidence >1.0
and support > 1.0
}

For Blood Type PHI from Record Zero (BloodType=BloodType0):
IF (BloodType=BloodType0, Patient_Fname=Patient_Fname, Patient_Lname= Patient_Lname)
THEN RETURN RECORD

Figure 20: GetQuery Schema and Association rules for the identification of patients' record based on PHI attributes meeting the criteria for Type 2-2 records.

Given the above reasoning, the GetRules operator disallows any "where" clause conditions on pure attributes of the source class C. These can always be performed by creating a view on C with the appropriate selections/projections and then using GetRules on that particular view. The only conditions allowed are the ones on rule components: Body and Consequent. Note that the evaluation of GetRules internally may involve selecting/projecting the data for efficiency, but it will preserve the query semantics.

The creation of these sub-queries allow the framework to cope with unexpected delays that arise when processing distributed queries in a wide-area network such as the virtual data depository of healthcare data. These sub-queries emulate Query Scrambling which uses two basic techniques to cope with unexpected delays:

1) It changes the execution order of operations in order to avoid idling, and

2) It synthesizes new operations to execute in the absence of other work to perform.

In this framework, the scrambling process is driven by the fact that the sub-queries are running concurrently across the virtual data depository looking for the same patient information based on different criteria. As in Query Scrambling, an initial query plan is chosen by a traditional System R-style optimizer. After every blocking operator in that plan, the remainder of the plan is re-optimized with the knowledge of the size of the intermediate result generated thus far. In essence, interoperability is a long-postponed union of the Ingres and System R optimization schemes: like Ingres, it takes advantage of the cardinality information in materialized sub-results; like System R it uses cost-based estimation of unknown work to be done. These schemes adapt at an inter-operator frequency, with arbitrary effects on the remaining steps after a block in a query plan; the extent does not go beyond the rest of the query. The framework allows fast implementation of most of the applications listed in the in phase 1 and phase 2, It still remains to be seen whether our set of primitives is sufficient to mine the vast amount of data within the virtual data depository.

5.5 FIRD Framework Data Discrepancy

Now that phase 1 and phase 2 have gone through the virtual data depository based on the initial biometric query returning records based on the set criteria. The framework has one last task to perform. That tasks is to correct the errors in the identifying attributes both biometric and PHI in the records belonging to the patient. In order to do so the system must use a classifier fusion algorithm based on the DS theory. This DS theory must be applied to two different types of errors.

5.5.1 Application of Phase 3 to Type 1-2 Records

As stated in Chapter 4, DS theory as it is applied to combine the output of individual fingerprint recognition algorithms to improve the verification performance. Using the underlying concept to DS theory and basic belief assignment, classifier fusion is performed using minutiae based fingerprint recognition algorithm [Jain, 1997], ridge based recognition algorithm [Marana and Jain, 2005] and finger code based recognition algorithm [Jain, 1999]. For every input fingerprint image, each classifier assigns a label true or 1 to proposition i, $i \in \Theta$ and the remaining classes are labeled as false or 0. Thus there are two focal elements for each fingerprint recognition algorithm i and $\neg i = \Theta - i$. i is for confirming and $\neg i$ is for denying the proposition for mass assignment in the DS theory. For each fingerprint recognition algorithm, we compute the respective predictive rates used to assign their basic belief assignment. For a c class problem, let us assume that an input pattern belonging to class j ($j \in c$) is classified as one of the k ($k \in c + 1$) classes including the rejection class, i.e. $(c + 1)^{th}$ class. So, the predictive rate of a

classifier P_k for an output class k is the ratio of the number of input patterns classified correctly to the total number of patterns classified as class k where input patterns belonging to all classes is presented to the classifier.

In this example, when a fingerprint recognition algorithm classifies the result $k \in c + 1$, it is considered that for all instances the likelihood of k being the actual class is Pk and the likelihood of k not being the correct class is $(1 - Pk)$. The predictive rate is used as basic belief assignment or mass m(k) and disbelief is assigned to m(¬k); with $m(\Theta) = 1$.

Further, multiple evidences are combined using the Dempster's rule of combination. Let A and B be used for computing new belief function for the focal element C, Dempster's rule of combination is written as:

$$m(C) = \frac{\sum_{AnB=C} m(A)m(B}{1 - \sum_{AnB=\phi} m(A)m(B}$$

Let m1, m2 and m3 be the mass computed from the three fingerprint recognition algorithms or classifiers which are combined recursively as shown:

$$m_{final} = m1 \oplus m2 \oplus m3$$

where $\oplus$ shows the Dempster's rule of combination. Final result is obtained by applying threshold t to m_{final} ,

$$result = \begin{cases} accept, & \text{if } m_{final} \geq t \\ reject, & \text{otherwise} \end{cases}$$

5.5.1.1 Update Rule for Calculating Belief Assignment

In most cases, it is required to update the belief based on new evidences or data. Let $E \subset \Theta$ and E_v be the evidence which states that the actual world is not in $\neg E$. Now suppose that the new data or evidence provides the exact value of E_v. Belief function is revised using the Dempster's update rule,

$$Bel\,[Ev\,](A) = Bel\,(A \cup \neg E\,) - Bel\,(\neg E\,)$$

This rule would be used to update the basic belief assignment associated with each fingerprint algorithm when a new training data is added. With this rule, only new basic belief assignments would be used to update the classifier. The time required for updating is significantly less as it is not required to train the complete classification algorithm when new training data is added.

5.5.2 Application of Phase 3 to Type 2-2 records

As far as using the DS theory as it is applied to the PHI is a little bit different. As stated earlier the threshold concept from phase two allows for a finite number of possible outcomes thus the recognition algorithms to improve the verification performance are enhanced. Using the underlying concept to DS theory and basic belief assignment, classifier fusion is performed. For every input blood type attribute, each classifier assigns a label true or 1 to proposition i, $i \in \Theta$ and the remaining classes are labeled as false or 0. Thus there are two focal elements for each blood type recognition algorithm i and $\neg i = \Theta - i$. i is for confirming and $\neg i$ is for denying the proposition for mass

assignment in the DS theory. For each blood type recognition algorithm, we compute the respective predictive rates used to assign their basic belief assignment. For a c class problem, let us assume that an input pattern belonging to class j ($j \in c$) is classified as one of the k ($k \in c + 1$) classes including the rejection class, i.e. $(c + 1)^{th}$ class. So, the predictive rate of a classifier P_k for an output class k is the ratio of the number of input patterns classified correctly to the total number of patterns classified as class k where input patterns belonging to all classes is presented to the classifier.

In this example, when a blood type recognition algorithm classifies the result $k \in c + 1$, it is considered that for all instances the likelihood of k being the actual class is Pk and the likelihood of k not being the correct class is $(1 - Pk)$. The predictive rate is used as basic belief assignment or mass m(k) and disbelief is assigned to m(¬k); with $m(\Theta) = 1$.

Further, multiple evidences are combined using the Dempster's rule of combination. Let A and B be used for computing new belief function for the focal element C, Dempster's rule of combination is written as:

$$m(C) = \frac{\sum_{AnB=C} m(A)m(B}{1 - \sum_{AnB=\phi} m(A)m(B}$$

Let m1, m2 and m3 be the mass computed from the three blood type recognition algorithms or classifiers which are combined recursively as shown:

$$m_{final} = m1 \oplus m2 \oplus m3$$

where $\oplus$ shows the Dempster's rule of combination. Final result is obtained by applying threshold t to m_{final} ,

$$\text{result} = \begin{cases} \text{accept}, & \text{if } m_{final} \geq t \\ \text{reject}, & \text{otherwise} \end{cases}$$

5.5.2.1 Update Rule for Calculating Belief Assignment

In most cases, it is required to update the belief based on new evidences or data. Let $E \subset \Theta$ and E_v be the evidence which states that the actual world is not in $\neg E$. Now suppose that the new data or evidence provides the exact value of E_v. Belief function is revised using the Dempster's update rule,

$$Bel[Ev](A) = Bel(A \cup \neg E) - Bel(\neg E)$$

This rule would be used to update the basic belief assignment associated with each blood type algorithm when a new training data is added. With this rule, only new basic belief assignments would be used to update the classifier. The time required for updating is significantly less as it is not required to train the complete classification algorithm when new training data is added.

5.6 Results of the Scenario Based Analysis

This section of the dissertation discusses the results of the Scenario Based Analysis and resolutions to scenarios that left the FIRD framework ineffective. Using the synthesize dataset created for testing purposes. The research designed a series of

scenarios to test the functionality of the framework. The scenarios ranged from one biometric identifier missing to a combination of missing identifiers both Biometric and PHI. The scenario based analysis of the framework provided validation to the functionality of its design. In most of scenarios based tests the intertwined and iterative process of the 3-phase system returned a 1:1 match from the synthesize dataset. However, the research discovered instances in which the framework mechanism was unable to correctly identify all of the records belonging to a particular patient. Each phase of the framework has drawbacks in certain scenarios where the resulting records cannot be classified belonging to the patient of query or the patient of query cannot be identified.

5.6.1 Phase 1 Drawbacks and resolutions

By design Phase 1 was constructed to be both a multi layer and multi modal biometric identifier system with scalable features to adapt to the changes in technology. As a minimal requirement the use of fingerprints biometrics needed for the framework to function. Also the assumption made in Chapter 4 that the patient's first and last name would be present was also a requirement. Due the fact that in the real world assumptions and or requirements do not always exist lead to the discovery that the FIRD framework cannot compute a proper identification of a patient in the event that only his or her fingerprint is the only biometric identification used and the absence of the first and last name. This is due to the fact that of all of the biometric identifiers fingerprints have the highest error rate (1 out of 500). So in the scenario where only fingerprints are the only form of identification the framework does not have enough of the identification resources

to match an Record Zero (enrollment record) to the patient of query. This research coined this phenomenon the “John”/”Jane” Doe scenario. The framework’s resolution to this is the additional layers of biometric identifiers (Iris, Retina or DNA) that allow for a more precise collection the patient query record zero. This would in turn allow for the healthcare facility tending care to identify the patient and their next of kin.

5.6.2 Phase 2 Drawbacks and Resolutions

This research designed the second phase of the framework as a safety net for the first phase. It purpose was to clean up any mistakes cause by the use of less accurate biometric identifiers and to retrieve records that truly belong to the patient of query yet did not have a biometric identifier associated with it. However upon analysis of the framework it was discovered that there was no safety net design for the second phase. Scenarios where a record does not contain any of the biometric identifiers required for phase 1 and only contains a subset of the PHI attributes caused a problem within the framework’s functionality. One such case occurs when records are absent of the biometric identifiers required for phase 1 and the PHI attribute blood type required in phase 2 could not be classified as belonging to the patient of query. Another such case was records belonging to twin patients with only the DNA biometric identifier present and PHI attributes which resulted in twice the number of records for a person of that age. The resolution to these types of scenarios were planned for in phase 2 that is why its design was constructed with the requirement that all four be present and not null.

5.6.3 Phase 3 Drawbacks and Resolution

As stated in the previous chapter the third phase of the framework is vital to the unification process in that it corrects omissions and discrepancies within the patients' records that are the results of errors or missing values. The underline assumption in data mining and data fusion is that the data to be mined is complete. Healthcare data is not complete do to human error or oversight some records for a particular patient may contain errors or omissions. This area is often overlooked in data management. Electronic format or not, data integrity within medical records is essential for error free high quality healthcare. This framework addresses this problem in the data to ensure the maximum amount of patient information is included in their complete healthcare history. The FIRD incorporates the concept of belief theory to allocate a value to inconsistent data within the collected dataset which is in this case a patient's electronic medical record. The belief theory phase of the framework takes the returned records that are incomplete or do not meet the criteria of Type 1-1 records (perfectly matched records) and analyzes the discrepancies within to see if a probability theory model can with high confidence correct those same discrepancies to determine if any of these records can be transformed into Type 1-1 records. This is achieved by either putting in place the biometric identifier for record were no biometrics identifier exist and or by correcting or inserting the 4 PHI attributes in those patient records that do not have any. The underline assumption is at this stage the framework has identified the record zero of the patient query. The

framework has undergone both phase 1 and phase 2 and that there exist a set of records that are either type 1-2 (it is missing one, two or three of the four biometric identifiers) or type 2-2 (it is missing one, two or three of the four PHI attribute identifiers). In this case Phase 3 would utilize the belief theory in section 5.5 to correct the errors. Although the assumption is HL7 would standardize the format in which the electronic medical record exist within the data depository. Also the functionality of the interoperability would exist for the creation of the virtual data depository. The inputting of medical data involves human interaction which leads to human error. This error can cause the data correction phase to not function. It is the purpose of this phase to correct mistakes but there are certain cases where this functionality cannot happen. These researches took these rare cases into account and develop a threshold that would not attempt to use belief theory if this scenario was true. A percent discrepancies formula was develop to provide a threshold when values that need to be corrected are not null but are incorrect. This threshold uses record zero as the baseline. Since all identifiers are suppose to equal the identifiers within record zero of the patient query. This allows for patient records with more than one type of incorrect value within a given identifier attribute to be sent out for review if the difference between the value within record zero (weighted @100%) and the percentage of the ratio of total incorrect values for that given attribute over the total number of returned values is greater than 30 %. Nevertheless, in all of these cases the query patient records would be included as a separate report called an exception report for human verification in an attempt to determine which records belong to which patient.

CHAPTER 6 CONCLUSIONS AND FUTURE DIRECTIONS

6.1 Summary of Main Contributions

This dissertation presents a comprehensive framework for accessing and unifying a patients' electronic medical record from various sources with the virtual data depository. The framework was developed to solve the dilemma of uniquely associating of a patient to all of his or her medical data distinctively. The framework would allow a patient to have real time access to all of his or her recorded healthcare information electronically whenever it is necessary, securely with minimal effort, greater effectiveness, and ease. The creation of a standardize nationwide electronic healthcare record system in the United States would require a way to identify individual patients out of about 300 million people and then link all of his/her records which could contain hundreds or even thousands of pages of information to a 1:1 match. The technology exists for the migration of healthcare data from its archaic paper-based system to an electronic one. The technology also exist for the new digitized healthcare data to be transported anywhere in the world in a matter of seconds. This would allow all of the healthcare industry to store and exchange all of it healthcare data with one another whenever it is necessary leading increase in quality of service/treatment and lower cost. A critical element of the functionality of this system is the ability to uniquely identify a patient and match them to all of his or her medical records regardless of location. Patient Identifiers are at the center of healthcare organization's day to day operations. Patient Identifiers in general play an integral part of the process of delivery of healthcare. Reliable Patient Identifiers are vital for proper identification for sensitive procedures,

such as blood transfusion, invasive testing, surgical procedures and medication administration. They are routinely used for various healthcare tasks from ordering and reporting the results of tests, procedures and medications, coordinating the quality delivery of patient care through multi-disciplinary channels and managing flow of all administrative functions, such as scheduling, billing, coordination of benefit, payment, etc.

The need for this type of universal identifier for electronic healthcare records is critical. The ability to accurately access all of a patient's medical data electronically will reduce the risk of errors because a physician will be more assured they are dealing with the correct patient and all of his or her medical information from allergies to correct and current medications unique to that person and their identifier. Government and private industry have developed possible solutions to the issue of a universal patient identifier for electronic healthcare records from the use of the Social Security Number to the Master Patient index. But the issue of security of patient information and well as implementation of a new system of identification creates complexities. This research proposed the use of a multi modal biometric system as a foundation for a universal patient identifier framework. Biometrics as the universal patient identifier presents a viable solution to the complexities of both security and implementation. The biometric identifier chosen from a person's physical attribute (fingerprint, iris, retina, DNA which all are a part the FIRD™ framework) becomes the patient passkey to access their medical data. The enrollment and acquisition of a patient's biometric sample is as easy a routine office visit or routine

physical. The multi modal biometric system makes it difficult for the patient passkey to be spoofed and since you are you it cannot be lost or forgotten.

The process of FIRD system to retrieve a patient's complete medical history is a transitional one. The process starts with the multi modal biometric system consisting of a combination of the four biometric identifiers with the greatest accuracy and the lowest error rate (Fingerprint, Iris, Retina, and DNA). All four together promote the best 1:1 match which is essential for the electronic healthcare record (EHCR) system to be effective at a global level capturing the entire population. However a key requirement of an effective universal patient identifier is it has to have scalability and be cost effective. The FIRD is a both a multi modal and multi layer biometric system which incorporated both of these criteria into its design. So for either technological or cost reasons the FIRD framework has other processes to compensate for the removal of up to three of the biometric identifiers. Once the FIRD retrieves a dataset based on the biometric identifier (single or multiple) it moves to its second phase of filtering and consolidation. The second phase of the FIRD framework uses data fusion algorithms based on four attributes of protected health information (Date of Birth, Ethnic Background, Blood type and Gender) for further unification of the data set to ensure accurate and precise data for the electronic medical record in question. This allows for the FIRD to address the fundamental assumption in data mining that the dataset to be mined is complete. We know that healthcare data is incomplete based on its origins from a paper based environment. The final phase of the FIRD framework processes the dataset returned from both the biometric and data fusion process to correct the data points of the patient's

dataset that contain data entry errors or omissions. The FIRD incorporates the concept of belief theory to allocate a value to inconsistent data within the patient's electronic medical record so the data fusion algorithm can reanalyze the dataset properly not that the fundamental assumption a complete dataset is met.

The research presented in this dissertation contributes to the field of Information Technology with an emphasis on Healthcare Information Systems. Focusing on the concept of universal identifying a patient to his or her records, this research has accomplished the following:

- Development and creation of universal patient identifier using multi layer and multi modal biometrics.
- The development and creation of the enrollment record named Record Zero
- Identification and primary development of the PHI attributes and their data mining algorithms.
- Identification of the attributes needed for belief theory equations.
- A process for extending the framework for growth as healthcare information systems become more robust that specifies a step-by-step layout of association rules to incorporate new constraints and assumptions within the data.
- A preliminary study attesting to the utility and functionality of the framework. The study classifies vulnerabilities present in specific scenarios which render the framework inefficient.

- An object model that represents an organized collection of patient electronic medical records based on either biometric or protected health information attributes (PHI) objects that can be present and exploited in a software process.
- A foundation for the basis of a common retrieval template for other aspects involved in the electronic medical record data system such as HL7 reporting conditions to interoperability of data through XML.

6.2 Future Directions

This dissertation presents a framework for the retrieval and unification of electronic medical records to a particular patient at the time of query. The framework only serves as a foundation upon which opens the possibilities of complementary areas of research to explore and build. This section presents those areas as future work.

6.2.1 Refining the Process of Using the Framework

Healthcare Information systems handle incredible amounts data. As Technology develops easier and transparent way to process data and improve the use of biometrics. It is important that the process of using the FIRD framework maintain currency in light of this dynamism. This can be accomplished by using the framework to test additional real-world healthcare data for true functionality. This testing can provide additional insights into the relationship between the framework and how it processes real world data with true errors and omissions. An understanding of this relationship will help refinement and evolution of the framework and its components.

6.2.2 Identification and Development of Tools

It is important to both identify and develop tools that simplify the process of using the FIRD framework. Such tools assist in and help reduce the time required for record retrieval, thereby increasing the usefulness of the framework to expand into more real-time areas such as the Emergency Room or Trauma units in hospitals. For example, tools that assist in identification of potentially useful techniques in massive data mining that processes objects concurrently for faster execution of queries are highly desirable.

6.3 Conclusion

Since healthcare information systems accumulate a tremendous amount data; A Electronic Healthcare Record (EHCR) system will facilitate the migration of data stored in antiquated methods (paper-based) to a more information exchange friendly environment. The EHCR system would format, transport, and properly describe medical data which is essential to healthcare information exchange and storage. The migration of patient healthcare data to electronic format dramatically increases mobility which leads to ways that data could be organized and shared to improve quality of care and decrease cost across many venues within healthcare. In recent years, software applications have been developed to address the migration of the data from paper based to electronic The ability to accurately access all of a patient's medical data electronically will reduce the risk of errors because a physician will be more assured they are dealing with the correct patient and all of his or her medical information from allergies to correct and current medications unique to that person and their identifier. This would allow all of the

healthcare industry to store and exchange all of it healthcare data with one another whenever it is necessary leading increase in quality of service/treatment and lower cost. This dissertations' framework not only provides the mechanism to achieve a functional foundation for record retrieval and unification but also support the other areas vital to the successful integration electronic healthcare date. The framework is both novel and distinctively different from the ones found in current literature.

REFERENCES

Agency for Healthcare Research and Quality (AHRQ). (2001). Making Health Care Safer: A Critical Analysis of Patient Safety Practices. Evidence Report/Technology Assessment: Number 43. AHRQ Publication No. 01-E058.

American Medical Informatics Association (AMIA). (1993). Position Paper on Standards for Medical Identifiers, Codes and Messages Needed to Create an Efficient Computer-Stored Medical Record.

American Standards for Testing and Materials (ASTM). (1995). ASTM E 1384-96 Guide for Content and Structure of the Computer-based Patient Record.

American Standards for Testing and Materials (ASTM). (1995). Standard Guide for Properties of a Universal Healthcare Identifier (UHID), Designation: E1714-95, Approved August 15, 1995, Published October 1995.

American Standards for Testing and Materials (ASTM). (1995). ASTM E 1762-95 Guide for Electronic Authentication of Health Care Information.

Agrawal, R., Imielinski, T., Swami A. (1993). Mining association rules between sets of items in large databases. Proceedings of the 1993 International Conference on Management of Data (SIGMOD 93), 1993:207–216.

American Standards for Testing and Materials (ASTM). (1996). ASTM E 1769-96 Guide for Properties of Electronic Health Records and Record Systems.

Appavu, S.I. (1997). Analysis of Unique Patient Identifier Options. A Final Report. Prepared for The Department of Health and Human Services.

Ashbourn, J. (2000). Biometrics: Advanced Identity Verification: The Complete Guide, SpringerVerlag, London.

Aspden, P., Corrigan, J.M., Wolcott, J., Erickson, S.M., editors. (2004). Patient Safety: Achieving a New Standard for Care. Committee on Data Standards for Patient Safety. Board on Health Care Services. Institute of Medicine. Washington, DC: The National Academies Press.

BioAPI Consortium (2001). "Members", January. Available at: www.bioapi.org/BioAPI_ home.htm (Accessed 22 January 2001).

Biomet.org (2000). "Interview with Richard E. Norton of the IBIA", 30 October. Available at: www.biomet.org/001029_ibia_interview.htm (Accessed 6 March 2001).

(The) Biometric Consortium (2001). "Introduction to biometrics", January. Available at: www.biometrics.org/html/ introduction.html (Accessed January 2001).

(The) Biometric Consortium (2001). "Standards", January. Available at: www.biometrics.org/ html/standards.html (Accessed 22 January 2001).

Booker, L., Goldberg, D.E. and Holland, J.H. (1989). Classifier systems and genetic algorithms, Artif. Intell. 40 , pp. 235–282.

Bouckaert, R.R. Properties of belief networks learning algorithms. (1994). Proceedings of the Conference on Uncertainty in Artificial Intelligence, pp.102–109.

Charniak, E. (1991). Bayesian networks without tears, AI Mag. 12 (4), pp. 50–63.

Chow, C.K. and Liu, C.N., (1968). Approximating discrete probability distributions with dependence trees, IEEE Trans. Inf. Theory 14 (3), pp. 462–467.

Clark, P. and Niblett, T. (1989). The CN2 induction algorithm, Mach. Learn. 3, pp. 261–283.

Biswas, G., Weinberg, J., and Fisher, D. (1998). “ITERATE: A Conceptual Clustering Algorithm for Data Mining”, in IEEE Records on Systems, Man and Cybernetics-Part C: Applications and Reviews, Vol. 28(2), pp.219-230. May, 1998.

Carpenter, P.C. & Chute, C.G. (1994). The Universal Patient Identifier: A Discussion and Proposal. Proceedings of the Seventeenth Annual Symposium on Computer Applications in Medical Care, 4, 49-53.

Cooper, G.F. (1990). The computational complexity of probabilistic inference using Bayesian belief networks, Artif. Intell. 42 , pp. 393–405.

Cooper, G.F. and Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data, Mach. Learn. 9, pp. 309–347.

Creating a Culture of Safety. (2004). VA National Center for Patient Safety (NCPS). Retrieved April 9, 2006, from VA National Center for Patient Safety (NCPS) website: http://www.patientsafety.gov/vision.html

Crossing the Quality Chasm: A New Health System for the 21st Century. (2001). Institute of Medicine. Washington, DC: National Academies Press.

De Jong, K.A, Spaers W.M. and Gordon, D.F. (1993). Using genetic algorithms for concept learning, Mach. Learn. 13, pp. 161–188.

Dick, R.S. and Steen, E.B. (1991). The Computer-Based Patient Record. Institute of Medicine. Washington, DC: National Academy Press.

Dubois, D. and Prade, H. (1987). Théorie des possibilités, applications à la représentation des connaissances en informatique.

Everitt, B. (1993). Cluster Analysis, 3rd ed. Hodder & Stoughton, London.

Everitt, B. and Rabe-Hesketh, S. (1997). The Analysis of Proximity Data, Wiley, New York.

Fayyad, U.M., Piatesky-Shapiro, G., Smyth, P. (1996). From data mining to knowledge discovery: An overview. AI Mag., pp. 37–51.

Fernando, B., Savelyich, B.S.P., Avery, A.J., Sheikh, A., Bainbridge, M. and Horsfield, P. (2004). Prescribing safety features of general practice computer systems: evaluation using simulated test cases. British Medical Journal, 328, 1171-1172.

Fogel, D.B. (1994). An introduction to simulated evolutionary optimization, IEEE Trans. Neural Netw. 5, pp. 3–14.

Fogel L., Owens A., Walsh M. (1966). Artificial Intelligence through Simulated Evolution. New York: Wiley, 1966.

For the Record, Protecting Electronic Health Information. (1997). National Research Council. Washington, DC: National Academy Press.

Ganti, V., Gehrke, J., and Ramakrishnan, R. (1999). CACTUS: Clustering Categorical Data Using Summaries. Knowledge Discovery and Data Mining, pp. 73-83.

Gibson, D., Kleinberg, J. and Raghavan, P. (2000). Clustering Categorical Data: an Approach Based on Dynamical Systems. Proceedings of the 24th VLDB Conference, Vol. 8(3/4), pp. 222-236.

Giordanam, A. and Neri, F. (1995). Search-intensive concept induction, Evol. Comput. 3, pp. 375–416.

Goldberg, D.E. (1989). Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, Reading, MA.

Goldberg, D.E., Richardson, J. (1987). Genetic algorithms with sharing for multi modal function optimization. Proceedings of the second International Conference on Genetic Algorithms, pp. 41–49.

Goodall, D. (1966). A New Similarity Index Based On Probability. Biometrics, Vol. 22(4), pp. 882-907.

Gostin, L.O. (1993). Privacy and Security of Personal Information in a New Health Care System. Journal of the American Medical Association, 270(20), 2487-2492.

Gowda, K. and Diday, E. (1991). Symbolic Clustering Using a New Dissimilarity Measure. Pattern Recognition, Vol. 24(6), pp. 567-578.

Gower, J. (1971). A General Coefficient of Similarity and Some of Its Properties. Biometrics, Vol. 27(4), pp. 857-871.

Guha, S., Rastogi, R. and Shim, K. (2000). ROCK: A Robust Clustering Algorithm for Categorical Attributes, in Information Systems, Vol. 25(5), pp. 345-366.

Gupta, S., Rao, K. and Bhatnagar, V. (1999). K-Means Clustering Algorithm For Categorical Attributes, Data Warehousing and Knowledge Discovery, pp. 203-208.

Han, E., Karypis, G., Kumar, V., and Mobasher, B. (2001). Clustering Based on Association Rule Hypergraphs. Proceedings of SIGMOD ’97 Workshop on Research Issues in Data Mining and Knowledge Discovery.

Han, J. and Kamber, M. (2001). Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, San Francisco.

Health Data in the Information Age - Use, Disclosure and Privacy. (1994). Institute of Medicine. Washington, DC: The National Academy Press.

Health Insurance Portability and Accountability Act of 1996 (HIPAA). (1996). (Pub.L.104-191, Aug. 21, 1996, 110 Stat. 1936). Standards for Privacy of Individually Identifiable Health Information (PIHI), Federal Register. (codified at 45 CFR §160, §164).
45 CFR. §164.502(b)(1).
45 CFR. §164.502(b)(2).
45 CFR. §164.512.
45 CFR. §164.508.

Heckerman, D. (1996). Bayesian Networks for Knowledge Discovery, chapter 11. Cambridge, MA: MIT Press, pp.273–306.

Heckerman, D., Geiger, D. and Chickering, D.M. (1995). Learning Bayesian networks: the combination of knowledge and statistical data, Mach. Learn. 20, pp. 197–243.

Heckerman, D. and Wellman, M.P. (1995). Bayesian networks, Communications of ACM 38, pp. 27–30.

Hefferman, S. (1999). The Role of Biometrics within Document Security, PISEC 99 Conference, Barcelona.

Herdman, R.C. (1997). The Department of Health and Human Services Guiding Principles for Choosing Standards For the record. Protecting Privacy in Computerized Medical Information, Report to the Senate Subcommittee on Federal Services, Post Office, and Civil Service, and House Subcommittee on Government Information, Justice, and Agriculture.

Herskovits, E., Cooper, G. (1990). KUTATO: An entropy-driven system for construction of probabilistic expert systems from databases. Technical Report KSL-90-22, Knowledge Systems Laboratory, Medical Computer Science, Stanford University.

Hippisley-cox, J., Pringle, M., Cater, R., Wynn, A., Hammersley, V., Coupland, C., Hapgood, R., Horsfield, P., Teasdale, S. and Johnson, S. (2003). The electronic patient record in primary care - regression or progression? A cross sectional study.
British Medical Journal 28(326), 1439-1443.

Hirano, S., Okuzaki, T., Hata, Y., Tsumoto, S. and Tsumoto, K. (2001). A Rough Set-Based Clustering Method with Modification of Equivalence Relations, in PAKDD 2001, D. Cheung et al eds., pp. 513-518, 2001.

Holland, J.H. (1992). Adaptation in Natural and Artificial Systems. Cambridge, MA: MIT Press.

Holland, J.H., Reitman, J.S. (1995). Cognitive systems based on adaptive algorithms. In: Waterman D.A., Hayes-Roth F., editors. Pattern-Directed Inference Systems. New York: Academic Press.

Hopcroft, J.E. and Ullman, J.D. (1979). Introduction to automata theory, languages, and computation. Reading, MA: Addison-Wesley, 1979.

Huang, Z. and Ng, M. (1999). A Fuzzy k-Modes Algorithm for Clustering Categorical Data. IEEE Records on Fuzzy Systems, Vol. 7(4), pp. 446-452.

Huang, Z. (1997). Clustering Large Data Sets With Mixed Numeric and Categorical Values. Proceedings of 1st Pacific-Asia Conference on Knowledge Discovery & Data Mining.

Imielinski, T. and Virmani, A. (1998). Association rules... and what's next? towards second generation data mining systems. In W. Litwin, T. Morzy,and G. Vossen, editors, Advances in Databases and Information Systems,Second East European Symposium, ADBIS'98, Poznan, Poland, Spetember 7-10, 1998, Proceedings, volume 1475 of Lecture Notes in Computer Science, pages 6–25. Springer, 1998.

Jain, A. K. and Ross, A. (2004). Multibiometric Systems, Communications of the ACM, Special Issue on Multimodal Interfaces , Vol. 47, No. 1, pp. 34-40.

Jain, A. K. and Zongker, D. (1977). Representation and recognition of handwritten digits using deformable templates. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(12):1386.

Jain, A.K. and Dubes, R. (1988). Algorithms for Clustering Data. Prentice Hall, New Jersey.

Janikow, C.Z. (1993). A knowledge-intensive genetic algorithm for supervised learning, Mach. Learn. 13, pp. 189–228.

Joachims, T. (1999). Transductive inference for text classification using support vector machines. In Proc. 16th International Conference on Machine Learning, pages 200 Morgan Kaufmann.

Kim, M.I. and Johnson, K.B. (2002). Personal health records: evaluation of functionality and utility. Journal of American Medical Information Association, 9(2), 171-180.

Kim, W.-Y. and Kak, A.C. (1991). 3-d object recognition using bipartite matching embedded in discrete relaxation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 13(3):224.

Robles-Kelly, A. and Hancock, E.R. (2004). A graph spectral approach to shape-from-shading. IEEE Transactions on Image Processing, 13:912.

Kondor, R. and Lafferty, J. (2002). Diffusion kernels on graphs and other discrete input spaces. Proc. 19th International Conference on Machine Learning, pages 315 Morgan Kaufmann.

Koza, J.R. (1992). Genetic Programming: on the programming of computers by means of natural selection. Cambridge, MA: MIT Press.

Koza, J.R. (1994). Genetic Programming II: automatic discovery of reusable programs. Cambridge, MA: MIT Press.

Lam, W. (1998). Bayesian network refinement via machine learning approach. IEEE Trans. Pattern Anal Mach. Intell.

Lam, W. and Bacchus, F. (1994). Learning Bayesian belief networks—an approach based on the MDL principle, Comput. Intell. 10, pp. 269–293.

Lam, W., Wong, M.L. Leung, M.S., Ngan, P.S. (1998). Discovering probabilistic knowledge from databases using evolutionary computation and minimum description length principle. Genetic Programming: Proceedings of the Third Annual Conference.

Larranaga, P., Poza, M., Yurramendi, Y., Murga, R. and Kuijpers, C. (1996). Structure learning of Bayesian network by genetic algorithms: A performance analysis of control parameters, IEEE Trans. Pattern Anal. Mach. Intell. 18 (9), p. 9.

Landro, L. (2004, January 29). The informed patient: consumers need health-care data. The Wall Street Journal, p.D3.

Larranaga, P., Yurramendi, Y., Murga, R. and Kuijpers, C. (1996). Learning Bayesian network structures by searching for the best ordering with genetic algorithms, IEEE Trans. Syst. Man Cybern. Part A: Syst. Humans 26 (4), pp. 487–493.

Leung, K.S., Leung, Y., So, L., Yam, K.F. (1992). Rule learning in expert systems using genetic algorithm: 1, concepts. Proceedings of the Second International Conference on Fuzzy Logic and Neural Networks, Iizuka, Japan, pp. 201–204.

Li, C. and Biswas, G. (1998). Conceptual Clustering With Numeric-and-Nominal Mixed Data-A New Similarity Based System. IEEE Transcript on KCE.

Macdonald, R. (2001). Commentary: A patient's viewpoint. British Medical Journal 322, 287.

Mandl, K.D., Szolovits, P., Kohane, I.S., Markwell, D. and MacDonald, R. (2001). Public standards and patients' control: how to keep electronic medical records accessible but private. British Medical Journal 322, 283-287.

Markwell, D. (2001). Commentary: Open approaches to electronic patient records. British Medical Journal 322, 286

McGuire, M. (2004). Steps Toward a Universal Patient Medical Record A Project Plan to Develop One. Boca Raton, FL: Universal Publishers

McMullen, W.L., (1994).Using Patient Identifiers from Legacy Systems for Healthcare Information Infrastructure. Presentation at the 10th International Symposium on the Creation of Electronic Health Record Systems. Washington D.C., March 24, 1994.

Medical Record Institute. (1996). Analysis on Patient Identifier. Toward an Electronic Patient Record.

Medical Record Institute. (1993). Position Paper 1: Patient Identifiers - Insurance Identification and Patient Identification in Healthcare.

Michalski, R.S., Mozetic, I., Hong, J., Lavrač, N. (1986). The multi-purpose incremental learning system AQ15 and its testing application to three medical domains. Proceedings of the Fifth National Conference on Artificial Intelligence, pp.1041–1045.

National Committee on Vital and Health Statistics (NCVHS). (1997). Health Privacy and Confidentiality Recommendations of the National Committee on Vital and Health Statistics, Approved on June 25, 1997, submitted to the Secretary of Health and Human Services June 27, 1997.

National Committee on Vital and Health Statistics (NCVHS). (1998). Assuring a Health Dimension for the National Information Infrastructure. A Concept Paper by the National Committee on Vital and Health Statistics, presented to the U.S. Department of Health and Human Services Data Council.

National Committee on Vital and Health Statistics (NCVHS). (2000). Report to the Secretary of the US Department of Health and Human Services on Uniform Data Standards for Patient Medical Record Information.

National Committee on Vital and Health Statistics.(2003). Recommendations for PMRI terminology standards.

Nigam, K., McCallum, A., Thrun, S. and Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. Machine Learning, 39(2/3):103-134.

Pawlak, Z. (1982). Rough Sets. International Journal of Computer and Informational Sciences, Vol. 11, no. 5, pp. 341-356.

Pawlak, Z. (1991). Rough Sets: Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dordrecht.

Prabhakar, S., Pankanti, S. and Jain, A.K. (2003). Biometric recognition: security and privacy concerns. IEEE Security & Privacy, 1(2), 33-42.

Ralambondrainy, H. (1995). A Conceptual Version of the K-Means Algorithm. Pattern Recognition Letters, Vol. 16, pp. 1147-1157.

Rashbass, J. (2001). The patient-owned, population-based electronic medical record: a revolutionary resource for clinical medicine. Journal of the American Medical Association, 285(13), 1769.

Rebane, G., Pearl, J. (1989). The recovery of causal poly-trees from statistical data. Uncertainty in Artificial Intelligence 3. Amsterdam: North-Holland, pp.175–182.

Rechenberg, I. (1973). Evolution Strategy: Optimization of technical systems by means of biological evolution. Stuttgart: Fromman-Holzboog.

Rundle, R.L. (2003, February 4). Big HMO plans to put medical records online.
The Wall Street Journal, p. D4.

Safran, C. (2001). Electronic medical records: a decade of experience. Journal of the American Medical Association, 285(13), 1766.

Sarawagi, S., Thomas, S. and Agrawal, R. (1998). Integrating association rule mining with relational database systems: Alternatives and implications. Proceedings of ACM SIGMOD Canference on Managemen~ of Data (SIGMOD '98), Seattle, Washington, June 1998.

Schneider, J.H. (2001). Online personal medical records: are they reliable for acute/critical care? Crit Care Med, (8 Suppl), N196-N201.

Schwefel, H.P. (1981). Numerical Optimization of Computer Models. New York: Wiley.

Singh, M., Valtorta, M. (1993). An algorithm for the construction of Bayesian network structures from data. Proceedings of the Conference on Uncertainty in Artificial Intelligence, pp.259–265.

Smets, P. (1990). The combination of evidence in the transferable belief model.
IEEE Pattern Analysis and Machine Intelligence, 12:447-458.

Smith, S.F. (1980). A Learning System based on Genetic Adaptive Algorithms. PhD thesis, University of Pittsburgh.

Smith, S.F. (1983). Flexible learning of problem solving heuristics through adaptive search. Proceedings of the Eighth International Conference On Artificial Intelligence. San Mateo, CA: Morgan Kaufmann.

Sneath, P. and Sokal, R. (1973). Numerical Taxonomy. Freeman and Company, San Francisco.

Spirtes, P., Glymour, C., Scheines, R. (1993). Causation, Prediction and Search. Berlin: Springer, 1993.

Sungur, E. Overview of multivariate statistical data analysis, http://www.mrs.umn.edu/~sungurea/multivariatestatistics/overview.html.

Szolovits, P. and Kohane, I. (1994). Against Simple Universal Health-care Identifiers. JAMIA 1(4), 316-319.

Tang, P.C., Hammond W.E. (1997). A Progress Report on Computer-Based Patient Records in the United States. Committee on Improving the Patient Record. Institute of Medicine. Dick, R.S., Steen, E.B., Detmer, D.E., (Eds.). The Computer-Based Patient Record: An Essential Technology for Health Care. Rev ed. Washington, DC: National Academies Press.

Terry, N.P. (2001). An eHealth diptych: the impact of privacy regulation on medical error and malpractice litigation. Am J Law Med. 27(4), 361-419.

Terry, N.P. (2002). When the “machine that goes ‘ping’” causes harm: default torts rules and technologically-mediated health care injuries. St Louis Univ Law J, 46, 37-59.

Terry, N.P. (2003). Privacy and the health information domain: properties, models and unintended results. Eur J Health Law, 10(3). 223-237.

The White House. (2004). Transforming Health Care: The President's Health Information Technology Plan. Executive Order 13335 of April 27, 2004. Incentives for the Use of Health Information Technology and Establishing the Position of the National Health Information Technology Coordinator. 69 Federal Register 24059, Sect. 3.

Tsai, C.C., Starren J. (2001). Patient participation in electronic medical records. Journal of the American Medical Association, 285(13), 1765.

United States Government Accounting Office. (2006). Health Information Technology: HHS is Continuing Efforts to Define a National Strategy. Report number GAO-06-346T.

Vaclav, M. and Zdenek, R. (2000). Biometric Authentication Systems. A technical report.

Walsh, S.H. (2004). The clinician's perspective on electronic health records and how they can affect patient care. BMJ, 328, 1184-1187.

Wang, K., Xu, C. and Liu, B. (1999). Clustering Records Using Large Items. CIKM. pp. 483-490.

Wong, M.L., Leung, K.S. (1995). Inducing logic programs with genetic algorithms: the genetic logic programming system, IEEE Expert 10, pp. 68–76.

Wong, M.L., Leung, K.S. (1997). Evolutionary program induction directed by logic grammars, Evol. Comput. pp. 143–180.

Zhang, Y., Wai-chee Fu, A., Cai, C., and Heng, P. (2000). Clustering Categorical Data. 16th International Conference on Data Engineering.