University of Windsor
Scholarship at UWindsor

Electronic Theses and Dissertations | Theses, Dissertations, and Major Papers

1-18-2016

SEMANTIC INTEROPERABILITY AND DATA MAPPING IN EHR SYSTEMS

Sreya Janaswamy
University of Windsor

Follow this and additional works at: https://scholar.uwindsor.ca/etd

Recommended Citation

Janaswamy, Sreya, "SEMANTIC INTEROPERABILITY AND DATA MAPPING IN EHR SYSTEMS" (2016). *Electronic Theses and Dissertations*. 5645.
https://scholar.uwindsor.ca/etd/5645

This online database contains the full-text of PhD dissertations and Masters' theses of University of Windsor students from 1954 forward. These documents are made available for personal study and research purposes only, in accordance with the Canadian Copyright Act and the Creative Commons license—CC BY-NC-ND (Attribution, Non-Commercial, No Derivative Works). Under this license, works must always be attributed to the copyright holder (original author), cannot be used for any commercial purposes, and may not be altered. Any other use would require the permission of the copyright holder. Students may inquire about withdrawing their dissertation and/or thesis from this database. For additional inquiries, please contact the repository administrator via email (scholarship@uwindsor.ca) or by telephone at 519-253-3000ext. 3208.

SEMANTIC INTEROPERABILITY AND DATA MAPPING IN EHR SYSTEMS

By

SREYA JANASWAMY

A Thesis
Submitted to the Faculty of Graduate Studies
through the School of Computer Science
in Partial Fulfillment of the Requirements for
the Degree of **Master of Science**
at the University of Windsor

Windsor, Ontario, Canada

2016

© 2016 SREYA JANASWAMY

SEMANTIC INTEROPERABILITY AND DATA MAPPING IN EHR SYSTEMS

by

SREYA JANASWAMY

APPROVED BY:

_____

Anne W. Snowdon, External Reader
Odette School of Business

_____

Imran Ahmad, Internal Reader
School of Computer Science

_____

Robert Kent, Advisor
School of Computer Science

Jan 11, 2016

DECLARATION OF ORIGINALITY

I hereby certify that I am the sole author of this thesis and that no part of this thesis has been published or submitted for publication.

I certify that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my thesis and have included copies of such copyright clearances to my appendix.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

ABSTRACT

The diversity in representation of medical data prevents straightforward data mapping, standardization and interoperability between the heterogeneous systems. We identify a specific problem, namely the need to achieve interoperability by applying a standard based data modeling approach to achieve a common platform that serves to improve the health data mapping of unstructured data and addresses ambiguity issues when dealing with health data from heterogeneous systems.

In this thesis, we proposed an original Hybrid algorithm that identifies the attributes of data in heterogeneous systems based on critical medical standards and protocols and then performs semantic integration to form a uniform interoperable system. Also, efficient data modeling techniques are introduced for improving data storage and extraction. We tested the proposed algorithm with multiple data sets and compared the proposed approach with traditional data modeling approaches. We found that the proposed approach demonstrated performance improvements and reduction in data losses.

DEDICATION

To my loving family:

Father: Prabhakar Janaswamy

Mother: Surya Kumari Janaswamy

Brother: Sravan Janaswamy

ACKNOWLEDGEMENTS

I owe many thanks to the people who guided me and supported me to complete this thesis. I would like to express my sincere thanks to my supervisor, Dr. Robert Kent for giving me the opportunity to get an exposure to the research field. I feel very fortunate for getting an opportunity to work under his guidance. At every stage, he gave me the encouragement and monitoring required and guided me towards successful completion of my thesis.

I would like to thank my internal reader Dr. Imran Ahmad for his support and positive attitude towards my approach. I am also thankful to my external reader Dr. Anne Snowdon for her suggestions and interest in the approach.

I give my sincere thanks to Mrs. Karen Bourdeau, graduate secretary, who always supported and helped me whenever I needed any assistance in various academic issues and in setting up last minute meetings.

Lastly, I would like to thank my family and friends for their utmost faith in me and for the support, encouragement and love that they always have towards me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS/SYMBOLS

API	Application Program Interface
CSV	Comma Separated Values
DB	Database
DOM	Document Object Model
EAV	Entity Attribute Value
EAV/CR	Entity Attribute Value with Classes and Relationships
EHR	Electronic Health Record
EMR	Electronic Medical Record
ESB	Enterprise Service Bus
HIS	Health Information Systems
HL7	Health Level 7
HSB	Health Service Bus
HTML	Hyper Text Markup Language
JDBC	Java Database Connectivity
LAN	Large Area Network
LOINC	Logical Observation Identifiers Names and Codes
MAN	Metropolitan Area Network
MB	Megabyte
MIN	Minutes
MML	Medical Markup Language
MPEG	Moving Picture Experts Group

MS	Milliseconds
MSH	Message Header
OBX	Observation
RDBMS	Relational Database Management System
RDF	Resource Description Framework
RELMA	Reginstrief LOINC Mapping Assistant
SEC	Seconds
SNOMED	Systematized Nomenclature of Medicine
SOA	Service Oriented Architecture
SQL	Structured Query Language
TCP/IP	Transmission Control Protocol/Internet Protocol
UML	Unified Modeling Language
WAN	Wide Area Network
XML	Extensible Markup Language
XSLT	Extensible Stylesheet Language

1 Introduction

1.1 Definitions

1.1.1 Health Informatics

Before the invention of computers, banking, business, agriculture, textiles, education, medicine and many other fields relied on the use of paper to record data of concern to decision making. In the modern era, computers and data storage techniques of many types have been developed and are being applied in all of these areas mentioned above.

Computer and Information Science is deeply influencing both research and development in the emerging field of Health Informatics [7, 14] (aka Healthcare Informatics). As per the authors Donabedian [1966], *"Health informatics can be regarded as being concerned with the structures and processes, as well as the outcomes involved in the use of information and communications technologies within health. The term 'e-health' has been coined to describe the application of these technologies in health and medicine."*

1.1.2 Electronic Health Record

According to authors Collen, M. F. [1999], Robert Ledley of United States was the first one to use digital computers in the field of medicine in 1950. From the early 1960s, Usage of computers for the laboratory/ clinical purposes gradually began. Hence, since, 1970s or so we started to use electronics in the field of medicine. The electronic format of storing medical data in a tabular or formatted manner is called as electronic health record. We can also say that, Electronic medical records are the information of patient health, stored in an electronic format instead of using a paper based approach. During the early phases of development, which is

around 1970s and 1980s programming language called MUMPS was widely used in storing medical records. Later, EHR/ EMR systems started to evolve. Häyrinen *et al* [2008] defined an EHR system as *"a repository or database which stores the patient information digitally, enabling secure data exchange and authorize only specific set users to access*".

1.2 Research Motivation

The main motivation for this research is obtained by observing different problems related to medical information or EHR data. We found that various research works emphasized on the importance of storage and data modelling techniques with respect to clinical data. Also, we observed that usage of medical standards can improve the representation of semantics and enable interoperability. While working with these medical vocabularies and repositories (like HL7 and LOINC), we found that, they can not only be used for enabling interoperability, but also for understanding the semantics and establishing a mapping between heterogeneous databases.

1.3 Problem Statement

Medical and Health data analysis involve data capturing, storage, processing, exchange, integration and interpretation of data from heterogeneous systems. The existing EHR systems store large volumes of medical records independently in various structures and formats. The structural differences could be different storage devices, databases, data types, attributes or syntax. The semantic differences may occur due to differences in vocabularies, usage of different representations of the same data (synonyms), language barriers, etc., generally need more human assistance in identification and interpretations. This creates a bigger problem in case of unknown environments when the data from multiple systems is unpredictable and is not in a pre-agreed format. The diversity of both the data and data models of existing system

obscures or prevent data mapping, standardization and interoperability between heterogeneous systems.

1.4 Thesis Contribution

This thesis concerns research which attempts to solve the existing problems of Health Information Systems (HIS) and various problems associated with integration of heterogeneous systems, specially interoperability issues and differences caused due to data representations. In spite of continuous efforts, limited access to health data still remains a major problem.

Hence, our aim was to develop a system which can firstly map and then merges the data which is stored in different data types, structures, file systems, databases and with vocabularies is considering a primary challenge. Once that is achieved, the next big challenge is to effectively use that data and be able to make it interoperable. Integration, expression and maintenance of secure data has always remained a challenging task for various researchers. An appropriate solution can be advantageous to real world situations as a realistic cost effective and efficient system which is not only useful in real time data acquisition but also resolving issues of interoperability. This data when applied through a series of principles and standards can be analysed and reused for medical systems as we represent the data in a standardized format.

1.5 Thesis Outline

This thesis is broadly divided into 7 chapters. In Chapter 1, we introduced a basic in problem in a broad way and explained the research motivation and contributions. In Chapter 2 concerns a survey of the related work on medical standards, metadata management and interoperability. Chapter 3 presents the details of existing works that tried to address the problems of data mapping and interoperability. This section consists of details of 10 research papers of which 4

are related to interoperability, 3 are related to data modelling, 2 focus on medical markup languages and 1 discusses about the importance of the data quality. Chapter 4 consists of a detailed explanation of our framework and our approach. Chapter 5 explains our experimental setup. In Chapter 6, we present all the results of our experiments and finally we conclude in Chapter 7.

2 RELATED WORK

This section tries to define the work done by researchers on interoperability or data mapping between electronic health record systems. The interoperability issues are further detailed into the different types or categories. Data modelling techniques are also discussed in this chapter. The use of various global medical standards can be considered as of the integral part of the work. Precisely, the required background for this thesis will be provided in this chapter.

All the relevant research papers were found by searching google scholar with the key words “Interoperability”, “Database mapping”, “Data access”, “HL7” and “Health data”. All the above mentioned keywords and author names were also used to search ACM digital library and IEEE publications for relevant works.

2.1 Electronic Health Record Data – Quality Assessment

Weiskopf *et al* [2013] presented a literature review regarding the data quality and quality assessment methodologies. They state that EHR data could be used for area of research but the quality of the data is a problem to be dealt with. They also state that, *“73% of previous research works considered only structured data and 22% considered combination of structured and unstructured data”*. They derived and presented five dimensions with respect to the quality of the EHR data as Completeness, Correctness, Concordance, Plausibility and Currency. These five dimensions can be represented with terms in literature as shown in table 1.

Completeness	***Correctness***	***Concordance***	***Plausibility***	***Currency***
Accessibility	Accuracy	Agreement	Accuracy	Recency
Accuracy	Corrections made	Consistency	Believability	Timeliness
Availability	Errors	Reliability	Trustworthiness	
Missingness	Misleading	Variation	Validity	
Omission	Positive predictive value			
Presence	Quality			
Quality	Validity			
Rate of recording				
Sensitivity				
Validity				

Table 1: Terms to describe 5 common dimensions **(Table 1 page 145 of [54])**

The authors state that, "*adopting consistent taxonomy of EHR data quality and integration of data quality assessments and systematic data quality assessments would help in the reuse of EHR data from which clinical research would benefit.*" In this thesis we are focused on using medical and health vocabularies for maintaining the quality and assist in re-use of clinical data. The specific approach we use is called Logical Observation and ***** (LOINC) which is discussed in detail in Chapter 4.

2.2 Interoperability

Interoperability and Integration of medical records of different hospitals or diagnostic centres are needed for various reasons, such as analysis, research, mergers and acquisitions. To put in simple words, when two or more similar entities are able to work together for a common purpose those systems are said to be interoperable. To be more specific, *"Interoperability can be defined as the ability of a system, organization or individual to be able to communicate and work together with other similar entities"* (Ide *et at* [2010]). Authors Geraci *et al* [1991], is defined interoperability as, *"the ability of two or more systems or components to exchange information and use the information that has been exchanged."*

Information systems can exhibit various forms of interoperability. Initially, a few decades earlier, the aim of various researches was to improve techniques for merging of data. But, later on the focus has shifted towards identifying related information with different forms of representations.

2.2.1 Levels of Interoperability

Various authors, including Ouksel and Sheth [1999] discussed about interoperability at a deeper level by dividing them into categories, also called as various levels of interoperability.

- **System Level or Machine Level Interoperability:** Every system should support the exchange of information right from the basic or lower level. System interoperability is important for facilitating interactions between distributed elements and agents.
- **Syntactic Interoperability:** Syntactical differences might be due to the use of various programming languages, data structures or data types for the exchange of information between heterogeneous systems. Low level to high level languages can be implemented, but we need to analyze which is the best amongst all. Thus XML is used as it's widely accepted and has an easy representation of data.
- **Structural or Organizational Interoperability:** This is considered to be one of the toughest levels. Agreements on various levels of interoperability is done by organizations. Technologies like RDF, KIF, OKBC, MPEG, etc. are used for the representation of multimedia, hypermedia, object oriented data and other forms of information.
- **Semantic Interoperability:** This is the most important level as not only data, but the meaning of the information is to be considered. The interacting systems need to agree on some common semantics for the exchange of information. Semantic representation requires contextual information and such kind of technologies are

still under development. MPEG-7 can be considered as one of the upcoming technology.

Ouksel and Sheth [1999] identified the following to be the key facilitators of the semantic interoperability.

- **Language Transparency**: This will provide the user with the freedom of choosing his/her won ontology.
- **Context Sensitive Data**: Filtering of information is done before returning the results to the user based on the context. This provides appropriate data to the requesters.
- **Rules for Interaction**: These rules specify the format in which the data types and messages are transferred without any violations or protocol issues.
- **Semantic Correlation**: Semantically identified data are to be represented in spite of issues of heterogeneity. This would allow development of ambitious applications over shared data sources.

Ouksel and Sheth [1999] state that, "*Metadata [40], Contexts [14] and Ontologies [40] are identified as the three key components of solution.*"

2.3 Metadata and Interoperability

Nogueras-Iso [2004] provided an overview of the problems of interoperability and huge data collections in case of geographic information systems. The case seems to be the same with Electronic Health Record Systems. When a huge amount of data is collected across multiple systems of different organizations, there is a high probability that the data is incompatible. In order to overcome these compatibility issues, certain systems adopted importing and exporting of the information rather than direct exchange. Unfortunately, those systems also were not successful as there were huge data loss issues during such cases. All such problems can be categorized as synchronization issues, as the problem is not about insufficient data, but about

inefficient utilization of the existing data. A proper documentation is necessary for management of the information in a structured manner. Metadata, which is the *"data about the data"* or *"information about the data"*, has a key role in this aspect. It has a procedure to organize and utilize the data for future reuse in several applications. With the improvement in the metadata the performance improvements can be observed. Heterogeneity of data or information existing in various systems always remains as the biggest obstacle for interoperability. Syntactic interoperability can be achieved by implementation using platform independent languages like HTML, UML, XML, etc. But this serves well only for syntactic Interoperability, which is insufficient for metadata description. Semantic interoperability is also equally important. Moreover the metadata interoperability need not be a cross domain issue.

By default the exchange of metadata is done by using XML based on XSLT. Nogueras-Iso [2004] state that, "construction of crosswalk specifies the mapping between two related standards, thus enabling communities that use one standard to access the content of elements defined in another one". But such a construction is tough and easy for error generations. Maintenance of such crosswalks is also a tedious task which requires special methods and additional methods for adjustments of historical data. Harmonization is necessary for the development of crosswalks. The author in this paper described a process that was implemented for achieving crosswalk which enable interoperations across few standards. The problems and solutions of metadata interoperability can be broadly divided into two approaches.

- Ontology based
- Specific crosswalk creations

2.3.1 Ontology Based

As per Nogueras-Iso [2004], *"ontology is defined as an explicit specification of some shared vocabulary or conceptualization of specific subject matter, and it seems to be an inadequate methodology that helps to define a common ground between different information communities"*.

Resource Description Framework: Most of the ontology based approaches depend on RDF technologies for semantic interoperability and data exchange. RDF is a w3c recommendation for modelling and metadata exchange. As per Manola *et al* [2004], *"Resource description framework (RDF) is a W3 recommendation for modelling and exchanging metadata, but the biggest disadvantage of RDF is its flexibility"*. It can be treated as a model that has independent or combination of metadata schemas. It describes the relationships between models in terms of properties/keys and values. RDFS, a description framework schema or Resource description framework Vocabulary Description Language (Brickley *et al* [2004]), provides a set of constraints for interpretations. This is used for describing the semantic meaning of the metadata contents for metadata schema. These documents enable reuse of other systems available on different sites. The RDFS techniques are combined with XML for appropriate representation of elements' outputs (standards) with respect to their inputs (names). The majority of the ontology based resolutions focusses on a unary interface for searching through heterogeneous descriptions of metadata. As per Hunter [2001], *"the wider the targeted scope of interoperability, the more difficult it is to achieve accurate and precise mappings"*.

2.3.2 Specific Crosswalk Creations

Interoperability of heterogeneous databases is similar to the metadata interoperability. Semantic heterogeneity is very essential for information exchange across various systems. The source schema and the target schema must be in sync without any ambiguity. As per Ceri and Widom [1993], there are four kinds of semantic conflicts that might occur.

- **Conflicts with names:** There might be aliases of the same data in with various database names.
- **Conflicts with domains:** Same concept can be expressed through different examples or values in different databases.
- **Conflicts with metadata:** Representation might differ in the levels (instance level/schema level).
- **Conflicts with structures:** Though the concept is same, databases might differ in organization.

Nogueras-ISO [2004] stated that, crosswalk tries to minimize all the above mentioned confits by establishing certain series of implementations using formal specifications. The steps of the entire process can be divided into four categories.

- **Harmonization:** In defining certain elements various metadata standards might use the same properties in a different ways. If there is a synchronized and formatted way to express similar properties, then metadata standards can also be presented in a similar manner. Thus, at the end of this step we get a harmonized specification for all standards.
- **Semantic Mapping:** In this step contents of elements in both the source and target standards is mapped. This is a very tedious task and as per Ceri and Widom [1993], metadata conflicts are detected and conflicts with names are resolved by the end of this step.

- **Conversion Rules:** When there are conflicts with metadata, i.e., schema or instance level representation issues or other hierarchy level issues, meta- data conversion rules are additionally required to solve such conflicts. As per Ceri and Widom [1993], it is expected that all the structural and domain conflicts are resolved by the end of this step.
- **Implementation of Mapping:** The crosswalks are implemented by using XML and XSLT as those are widely accepted and used. Thus, different standards, maintain only a single metadata standard by the use of automated crosswalks.

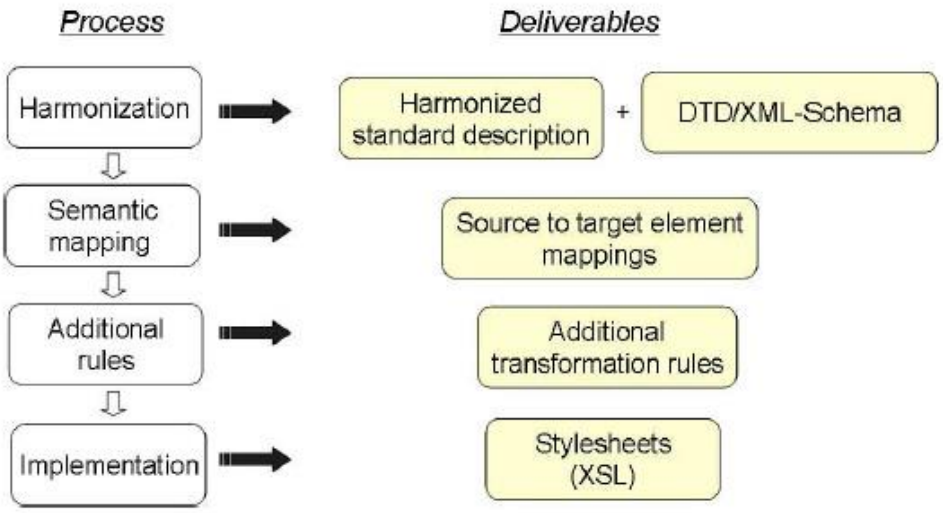

Fig. 1: Process Steps (Figure 3.3 page 118 of [40])

2.4 Multimedia Data and Metadata

Most of the research works have been focusing on heterogeneous data sharing methods and interoperability issues. Hossain and Masud [2014], in their work concentrated on the importance of multimedia database systems. That is, medical data can be represented in either text, pictorial, audio or video forms. All this data needs to be stored and shared amongst various hospitals, medical organizations and other institutions. Multimedia data need to be represented as metadata structures for querying the information. This is because the metadata describes

multimedia contents in a better manner resulting in faster data acquisition and also maintaining the accuracy. Identification of service providers is also a tedious task. Logical integration of patient data is encouraged for proper maintenance of records in distributed environments. There are two levels of data sharing.

- Schema level
- Data level

2.5 Metadata Management Systems

The authors Hossain and Masud [2014] state that, their main aim of the research was not only to transfer the data, but also allow communication of the various systems, enabling internal and external data querying from service providers. In order to achieve this, they proposed a *"metadata management system based on a distributed query processing mechanism".* The proposed distributed system architecture that enables multimedia data storage and access to it from different smart devices. Multimedia queries are different from relational or Boolean database queries. As the search techniques are more complex and needs to be improved. An image can be searched not only by its descriptions or subject, but also by using its name or the related patient name. Hence, in order to maintain such efficient systems, indexing and catalogues need to be employed.

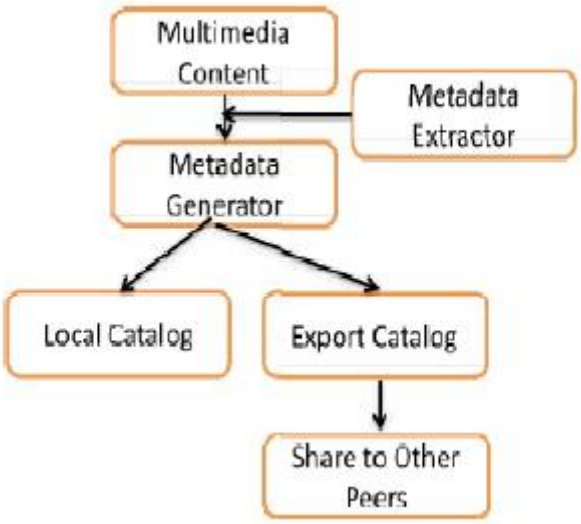

Fig. 2: Process of sharing metadata (Figure 1 page 2 of [23])

The system proposed, was based on dynamic decomposition of requests, queries and intermediate processing, data transmissions, etc. For generating multimedia contents through its peers certain steps need to be taken.

- Metadata structure is generated by using the metadata extractor.
- Local catalogues, storages and exports catalogues are used to store multimedia resources, metadata and subset of metadata respectively.
- Multimedia content is searched over the network.

The requested query is processed by using a two phases.

- **Matching and Finding:** Metadata strategies are used for this process. Most relevant data resources are fetched in this step. Thus, all the unnecessary data are neglected reducing the overload.
- **Collection and Execution:** The user or the requester will be able to choose the one best resource out of the most relevant data shortlisted through the previous phase and those nodes are accessed. Thus the results from the desired location are fetched.

The authors claimed to have tested their system by considering a huge metadata sample of multimedia contents. Set of classification and respective virtual resources were initially considered. In the second step couple of semantic descriptors were generated. Then precision, recall and accuracy for those were calculated. The authors claimed to have received more resources, but with lesser accuracy. In spite of this limitation, this has a positive impact as reduces irrelevant data transmission over the network and it expedites the entire process. The authors state that, they would consider *"development of a service oriented architecture with an agent based techniques for multimedia service composition, allowing more scalable and robust multimedia delivery"* to be their potential future work.

2.6 Data Mapping

The authors El-Sappagh, S., *et al* [2012], defined Electronic health records as the patient information that is stored in a longitudinal records. The broad structure of EHR systems (at a higher level) includes the database, data types, structures, authentication, authorization, network, architecture, workstations and so on, whereas the data (at a lower level) might include clinical observations, demographics, symptoms, basic personal details, billing, history, laboratory reports and related items.

Firstly, we will consider the situation at the higher level. Different organizations store the information in various databases (e.g. MySQL, DB2, Oracle, Access, SqlServer, etc.) Or independent file systems. The data stored in the databases might be again of different formats and data types. Number of attributes may vary and type of attributes may vary.

Secondly, at the lower level, when we analyze the data, an attribute might be addressed with a particular naming convention in one system and the same attribute might have a different name in another system. Typically the entity is unique, but there is an alias name (synonyms,

short names, etc.). For example, the name of the patient can be represented as a “Name”, “Given Name”, “Patient Name”, etc. So this kind of issues deal with the semantics. Once the data is mapped and integration is successful, interoperability is also very important in EHR systems.

When we want to merge the data from heterogeneous systems we need to observe the data and understand the similarities and need to perform the data mapping. Once the mapping is established, we can import the data into a uniform single system In case of medical databases, we might need to use the metadata or clinicians need to identify the vocabulary differences manually and only then these differences can be overcome. This is a very big disadvantage as in spite of multiple revisions a human prediction can still not be reliable as these are error prone.

Data mapping can be simply explained as an identifying and analysing the attributes and establishing a common platform for integration of data from heterogeneous systems. Mathematically, it can be defined as*, “for any attr∈ATTR (attr is an attribute and ATTR is a set of all relevant attributes), through mapping rule f, get o=f (attr), we call this a data mapping relation from attr to o where o is a mapping relation entity”* (Zhao *et al* [2007]). In this definition, “attr” represents the attribute, “o” represents the mapping relation and “f” can be identified as the function or the mapping rule that is applied over the attributes. The attribute could be identified as a database names, constraints, table names, filed names, database IP, data types or any other information depending on the system and mapping rule.

The attributes in the input systems might either have exact same representations or equivalent representations. The mapping might need to address various criteria, human interventions or identifications and also data cleaning to some extent.

2.7 Medical Standards

This section deals with the importance of the use of medical standards and metadata management with the help of medical standards. It also introduces the works of Lewis *et al* [2008], Rao and Makkithaya [2013] and Hossain and Masud [2014], in which various health standards and data management techniques have been discussed along with their implementations.

2.7.1 Use of Standards

In most of the domains, standards act as one of the major units of reliability as they represent the quality and significance. The same is the case with healthcare sector. By migrating to the higher level or version of a health care standards, providers then assume to have achieved interoperability. But that is not the case. Usage of standards in an appropriate format has to be done. However, there is an improvement in quality of data and accuracy.

Lewis *et al* [2008] presented few limitations for not achieving interoperability and strategies to minimize their effect. Most of the standards only support interoperability between machines. The authors state that, *"standardization of web services by implementing SOA architecture was proposed by the World Wide Web Consortium (W3C) [http://www.w3c.org] and OASIS [http://www.oasisopen.org]. These aim at attaining interoperability by providing standards and loose coupling of systems"*. But this is restricted only to machine to machine exchanges. In reality, especially the areas like health care demand for information transfers from institutions, systems, machines and people. This is because of the importance of all these actors and the information can be obtained from any of these sources. Thus standards which allow and encourage data channelization are required. There are certain important factors which are to be considered here.

- We need to understand the various levels of interoperability and what is the significance of each kind is.
- What is the role of standards and their importance
- Limitations of the standards
- If the standards are sufficient or not
- Are all the standards perfect
- How to overcome the issues

All these factors are needed to be considered and understood in detail.

- **We need to understand the various levels of interoperability and what is:** the significance of each kind: To understand the levels of interoperability, we need to know the basic classification of the interoperability standards (organizational, system, syntactic, semantic). A detailed explanation is presented in the previous sections.
- **What is the role of standards and their importance:** When welcome to the second point, that deals with the importance, Lewis *et al* [2008] have stated the example of various internet based on standards such as TCP/IP, UTF-8, SMTP, XML, and HTML, demonstrates the importance of standards. The standards basically reduce the overload as they provide the base level protocols. Thus the functionality is the only key aspect that needs to be concentrated by the administrators. These standards provide level one and level two interoperability and thus they help in marketability and quality assurance and grabs the belief of the customers.
- **Limitations of the Standards:** There are various advantages and enhancements in the levels of standards. However, in spite of the release of higher versions, there are limitations within few semantic and organizational standards. This is because it is very difficult for usage and implementation of a particular language to represent entire domain knowledge and also because of the human interference. Semantics deals with

the meaning of the data which might vary depending upon the perspectives and human interferences. Similarly organizational deals with domain specific representations and usage of various languages like BPEL, OWL, BPML, etc. There is a high chance that there might be different from the expectations and reality and this is always increasing. Thus the workflows, demands, implementations and requirements of the organizations keeps changing constantly which causes major limitations.

- **If the Standards are Sufficient or not:** In the most idealistic case, all the standards should be similar or exact and thus they should support interoperability with all other implantations. But in reality, this is not achieved because all the organizations modify the usage of the standards by adding some extensions or performing customizations as per the requirements. There might also be issues with some default standards. The standards that were initially laid are treated to be the best or the default standards and migration from them to the new ones wouldn't be a desirable option for many organizations.
- **Are all the standards perfect:** Not all standards are considered to be perfect? There are certain standards which are bad or not to be considered. One of the key reasons for not selecting a standard deal with its specification. Standards might me underspecified, over specified, inconsistently specified, not stable or irrelevant. Such kind of standards are definitely not preferred. There might be issues with conflicting standards also where in few conflicts might overlap with each other, exclusive to limited set of standards or inflexible. All such standards are not considered or preferred.
- **How to overcome the issues:** After understanding the types of standards, the good and bad standards, various levels and limitations, the interoperability levels can be identified easily. It is important to know that identification of levels of interoperability is not important for solving the interoperability issues. Understanding the existing

standards and analysing the issues with the existing standards is very important. Once that is done, we need to overcome the issues identified in the available standards by certain modifications. A general understanding of standards is necessary for interoperability. It is not possible to attain ideal solutions all the time. However, we can always try for the best solution.

2.7.2 An Informal Method

As per Lewis *et al* [2008], *"A standard is established by consensus among stakeholders and is approved by a recognized body that provides rules and guidelines for activities and adoption of standards is the only realistic way to achieve interoperability. Also, the standard should not be underspecified, unstable, inconsistent or irrelevant"*. Updating to a higher version of standards doesn't ensure interoperability. A thorough analysis of standards needs to be done for choosing the appropriate standard.

Rao and Makkithaya [2013] stated that, after the identification of the standards, tools for document model design and implementations are lacking. The authors presented a series of steps which would help in identification of the appropriate standards for data sharing. They claimed to have considered a real case where they considered an organization called "RMCWH" with fifty thousand patients. The patients need to be redirected to other physicians as per the requirements and thus their medical records were to be shared. An informal method was adopted for the identification of the appropriate standards have been presented and finally, after the identification the results have been analysed.

According to authors, software development life cycle should also be considered for achieving interoperability. They analysed various literature works on standards and interviewed medical practitioners and nurses. Then use cases were developed accordingly.

Then a series of steps were identified and were applied to maternal data in attaining syntactic interoperability. The steps presented by Rao and Makkithaya [2013] are as follows.

- **Requirement Analysis:** Initially the data that needs to be transferred and corresponding levels of interoperability are to be identified. Initial aim would be to obtain syntactic interoperability. Data Set Identification: The minimal data set for the data which is suitable for UML representation is to be identified. Then a UML diagram representing various entities, their members and relationships is to be developed. Data type identifications also have to be made in this step.
- **Identifiers:** The next step would be to define identifiers of all the described entities. A unique identifier is to be chosen which should be acceptable throughout all the systems available. Entire patient information would be available through these identifiers. Semantic interoperability is achieved by allocating IRIs (Internationalized Resource Identifiers).
- **Vocabulary Adoption:** In the field of medicine, vocabularies represents unambiguously defined medical terms are consistent and can be used in communication. This step needs to be done before adopting any particular standard.
- **Standard Evaluation:** Identification and awareness of standards are one of the key responsibility of the medical individuals. The standards should support various products and vendors. The authors also present a table format where the data and attributes are examined and matched with most suitable standards in order to obtain the best fit.
- **Implementation:** Finally in this step the selected best standard is implemented and depending on the outcomes, it is concluded if the standard is a good or not. If there isn't any suitable standard that is found, then a new standard can be defined.

- **Graphical User Interface:** Once the entire data is obtained, it has to be presented in an appropriate presentable format so that the medical practitioner can understand and analyse the information in a quick and efficient way. Most of the practitioners prefer graphical, pictorial or tabular representation over the information presented in words and characters.

As the aim of the authors was to transfer information collected from various patients of RMCWH to a public domain in an accessible format. For this XML was chosen as it is both machine readable and understood by the humans. XML makes sure about the syntactic interoperability. Thus, in the process of data sharing, the syntactic process deals with obtaining the information and semantic process deals with understanding the meaning and enable future interpretations.

2.7.3 HL7

As per authors Häyrinen *et al* [2008], Open Systems Interconnection (OSI) Reference Model was developed by ISO (International Standards Organization), and only after its discovery with 7 different layers viz., Physical layer, network layer, transport layer, datalink layer, session layer presentation layer and application layer, the application layer was identified as a medium which allowed the data exchange between various application processes further facilitated the discovery of health level seven standards (HL7).

Diagnostic systems or hospitals have different mediums of storing the patient data. These systems not only store test results but also include information about billing, symptoms, medicines, methodology of treatments, etc. Multiple systems may communicate with each other for analysis, transfers, research and other purposes. HL7 acts as a medium for serving the purpose of providing this flexibility across different platforms.

As per authors Beeler [1998], Health Level 7 (HL7) began in 1987 as a consortium founded at the instigation of a group of health care providers, who set out to develop a protocol for the exchange of healthcare information in clinical settings. The key features of HL7 can be identified as below.

- HL7 is a non-profit oriented.
- All the standards developed or released are ANSI accredited.
- HL7 standards are globally accepted and already being used in many continents like Asia, Africa, Europe and North America.
- HL7 follows a pragmatic approach unlike other static standards.
- The mission of HL7 was to achieve Semantic Interoperability.
- As of now there are various versions of Messaging Standards and also standards for representing clinical documents.

The messaging standards act as the most important part of HL7 leading the core concept behind HL7 to be, *"when an event occurs and is recognized by the healthcare computer application, then a message is sent back to one or more recipients as a response"*. The structure of one of HL7 version 2.3 messages can be represented as shown in Figure 3.

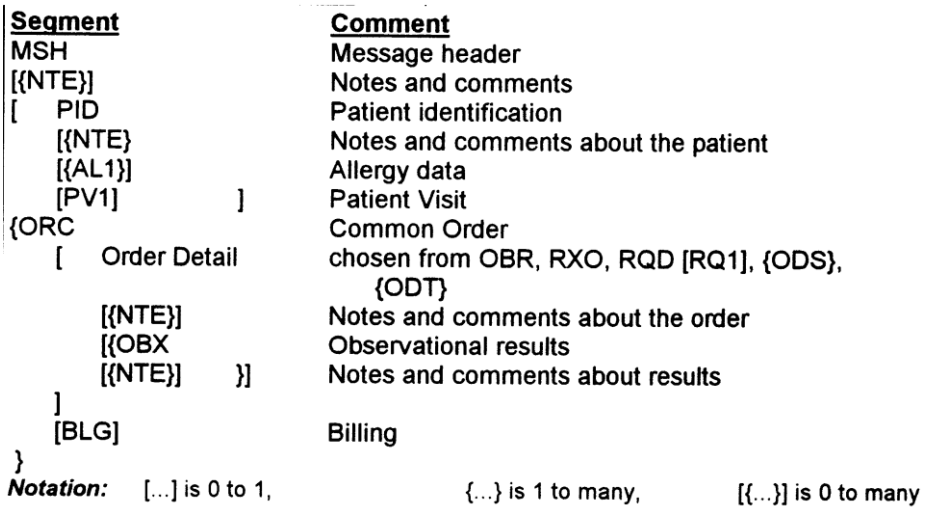

Fig. 3: Segment diagram for one of the HL7 Version 2.3 messages (Figure 1 page 2 of [4]))

The messages have ASCII strings which are divided into segments and further into fields. Every filed or segment represents information about a concept belonging to the clinical domain. Every part of the message acts as a place holder and have a specific meaning, which is identified by the systems that are using HL7 as a medium of information exchange. Also, the upward compatibility of these standards is of great advantage in case of real time data management systems in both distributed and centralized environments.

The HL7 messages are generally divided into segments, separated by delimiters. The message can be a string or stored in a file and is transferred by using a TCP/IP (Transmission Control Protocol/Internet Protocol). For every request that is generated, a message is generated as a response and is transferred and then the requester sends back and acknowledgement. Example structure of the message can be represented as shown below (by authors **Liu, J. [2012]**).

"MSH|^~\&|EB^EB^GUID||DPS^DPS^GUID||199601061000||ACK^A02|

C7E7-85-11-A5-004005|P|2.3|AL|NE

MSA|AA|000002"

The above example structure is from page 29 of Liu, J. [2012]. Each component has a specific meaning and can be clearly understood from table 1.

Notation	**Element Name**	**Element Meaning**
|	Field separator	Hardcoded
^~\&	Encoding characters	Hardcoded
EB	Namespace ID	Sending application name
GUID	Universal ID type	Hardcoded
DPS	Namespace ID	Message application name
199601061000	Date/Time of message	Hardcoded
ACK	Message type	General Acknowledgement
A02	Trigger event	Trigger event of acknowledged message
C7E7-85-11-A5-004005	Message control ID	Hardcoded
P	Processing ID	Hardcoded
2.3	Version ID	Hardcoded
AL	Accept acknowledgement type	Always
NE	Application acknowledgement type	Never

AA	Acknowledgement code	Application Accept
000002	Message control ID	Hardcoded

Table 2: HL7 message standard notation **(Table 2.1 page 29 of [28])**

2.7.4 Logical Observation Identifiers Names and Codes (LOINC)

According to authors McDonald *et al* [2003], *"LOINC has been identified as a universal code system for identifying clinical observations and laboratory terms. These terms when used in HL7 messages clinical and research clients can easily integrate the results data across various repositories"*.

The main purpose of developing LOINC was to provide a unique identifications for the observations that are used in HL7 messages. Most of the clinical/ diagnostic information is exchanged using HL7 messages in many electronic health record systems. Hence the development of LOINC reduces the ambiguities and manual interpretations in HL7 messages. These HL7 messages have independent records for every unique medical observation. In the HL7 message structure, *"the filed carrying the observation identifier is referred as OBX-3, and respectively, field that carries the same observation's value is called OBX-5"* McDonald *et al* [2003]. Until recently, most diagnostic centres and laboratories used their own convention for representing and transferring the information using HL7 messages. But this leads to an overhead of decoding the textual information and clinical representations of independent systems by other end systems. The issue is not only with the textual interpretations of same language, sometimes the language differences might also create a big issue for exchange of data or interoperability. Hence we can say, the OBX-3 codes might either be local representations or idiosyncratic codes, which need to be further processed, generally by a human operator or a medical administrator, creating a huge task with lots of manual efforts.

Hence, obtaining a unique identifier which differentiates clinical terminology was the main aim of LOINC. Thus, it provides a coding system for the observation or clinical identifier fields in HL7 messages. The biggest advantage of LOINC is, it also supports internationalization or addresses the language barriers across heterogeneous systems. If we consider any observation with a set of questions or attributes and a set of answers or values the initial part (questions or attributes) are identified with a specific code called LOINC code. LOINC database is open source and available for usage. If we identify any new observations that need a new code, we can request LOINC and propose for a new term. Hence it is a very useful and upcoming standard vocabulary useful for clinical terminology identification as it supports HL7 messaging system. Initially, LOINC was used only for clinical observation information exchange, but now, *"LOINC is also used in the areas of Communication and Digital Imaging of health data (DICOM) ultrasound messages and in Clinical Data Interchange Standards Consortium (CDISC) pharmaceutical industry, messages identify clinical and laboratory observations, respectively, and could well be used in clinical and research databases for the same purpose"* (McDonald *et al* [2003]).

Another advantage is that, the codes are associated with detailed meaning of the entities and related information required for identifying the entities. Some of the examples of LOINC codes are as shown in the table 2.

CODE	COMPONENT	PROPERTY	TIME	SYSTEM	SCALE	METHOD
8302-2	BODY HEIGHT:	LEN	PT	^PATIENT	QN	
3140-1	BODY SURFACE:	AREA	PT	^PATIENT	QN	DERIVED
8331-1	BODY TEMPERATURE: TEMP	TEMP	PT	MOUTH	QN	
8632-2	QRS AXIS:	ANGLE	PT	HEART	QN	EKG
8642-1	PUPIL DIAMETER:	LEN	PT	EYE	QN	AUTO
21611-9	AGE:	TIME	PT	^PATIENT	QN	ESTIMATED
19867-1	CAPACITY.VITAL:	VOL	PT	RESPIRATORY SYSTEM	QN	

9279-1	BREATHS:	NRAT	PT	RESPIRATORY SYSTEM	QN	
11882-8	GENDER:	FIND	PT	^FETUS	NOM	US

Table 3: Example LOINC terms and names (Table 3 of page 4 of [35]))

Each code or LOINC term is associated with six other sub components which provide the details on the term. As we have seen in Table 2, the six identifiers or parts associated with the code can be identified as shown below.

- Component: e.g., height
- Property measured: e.g., length
- Timing: i.e., at a point, 24 hrs
- System: i.e., entity or object
- Scale: e.g., Quantitative, nominal, narrative
- Method used to produce the observation: e.g., recorded, estimated

LOINC repository is a Microsoft Access database and is available to everyone free of cost and can be downloaded from the official LOINC Regenstrief website (*https://loinc.org/*). It basically has two important components, the first is the master table storing all the codes with their associated components and the second table stores the information about all the mapped terms or codes with mappings. For all the users who are not comfortable with using the Microsoft access version of the database, the Regenstrief organization also provides a desktop application which acts a graphical user interface based tool for browsing and accessing the LOINC repository. This tool is named as Reginstrief LOINC Mapping Assistant (RELMA). This is also an open source software that can be downloaded for all research purposes.

LOINC terms can be searched with the codes, short descriptions, long descriptions, unit of measurements or any other information. We can also specify a file to be imported with a set of terms to be searched for. The parsing of local names into identifiable words is done by the software for mapping and with the specifications, the search is conducted. It returns all the possible potential matches with the search string and the search can be further processed with more specific criteria. If there are any new terms that are not represented in the repository, there is also an option to create new LOINC terms by requesting the Reginstrief organization. It the term is approved, then the update is made in the repository globally. But the disadvantage with this RELMA application is that it is only desktop application. There are no Android/ iPhone compatible application versions of the same software. Figures 4 and 5 represents a sample search tabs and strategy specification screens. LOINC doesn't have the ability for direct representation of multimedia data but it stores metadata information of such data. Hence an Interpretation is required for recording such data.

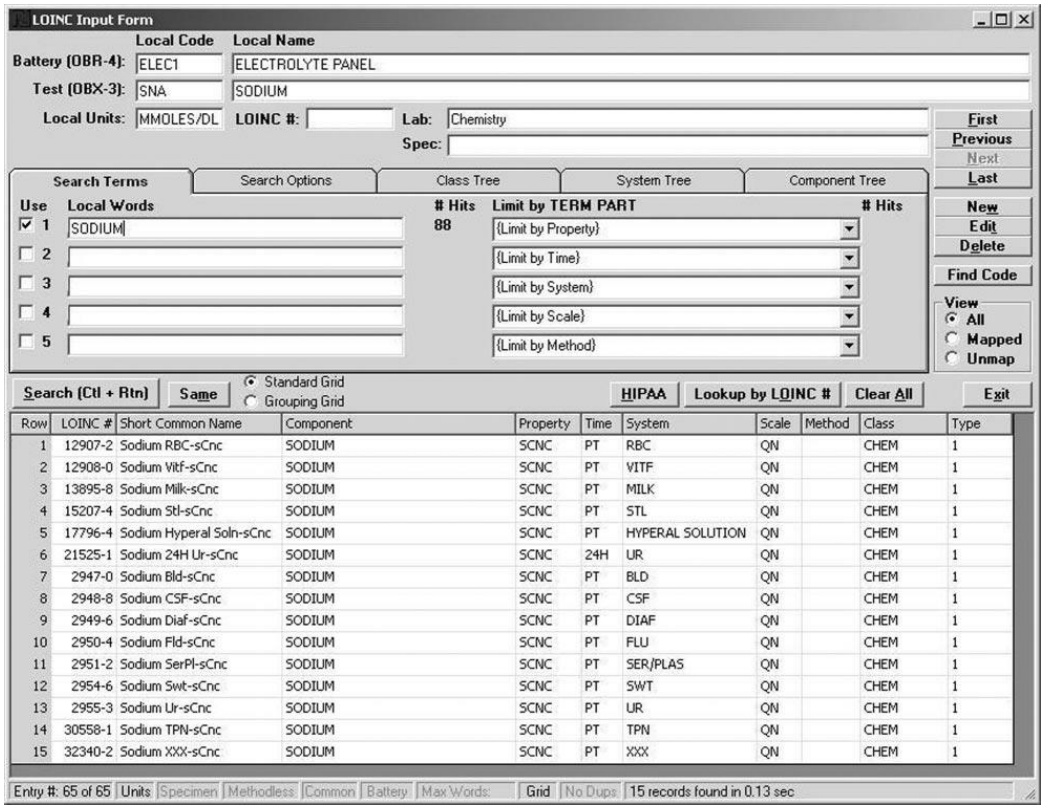

Fig. 4: LOINC terms Search screen (Figure 1 page 5 of [35]))

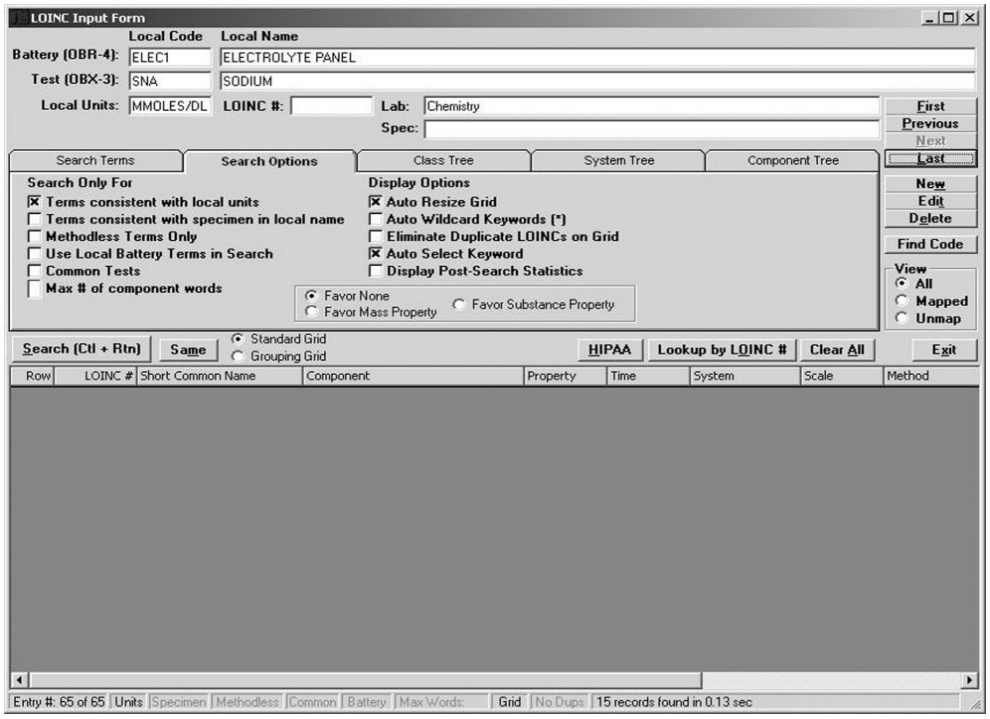

Fig. 5: LOINC strategy specification (Figure 2 page 5 of [35]))

Clearly, reliance on humans to decide about specific data that goes into LOINC cannot be sustained. Into the future, there should be increasing use of automation, supported by semantics, machine learning, ontology creation and ontology alignment to achieve the goals of an up-to-date, well managed LOINC. These approaches are beyond the scope of this thesis, but will be addresses briefly in our discussion in later chapters.

3 Existing Work

Considering the ideas and strategies explained in the previous chapter, there is an increasing need to develop a standardized data mapping framework while addressing the problem of interoperability. Various researchers tried to solve the above mentioned problems. The remainder of this section tries to present an overview of previous works that tried to solve the problems of data mapping, efficient data modelling and interoperability in electronic health record systems. Most of the researches used XML based standardization techniques for achieving data mapping and several data modelling techniques like relational database

management systems, entity attribute value modelling, object oriented database management models, etc. This section presents a detailed explanation of all such key ideas in this area.

3.1 Overview of Interoperability and MML

This sections deals with review of work of Dogac *et al* [2005]. The existing Health Information systems (HIS) were not very effective as the information is presented in several proprietary formats. There is also a huge problem because of the availability of a multitude of medical information systems. Thus, it leads not only to interoperability issues, but also differences in health record standards that are observed in various organizations. This restricts the reuse of information as it limits the exchange of information between various systems.

A proper solution to this problem would contribute to efficient patient care as the reduction in interoperability issues and provide multimedia support for the existing health systems maintaining the security of the data. This will help the systems to become more effective by interacting with various systems or sites that are available and gather patient information from various places.

The authors Dogac *et al* [2005] state that, the medical information is stored in various proprietary formats which include RDMS systems, hard copies, structured and unstructured documents and various other forms. This is the major reason for the interoperability in the domain of medicine or health care. Interaction and information exchange of health care data across various health care systems would help in faster access and analysis of data along with the reduction in duplication. Interoperability is used with various meanings. We have already seen the definition of interoperability in Chapter 2. Authors Brown *et al* [2000], also described "*Interoperability with regard to a specific task is said to exist between two applications when one application can accept data (including data in the form of a service request) from the other and perform the task in an appropriate and satisfactory manner (as judged by the user of the*

receiving system) without the need for extra operator intervention." It is observed that interoperability issues are not just limited to one level instead they are existing are at various levels. The authors Dogac *et al* [2005] state that various approaches have tried to solve this issue but weren't successful completely. Most of the approaches were limited to syntactic levels but were not successful in attaining interoperability at schema level and data level. In order to address these issues certain health standards which try to structure the clinical data and support exchange of information are been developed. The authors presented an analysis of various electronic health record standards also addressed few issues. The key points based on which the standards are analysed are as follows.

- The interoperability levels
- If it can support multimedia data and specific data
- If it permits the combination of various standards
- If the standards are acceptable in the market

Based on various parameters, various electronic health record standards have been discussed and the authors state that most of the standards are specific to certain services. Thus an optimal solution would be to use combinations of the electronic health record standards for better performance.

3.1.1 Medical Markup Language

A special language called MML (Medical Markup Language) (Araki *et al* [2000] and Guo *et al* [2004]) has been introduced. As per the authors the main purpose of developing MML (Medical Markup Language) by Electronic Health record research group was to facilitate a standard way for exchanging the medical information across several systems. Exchange of MML documents can be done by using any electronic communication or by using HL7

messages. They use XML based languages with special headers and markup sections. But this was developed and used mostly in Japan.

The authors Dogac *et al* [2005] have stated that*, "in any case conformity to any or combination of all of the standards would not be the solution to the interoperability issues, as there might be exceptions when certain institutions use incompatible standards. Hence interoperability of the electronic health record systems can be addressed by using semantic interoperability, which demands the data description in domain specific formats, especially in fields like health and medicine".*

3.2 Semantic Interoperability and Frameworks

This section focuses on works done by Ryan *et al* [2007], Ryan and Eklund [2008], Grechenig *et al* [2008] and Xiao-guang *et al* [2009], These authors proposed various frameworks that work efficiently to support semantic interoperability.

3.2.1 Semantic Interoperability and HL7 Standards

One of the biggest problem with healthcare data exchange is due to unambiguity that occurs due to multiway representation of same information. If the data representation is done in a standardized way, it can be reused in future. This kind of standardization avoids duplication. In simple words *"reinvention of a wheel is not a good idea".* Thus the primary level patient data can be analysed and can be used at secondary level decision support systems and health recording systems.

Ryan *et al* [2007] stated that, HL7 version 2 has wide implementations, but it lacks incompetence with computational systems which require interoperable healthcare services. The later versions integrates new features for supporting data representations and messaging. HL7

framework isn't very efficient in representing clinical concepts. Thus SNOMED CT standards have been introduced by the authors. They provide a common platform for sharing and accumulation of data present across the medical systems and available through internet.

This terminology uses the knowledge of healthcare whenever required as per the users. Ryan *et al* [2007] conducted an experiment collecting eight clinical observations. These eight observations were taken into considerations for all the patients and then transferred from a common database to a PDA and again from PDA to database. The observations taken into considerations were as follows.

- Weight Pulse
- Temperature
- Blood Pressure
- O2
- Saturation
- Blood Sugar
- Levels
- Urinalysis
- Respiration

Codes for each of the above mentioned observations were determined by the physicians or clinical representatives. All the eight observations were represented in HL7 model. It is observed that there are issues when different people use different representation methods. All the information from the findings can be represented in a code-value formats and then used in decision support systems.

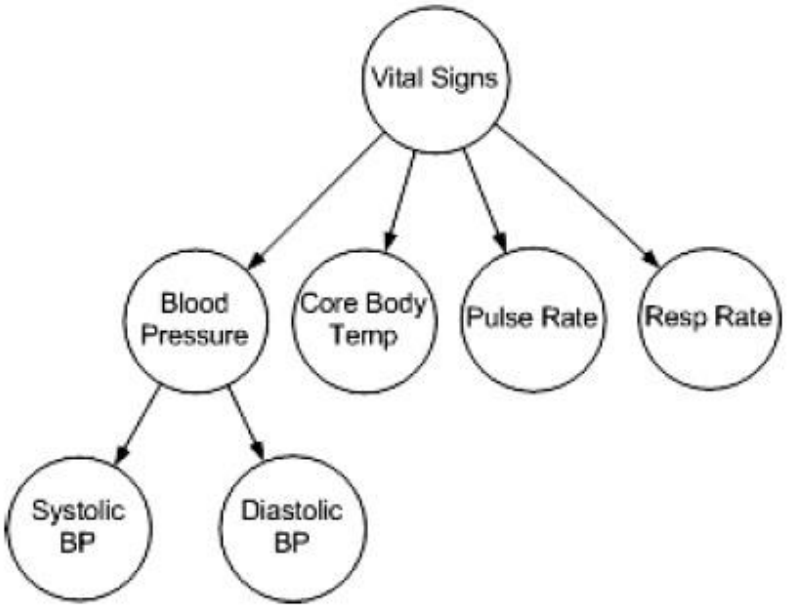

Fig. 6: Codes used in Clinical Observation (Figure 1 page 5 of [47]))

3.2.2 A Health Service Bus Architecture

When there are huge collections of data from heterogeneous systems, demanding interactions between themselves, their semantics come into play and it is necessary for them to have semantic interoperability. The death rates have been constantly increasing because of the improper usage of existing information. Intelligent health care systems which can analyse the existing information and provide appropriate predictions and suggestions are to be developed in order to improve health care of the patients.

As per Chappell [2004], *"Enterprise Service Bus (ESB) is a term used to describe a middleware software architecture with a standards-based messaging engine, which is event-driven and provides foundational services for more complex software systems"*. It works on cross platform and cross language environments. It uses various programming languages for managing applications but XML is used as a common means of communication.

Ryan and Eklund [2008] proposed a health service bus (HSB) based on the above ESB thought. HSB acts a communication media between heterogeneous medical systems and other

soft wares and hardware and provide services. As we have discussed in previous subsection, SNOMED CT is a small subset of research that is been done. The main idea behind the entire process is the conversion of different viral signs or observations that were discussed in last subsection into XML format and store into an XML database structure. HSB has been built on SOA (Service oriented architecture) for providing a secure distributed system environment. Various PDAs, independent systems communicate with the HSB in order to fetch the information from medical databases and servers.

This entire architecture can be explained in Figure 7. In the system proposed by the authors, the software implements HL7 messages depending upon the users' interests. The physicians enter the patient details on the front end and that information is stored. New observations made can be appended to the stored information by the physicians. For every entry there is a respective SNOMED CT element along with values and textual data.

Finally a dynamic model is generated and the outputs can be observed in an XML format. Thus the HSB model enables semantic interoperability in healthcare and enables messaging environments of large hospital systems. Translation services implementation has not been done and it can be considered as a potential future work.

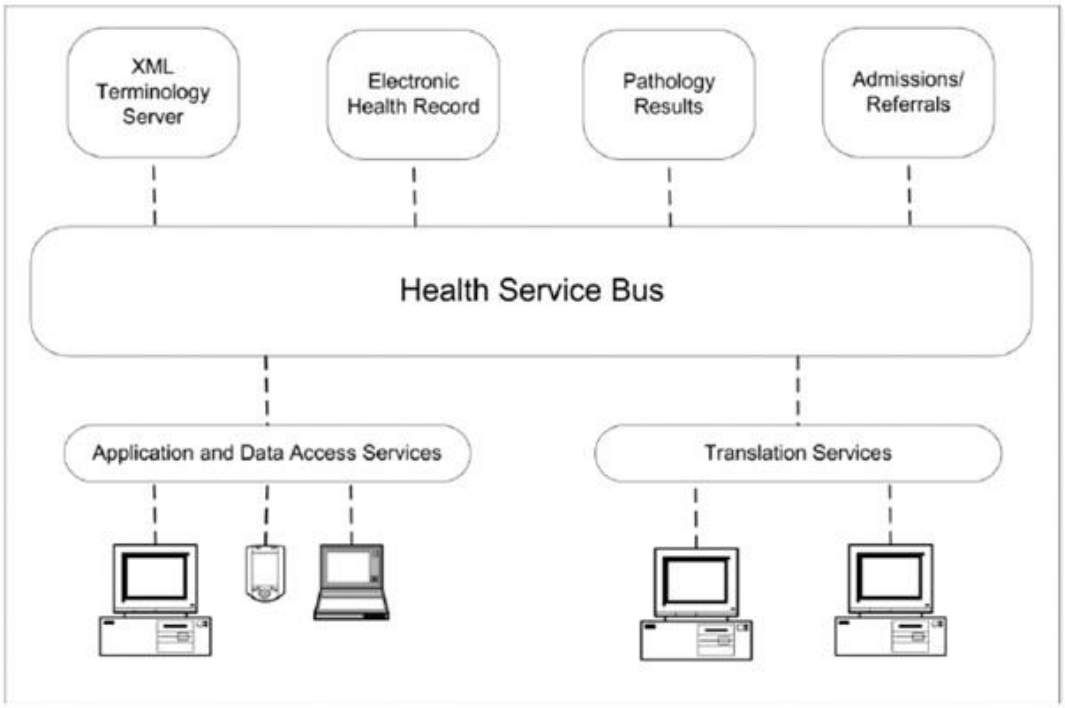

Fig. 7: Architecture of HSB (Figure 1 page 3 of [46]))

3.2.3 Interoperability in Health Care Network Designing

There are certain design criteria with respect to e-Health, which are to be taken into consideration for developing an efficient network design. As per Grechenig *et al* [2008], the following are few of the important criteria.

- Interoperability
- Security
- Tolerance
- Flexibility
- Reusability
- Availability
- Maintainability
- Openness

- Performance
- Scalability

Interoperability acts as one of the crucial criteria. Data from various medical entities, hospitals and health centres are collected and are related for reusability in future. In the design proposed by the authors, all the peers have to accept to use a single TCP-IP transport mechanism and the single connection is shared by all the nodes that are connected. The authors claim that, *"they have used TCP-IP as Layer 3 networking protocol, but can be used as Layer 2 based interoperability".*

Grechenig *et al* [2008] state that, *"none of the existing networks provided a 2X10 GB bandwidth/second as per the design of the health network design proposed. Most of the networks have been following out-dated technologies".* In order to satisfy the state of requirements a MAN is preferred over LAN and WAN. Huge bandwidths networks are required in order to maintain health systems and it is also necessary that multiple nodes of similar bandwidths are used for connections.

3.2.4 Interoperable Health Information System Based on SOA

As mentioned in all the previous sections one of the biggest challenge in health care systems in interoperability. Extended and efficient practices and interaction is very essential for developing a high quality, specialized distributed health systems. In spite of various institutional limitations the medical treatments have to be done in a dispersed manner. Limitations to such kind of treatments are caused due to interoperability issues. There are wide range of complex heterogeneous heath care systems across various hospitals or institutions. Integrating the data spread across all these systems can be done by collecting all the data from various systems and storing them in a uniform management system. But this procedure is not

recommended as it requires large storage spaces, management and maintenance complexities and other problems of updates and migrations. The second approach is to establish a distributed system, which is not cost effective.

Xiao-guang *et al* [2009] presented an SOA based interoperable information system which provides an effective solution for the integration of such heterogeneous distributed systems. Each and every function is defined as an independent service with a respective interface and its implementation in a Service Oriented Architecture. Thus various services are split and combined as per the necessity in SOA based implementations. The design of the proposed Health/Medical information system broadly has the following parts.

- **System Server:** This consists of patients, hospitals, clinics and other institutions. It provides authentication, authorization and other special services to the users.
- **Interface Server:** Each hospital or an institute will have an independent interface server. The level two authentication can also be done here. The information here is stored in XML format. Information from these interface servers are then transfer to the system server later.
- **Broadband Network:** This is essential for establishment of the connection.
- **Service Providers:** Creation of various web services and registrations are done by providers.
- **Service Brokers:** It's a platform where registration and classification of services is done.
- **Requesters:** The requesters can be either the hospitals or the patients who need to utilize the services.

The system designed is based on service oriented architecture and thus various functions are represented as services which can be utilized independently or grouped together. There are various layers that are designed with specific tasks assigned.

- **Portal Layer:** The portal interfaces are categorized into various levels depending upon the user (i.e., doctors, administrators, patients, etc.)
- **Connection Layer:** The connection interfaces supports various databases and transport layer protocols depending upon the network (e.g.: sctp, http, soap, ftp, etc.)
- **Business Process Layer:** The service interface is operated by the business process engine in order to serve requests from various requesters. Thus services are managed dynamically according to the business process alterations.
- **Service Layer:** This layer provides data services, log services and security services. Data can be accessed in both pictorial and theoretical formats. Log services ensure there is a record of all the events or updates that are run on the system. Security and confidentiality are key factors in maintenance of health systems. This layer ensures the authentication and authorization of the users connected to the systems and also maintains corresponding records and log files.

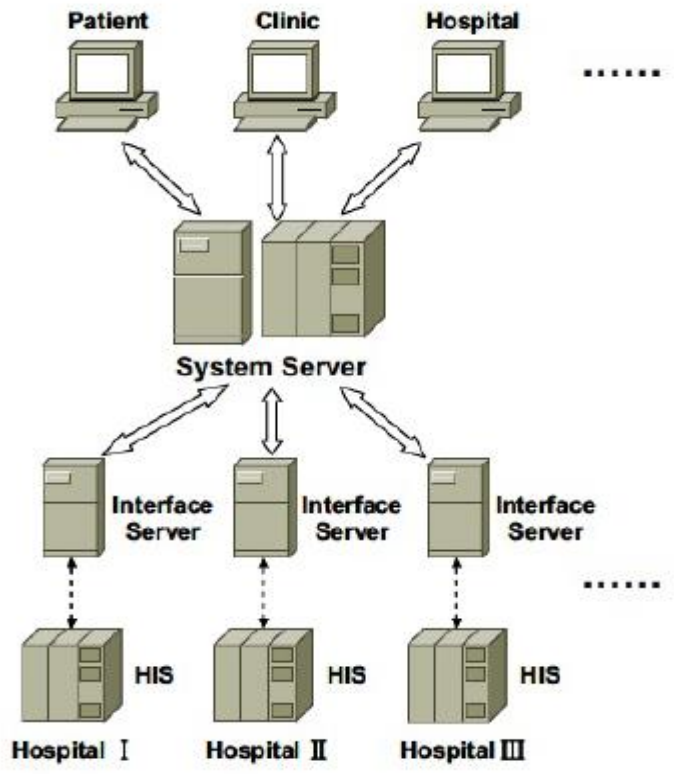

Fig. 8: Interoperable medical Information System (Figure 1 page 2 of [52]))

Thus each layer has specific tasks and responsibilities and they all run in parallel for delivering an efficient information system. Xiao-guang *et al* [2009] claimed to have used .NET as their development platform and key technologies like WSDL, SOAP, UDDI, BPEL, XML, etc. can be used in web service technologies. The use of SOA here will minimize the maintenance costs and maximize the services offered. Such kind of medical systems not only enable the patients to access their diagnosis information but also help the doctors or the medical practitioners in understanding the diagnosis history and better treatment.

3.3 Importance of Data Quality

This section describes about the works of Halevy [2011]; this paper introduces the importance of quality of data that is being managed. There is huge amount of data available in healthcare but the quality of data is important. Efficient utilization and access of existing data, (instead of read only purposes) is a crucial task. Health systems that are limited to a single organizational data are not sufficient for the analysis of data and improvement of the treatments. The patient data needs to be up to date and accessible to physicians from different hospitals in order to maintain highest level of quality treatments. Data security is also an essential factor as erroneous data storage has to be avoided. There is huge amount of data available with various health systems and this has to be harmonized to generate much cleaner information.

Most of the organizations are trying to migrate from paper based medicine to electronic medicine but they still lack in combining data sets and generating meaningful analysed reports. The systems must be not only able to access the historic data but also integrate data collected over the periodic intervals for better analysis. Health standardizations and engines are not sufficient for connecting the medical applications. The Health information systems must be able to use the static data analyse the patterns and provide an intelligent report that might be

useful for future predictions and improvements in the treatments. As per Halevy [2011], Harmonization of data occurs in two levels.

- All the information is converted into a common format accepted across various systems and this is done by interoperable systems.
- The information is semantically organized making sure that there is no loss in both the data and its meaning.

Thus physician's productivity and quality of the treatment can be improved by semantic interoperability. When this process in employed, all the duplicate and outdated information is erased and only valuable information which is semantically harmonized is stored. This reduces the work of the physician and enables a better understanding of the diagnosis and medical history. A semantic translation of information is done but still the integrity is not disturbed. The systems require next level of intelligence where in differentiation between useful and not useful data can be done automatically by background processes should be possible. Healthcare systems requires semantic harmonization of medical information which is not only read only but also serviceable.

3.4 Row Modelling Approach for Structured Data Integration

Medical data is recorded in multiple environments in heterogeneous systems. The data recorded may be of different types including billing information, clinical history, tests, notes, examinations, medical or laboratory reports, etc. Information entry might be done once or might be updated in a timely manner depending upon the system. Also the details of the information might be recorded by different physicians in a different manner. Hence there is a variation in data details, data types, and time of recording, quality and size of the data. Hence

row based methodology adopted by authors Los Renske K., *et al* [2004] is very useful for data sets that are evolving and handling variations caused by multiple sources.

The row modelling technique stores the attribute information to be stored in a row based or modelled table unlike the usual column model where the attributes act as column headings. Hence we can say that column to row transformations is the key idea behind the row modelling methodology. In this model each table contains only three columns each representing the identifier information, attribute name and its respective attribute value. The metadata information in this model is stored separately in some other tables, instead of storing the metadata or data definitions in the single table along with the attribute information.

In regular column based approach each record holds set of facts or detail information about a patient, where as in row model approach multiple records may be used to store the information of a single patient. The biggest advantage of this being, attributes who have some value assigned are only stored, if any attribute doesn't have any details recorded then it is not stored in the row model unlike the column model where a null value is assigned in the dedicated column. The authors also claim that, *"there is an advantage of separating the metadata from the physical data schema because it eliminates the need for changing the physical data structure in case the data set changes, only the metadata content needs to be altered. Whereas, a conventional column-modelled approach would hold metadata in table definitions and relations between tables."* This would cost a lot because, whenever there are changes in editing, addition or removal of columns in a relational or column based approach entire structure is edited. Hence it is difficult and expensive to change in the structure of the database itself.

In this row based approach any kind of changes would cause change in the content of the database tables not effecting the structure. As both attributes and values are stored in rows independently any change is values will change the content of a cell, any deletion of attributes

or values will cause the deletion of rows, not deletion of columns and finally any additions of values or attributes will cause addition of rows to the table wish is possibly very simple task and not expensive when compare to column changes or column deletions and column additions, which also involving storing null values for certain records.

The authors Los Renske K., *et al* [2004] developed an OpenSDE model based on the row modeling data management technique. They developed an interface for data entry and data storage at the back end is following the row model instead of the traditional relational model. Also as an extension to the row model the authors used an additional column to store the data type of the attribute. Which means, the row model has four columns instead of three representing the patient identifier, attribute name, attribute value and the data type of the attribute.

To summarize, Los Renske K., *et al* [2004] Implemented Row Modeling Methodology for storing heterogeneous data sets and then OpenSDE was developed to support structured data entry only and it doesn't model ontology. Finally, OpenSDE had an additional data type reflection and the descriptions of multiple occurrences was done. Thus the data mapping was done successfully but interoperability was not addressed.

3.5 XML based Framework for Interoperable Healthcare Systems

Achieving interoperability between heterogeneous information systems was a huge challenge as we have discussed in the previous sections. The complexity is even higher in case of medical data as the diversity and variations are even more. As a solution to this problem of data integration and interoperability, authors Kumar *et al* [2010] introduced a framework for exchange of medical records or data across two or more systems using XML as the key concept.

The two important aspects the work tried to address was the need for providing easy access to data for the clinical decision support and provide patient safety. For this, exchange of electronic health records (EHR) was the key task. Hence the main aim of the work was as follows

- To identify the challenges and impact of enabling interoperability.
- Identify different methods of interoperability.
- Provide secure data access from healthcare systems.

The solution provided by the framework proposed can be divided into different steps as shown below.

- Fix a standard object structure, by both the systems as a medium of exchange.
- Send the EHR to other system in an XML based format.
- Map the XML format to the agreed data object structure.
- On the receivers end, collect the information.
- Store the data into the required format by analyzing the mapping.

The first and foremost step is to capture the data into a single electronic health record and creation of EHR involves many components as shown in Figure 9.

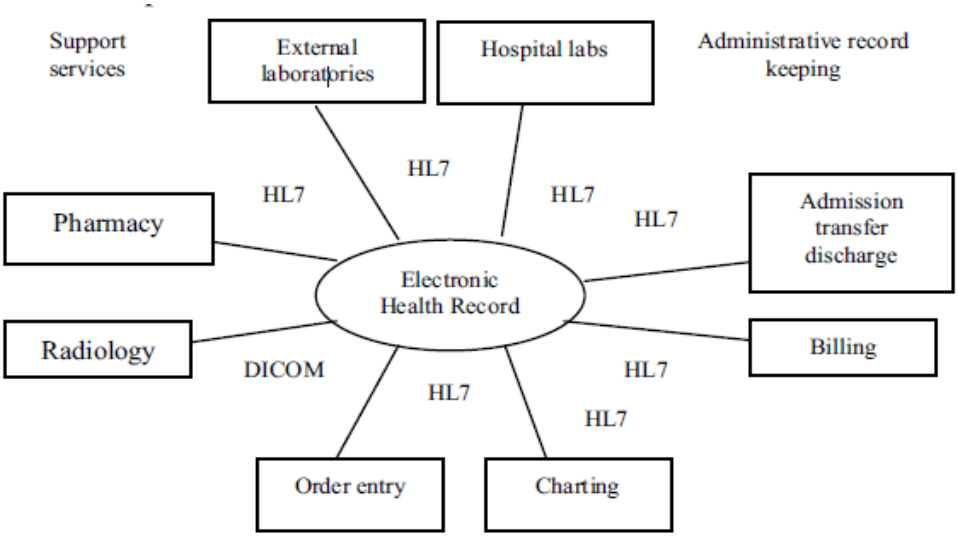

Fig. 9: EHR creation (Figure 1 page 2 [26]))

Once the EHR data is created and is ready to be exchanged between multiple systems, it involves three main phases, i.e., data parsing, data transportation and data reception. As we can see in the Figure 10, in the framework proposed, data from database is fetched and sent to the RIM interface which parser the information using DOM parser and composes a message that is ready to be transmitted. Once the message is created, it is passed to the receiver's end by using the serialization process over the transport layer. Finally the message is received from the destination end or the receiver and it is sent to the parser (which is a DOM parser according to the model) and further sent to the RIM interface which then directs it to the database. Finally the receiver will store the data received into the database in the desired format.

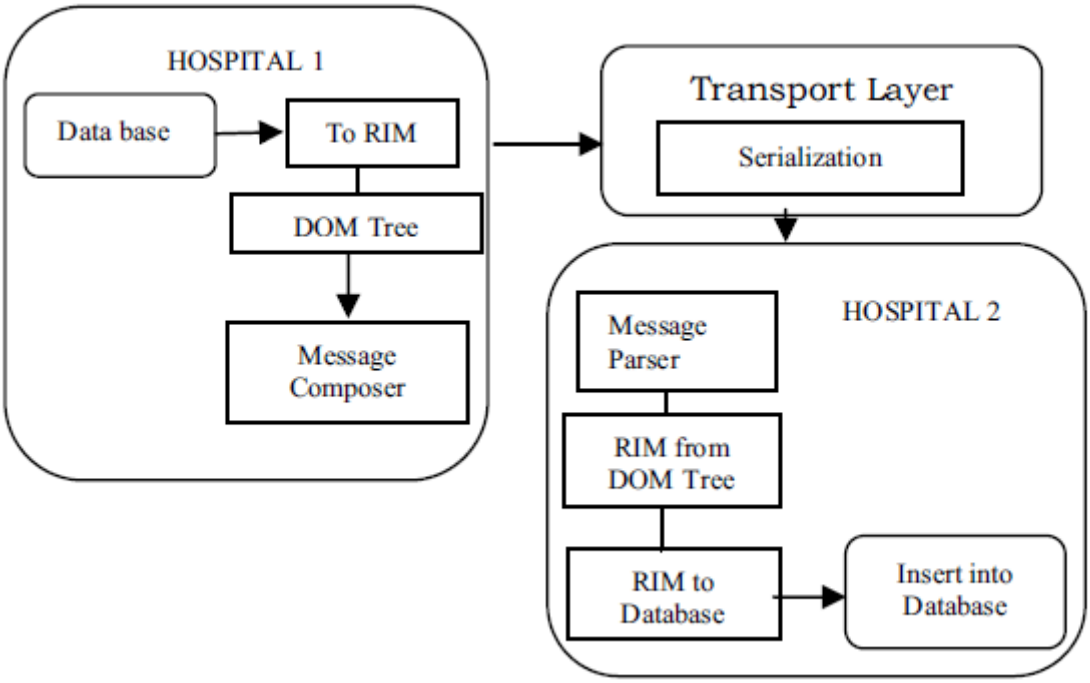

Fig. 10: Message Exchange Model for EHR (Figure 2 page 2 [26]))

The EHR data is stored in an XML based document as a DOM object. For each object that is created, there are set of patient identifiers (name, id, etc.) as root nodes, followed by set of attributes and respective values stored as child nodes of the XML object. An example tree representation of an EHR can be as shown in Figure 11.

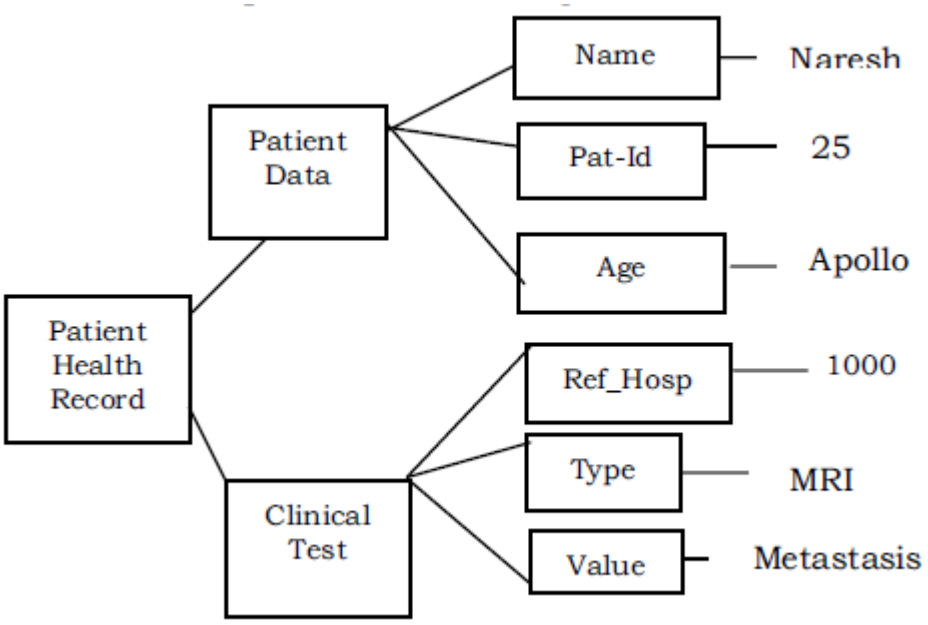

Fig. 11: EHR's Tree representation (Figure 3 page 3 of [26]))

The authors proposed a XDB algorithm for storing the medical records into relational database model and a DBX algorithm for extracting the data from database.

The DBX algorithm reads the attributes or set of attributes as input, connects to database, fetches the data, creates an XML document tree object where the first element is the root and each row is added a child element and then returns the final XML document.

The XDB algorithm then accepts and reads XML document, parses the messages, identifies the root and child elements, maps the information with the target format and then stores it into the database as an electronic health record. The two algorithms, i.e., DBX and XDB can be explained with Figure 12.

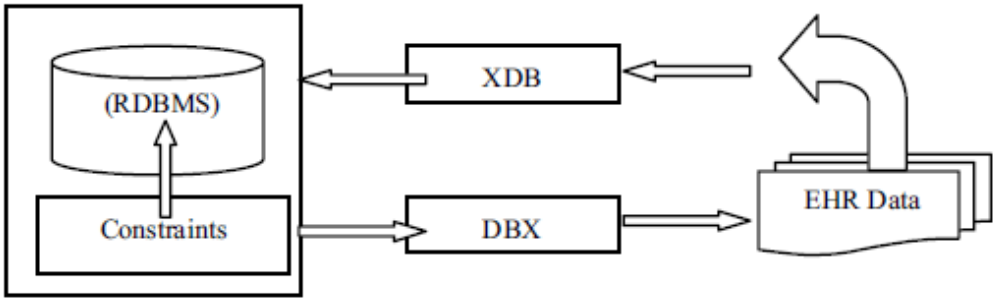

Fig. 12: XDB and DBX algorithm (Figure 6 page 3 of [26]))

The most important thing to note here is that the data element identified will be a unique and of a specific format only. A standard data structure is used for data sets that are structured. Both the systems are agreeing upon a specific structure. This implies that there should some sort of manual analysis of the data on both the ends and it has to be in perfect sync with both the ends. The mapping of the data is identified maybe manually and then a structure is set and it is being followed. The advantage here is that the data exchange is being done and it supports interoperability, but the biggest disadvantage is that it cannot handle any unstructured data and without manual interpretations and mapping. Moreover, it is not followed and globally accepted standards so again in future if we want to integrate with another system we should follow the same procedure with the same data structures. Another issue is that, this model is not having any databases, hence we cannot apply analysis on the data stored in XML format directly without pre-processing. So there is an additional step involving extra query processing time and complexity.

3.6 EAV and EAV/CR data model

The authors El-Sappagh *et al* [2012] proposed a framework that tries to model patients' clinical events. They implemented an EAV/CR which is an entity-attribute-value model with class relationships, where data and metadata are stored in an Object-Relational data model. EAV/CR model is an extension to the existing EAV model, where the attributes maybe also linked to their substructures facilitating a complex structure. Hence, by the addition of classes and relations, it becomes an object oriented framework. The authors also state that the schema is patient problem oriented, where the structure provides a meaningful depiction of the problems of the patient and all relevant clinical entries. Also, the model focussed on collecting as many

identifiers as possible for each patient and used a special table for identifiers. All the problems are stored as independent tables in an extended row based approach as shown in table 3.

Problem ID	Name	Status	Link	Start	End
1	P1	A	1/1/2012	1/3/2012	
2	P2	A	1	2/3/2012	null
3	P3	A	2	2/3/2012	null

Table 4: Example Problem Table (Table 5 of page 8 of [16]))

The EVAR/CR model proposed by the authors had a very special importance for the time of the medical recordings. They used a temporal database for storing either the transaction time, time of validity or both of these. Therefore, all the attributes and values are stored as events and relations to these events the reference times or calendar notes. The database should be allowed such relations and events to be recorded as shown in Figure 13.

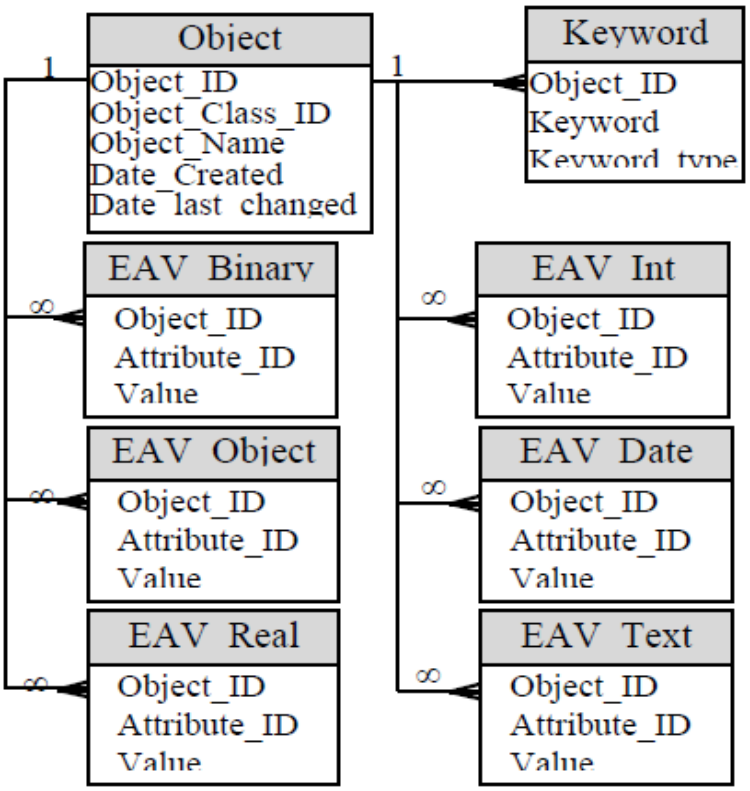

Fig. 13: EAV/CR data model (Figure 9 page 6 of [16]))

The authors **El-**Sappagh *et al* [2012] state that they have used various types of relations in there model, some of the examples are as stated below.

- Inheritance
- Composition
- EAV and conventional relations
- Temporary relations

The results show that this model was better than traditional row based model. They tried to achieve a mixed design model with varied information on different data types, related by using an object relational data model. The framework was useful for the integration and collection of data from EHR systems for the purpose of medical decision making, but was not interoperable and consider application of data mining techniques to be a key future work.

4 Methodology

4.1 Introduction to the Problem

Medical and Health data analysis involve data capturing, storage, processing, exchange, integration and interpretation of data from heterogeneous systems. The existing EHR systems store large volumes of medical records independently in various structures and formats. In spite of continuous efforts, limited access to health data still remains a major problem. There are various reasons for the limited access to health data. Most of the records in various hospitals either remains paper based or HIS have too old data. Health care has become more complex and multidimensional social contract as it has to deal with the various associations of citizens from heterogeneous sectors, to provide a safe and effective health care in a timely manner. In order to satisfy the requirements of these social contracts, effective decisions based on solid evidences which ensure the quality of care for all the patients are needed. When issues of interoperability are addressed and support for the data mapping is provided, it helps the systems to become more reliable.

The structural differences could be different storage devices, databases, data types, attributes or syntax. The semantic differences may occur due to differences in vocabularies, usage of different representations of the same data (synonyms), language barriers, etc., generally need more human assistance in identification and interpretations. This creates a bigger problem in case of unknown environments when the data from multiple systems is unpredictable and is not in a pre-agreed format. The diversity of both the data and data models of existing system obscures or prevent data mapping, standardization and interoperability between heterogeneous systems. As we have seen, Interoperability is very important, especially in case of medical data. It can be explained with an example as shown in Figure 14.

Let us consider a scenario where a patient is undergoing a surgery in a hospital in Toronto and the surgeon needs to contact another specialist who is in a different hospital in a different city or country for some information. In such case the surgeon can contact the specialist and

transfer the necessary clinical reports via email, phone or any other medium of communication. But all of these will involve another step of extracting the information and interpreting it and even after that we cannot guarantee if the knowledge transfer is hundred percent as expected. Instead, if both the hospital systems are interoperable, the exchange of any kind of information or clinical reports can be done directly through the electronic systems. The later interoperable way assures reliability, security, faster exchange of data and also is trackable for future references.

Fig. 14: Illustrating the need for Interoperability

4.2 Research Objectives

The main objective of our research is to develop a single solution for addressing the issues of data mapping, integration and interoperability between heterogeneous EHR systems with unstructured data, by using standard vocabularies and minimal human interpretations.

The key objective could be further divided into below mentioned set of objectives.

- Achieving data mapping between heterogeneous hospital/diagnostic databases or file systems of any form/structure/design.
- Usage of standard vocabularies and promote interoperability.

- Establishing a dynamic data model with efficient storage allocations and nominal redundancies.
- Making the best use of knowledge base without data losses for analysis and predictions.
- Addressing the issues of privacy and security.
- To be able to reduce manual or human interpretations of data/metadata without losing their semantics.

4.3 Hypothesis Statement

A standardized framework based on an efficient data model will improve the quality of healthcare data acquisition and exchange.

- In this context, "quality" may refer to reducing the percentage of data losses, the time for query processing, reducing the work burden to human operators (e.g. sql experts, health experts).
- The quality measure is based on a phenomenological viewpoint that we intend to establish as an objective measure.

4.4 Entity Relationship Diagram

- ER (Entity Relationship) diagram broadly represents the entities present in the framework (Hospital & Patient), the various attributes associated with each entity (system id, name, etc.) and the relationship between the entities (HAS A relation).
- Table for higher level entity will be created first (E.g. Hospital information table) with a primary key.
- Table for the lower level entity will be created (E.g. patient details) with another primary key.

- On the lower level, table we will declare the primary key of higher level table and use it as a foreign key.
- This maintains the relationship at its best.

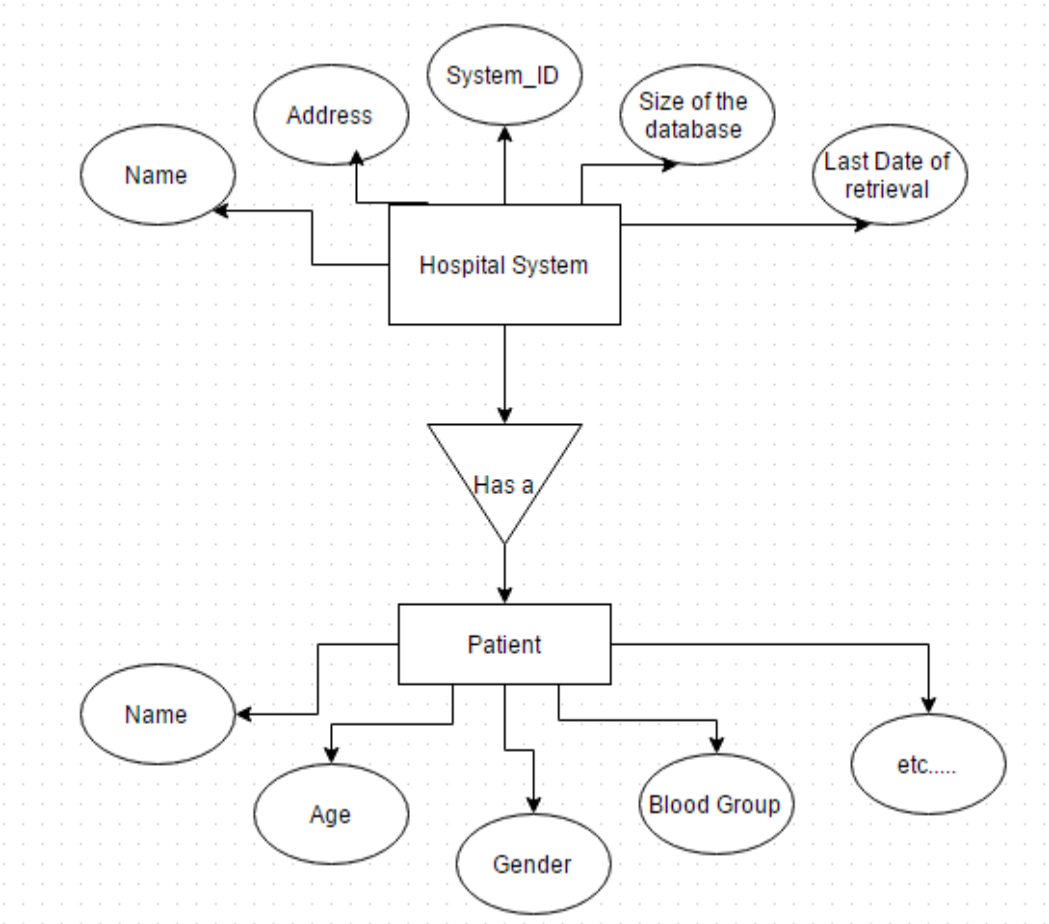

Fig. 15: Entity Relationship Diagram

4.5 Algorithm

4.5.1 PSEUDOCODE:

PROCEDURE: *INTEROPERABLE HYBRID MODEL*

INPUT: *EHR source systems*

OUTPUT: *Integrated database implementing Hybrid data model*

BEGIN

Read inputs sys1, sys2, etc. as the EHR source systems (SYS represents the set of systems)

FOR *(each sys ∈ SYS)*

IF *data in relational or column based format*

FOR *(each attr ∈ ATTR)*

Identify the attribute

Identify the LOINC code of the attr

Analyze the codes and identify intersection and independent attr

END FOR

Implement Hybrid data model (create database) and create HL7 messages as response objects for all requests

ELSE

Convert row into relational model

END IF

END FOR

END

4.5.2 Details

The proposed approach will be to use programming techniques along with standard vocabularies (HL7, LOINC), for analyzing the semantics of the data in the systems and achieving the data mapping. This can be broadly divided into three stages.

- Analysing the attributes using standard Vocabularies.
- Implementation of Hybrid data model for data integration.
- Achieving the data mapping and interoperability.

AIM: To map the unstructured data, perform semantic integration and form a uniform interoperable system.

PROCEDURE: Hybrid Data Modeling.

INPUT: EHR data from independent heterogeneous systems in different formats.

OUTPUT: Database implementing Hybrid model with collection of data from heterogeneous systems.

EXPLANATION:

The algorithm can be broadly divided into six important steps.

Step 1: Identification of the attributes present in two systems independently.

In this step, the inputs, which could be the independent heterogeneous EHR systems are read. Both the systems might be having any kind of structural, syntactic and semantic differences. The databases could be some flat file systems, CSV files, Relational Databases, or any other database systems. The sub steps required in this stage are to verify if the data is stored in columnar model. If not, then the data need to convert into appropriate manner followed by the attribute extraction process. This is to be repeated for all the set of inputs considered in the integration and independent sets of attributes is the result of this step. (E.g.: Patient Name, Patient Age, etc.)

Step 2: Identify and extract the LOINC Codes for the attributes.

Once all the input systems are read and attributes are identified, LOINC repository is used as a reference. Identified attributes are considered as search terms and respective LOINC

representations of the attributes is done. (E.g.: Patient Name = 54125-0, Patient Age = 21612 -7, etc.)

Step 3: Analyze the codes and identify the function type.

Once the codes of the attributes are identified, they are compared against each other and identified in either of the two categories, intersection or independent.

Step 4: Implementation of Hybrid data model

This is the most important step where the database and the tables are created. There are different tables that are created based on the function types. The intersection attributes will follow a columnar model with each attribute representing a specific attribute in both the input systems and independent attributes will follow the traditional row based approach with each row having an identifier, attribute and the attribute's value. All the EHR data is loaded into the new tables this manner.

Step 5: Interoperability using HL7.

When either of the source system request for any patient information based on available fields, the framework will extract the information from the database and generates an Hl7 message as a response.

Step 6: Import results into the requestor's system.

The Hl7 message which is generated as a response, can be imported into the requestor's system by using parsers and http request response objects.

4.5.3 Hybrid Data Model

There are various types of data models like relational database model, entity attribute model, object oriented model, NoSQL databases, etc. for storing the data. As per authors Homan *et al*

[2009] and Thakur *et al* [2014], Relational database model have some advantages such as it is more organized, efficient data storage, we can apply various data analysis tools, the cost of updating the attributes is higher, data extraction is easy, less processing time, etc. and disadvantages such as it stores null values, manual interpretations, designations of columns, new columns might have to be added sparseness. Where as in case of row model or entity attribute value model, there are some advantages like volatility, no new columns, only row additions, easy logic, no null values and disadvantages such as high data storage, only one data type, we cannot apply any data analysis tools, memory shortage, different interpretations, network congestion, etc. Also authors Thakur *et al* [2014], state that the dynamic generation of tables is better than static tables in case of medical or health related data storage.

We propose a hybrid data model, which tries to make the best use of both relation and row based model based on the properties of the data. Dynamic tables are then created based on the function types identified for the attributes. The intersection attributes will follow a columnar model with each attribute representing a specific attribute in both the input systems and independent attributes will follow the traditional row based approach with each row having an identifier, attribute and the attribute's value. The master table is following a relational approach as the master table consists of all the common attributes of the multiple input systems. As the data is common and number of null values in this are very less. Most of the analysis algorithms are applied on the combined data sets and thus the query processing time is low. All the EHR data is loaded into the new tables this manner. Hence we are making the best usage of Relational Model here as for the data set EAV would be talking a lot of space and analysis of data is not possible with EAV model. The second table (miscellaneous attributes) of all the uncommon attributes of the input systems. As this data is not present in all the input repositories, we will have lots of null values if we maintain these in relational form. Also, if there are any duplicate data from the source, we can easily identify eliminate them. This data

will not disturb any analysis we want to perform and thus results will be accurate. In case we have any new attributes or unidentifiable attributes because of any erroneous representations, it is very easy to store such data. In case where access to individual data based on separate hospitals is requested, it is easy to apply views, indexing and other techniques to fetch the data along with optimization of query processing times.

4.6 Application

Practically, we can say that the proposed approach would provide a cost effective solution for support and analysis of patient level data and attribute based analysis reports across different hospitals. In many cases, clinicians store the attribute information in different representations, which might affect communications between the application interfaces for different storage systems. The severity of the problem due to the difference in representations of clinical data is not constant across every country, as it depends on the laws and constraints from the government or the constitution. However, sufficient differences between medical record systems may arise leading to dependencies on repositories like LOINC. For instance, in Canada and similar jurisdictions, where there exists greater homogeneity of data storage, structure and metadata, LOINC cannot be applicable to patient level; rather, it would be more applicable to other services or cost effective analysis purposes.

5 Experimental Setup

We designed a framework based with the intension of solving both the problems of data mapping and interoperability. For this, we chose to use JAVA software platform. We have also used Microsoft access database, MySQL database, CSV files as the medium of storage for the data. We also used JDBC (Java Database Connectivity), as it is a standard Java API for database connectivity between Java programming language and the databases. The

5.1 Dataset Description

For testing the framework we have used the clinical data sets with about 12000 patient details was generated and used as input data sets. The attributes in the datasets were clinical attributes such as (platelet count, urinalysis, metabolic parameters, etc.). For our convenience, we are trying to explain with a small example here. Considering the case where there are about 12000 records and the number of attributes are 200 (i.e., 80 attributes in one input system A and 120 attributes in input system B), stores in column based manner. The example details of the systems are as pasted in table 4 and Figure 16.

System	No. of Records	No. of Attributes	No. of Common Attributes	No. of Unique Attributes	Redundancies?
A	5177	80	70	10	Yes
B	6823	120	70	50	No
Final	12000	200	70	60	No

Table 5: Example data sets

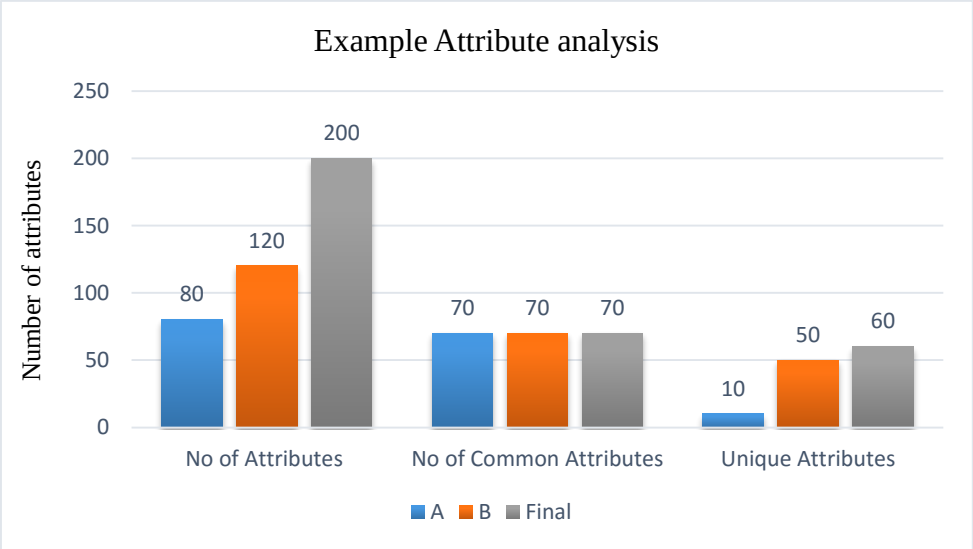

Fig. 16: Example data sets attribute analysis

In the traditional column based approach, if the number of records in the system A are 'a' and number of records in system B are 'b' then the number of records in final integrating system

will be equal to ‘a+b’ as shown in Figure 17. In case of row based approach the number or records will be linearly dependent on the number of attributes of the systems. If the number of attributes in System A are ‘attr1’ and System B is ‘attr2’ the number of records (N) in the integrated system can be calculated (example as shown in Figure 18).

$$N = (a^{*}attr1) + (b^{*}attr2)$$

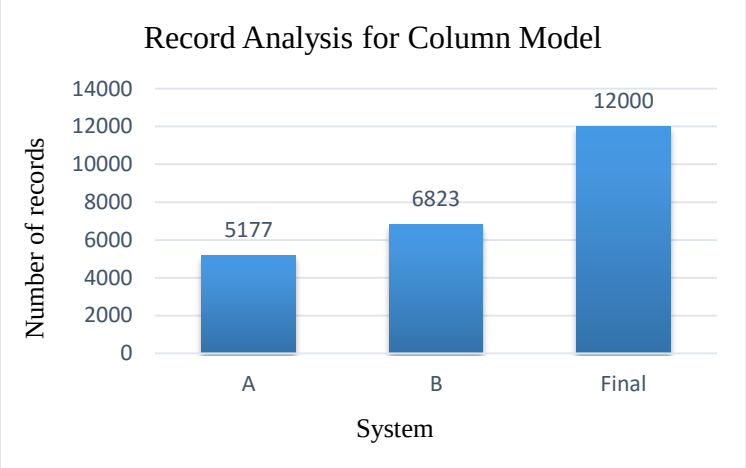

Fig. 17: Example record analysis for column based approach

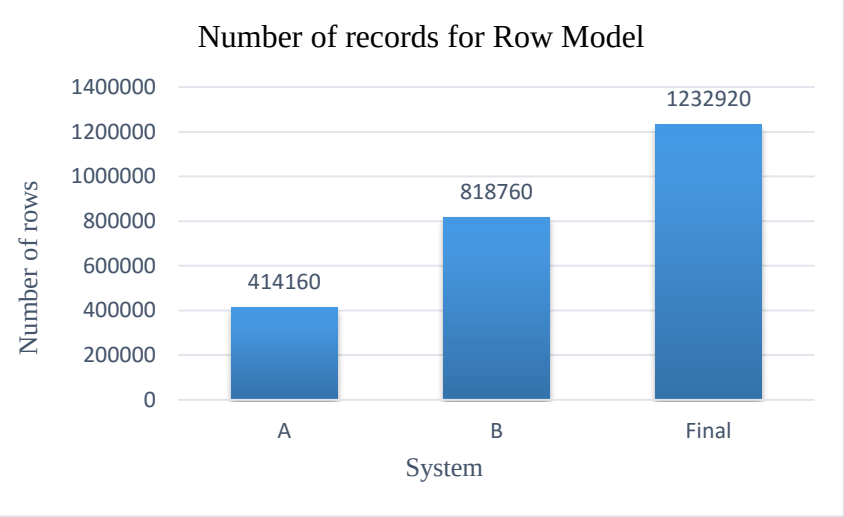

Fig. 18: Example record analysis for row based approach

The proposed hybrid model which does the attribute analysis and splits the data into both row and column models. Considering the above example in table 4, in hybrid model details are

as shown in table 5. The number of records varies according to the number of attributes in common (70 according to our example).

Hybrid Model	Columns	Rows
Common Attributes	70	12000
Unique Attributes	3	720000

Table 6: Example Hybrid model

In the actual experiment, analysis was conducted with multiple systems with different set of attributes. We also varied the number of common attributes in different sets in intervals and observed the behaviour of the data on various parameters. The behaviour in different cases is discussed in detail in the Results Chapter.

5.2 Experimental Details

Table 6 describes the experimental setup in detail.

Input	Different EHR systems	We considered 2 source EHR data repositories; one in CSV format in non-relational format and another as a single Database file in RDBM model.
Output	Integrated database, implementing Hybrid data model.	A single DB file that can be imported and also an interoperable system.
Parameters	a) Multiple sets of inputs.	Number of data sets were altered.

	b) Number of iterations of data load.	For each set of inputs the number of common attributes was varied and tested.
Variables	a) Size of the data. b) Processing time c) Amount of data losses recorded.	Size of the independent and integrated databases Processing Time for various activities like insert, delete, search, etc. Measure the amount of data losses if a standard data object was set.

Table 7: Experimental Setup

5.3 Program Logic

The algorithm was implemented using Java language and the program logic was divided into four modules. The first module analyses the input data, for the structure and data models and identifies if all the inputs are in Row model or columnar model. If any or all of the input systems are in row model then data is reorganized into columnar model by eliminating the duplicate entries of the attributes. Hence by the end of this step all the input data will be in columnar or tabular format where the first rows of each database will indicate the attributes of the respective data set.

Once the above step is done, which is also called as data clean up phase, a call to second function is invoked, where the attributes are read from each of the input systems and stored in a list structure. These attributes are given as inputs to the LOINC repository and respective LOINC codes are extracted. All the attributes with successful match will be assigned with their respective LOINC codes and a match for both sets of attributes will be done. In case there is match all such attributes will be marked as intersecting attributes, and attributes with no match are marked as independent attributes. If search results are unsuccessful for any of the attributes

inputted, the attributes will be marked as independent attributes and a call to third module is triggered.

In the third phase, Database tables are created as per the hybrid model approach by analysing the return values of previous step and further data load is done by transferring the contents of the inputs (both CSV files and DB files) into the new hybrid model. This completes the first requirement of the experiment which is the data mapping.

The last step is for data search and extraction. Any patient details from any of the input systems can be searched with any attribute details or values. The results will then be returned accordingly from the integrated system. If either of the input systems wants to exchange or extract the information, the data is generated as a HL7 message (where the attributes are represented using their respective LOINC codes.) As expected, for the unsuccessful search results the algorithm will recommend to redefine the keywords used for the search.

5.4 Representations

The number of common attributes have been varied from 0 to the maximum number of attributes (i.e., 100%). For convenience, in all the graphs and tables of the results section, the normalized values of a number of attributes in common were presented instead of actual values. The normalization was done using the basic mathematical normalization rule.

$$X_i = (X_i^{meas} - X_{min}^{meas}) / (X_{max}^{meas} - X_{min}^{meas})$$

In the above equation, (X_i^{meas}) is the number of common attributes considered, hence it starts at 0 and ends at maximum number of attributes present in either of the systems. X_{min}^{meas} represents the minimum number of common attributes which is 0 in our case and X_{max}^{meas} represents the maximum number of attributes in common, which is maximum number of attributes presents in either of the systems. As we are staring at 0 and ending at maximum number of attributes in either of the systems, for simplicity, we can rewrite the above equation as pasted below.

$$X_i = (X_i^{meas}) / (X_{max}^{meas})$$

All the processing times are recorded in Milliseconds but for convenience they are either represented in seconds or milliseconds depending upon the nature of the values in the results section. The size of database is measured in bytes and then represented in MB for convenience. The processing times (which is represented on the Y axis of most of the graphs in results section) have been rounded off to the nearest 10^{th} of a second, reflecting the actual test measurements.

5.5 Implemented Data Models

We implemented three kinds of data models with same data sets and same attributes. The number of common attributes was also varied in the same intervals for all three setups. In the first case, we implemented in a complete row model followed by the complete column model and then finally we implemented the Hybrid Data model. The results section presents the differences observed in all three setups.

6 Results and Discussion

In this section, we explain the performance of a Hybrid model by comparing with the row model and also we compare the amount of data losses that might occur if we agree to store and exchange only a standard data object (based on common attributes) instead of collecting the complete details. So we present the comparisons in query processing times, the percentage of data losses and comparisons in the size of the databases.

6.1 Comparison for Database Size

We implemented the same datasets in row model, column model and hybrid model. In the first case, i.e., in row model, when the number of attributes are constant, as the number of rows increase the size increases proportionately. Table 7 and Figure 19 indicates the amount of

increase in the size as the number of rows are increased and graphical representation of same details respectively.

Total Rows in Table	size for table in bytes
10000	1589248
15000	1589248
20000	1589248
25000	2637824
30000	2637824
35000	2637824
40000	2637824
45000	3670016
50000	3670016

Table 8: Size for row model

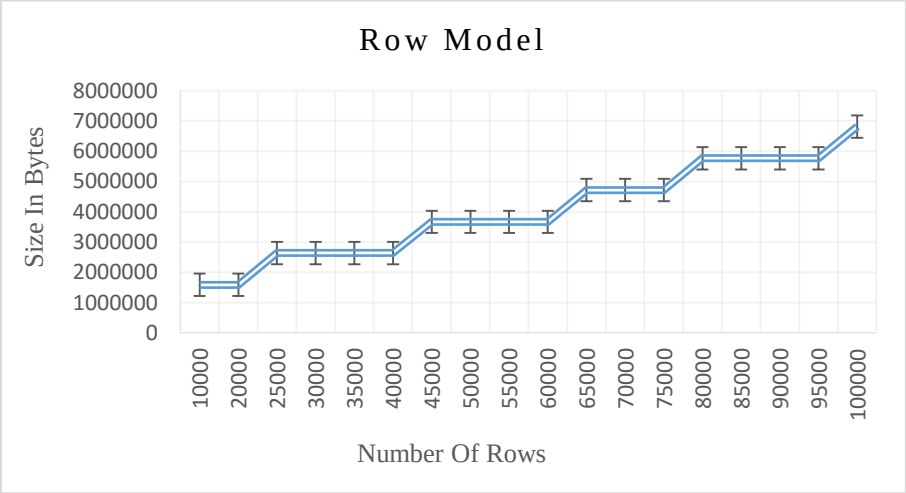

Fig. 19: Size for row model

We could observe a staircase kind of pattern in this case. On the other hand, in the column based model; when the number of rows is constant, as the amount of attributes increase, the size of the database increases in some intervals (instead of increasing for every column added) as represented in Figure 20.

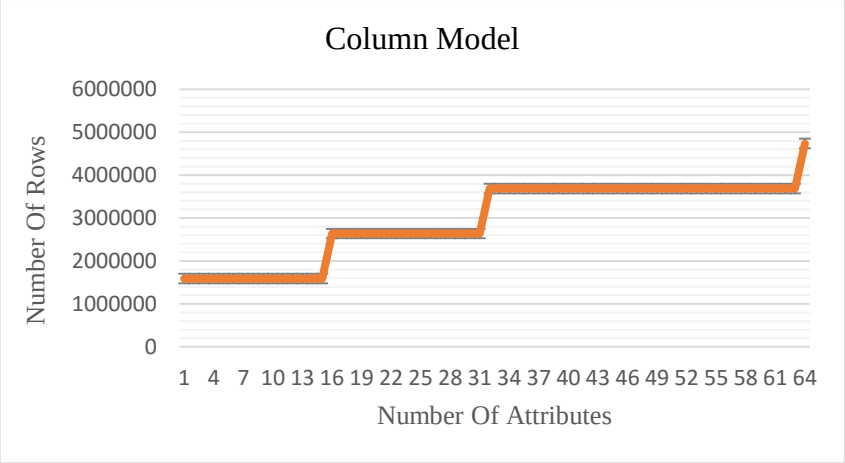

Fig. 20: Size for Column model

As it is explained in the previous chapter that for hybrid the number of records and number of columns in the database vary depending on the number of common attributes. Hence, for row model representation the size of the database remains constant irrespective of the similarity of attributes, whereas in case of hybrid model we observed that, with the increase in the number of attributes in common, the size of the database seem to be reducing as shown in table 8 and Figure 21.

No. of Attributes in common	Hybrid Model	Row Model
0	41.4	41.4
0.08	39.8	41.4
0.16	35.6	41.4
0.25	32.5	41.4
0.3	29.3	41.4
0.4	26.2	41.4
0.5	23	41.4
0.58	20.9	41.4
0.6	17.8	41.4
0.75	13.6	41.4
0.8	10.5	41.4
0.9	7.3	41.4
1	3.6	41.4

Table 9: Database Size: Hybrid Vs Row

As per our observation, the trend line for row model and column model it followed a growing staircase pattern. The Hybrid model is a combination of both row and column model. Hence we see slight fluctuations in the trend line as shown in Figure 21. But considering the error bars, we can say it is almost linear.

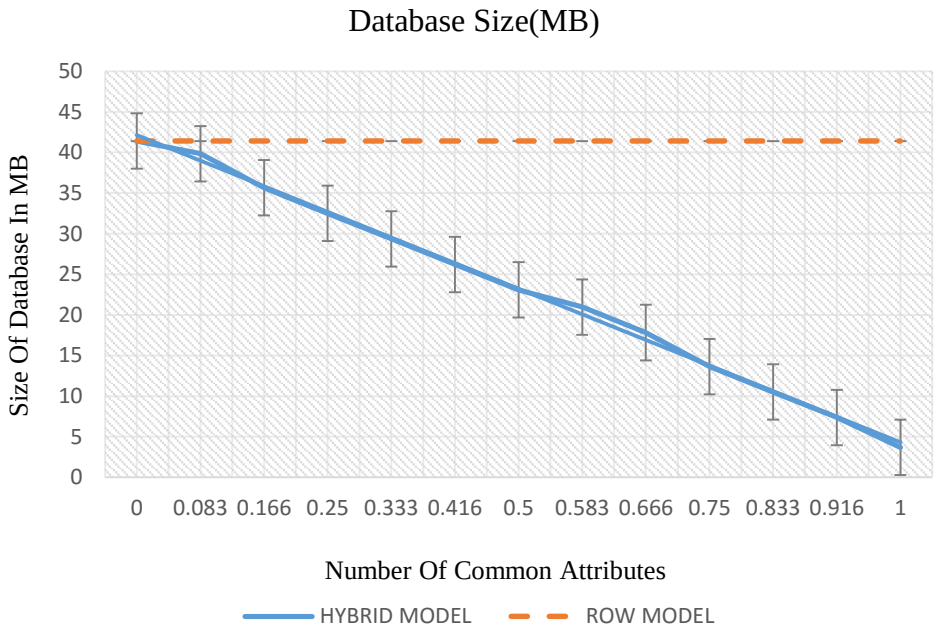

Fig. 21: Database Size: Hybrid Vs Row

6.2 Comparison of Row Model and Hybrid Model for Query processing time

6.2.1 Data Load

As it is explained in the previous chapter that for hybrid the number of records and number of columns in the database vary depending on the number of common attributes and the query processing time depends upon the size of the database to some extent. The recorded attribute intervals have been normalized and average of the load times are as per table 9 and Figure 21.

No. of Attributes In Common	Hybrid Model (Min)	Row Model (Min)
0	319.1	319.1
0.08	301.5	319.1
0.16	272.1	319.1
0.25	244.1	319.1
0.3	216.4	319.1
0.4	188.7	319.1
0.5	163.6	319.1
0.58	138.5	319.1
0.6	110.9	319.1
0.75	83.1	319.1

0.8	55.3	319.1
0.9	27.5	319.1
1	5	319.1

Table 10: Data Load Time in Min

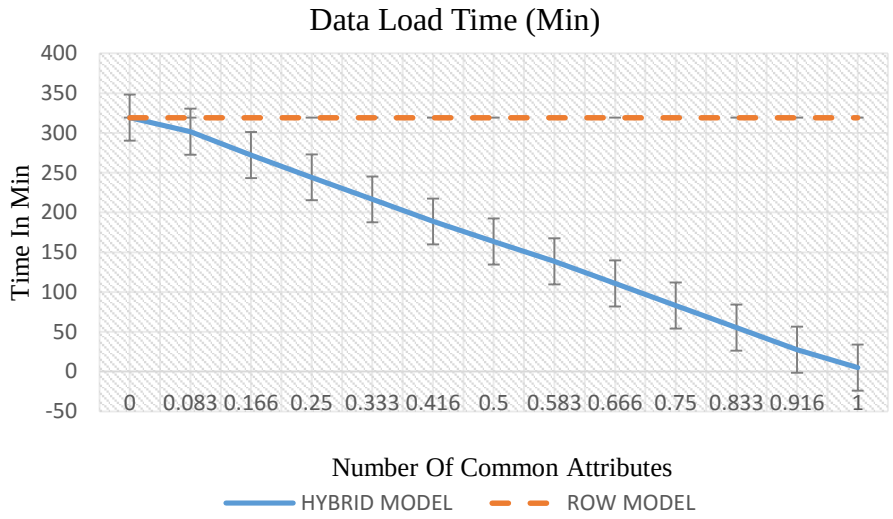

Fig. 22: Data Load Time: Hybrid Vs Row

6.2.2 Search and output

This is important to know that in hybrid model it is much faster to retrieve the details of common attributes when compared to independent or unique attributes. The search time in row model seems to be same irrespective of the attributes. We have tested for multiple runs and average times have been represented in table 10 and Figure 23.

No. Of Common Attributes	Best Case Search(ms)	Worst Case Search(ms)	Row Model (ms)
0.03	14	1577.3	1695.5
0.16	16	1418.9	1695.5
0.25	18	1272	1695.5
0.3	19	1039.8	1695.5
0.4	22	903.2	1695.5
0.5	23	779.6	1695.5
0.58	25	656.1	1695.5
0.6	26	520.3	1695.5
0.75	27	383.7	1695.5
0.8	29	247	1695.5
0.9	33	110.4	1695.5
1	36	36	1695.5

Table 11: Data Search Time in Milli Sec

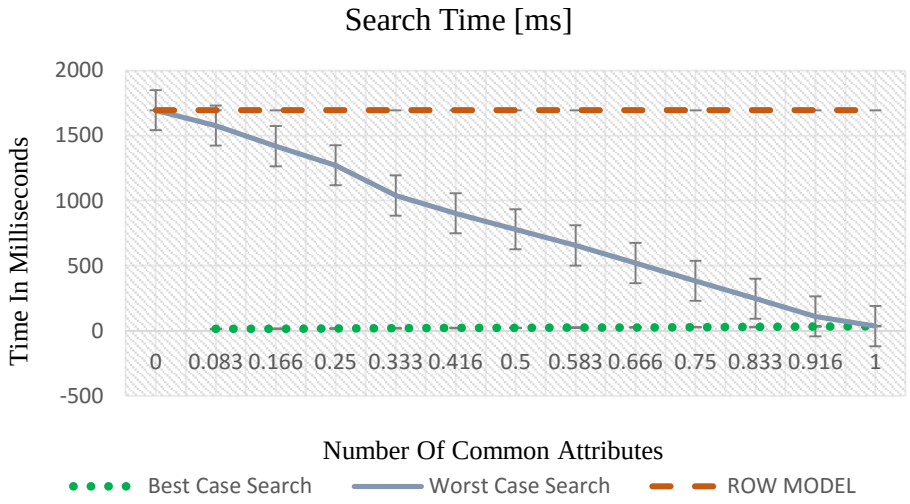

Fig. 23: Data Search Time: Hybrid Vs Row

6.2.3 Data Deletion

As the deletion time is directly proportional to database size, the maximum time is observed at most number of attributes and the minimum is recorded for least number of common attributes.

No. Of Common Attributes	Hybrid Model (sec)	Row Model (sec)
0	1068.5	1068.5
0.08	994	1068.5
0.16	894.2	1068.5
0.25	801.6	1068.5
0.3	708.6	1068.5
0.4	615.5	1068.5
0.5	531.4	1068.5
0.58	447.2	1068.5
0.6	354.7	1068.5
0.75	261.6	1068.5
0.8	168.5	1068.5
0.9	75.4	1068.5
1	0.2	1068.5

Table 12: Data Deletion Time in Sec

Again, irrespective of the number of attributes the data deletion time in row model remains constant. We have recorded multiple runs and average time has been represented here. The values on X-axis have been normalized.

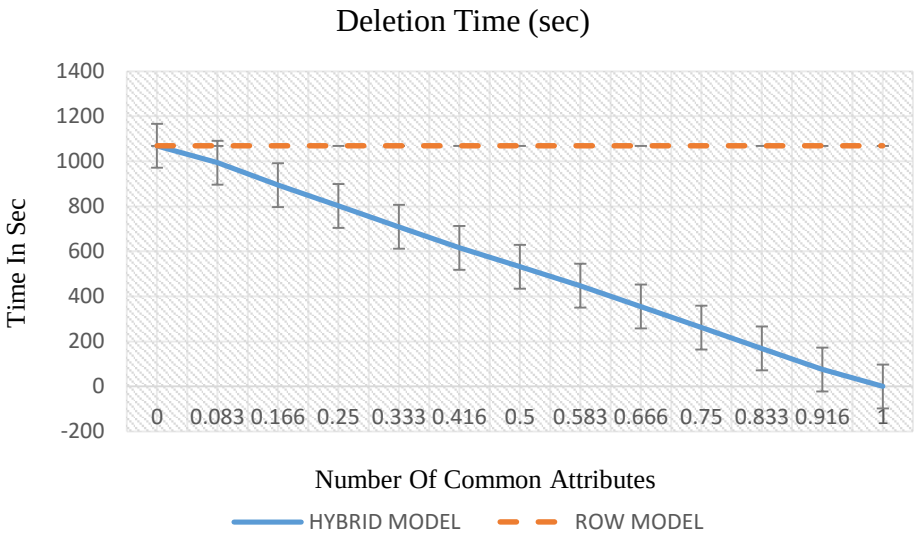

Fig. 24: Data Deletion Time: Hybrid Vs Row

6.3 Amount of data loss

Table 12 and Figure 25 represents the case where we recorded only information about common attributes in multiple systems; that is, we are agreeing upon a set of common attributes (a standard data object) and then recording the data. If any of the attributes is not present in all the input EHR systems, the corresponding data is discarded. Hence, this approach is good when there are exactly the same attributes in multiple systems; but, its performance is worst in the scenario where there are totally independent sets of attributes.

Number of common attributes	Size in MB
0	41.4
0.08	38.2
0.16	34
0.25	30.9
0.3	26.7
0.4	23.5
0.5	20.4
0.58	17.2
0.6	14.1
0.75	9.9
0.8	6.8
0.9	3.6
1	0

Table 13: Amount of Data Loss

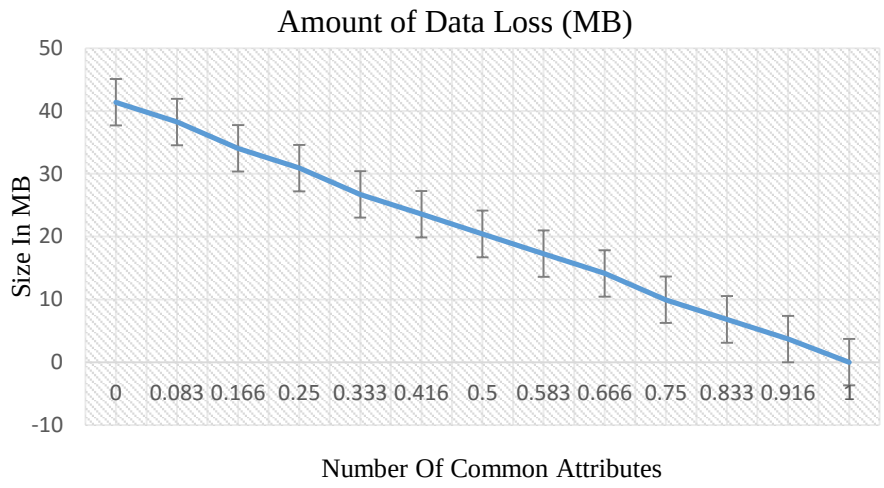

Fig. 25: Amount of Data Loss

6.4 Summary of Results

The results show that the Hybrid model is better than the traditional row based approach and key step here is the attribute identification using LOINC. Also this helps in interoperability by exchanging the data using HL7 messages. It is easier to generate these HL7 messages as the LOINC codes are stored as the column headers. The size and query processing time of row model is always constant whereas it is varying in case of hybrid model. We also found that the trend line of the hybrid graph is almost linear and can say the equation of the hybrid model is a linear equation.

It should be noted that, actual values of the slope (A) and intercept (B) are dependent on the various factors including data set properties and system and device properties. Thus, what is important to note in any replication of our experiments is that the linear behaviours must persists.

We found that the best performance is achieved when there are a maximum number of common attributes and the worst performance when there are independent attributes. In worst case scenario the algorithm works like the row model.

7 Conclusion and Future Work

7.1 Conclusion

In this thesis, we proposed an algorithm for solving the issue of interoperability and data mapping. We implemented the algorithm and succeeded in merging the data from multiple databases/file systems into an integrated single database which can be used for analysis or testing. Data losses, query processing time, storage space, management of redundancies, etc. were analyzed. We could say that usage of programming techniques and open source tools to solve the problems of both data mapping and interoperability between electronic health record systems using a single framework was achieved. Focus and use of Standard vocabularies (HL7, LOINC, etc.), efficient space management & no data losses can be a key contribution of this work.

The proposed algorithm that identifies the attributes of data in heterogeneous systems of data and then creates tables based on the attributes at the time of integration. We tested the algorithm with multiple datasets and observed that it is more efficient than the traditional row based approach and it reduces data losses when compared to the XML based approach of exchanging standard data objects. The results were analyzed for behavior over various parameters, such as database size, query processing time to select, delete, and insert queries. Also, we measured the amount of data losses relative to a standardized data object and recorded only a set of common attributes from multiple systems. We found that the best performance is achieved when there are a maximum number of common attributes and the worst performance

when there are independent attributes. In the worst case scenario the algorithm works like the row model. The biggest advantage of this approach is that there are no data losses in any case and interoperability is achieved.

7.2 Future Work

Data cleaning continues to be a significant issue for future work. Interoperability and data mapping is a key issue in many areas. We tried to address this problem in the area of health care. A similar approach can be applied to non-medical data. In our approach, attribute identification is dependent on medical vocabularies and repositories in the proposed algorithm. If the attribute representations are incorrect or erroneous, although the current system stores that information, it will not be able to identify or predict the exact medical representation without human interpretations. Handling such issues of ontology construction and alignment are also part of future work.

References

[1] Araki, K., Ohashi, K., Yamazaki, S., Hirose, Y., Yamashita, Y., Yamamoto, R., Yoshihara, H. (2000). Medical markup language (MML) for XML-based hospital information interchange. Journal of Medical Systems, 24(3), 195–211.

[2] Atkinson, M., Moody, K., Leslie, I., Storey, T., & Atkinson, M. (2002). UK Role in Open Grid Services Architecture UK e-Science Core Programme Draft : Please do not Circulate Prepared by The UK e-Science Architecture Task Force for The UK e-Science Technical Advisory Group, (March), 1–49.

[3] Balaji, V., Redler, R., & Budich, R. (2013). Earth System Modelling - Volume 4, 4, 49–51.

[4] Beeler, G. W. (1998). HL7 Version 3—An object-oriented methodology for collaborative standards development1Presented at the International Medical Informatics Association Working Group 16 Conference on Standardisation in Medical Informatics—Towards International Consensus and C. International Journal of Medical Informatics, 48(1-3), 151–161.

[5] Begoyan, a. (2007). An overview of interopberability standards for electronic health records. Society for Design and Process Science, 1–8.

[6] Berman, J. J., & Bhatia, K. (2014). Biomedical data integration: using XML to link clinical and research data sets. Expert Review of Molecular Diagnostics. Retrieved from http://www.tandfonline.com/doi/abs/10.1586/14737159.5.3.329

[7] Brender, J., Nøhr, C., & McNair, P. (2000). Research needs and priorities in health informatics. International Journal of Medical Informatics, 58-59, 257–289.

[8] Breu, F., Guggenbichler, S., & Wollmann, J. (2008). Data Management Guide for Public Participation in Scientific Research. Vasa.

[9] Brickley, D., & Guha, R. (2004). {RDF Vocabulary Description Language 1.0: RDF Schema}.

[10] Brown, N and Reynolds, M. (2000). Strategy for production and maintenance of standards for interoperability within and between service departments and other healthcare domains. Short. TC, 251, N00–047.

[11] Ceri, Stefano and Widom, J. (1993). Managing semantic heterogeneity with production rules and persistent queues.

[12] Collen, M. F. (1999). A Vision of Health Care and Informatics in 2008. Journal of the American Medical Informatics Association, 6(1), 1–5.

[13] Donabedian, A. (2005). Evaluating the quality of medical care. 1966. The Milbank Quarterly, 83(4), 691–729.

[14] Dupplaw, D., Dasmahapatra, S., Hu, B., Lewis, P., & Shadbolt, N. (2009). A distributed, service-based framework for knowledge applications with multimedia. ACM Transactions on Information Systems, 27(4), 1–29.

[15] Eichelberg, M., Aden, T., Riesmeier, J., Dogac, A., & Laleci, G. B. (2005). A survey and analysis of Electronic Healthcare Record standards. ACM Computing Surveys, 37(4), 277–315.

[16] El-sappagh, S. H., El-masri, S., Riad, a M., Elmogy, M., & Arabia, S. (2012). Electronic Health Record Data Model Optimized for Knowledge Discovery. International Journal of Computer Science, 9(5), 329–338.

[17] Geraci, Anne and Katki, Freny and McMonegal, Louise and Meyer, Bennett and Lane, John and Wilson, Paul and Radatz, Jane and Yee, Mary and Porteous, Hugh and Springsteel, F. (1991). IEEE Standard Computer Dictionary: Compilation of IEEE Standard Computer Glossaries.

[18] Grechenig, T., Tappeiner, B., & Wujciow, A. (2008). Challenging interoperability and bandwidth issues in national e-Health strategies by a bottom-up approach: Establishing a performant IT infrastructure network in a Middle East State. HealthCom 2008 - 10th International Conference on E-Health Networking, Applications and Services, 148–155.

[19] Guo, J., Takada, A., Tanaka, K., Sato, J., Suzuki, M., Suzuki, T., Yoshihara, H. (2004). The development of MML (Medical Markup Language) version 3.0 as a medical document exchange format for HL7 messages. Journal of Medical Systems, 28(6), 523–33.

[20] Halevy, A. (2011). Game-changing interoperability for healthcare: Bringing semantically harmonized clinical information into provider workflows from disparate health information technologies. 2011 8th International Conference and Expo on Emerging Technologies for a Smarter World, CEWIT 2011.

[21] Häyrinen, K., Saranto, K., & Nykänen, P. (2008). Definition, structure, content, use and impacts of electronic health records: a review of the research literature. International Journal of Medical Informatics, 77(5), 291–304.

[22] Homan, J. V, & Kovacs, P. J. (2009). a Comparison of the Relational Database Model and the Associative Database Model. Context, X(1).

[23] Hossain, M. S., & Masud, M. (n.d.) (2014). Medical Data Management and Interoperability in E- Health Systems.

[24] Hunter, J. (2001). MetaNet - a metadata term thesaurus to enable semantic interoperability between metadata domains. Journal of Digital Information, 1(8), 234–253.

[25] Ide, N., & Pustejovsky, J. (2010). What Does Interoperability Mean , Anyway ? Toward an Operational Definition of Interoperability for Language Technology.

Proceedings of the Second International Conference on Global Interoperability for Language Resources.

[26] Kumar, C. S., Rao, C. V. G., & Govardhan, a. (2010). A framework for interoperable healthcare information systems. 2010 International Conference on Computer Information Systems and Industrial Management Applications, CISIM 2010, 604–608.

[27] Lewis, G. a., Morris, E., Simanta, S., & Wrage, L. (2008). Why standards are not enough to guarantee end-to-end interoperability. Proceedings - 7th International Conference on Composition-Based Software Systems, ICCBSS 2008, 164–173.

[28] Liu, J., Fensli, R., Trinugroho, D., & Technology, C. (2012). Computer-Supported Cooperative Work in Tele Home Care - Architecture Design, Implementation and Evaluation.

[29] Los, R. K., Van Ginneken, A. M., De Wilde, M., & Der Lei, J. Van. (2004). OpenSDE: Row Modeling Applied to Generic Structured Data Entry. Journal of the American Medical Informatics Association, 11(2), 162–165.

[30] Manion, F. J., Robbins, R. J., Weems, W. A., & Crowley, R. S. (2009). Security and privacy requirements for a multi-institutional cancer research data grid: an interview-based study. BMC Medical Informatics and Decision Making, 9(1), 31.

[31] Manola, F., Miller, E., & McBride, B. (2004). RDF primer. W3C Recommendation, 10(February 2004), 1–107.

[32] Manuscript, A. (2012). NIH Public Access. Changes, 29(6), 997–1003.

[33] Marovets, J. L. (2014, November 25). System, method, and apparatus for storing, transmitting, receiving, and using structured data using un-structured text message bodies. Google Patents.

[34] McDonald, C. J. (1997). The Barriers to Electronic Medical Record Systems and How to Overcome Them. Journal of the American Medical Informatics Association, 4(3), 213–221.

[35] McDonald, C. J., Huff, S. M., Suico, J. G., Hill, G., Leavelle, D., Aller, R., Maloney, P. (2003). LOINC, a universal standard for identifying laboratory observations: A 5-year update. Clinical Chemistry, 49(4), 624–633.

[36] Mena, E., Kashyap, V., Sheth, A. P., & Illarramendi, A. (1996). OBSERVER: An Approach for Query Processing in Global Information Systems based on Interoperability between Pre-existing Ontologies. Conference on Cooperative Information Systems, 271, 14–25.

[37] Milosevic, Z. (2006). Addressing interoperability in e-health: An Australian approach. Proceedings - 2006 10th IEEE International Enterprise Distributed Object Computing Conference Workshops, EDOCW2006, 1–7.

[38] Nadkarni, P. M., Marenco, L., Chen, R., Skoufos, E., Shepherd, G., & Miller, P. (1999). Organization of Heterogeneous Scientific Data Using the EAV/CR Representation. Journal of the American Medical Informatics Association, 6(6), 478–493.

[39] Neitzel, L. (2004). OPC unified architecture internals. Proceedings of ISA Conference on 4th Annual …, 1051–1063.

[40] Nogueras-Iso, J. (2004). Geographic information metadata for spatial data infrastructures. Information Retrieval.

[41] Ouksel, A. M., & Naiman, C. F. (1994). Coordinating context building in heterogeneous information systems. Journal of Intelligent Information Systems, 3(2), 151–183.

[42] Ouskel, A., & Sheth, A. (1999). Semantic Interoperability in Global Information Systems. A brief Introduction to the Research Area and the Special Section. SIGMOD Record, 28(1), 5–12.

[43] Raminhos, R. F. (2007). Extraction and transformation of data from semi-structured text files using a declarative approach.

[44] Ramos, A. (2007). Collaborative software architectures for interactive biomedical applications.

[45] Rao, R. R. (2013). An informal method for Identifying Standards to enable Meaningful Exchange of Public Health Records. 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2037–2042.

[46] Ryan, A., & Eklund, P. (2008). A framework for semantic interoperability in healthcare: a service oriented architecture based on health informatics standards. Studies in Health Technology and Informatics, 136, 759–764.

[47] Ryan, A., Eklund, P., & Esler, B. (2007). Toward the interoperability of HL7 v3 and SNOMED CT: a case study modeling mobile clinical treatment.

[48] Safran, C. (2002). Health care in the information society. International Journal of Medical Informatics, 66(1-3), 23–4.

[49] Stell, A., Sinnott, R., Ajayi, O., & Jiang, J. (2009). Designing privacy for scalable electronic healthcare linkage. Proceedings - 12th IEEE International Conference on Computational Science and Engineering, CSE 2009, 3, 330–336.

[50] Thakur, M., Patel, D., Kumar, S., & Barua, J. (2014). NewsInstaMiner: Enriching News Article Using Instagram. Big Data Analytics.

[51] Turner, M., Zhu, F., Kotsiopoulos, I., Russell, M., Budgen, D., Bennett, K., Rigby, M. (2004). Using Web service technologies to create an information broker: an experience report. Proceedings. 26th International Conference on Software Engineering.

[52] Zhang, X., Li, J., Zhou, T., Yang, Y., Chen, Y., Xue, W., & Zhao, J. (2009). c2009 IEEE International Symposium on IT in Medicine & Education, 1, 1074–1078.

[53] Zhao, H. Z. H., Zhang, S. Z. S., Zhou, J. Z. J., & Wang, M. W. M. (2007). Semantic Model Based Heterogeneous Databases Integration Platform. Third International Conference on Natural Computation (ICNC 2007), 5(Inc.).

[54] Weiskopf, N. G., & Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. Journal of the American Medical Informatics Association : JAMIA, 20(1), 144–51.

VITA AUCTORIS

NAME:	Sreya Janaswamy
PLACE OF BIRTH:	Hyderabad, India
YEAR OF BIRTH:	1991
EDUCATION:	Keshav Memorial Institute of Technology, Hyderabad, India, 2012 University of Windsor, M.Sc., Windsor, ON, 2016