APPROVAL SHEET

Title of Dissertation: Development of a Collaborative Data Quality Improvement Approach for Healthcare Organizations

Name of Candidate: Yili Zhang

Doctor of Philosophy, 2018

Dissertation and Abstract Approved:

Güneş Koru

Department of Information Systems

Date Approved:

Title of Document: DEVELOPMENT OF A COLLABORATIVE DATA QUALITY IMPROVEMENT APPROACH FOR HEALTH CARE ORGANIZATIONS

Yili Zhang, Doctor of Philosophy, 2018

Directed By: Associate Professor, Güneş Koru, Department of Information Systems

In the United States (US), the large volume of health data maintained by healthcare organizations holds tremendous potentials to support healthcare operations and decision making, which can improve health services, particularly for socioeconomically disadvantaged and under-served populations in an effective and efficient fashion. Unfortunately, for various reasons, there have been substantial problems with the quality of data maintained by healthcare organizations which reduces its usefulness to support day-to-day healthcare operations and serve decision making purposes. This research contributes to the body of knowledge by developing a novel collaborative approach to organizational data quality improvement. It has the following three interconnected aims: (i) identifying a taxonomy of data defects; (ii) identifying the challenges and opportunities for organizational data quality improvement and developing a software prototype which automates defect detection in big data sets and fosters communication among participating actors to correct data problems; (iii) implementing the approach as a pilot for four data quality improvement teams and continuously refining it through various assessments performed during the implementation. The research adopted qualitative methods to collect and analyze rich contextual data and various iterative software development activities resulting in a software prototype, which is a multi-user client-server solution. This prototype played a critical role in implementing and refining the novel data quality improvement approach. By doing so, this dissertation research developed a blueprint for data quality improvement initiatives in healthcare organizations, which can potentially benefit patients and their families through the utilization of high-quality health data in the future.

DEVELOPMENT OF A COLLABORATIVE DATA QUALITY IMPROVEMENT APPROACH FOR HEALTH CARE ORGANIZATIONS

By

Yili Zhang

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, Baltimore County, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2018

Advisory Committee:
Güneş Koru, PhD - Doctoral Advisor and Committee Chair
Dongsong Zhang, PhD
Lina Zhou, PhD
George Karabatis, PhD
Jennifer Callaghan-Koru, PhD

Acknowledgements

First and foremost, I would like to thank my mentor Dr. Güneş Koru, whose intelligence and professions led me to a new level of scientific research. His patience and integrity will effect me through my rest of life in academic. His spirit of no giving up and facing the challenge inspired me a lot during the journey. It was my honor and pleasure to have worked in Health IT Lab and advised by him.

I am also grateful to committee members of my dissertation: Dr. Dongsong Zhang, Dr. Lina Zhou, Dr. George Karabatis, and Dr. Jennifer Callaghan-Koru. They contributed their valuable expertise, opinions, and time on this work to make it good shaped. I want to show the gratitude to the sponsor, the healthcare agency of one of the State in the U.S., without the support of them the research would not have been conducted, and the data would not have been available in this research.

I want to thank my big family and friends. I can always feel the support from bottom of their heart. They encouraged me to get through the most hard time of my PhD life, and shared the happiness of achievements I made in research.

Nevertheless, my boy friend, Bai Xue, with whom I built confidence for myself in research and everything else. He is the best therapy to me when I am upset or confused. Thanks for his companionship, which made me become a better myself.

The last but not the least, my parents, Weiguo Zhang and Xiangzhi Zhao, who always give me the most selfless love. They are the best model to teach me in my personal life, I appreciate their backing for my dream, and I'm proud to be their child.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Healthcare System in the United States

The healthcare system in the United States is a highly decentralized system influenced by many stakeholders such as governments, insurance companies, provider organizations, and patients and families. In this system, providing quality care, improving health outcomes, and reducing overall healthcare costs have been persistent, yet hard-to-reach goals.[1] The system was even characterized as "the world's most expensive, yet least effective (health system) compared with other nations".[2]

In 2011, nearly 18% of the gross domestic product (GDP) was spent on healthcare alone; this percentage is expected to increase to 20% by 2020.[3] Centers for Medicare & Medicaid (CMS) reported that in 2012, $2.8 trillion, or $8,915 per person was spent on healthcare alone.[4] This continued annual increase in health expenditures underline the urgency of improving healthcare productivity and preventing waste, fraud, and abuse. Although there are plenty of research studies, and plans developed to reduce costs in different aspects of healthcare,[5–7] such as diabetes and comorbid mental health disorders,[8] teriary psychiatric service,[9] bariatric surgery,[10] and so on; there is still a large amount of room for reducing healthcare costs.[11] If such an unsustainable trend continues, everybody in the society will suffer dire consequences because there will be limited access to

affordable and quality care.[12]

In addition to the high costs, healthcare quality and safety are constant concerns:[13–16] Medical mistakes are unacceptable in modern medicine[17] because they threaten our lives and interfere with our social and productive endeavors.[18] In the last two decades, healthcare quality and safety have become the first concern among different organizations.[19] Such as United Nation's World Health Organization,[20] European Union,[21] and other regions.[20] Most certainly, healthcare quality and safety draw attention from healthcare professionals, patients and their families, researchers, payers, planners and educators in the US as well.[22]

1.2 Crucial Role of Data in Healthcare

The day-to-day use and analysis of quality healthcare data holds tremendous potentials to alleviate some of the operational and decision making problems in the US health care system.[23] Health data is used for various purposes such as paying reimbursement claims, treating patients at the point of care, discovering the patterns and trends for diseases, health, and well-being,[24] finding and tracking policy violators,[25] identifying the effective treatments and best practices,[26] and detecting fraud and abuse.[27]

Consistent with Moore's and Kryder's laws of exponential increase of computational power and information storage, healthcare data has seen a rapid growth.[28] In addition, with better methods of extracting information, translating information to knowledge, and deriving appropriate actions, the value of healthcare data is expected to increase even more rapidly in the near future.[29]

For example, the substantial increase in the use of electronic health records (EHRs) in the last decade resulted in an accumulation of health data which presents opportunities to provide better care.[30] EHRs are valuable tools in documenting the patient health history and communicate it to the care teams.[31] Healthcare data solutions that provide clinical

decision support were shown to provide positive results in providing better care.[32] For example, using data stored in EHRs can potentially help providers with early recognition of medication errors, avoid potential drug-drug interactions, receive and examine lab results early, and provide recommended care based on accepted clinical guidelines.[33]

With the emerging technologies, there are also other kinds of health-related data collected from multiple sources such as sensors, images, and text such as the scientific literature.[34] As a result, the dimension, complexity, and magnitude of healthcare data are increasing.

1.3 Big Problem: Poor Data Quality

Scope of the Problem: As noted, health data analytics present important potentials. Generally, in healthcare, early detection and intervention of negative trends is key to provide better care, improve outcomes, save lives, and lower the healthcare costs.[35–37] However, a basic requirement in all cases is readily available high-quality data.[37–39] Only through the common and routine use of high quality healthcare data, can we reliably detect problems and intervene in developing threats such as the lead poisoning cases in Flint, Michigan,[40] the pill-mill pharmacies in West Virginia,[41] or the sudden HIV outbreaks in Indiana.[42] Healthcare data also have served as a useful resource for a various healthcare research activities such as population health research, disease control, healthcare improvement, patient and provider action study, and so on.

However, as the use of information system in healthcare increases, data problems are encountered at an increasing rate in healthcare organizations.[43] Despite the availability of a large volume of complex healthcare administrative data currently maintained by healthcare payers, providers, and governments, there are substantial wrong data items, i.e., inadequacies for intended purposes, which reduce the financial value that can be driven from data and constitute barriers to effectively leveraging it to improve health.[44–47]

In many instances,[48] the deficiencies in the data, due to lack of documentation,[49,50] lack of user training,[51,52] or delays in system updates,[53] decreased the quality of data and increased the difficulty of utilizing data effectively and efficiently. It has been stated that clinical data does not receive the same care of research data receives when it is recorded,[54] which is generally accepted as a norm because of the differences in priorities between clinical and research settings. Introduction of HIT like EHRs has not led to any improvement in the quality of the data being recorded, but, according to some, it rather led to the recording of a greater quantity of bad data.[55] Concerns specifically against the reuse of clinical data for research were expressed due to quality problems.[56]

Generally, a lack of data quality can result in imprecise, useless, or even misleading results which detract from the quality of reports produced and decisions made.[57,58] The deficiencies in health data often increase the difficulty in obtaining useful results.[48] For example, according to a study from oracle, healthcare providers lose on average of $70.2 million annually, or 15% of additional revenue per hospital, because they cannot interpret and translate the information into actionable insight due to the poor quality and garantuan size of data they collect.[59]

A motivating real life example: A specific example is the poor quality of the data stored in the Medicaid Management Information Systems (MMISs) adopted by the states. Typically, an MMIS deployed in one of the states in the US helps manage an integrated group of procedures and computer processing operations (subsystems) developed at the general design level to meet principal objectives. MMIS objectives include the Title XIX program[60] control and administrative costs; service to recipients, providers and inquiries; operations of claims control and computer capabilities; and management reporting for planning and control.

MMIS data has been utilized successfully in a small number of studies, such as improving the quality of myocardial infarction care,[61] as a resource for epidemiologic studies,[62] and estimating prevalence and medical care costs for diseases.[63] However, these

uses have been limited compared to the large amount of data stored in MMIS. Unfortunately, there are substantial problems with the quality of the existing data in the MMIS which reduces its usefulness for various data analytic purposes.

The reasons behind poor data quality are both accidental and essential problems that are rooted in software development, maintenance, evolution, and adoption phases which take place over decades in an organization. For example, the MMIS in one state has a user interface for data management with no data validation feature in the system. The manually inputting operation of MMIS increases the potential incorrect values in it. There could be provider service begin dates be later than provider service end dates, which would cause difficulty in identifying provider service valid period. In addition, there could be some empty cells of procedure code missing in Procedure subsystem, which is not only a required code, but also the primary key of Procedure subsystem. Furthermore, mismatches may happen between address information of healthcare providers, i.e., zip code, state code, county code, etc.

A straight forward example is that one of the provider has a county code belongs to one state but his/her state code indicates another state. If data problems are not corrected, it will cause bias when, for example, a heat map of healthcare providers in the state is produced. In this particular case, the deficiencies in the Medicaid data tremendously increased the difficulty of achieving useful results. Thus, effective data maintenance and cleaning becomes crucial to improve the quality of data in the subsystems of Medicaid Management Information Systems.

Both accidental and essential problems: It is highly important to note that accidental problems are, in fact, easier to deal with compared to the essential problems. Some accidental data quality problems can be reduced by adopting better data capturing mechanisms.[64] On the other hand, the essential problems are related to various other factors associated with successful systems and software evolution tied to the policy changes,[65] and the subsequent changes in software requirements and software tools,[66] and acciden-

tal problems,[67,68] particularly in big data.

Therefore, in the current state of health organizations, it is important to develop strategies to help improve data quality in an effective and efficient manner to reduce *defects* in data. In this context, a data defect refers to a discrepancy between actual and expected states of data item which requires a corrective change. Higher numbers of data defects are associated with poor data quality, and the lower numbers of data defects are associated with high data quality.

Due to the inherently difficult and human-intensive nature of understanding and resolving data defects, the most feasible approach for improving data quality seems to be about creating systematic, effective, and efficient organizational processes supported by tools. For this purpose, it is important to understand the actors, tasks, information needs, and information flows.

1.4 Knowledge Gap in Addressing Health Data Quality Issues

While healthcare and health data are uniquely complex, with an increased use of data in various domains, there is an increasing attention paid to data quality problems. There are some research studies defining the dimension of data quality,[69–71] and exploring the impact of data quality.[72–74] And there are a number of studies[75–82] about how to detect anomalies in data and improve data quality.

So far, there has been no systematic solution proposed by researchers or adopted by practitioners for data quality improvement in real life healthcare organizations.

In day-to-day operations, many health systems collect data from disparate systems often updated with hastly designed software patches and user interfaces; or from no software other than spreadsheets.[83–86] When data quality improvement is concerned, the first issue is how to understand the adopted business rules should place constraints on

the data (e.g., all providers licensed after 2012 cannot be associated with a certain procedure code removed in 2012). The storage, versioning, and the utilization of business rules are also problematic.[87] In addition, methods and processes for data defect detection on large datasets should be developed.[86,88] Furthermore, tracking and resolving the detected data defects require orchestrating the activities among multiple actors who should collaborate towards improving data quality.[88–90] Research in these areas are seriously lacking, resulting in a big knowledge gap to address health data quality issues effectively and efficiently.

Under these circumstances, for example, a healthcare expert in a state office who need to analyze existing data about teenage pregnancy rates in an urban neighborhood finds that data quality needs improvement because the addresses are either mal-formatted, or missing, or wrong. The data of interest can be retrieved from a single source or from multiple data sources to be imported in system. This health expert, perhaps lacking technical knowledge, could only write the business rule narratives about the collection and use of the data along with the constraints that can be verbally described. But then, a problem arises: Who will interpret this constraint at a programmatic level that can be further implemented to fix data problems? A systematic organizational approach is needed to improve the poor quality of data, which should govern extracting data from data sources, detecting data defects in an efficient manner, correcting wrong data, checking whether corrected data is indeed correct (validation), and manage different versions of data dumps, among other important activities.

Currently, in most health organizations, there is no such systematic approach. Furthermore, data quality improvement requires different levels of health or IT professional knowledge, different roles should take charge of these processes in the usecases. Hence, systematic solutions are needed to help, for example, (*i*) Documenting and versioning data constraints at different abstraction levels from business rule narratives to structured business rules, and to formal data specifications;[91] (*ii*) Automatically detecting data de-

fects based on the data specifications in an accurate and fast manner;[92] (*iii*) Monitoring and documenting the improvements achieved in data quality in a continuous fashion while also versioning the data.[93]

1.5 Research Objectives

The overarching goal of this research is to contribute to the body of knowledge by developing a novel approach to data quality improvement. It has three interconnected aims:

1. Identifying a taxonomy of data defects;

2. Understanding the requirements for improving data quality in a collaborative fashion in healthcare organizations and developing a novel approach to respond to those requirements;

3. Implementing the approach and refining it through various assessments

The research was applied and interdisciplinary in nature, and it collected and analyzed qualitative data from the human subjects in a real health organization. The researchers obtained an Institutional Review Board approval from the university and the health organization to conduct this research. The data subject to the quality measurement in the health organization is Medicaid data collected from the actual healthcare providers, Medicaid billing managers, as well as the organization itself.

The research also included various software development activities. These development activities resulted in a fully functional software prototype. This prototype is a multi-user client-server software toolkit, called Data Quality Toolkit (DQT). DQT played a critical role in the implementation and refinement of requirements.

1.6 Research Context and Methods

Many health data agencies are struggling with bad quality of health day-to-day. For example, a state agency in health organization makes reports about population health by analyzing the data from this organization. One day, when he is asked to create a new report about surgery in last five years, he discovers a number of suspicious entries entered by data operators of this health data system and verifies the problem spans back many years. For instance, the surgery codes in some entries are not valid since 2005, which appears in records of provider who registered after 2010. These issues obstacle the agency to produce a accurate and convictive report and must be fixed. However, the agency does not have all constraints between surgery code and registration date. Moreover, he needs a efficient way to find all invalid records in past-five-year data. With certain difficulties intercepted, the invalid data cannot be fixed in a short time and the report is not produced on time.

Considering about and facing the problem described above, three aims of this research was achieved in three phases. (*i*) A comprehensive and refined taxonomy of data quality was constructed via literature review as well as document analysis and descriptive analysis conducted on the real-life artifacts and data, respectively. (*ii*) An approach for data quality improvement was developed by using the taxonomy as the base knowledge. The development of this approach involves (a) identifying the organizational needs, such as the challenges and opportunities, through qualitative research (b) developing an approach and fully functional client-server software tool, DQT, to support the approach in which multiple actors play a role. (*iii*) The approach was implemented in an educational setting through continuous assessments of perceivable advantages provided to users, problems and their resolutions, usage data, and measurable improvements in data quality. Rather than create a generalizable model to be adopted in various types of data, this research leveraged qualitative approach for an in-depth investigation of data quality issues and solutions in educational settings.

1.7 Research Contributions

Overall, the evidence obtained in this research shows that the approach to data quality improvement was successful and met important needs and priorities for data quality improvement. By developing an approach and accompanying DQT, the research created a blueprint which can inform data quality improvement initiatives in healthcare organizations.

Following this blueprint in a health organization will facilitate checking and improving the quality of data, communicating among multiple units, and tracking data problems to their resolution. DQT, which is the toolkit automating these high level business processes is an important enabler. Using the approach and DQT created in this research, the data quality improvement initiatives will potentially achieve higher degrees of effectiveness and efficiency. As a result, there may be potentially smoother operations in health organizations providing a number of benefits such as a smaller number of denials for rightful reimbursement claims, avoiding the loss of profits, and preventing accidents in healthcare. In addition to operational benefits, better quality data will improve the accuracy of the decisions made as a result of various data analytics and mining activities. Therefore, improving the data quality will help health organizations to increase profits, produce a better healthcare for patients, and the most important, improve the result of healthcare.

The research was carried out by a competent PhD student who has a professional background and experience in the HIT domain as well as academic knowledge of research and its rigor. The student has a wealth of experience applying the research methods and techniques in prior work specifically in health data cleansing. In the following sections of this doctoral thesis, methods and results of three aims will be presented. After that, limitations and conclusions of this research will be discussed. Finally, details about the research are listed in appendix.

Chapter 2

Background

2.1 Related Work

A process[67] was developed for data quality improvement with five phases shown in figure 2.1: data analysis, definition of transformation workflow and mapping rules, verification, transformation, and backflow of cleaned data. As the most of research about data quality improvement emphasis on data cleansing, data cleansing methods was studies and categorized in statistical, clustering, and association rules.

2.1.1 Statistical

A method for automatically detecting anomalies and outliers in data is statistical approach, which assume that data or its features can form statistical distributions.

An application[75] adopted statistical approach by using mean, standard deviation or range of records values to analyze data, and detect outliers and anomalies which are beyond confidence intervals. In one of the experiment, fields were constructed by using mean and standard deviation as μ_i+$\varepsilon\sigma_i$, in which μ_i and σ_i indicate mean and standard deviation of ith record correspondingly, and ε is defined as confidence interval of ith field, which can be set using Chebyshev's theorem or manually. Those detected outliers and anomalies can then be indicated as invalid tuples or error data.

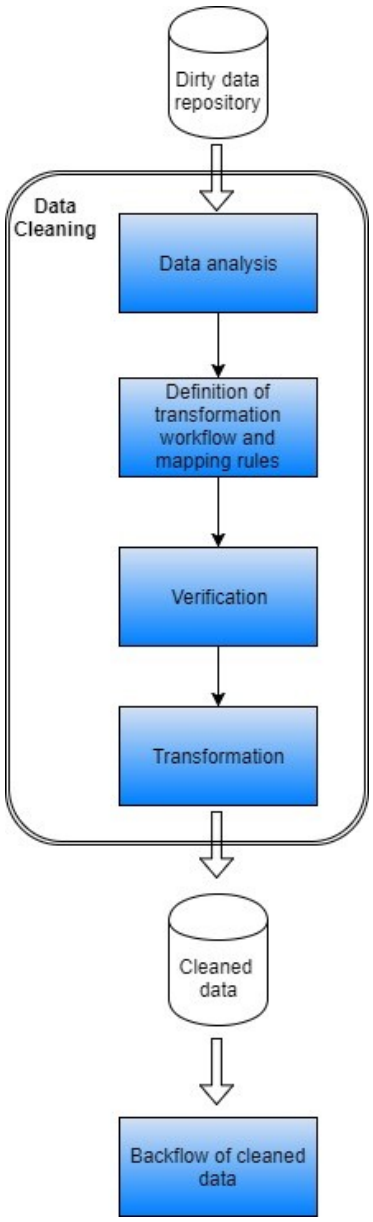

Figure 2.1: Data Quality Improvement Process

Outliers and anomalies were analyzed[76] in data set before produce data analysis processing in their work. Instread of using conficdence interval, theysphere the data set by global mean and Monte Carlo method to determin a threshold value.

About dirty data,[94] three techniques were developed for the query processing of these tools. The first is the similarity based operator. This was created because of the selection and the join operators. In the indice part,to process entity similarity search and join effi-

ciently, Fgram-Tree and bi-layer prefix indices were designed for entity similarity search and entity similarity join, respectively. The last part was query optimization.

Other research[95] addresses different methods available for treating and analyzing missing values. The complete case analysis uses only the data of variables observed at each time point for analysis after removing all missing values. this is used for smaller data sets. The available case analysis deals with the data available at each analysis. The sample size for this is larger than the complete data analysis. The last method used is Imputation analysis which involves replacing missing values with substituted values obtained from a statistical analysis to produce a complete data set without missing values for analysis. For determining outliers, the distance between a data point and the center of all data points is measured.

A paper concentrates on data quality[96] studies effective methods for improving both data consistency and accuracy. They employ a class of conditional functional dependencies (CFDs) proposed in to specify the consistency of the data, which are able to capture inconsistencies and errors beyond what their traditional counterparts can catch. To improve the consistency of the data, the authors propose two algorithms: one for automatically computing a repair that satisfies a given set of CFDs, and the other for incrementally finding a repair in response to updates to a clean database. They show that both problems are intractable. Although our algorithms are necessarily heuristic, the authors experimentally verify that the methods are effective and efficient. Moreover, they develop a statistical method that guarantees that the repairs found by the algorithms are accurate above a predefined rate without incurring excessive user interaction.

Data quality evaluation[97] uses two sets of methods to determine and improve the quality in files and group of files. The first method is data editing that verifies that data values satisfy predetermined restraints. The second method is record linkage that is based on a statistical model due to Fellegi and Sunter.

2.1.2 Clustering

Unlike statistical approach above, the pattern based approaches were assumed data are forming patterns instead of statistical distributions.

Using identification procedures to identified patterns inside data, and the record beyond patterns' field are identified as outliers and anomalies.

Several pattern recognition algorithms are test in a research.[75] A combination of multiple group-average clustering algorithm[77] is tested, and consider Euclidean distance between each record in data set. The constructed clusters are generated in terms of certain similarity or connection. Therefore, it is high likely for the algorithm generated clusters forming patterns.

And in several experiments,[75] according to the distinguishing features of tested data, the Hamming distance is introduced to deal with the similarity measurement of strings elements inside records, and shows highly efficient. However, since group-average clustering algorithms are not performance globally but a fixed size cluster contained selected samples. and size of cluster should be manually setup previously. This prior information is hardly can be estimated and the selection of parameters is empirical. Therefore, it has high computation complexity deal to its high quantity of records selection combination, and it has difficult implement issues in terms of its empirical parameters selection.

Another pattern based algorithm is tested is k-means algorithm.[78] In the experiments, equal distance as initial condition. Hamming distance for strings tuples in record is considered as well. In this case, initial parameters set for the algorithm is important. the number of clusters must be estimated priory.

2.1.3 Association Rules

Associate rules approach forms a subset from original data set in which samples follow a certain rule, in this case, records which are not follow the rules are considered as outliers and anomalies.

The association rules approach was first introduced[79] into data mining, and other researchers[80] describe studies cover various topics in mining association rules approaches area. The most popular approaches includes: quantitative association rules,[81] which build categorical events from the quantitative data by considering range of values in records, the ratio-rules,[82] which build categorical events based on the values in records inside certain value intervals, and ordinal association rules,[75,80] which considers items in each record have same numerical domain, and exist relationship between items in their domain.

Stream data cleansing,[98] adopt anomaly detection as a method to find the incorrect data in the data set. The overall steps is to first input the data, then detect the error and process to the cleaning.

The baseline algorithm[99] was also introduced. The first phase of the algorithm is initialization where a matrix is generated with the location of each tagged object. Update generates a neighbouring state by applying uniform distribution. In the selection state, probability is used to find out if the neighbouring state is accepted or not.

Rule collector,[100] the tool collects user specified rules like deduplication for example. Then in the core the rule compiler is used to compile heterogeneous rules into homogeneous constructs to enable the development of a default complete data cleaning algorithms; and the metadata management and Data quality dashboard modules deal with maintaining and querying various metadata for data errors and their possible fixes.

Steps[101] taken were also discussed. In insertion, batch computation talks about tuple comparison in horizontal partitioning, or shipping tuples in vertical partitioning. In incremental computation, data is shipped in vertical computation but not in horizontal computation. Then there is the deletion process where violations are found and deleted. This paper establishes the complexity bounds and provides efficient algorithms for incrementally detecting the violations of CFDs in fragmented and distributed data, either vertically or horizontally.

2.1.4 Duplicate Data Treatment

Bigdancing[102] consists of five logical operators strong enough to clean a large amount of data. Scope removes the unrelated data and outputs. It allows data units to be replicated but only outputs the data based on the rule given. Block groups the data that has the same blocking key. It helps narrow down the data on which the violation might occur. Iterate defines a rule to combine the data that performed a violation and puts them in sets of two. Detect finds the sets from iterate that are not exactly identical. GenFix computes couple fixes for each set.

A semantic querying for duplicate data[103] was proposed. They suggest a rewriting technique for a large class of select-project-join queries and the answer of the queries are collected in a table according to the semantics that they defined for dirty data. Then a method is proposed find the probability of potential duplicates. Using real data the authors argue that argue that the computed probabilities have an intuitive semantics and can be computed efficiently.

2.1.5 Pre-process Data before Sorting

A research about EHR data quality[104] was conducted in four general practices participating in the Fairfield neighborhood electronic Practice Based Research Network (ePBRN). Data were extracted from their clinical information systems and summarised as a SDQR (structured data quality report) to guide feedback to practice principals and managers at 0, 4, 8 and 12 months. SDQR data was deployed by Microsoft SQL Server. Its query output was cross-checked with IBM SPSS Statistics V20. Data quality (DQ) metrics included completeness, correctness, consistency and duplication of patient records. Information on data recording practices, data quality improvement, and utility of SDQRs was collected at the feedback sessions at the practices. The main outcome measure was change in the recording of clinical information and level of meeting Royal Australian College of General Practice (RACGP) targets.

Inconsistent database[105] was applied by a logical characterization of consistent query answers in relational databases that may be inconsistent with the given integrity constraints. Intuitively, an answer to a query posed to a database that violates the integrity constraints will be consistent in a precise sense: It should be the same as the answer obtained from any minimally repaired version of the original database. they also provide a method for computing such answers and prove its properties. On the basis of a query Q, the. method computes, using an iterative procedure, a new query Tw(Q) whose evaluation in an arbitrary, consistent or inconsistent, database returns the set of consistent answers to the original query Q.

2.2 Comparison of Tool

Currently, there are some existing tools for data quality improvement. Some of the tools are free and some are commercial tools which only provide a tutorial video online. To compare the existing tool with DQT going to be developed, researcher did a tool review before the development begins.

Following are some tools used for data cleansing and the specifics of what each of them are used for. All these tools have a common goal which is to make sure that all the companies using data in their daily activities have a clean data in a limited amount of time if not systematically.

Acme data is a data cleansing tool used for used to clean data, find the addresses, remove duplicates records from the batch in real time. Trifacta is a tool used to prepare data for analysis. This tool is designed to enable analyst to explore, transform and enrich raw data. Openrefine is a tool installed on a desktop for data cleanup and data transformation into other formats. Paxata is a tool used to combine data from different sources and makes sure that the data does not have any duplicates. Alteryx Analytics provides analysts with the unique ability to easily prep, blend and analyze all of their data using a

repeatable workflow, then deploy and share analytics at scale for deeper insights in hours, not weeks.

Trifacta was a research project from the Stanford/Berkeley visualization group. The tool was designed to save companies time from manipulating and preparing the data. The tools allows the users to export data for use in excel, R or Tableau. This tool is interactive with a visual representation of what the user is working on. For example when a user select things, they have visually represented options of what they could do with the data selected. Every wrangling step by the user is tracked and and automatically compiled down into the appropriate processing framework without the user having to worry about every step.

DataWatch is a data preparation tool used for handling unstandardized and inconsistent data. The data can be automatically extracted from webpages and other forms. This tool does Data analysis, Creating an intuitive workflow, Validation, transformation, and backflow of cleaned data.

ActiveClean is a tool created by students who collaborated from the University of Berkeley, Columbia university and Simon Fraser University. The tool is mainly based on machine learning algorithm because they believe that some mistakes will be reduced by taking out humans out of the equation.

Overall, these products have more pattern repository than DQT but none of it has semantic validation feature. Rules in DQT rely on user customization, which provide a more flexibility to detect more types of data defects in certain domain. And the semantic validation in DQT give user the ability to detect data defects across columns, that no other tool has it.

To have a better comparison of DQT with other tools, figure 2.2 listed several important criteria of data quality improvement, and check function completeness of DQT and selected tools.

There are 11 criteria, In the figure 2.2, which are unique validation, syntax validation,

Tool	Unique Validation	Syntax Validation	Semantic Validation	Rule Custo-mization	Split Data	Business Rule Versioning	Data Versioning and Comparison	Text Correction	Defect Meta-data Storing	Defects Grouping	Quality Profile Gene-ration
DQT	√	√	√	√	√	√	√		√	√	√
Data Ladder	√	√				√		√		√	√
DataCleaner	√	√		√		√		√		√	√
Win Pure		√						√			√
BackOffice Associates		√									√
Innovative systems	√	√									√
Uniserv	√	√									
REdPoint	√	√									
Neopost	√	√									
Talend	√	√									
Ataccama	√	√						√			√
Omni-Gen		√		√							
Experian		√									√
Pitney Bowes	√	√		√							√
Oracle-EDQ		√	√	√							√
Trillium Software	√	√		√							
SAS		√	√				√	√			√
SAP	√	√									
IBM	√	√		√				√			√
Informatica	√	√	√	√							√

Figure 2.2: Feature Comparison of DQT and Other Selected Tools for Data Quality Improvement

semantic validation, rule customization, split data, business rule versioning, data versioning and comparison, text correction, defect metadata storing, defects grouping, and quality profile generation. Here is detailed description of these criteria:

1. Unique validation is the function to check whether all data tuples are unique in dataset, i.e., if there is any duplicates in dataset. Most of tools has this function.

2. Syntax validation is to check the whether the data follow the rule of itself. In other words, it is the function to check syntax violation that will be mentioned later. Every software has this function. Most of tools check only syntax rules for contact information, such as format of address, telephone number, zip code, and so on. Some tools check other rules of data; for example, range of heart rate.

3. Semantic validation is to check whether data follow the logic of its own meaning. In other words, it is the function to check semantic violation in the taxonomy identified. Tools that this function applies to support to check inconsistencies between geographic information such as whether address and zip code matches.

4. Rule customization is the function to give user the access to modify the rule to detect violations. Less than half of the tools have this function. Those tools give user only the ability to change parameter in constraints, but not any constraint that user need. While in DQT, user can set any constraint or rule that can be performed by SQlite or regular expression, which enlarge and enhance the scope of detection.

5. Split data is the function to split large data set into pieces and process data pieces in parallel to reduce time of execution. DQT is the only one tool having this function, which allow DQT to process dataset with large number of columns and records.

6. Business rule versioning is the function to manage different versions of business rules. Only three tools including DQT provide this function. The function give user the availability to use business rule already set up and to check the change history

of each business rule. Only three tools including DQT have this function. Other tools only provide instant business rule setting and using.

7. Data versioning and comparison is the function to keep track of different versions of data during the correction process. With this function, user could roll back and forth to get newer or older version of data in a long term of data dump correction process. This function also allow user to compare two different versions of data to check whether corrected part is as expected.

8. Text correction function is to make change to data dump in the tool, and export changed data out as a new version. This is the only function that DQT is missing compared with other tools. The reason of this function not being developed is that DQT will still put data on user's side to keep high security of data to minimize risk of corrupt data dump in the system.

9. Defect metadata storing is the funtion to keep information of defects, which includes value of defects and the location of defects. With this function, DQT can provide a more intuitional review of data defects, and this function also facilitate the grouping function.

10. Defects grouping function is to get a group of defects from a large amount of it by a certain filter. This function give user a better approach to increase efficiency when they correcrt data defects becasue in this method they can focuse on a certain type of error.

11. Quality profile generation is to generate a quality report of data to give user an overview of data in multiple criteria. For example, in Data Cleaner, it will generate the percentage of missing data in each column, percentage of duplicated data in each column, and so on. This function is useful because through an over look, the user will get an idea about whether the data dump has enough good quality to be used for analyzing.

2.3 Scientific Knowledge Gap

Although statistical approach provides a solution in a relatively interpretable fashion for detecting some problems with data, it is not a sufficient method to understand and solve problems of data quality. For example, it is time-consuming to find optimal workflows simply by looking at the statistics obtained from the data subject to quality improvement. When there is no distribution in data that can be detected by statistical approaches, clustering is a good alternative because in this unsupervised approach, data can be clustered to observe discrepancies. However, result of clustering is hard to interpret because there is no variable set in data could explain the result. Association rules is implemented by setting rules to data. Data follow the rules are normal data and data does not follow the rules will be recognized as anomalies. With rules are set, precises result can be high as long as constraints are exact and complete. But result will miss anomalies with no rules are set of found. This approach cannot be used when rules can not be collected completely in data.

So far, no study has addressed the problem of developing a collaborative approach to data quality improvement supported by effective workflows and software solutions. The related work presented in this chapter only deals with some of the sub problems of this large problem, and with limited effectiveness and efficiency.

Chapter 3

Aim 1: Taxonomy of Data Defects Identification

To identify a taxonomy of data defects, this study classified and detected data defects in the Procedure and Provider subsystems of a Medicaid Management Information System (MMIS), adopted and maintained by one of the states in the US in order to support the principal objectives of the Medicaid program. In this context, a *data defect* refers to a discrepancy between the actual and expected values of the related healthcare administration data.

The objectives were (i) Developing a comprehensive taxonomy for data defects; (ii) Develop and examine a program toolkit that facilitates and automates data defect detection based on the taxonomy.

Consequently, the study provides the most comprehensive taxonomy for data defects reported so far in the literature. The program successfully identified a large number of defects in various categories of this taxonomy, which underlines a substantial need for quality improvement in Medicaid data as a prerequisite for utilizing it for better decision support. The findings also suggest that augmenting data governance policies with an emphasis on data quality is likely to provide useful results.

3.1 Methods

This study first identified a taxonomy of Medicaid data defects and then developed the program to detect data defects automatically in an efficient and accurate manner, finally defects in data of Procedure and Provider subsystems were checked by using the program.

3.1.1 Identifying a Defect Taxonomy for Medicaid Data

Identifying a taxonomy contributes to the knowledge about data defects in this particular problem domain. In addition, the taxonomy facilitates the tasks related to implementing the program tools for automatic defect detection. For this purpose, all available documents related to data were examined. Some documents define certain constraints on data items were concluded as business rules for defect detection. Then, a descriptive analysis was performed to observe common data trends and abnormalities in order to discover existing and potential violations in data. This process resulted in a basic taxonomy with different types of violations. Finally, the data defect taxonomy was supplemented and completed by conducting a comprehensive literature review.

Reviewing Related Documents and Understanding Data

Table 3.1 shows the basic information about Procedure and Provider subsystems in Medicaid datasets. The subsystem, table name, number of columns, and number of rows for each data table are shown. Other documents such as user guides and value description files were collected and reviewed to understand the data and the associated business rules. The user guides helped us obtain information about the column descriptions and with three display screens for Procedure subsystem and seven display screens for Provider subsystem. Value description files consist of information about the full names and valid values about the variables in the datasets. Knowledge about business rules and

Table 3.1: Basic Information about the Medicaid Datasets Studied

Subsystem	Table Name	Number of Columns	Number of Rows
Procedure	Claim Type	3	242
Procedure	Coverage Group	3	67,465
Procedure	Master	55	35,076
Procedure	Modifier	3	184,378
Procedure	Place of Service	3	38,564
Procedure	Price	8	371,177
Procedure	Provider Type	3	39,093
Procedure	Specialty	3	2,052
Provider	Address	9	241,875
Provider	Category of Service	12	375,541
Provider	Enrollment Period	4	342,877
Provider	Group	6	113,723
Provider	Lab Classification	5	18,408
Provider	Master	97	165,036
Provider	Receiver	7	26,812
Provider	Specialty	6	189,100
Provider	Supplement	4	79,151
Total			**2,290,570**

valid values are extracted and concluded from these documents.

Performing a Descriptive Analysis to Identify a Taxonomy of Data Defects

Data tables were imported into the statistical environment, R,[106] by paying close attention to the data format. After importing the data, descriptive results were generated for each table which included the number of rows, number of missing values, number of unique values, lowest values, and highest values.

Each variable in the dataset was examined to identify potential violations by recognizing no-value (NA); unexpected symbols, such as a comma or a period in name field; and deviant values, such as "01/01/1901" or "12/31/9999" in date field. From the observations of potential violations, it can be concluded that one column can have multiple types of data discrepancies. What's more, some rules between multiple columns could also be discovered. For example, begin date must be earlier than end date in a date pair,

previous provider number must be different with new provider numbers, and so on. This examination refined the types of violations in Medicaid data and resulted in a basic data defect taxonomy.

Conducting a Literature Review to Supplement Categories of Taxonomy

Researchers have tackled several problems in data quality, some of which include, dealing with missing data, erroneous data,[107] duplicate dataset,[67,107–110] quality of metadata,[111] and data repair.[112–114] Research has also been conducted to build data quality framework and data quality principles,[71,115–117] solve the problem of merging multiple databases, and eliminate redundancy in data.[118]

In order to achieve a more comprehensive categorization of data defects, a literature review was conducted by searching relevant keywords including data quality, data cleansing, dirty data, data defect, data profiling, and data repair. Combining knowledge obtained from this literature review and descriptive analysis, a taxonomy of defects with six major categories and seventeen subcategories was created.

3.1.2 Developing a Program for Automatic Defect Detection

The next stage of research was to develop the program to detect defects within the identified categories automatically in an effective and efficient manner. The program was mainly developed using the Tool Command Language (Tcl)[119] and stores the data in the SQLite database.[120] Rules identified in document review were coded into the program, the operation of detection can be informed by the defect taxonomy identified.

For performance improvement, threads were adopted to process the data in parallel, which substantially decreases the running times. The program includes functions of detection that could be invoked through both command line and the graphical user interface (GUI). Data about each defect is stored in a database for reporting and visualization purposes. The program performs in an efficient and intuitive way to detect defects in MMIS

data, which provides the necessary first step to improve data quality.

3.1.3 Detecting Data Defects in Procedure and Provider Subsystems

Tables in Procedure and Provider subsystems were imported into a SQLite database by the program as a data quality improvement initiative. Every column associated with a rule was detected based on the type of data defect. The program stored information about each data defects in a defect table within database, which includes table name, column name, defect value, data defect type, and row number. The defect table can be queried to collect statistical information of data defects in Procedure and Provider subsystems.

3.2 Results

3.2.1 Taxonomy for Medicaid Data Defects

Figure 3.1 shows six categories and seventeen subcategories of the taxonomy created for Medicaid data defects in the form of a tree structure. The square symbol and triangle symbol mark the subcategories discovered in document and descriptive analysis and from literature review, by which should be noted that conditional required value missing and multiple attribute value are contemporary subcategories discovered from this study.

Missingness[67,121–125] indicates absence of fields in the columns which are supposed to be filled. There are some columns which must not be empty under any circumstances, such as provider base number, provider number, and columns which are primary keys in dataset. Records not presented in such mandatory filled columns are referred as required value missing.[67,121–125] Conditional required value missing subcategory refers to the absence of values in such columns whether values should be filled or not depending on values in other columns. For example, Health Maintenance Organization (HMO) type must be filled when provider type is "HMO" and provider location is "00" in same record.

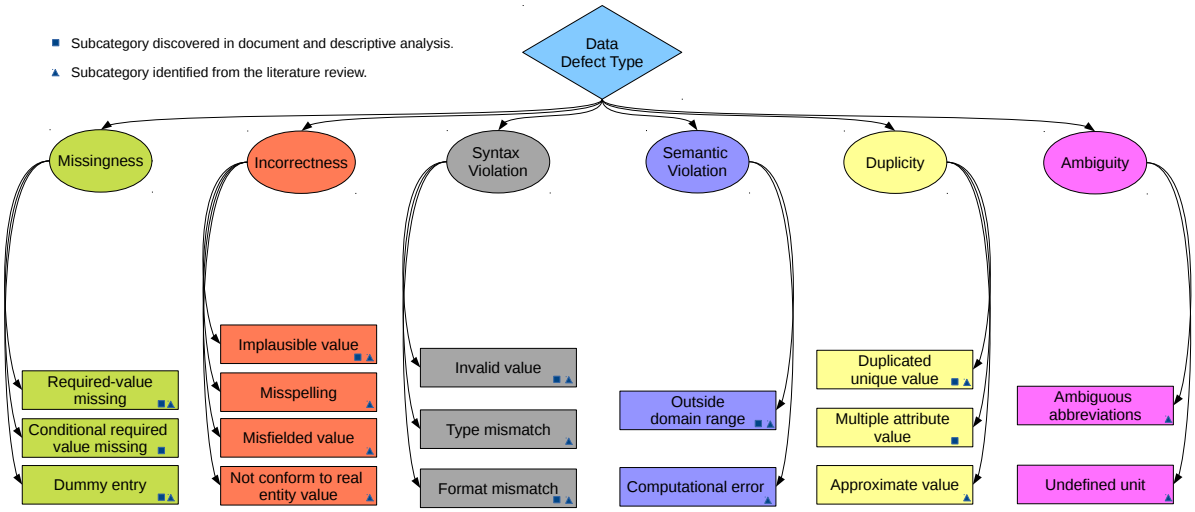

Figure 3.1: Taxonomy Tree for Medicaid Data Defects

Dummy entry[67,125] is the value with no actual meaning. For example, the presence of value "000000000" in provider social security number field.

Incorrectness[67,121–123,125] is about values that are not identical with real record values. Implausible value[67,121,122,124] indicates values are very high or very low compared with normal value, such as "1901-01-01" in provider service begin date, which is obviously inconsequence. Misspelling is the value with spelling errors or typos. Misfielded value[67,108,122,123,125] means value shifted to other columns by input mistake or programming bug. And there are also values not included in any of the subcategories above but do not match the real record, this type of error is value that not conform to real entity.[122,124,126]

Syntax Violation[121,122,124] refers to a deviation from the required data form. Some columns should follow a list of valid value, values entered beyond this list will be invalid value.[67,108,110,112] For example, in provider state code, there are 50 valid codes for 50 states, but the presence of "OO" in this column is invalid because it does not present any state in the US. Type mismatch[122–125,127] refers to value does not fulfill requirement of data type, such as numeric appears in provider name column, which is suppose to be text

only. Some columns also have constraints about number of digits and combination of alphabet and numeric, the value violates those constraints are format mismatch.[122–124,127] For example, provider base number must be seven digits, but provider base numbers with five digits, four digits, or even three digits were found in dataset.

Semantic Violation[67,107,113,114,121,124,127] is related to the inconsistencies of information within and across columns. Outside domain range subcategory[67,114,121,122,124–127] indicates that the value is not in the value range restrained by other columns. Such as each provider specialty code determines a group of valid value of provider type, therefore, provider type will be restricted by provider specialty code in one record. There are other types of outside domain range cases. For example, there are provider service start dates later than provider service end dates in same records in provider enrollment period table. In these two example, two columns are restricted to each other and the rules between them are violated. The computational error[124] refers to the value does not follow the computational relationship with other columns.

Duplicity[67,108,122–124,127] indicates same or near same (approximate) values for two or more records in a dataset. Unique value duplicate[67,108,108,122–124,126,127] subcategory refers to duplicate rows or duplicate primary keys, which are supposed to be unique in dataset. Multiple attribute value means same value in multiple columns of one same attribute. For example, in provider category of service table, there are eight columns for provider category of service code because each provider can provide at least one and at most eight categories of service. Thus, values in these eight columns must be unique. Also, for a particular record, approximately same values in all the columns across tables can be referred as approximate value.[67,108,110] The approximate can be few alphabet missing or mismatch, or alphabet case problem.

Ambiguity[108,113,123–125] is the value without accurate description or unclear meaning. Approximate values[67,108,122,125] are data with similar contents. Ambiguous abbreviation[67,108,123–125] refers to misleading abbreviation such as "Dr", which could be interpreted

as both "doctor" or "drive". Another example can be a misleading context such as "Miami", which does not indicate whether it is in Ohio or Florida. Undefined unit[67,108,109,122,125] indicates numeric in a measurement column without a measurement unit.

3.2.2 Detecting Data Defects by Developing and Using the Program

Rules of Medicaid data investigated from documents and description analysis were applied in the program for detection. Table 3.2 shows number of rules associated with different data defect types in each table. As one column can violate one or multiple type of rules, the number of rules in each table can be more than the number of column.

Table 3.2: Distribute of Different Type of Rules in Each Table

Table Name	**Missingness**	**Incorrectness**	**Syntax Violation**	**Semantic Violation**	**Duplicity**	**Total**
Procedure Claim Type	3	1	2	0	1	7
Procedure Coverage Group	3	1	2	0	1	7
Procedure Master	5	3	33	3	1	45
Procedure Modifier	3	1	2	0	1	7
Procedure Place of Service	3	1	2	0	1	7
Procedure Price	5	2	1	1	1	10
Procedure Provider Type	2	1	2	1	1	7
Procedure Specialty Code	3	1	2	0	1	7
Provider Address	6	1	4	2	1	15
Provider Category of Service	3	2	1	9	2	17
Provider Enrollment Period	2	1	2	0	1	7
Provider Group	3	2	1	1	1	9
Provider Lab Classification	4	2	2	0	1	10
Provider Master	19	16	27	9	2	74
Provider Receiver	1	2	1	1	1	7
Provider Specialty	4	1	2	0	1	9
Provider Supplement	2	2	1	0	1	7
Total	**71**	**40**	**87**	**27**	**18**	**243**

The program detects data defects in Medicaid data by performing regular expression and SQLite query. A domain specific language (DSL) was created in the program to op-

erate data defect detection by a "detect" function, followed by parameters such as type of data defect, table name, and column name. The DSL not only support detection in command line operation, also can be invoked by program to facilitate a GUI operation. The adoption of threads reduces execution time by 0.956% in maximum.

3.2.3 Defects Detected in Procedure and Provider Subsystems

Table **??** shows the number of defects detected by the categories outlined in the taxonomy. Feature to detect Ambiguity in the program is under developing therefore it is not shown in the table. The total number of defects is more than two million, which indicates the needs for a substantial quality improvement effort to be performed for the Procedure and Provider subsystems.

3.3 Discussion

It can be discovered that some tables have defects more than the number of records. Especially in procedure master table each record violates 4.86 rules and in provider master table, each record violates 5.24 rules. That is because in these two tables, there is a large use of Medicaid codes, which can cause misuse, inconsistencies and errors. Provider category of service table, with eight out of twelve columns as provider category of service code, which is one of the Medicaid code, also hold more than one violations on each record.

In total, 2,086,628 data defects were detected, i.e., on average 0.91 defect types per record violates in the Provider and Procedure subsystem among 2,290,328 records. This defect rate revealed the severity of quality in Medicaid data and emphasizes the necessity of improving quality of Medicaid data.

The results obtained through the use of the program show that most of data defects fall into the Syntax Violation. Syntax Violation takes about 64% of defects, it happened

Table 3.3: Data Defect Counts in Tables

Table Name	Missingness	Incorrectness	Syntax Violation	Semantic Violation	Duplicity	Total	Defect Density %
Procedure Claim Type	0	0	61	NA	0	61	25.2
Procedure Coverage Group	0	0	911	NA	0	911	1.4
Procedure Master	2	2	168,346	1,428	766	170,544	486.2
Procedure Modifier	0	0	87,818	NA	0	87,818	47.6
Procedure Place of Service	0	0	2,153	NA	0	2,153	5.6
Procedure Price	100,813	1	25,453	218	0	126,485	34.1
Procedure Provider Type	0	0	3,983	651	0	4,634	11.9
Procedure Specialty Code	0	0	324	NA	0	324	15.8
Provider Address	14	0	95,527	105,124	0	200,665	83.0
Provider Category of Service	9	8	75,095	307,188	0	382,300	101.8
Provider Enrollment Period	0	0	66,837	NA	0	66,837	19.5
Provider Group	0	1	24,275	26	0	24,302	21.4
Provider Lab Classification	0	5	1,526	NA	0	1,531	8.3
Provider Master	61,086	70,158	709,838	23,953	30	865,065	524.2
Provider Receiver	0	0	6,639	0	0	6,639	24.8
Provider Specialty	69,856	76	55,783	NA	0	125,715	66.5
Provider Supplement	0	0	20,644	NA	0	20,644	26.1
Total	**231,780**	**70,251**	**1,345,213**	**438,588**	**796**	**2,086,628**	**91.1**

because of the misuse of Medicaid codes and Medicaid indicators. It should be noted that all provider remittance media codes, all record codes, and more than 99.8% of Medicare part codes in provider master table are wrong, which is appalling because the misuse of Medicaid codes. There are 57 columns related to Medicaid codes and Medicaid indicators in procedure and provider datasets. Each column has a list of valid values, the number of values in the list varies from 2 to more than 100. Since there is no validating feature in MMIS, there is a high possibility that user input wrong values.

Nearly 21% of data defects are Semantic Violation, which mostly occurred due to the presence of mismatch of address information (zip code, county code, state code, out of state code, etc) in provider address table and provider master table, and mismatch between Medicaid codes. For example, each provider type code has a group of valid provider category of service codes. In this case, the provider type and provider service code not only need to follow the syntax rule, but also need to match with each other in a semantic manner. Usually, a value has Syntax Violation would also violates semantic rule if it is correlated with a value in another column.

About 11% of data defects fall in Missingness. Most of of Missingnesses belongs to a unique category of data defects – conditional required value missing. The violate cases are 69,856 missing provider specialty certificate number, which is required if there provider specialty certificate date is entered; and 100,811 missing procedure code modifier, which should be filled if procedure modifier indicator is filled. Other Missingnesses are dummy entries of numeric columns such as social security numbers, telephone numbers, and provider certificate numbers. There are also two missing procedure short names in procedure master table, which is a required value.

Incorrectness takes about 3% of defects due to a large number of implausible values. For example, there are many dates as "01/01/1901" or "12/31/9999", which are not acceptable. This could happen because those two dates were entered when user left start date and end date empty, or user entered it in system intentionally as "01/01/1901" indicates

"no start date" and "12/31/9999" indicates "no end date", in other words, "indefinite".

Less than 1% of defects are Duplicities. Most of Duplicities are records with same primary keys in procedure master table. Other duplicities are duplicate previous provider number and new provider number in provider master table. There is one rule applies to multiple attribute value subcategory, which is the eight columns of provider category of service code should not be equivalent with each other. Although there is no defects found in this rule, the discovery of this new data defects type can contribute to future Medicaid data quality improvement process.

Feature to detect Ambiguity is under developing phase, Ambiguity records or values can be detected in future work. Overall, the data quality problem of Provider and procedure subsystems in MMIS is sever and needs to be solved urgently. Syntax Violations in Medicaid code caused by lack of documentation, lack of user training, or delays in system updates required attention to eliminate data defects in an efficient manner. By identifying the problem and location of data defects, users of MMIS can correct and clean it in the system more easily.

3.4 Limitations

The research achieved its objectives, but there were few limitations. It is possible that few data defect categories are still left undetected. Since the documentations for MMIS data are limited and not sufficient, business rules collected from user guide and value description file only cover a portion of overall Medicaid code constraints. However, deriving the data defects both from data and literature review helped us obtain a sufficiently comprehensive coverage.

Also, data collected from different sources varies in quality and format, which requires pre-processing before it is ready for analysis. For example, if date format in all tables are not consistent, they need to be transferred and unitized before imported in to the

program. In this research, format of each column has been identified and conformed to make result more reliable.

3.5 Conclusion

This research provides detailed knowledge about the categories for data defects and the methods to detect them. A comprehensive taxonomy of defects was created as a result of descriptive analysis and literature review, which categorizes data defects into six major categories and seventeen subcategories and is comprehensive enough and generalizable to other dataset. Conditional required value missing and multiple attribute value duplicates are unusual data defect types discovered in this research, which have not been mentioned in other study. The program was developed in an intuitive manner to identify data defects, based on the defect categories, which can be operated easily by data agencies in healthcare organization. The program detected more than two million data defects in Procedure and Provider subsystem of the Medicaid data examined.

Data quality is a vital problem to be considered. Improvement of data quality can be useful in many areas in healthcare from decision support to fraud detection. Data sharing policies of healthcare organizations can be revised to include steps to draw attention to data defects and to encourage organizations to reduce them as much as possible to achieve higher levels of data quality and utilization.

Chapter 4

Aim 2: Understanding the Requirements for Data Quality Improvement and Software Toolkit Development

Based on the taxonomy of data defects created, requirements should be collected by understanding challenges and opportunities of data quality improvement in healthcare organization. Rich contextual data about the challenges and opportunities of data quality improvement were collected from a qualitative study in order to design the high level business process. After that, a fully functional software prototype, DQT, which is a client-server solution with GUI was developed based on the requirements driven from the process. DQT supports multi-user operation on large datasets and performs in a fast and efficient manner.

4.1 Methods

A qualitative research approach was taken in understanding the challenges and opportunities for organizational data quality improvement. To identify that, human experiences and observations are important, therefore, qualitative research was chosen to col-

lect subject materials. Nine semi-structured phone interviews were conducted with the health professionals who take charge of data governance in health organizations. Health professionals who take charge of data governance are a group of people managing and manipulating healthcare data in their daily life, who are the best objective to be interviewed. Questions designed according to established system analysis strategies[128] were asked, and each interview was recorded and transcribed verbatim to ensure accuracy. Framework method,[129–131] was adopted for data analysis, which is flexible to analyze quantitative data and extract conceptual result.

The design of high level business process and the development of DQT adopted custom development, which provides the necessary flexibility in developing uniquely needed features in this particular problem domain. DQT development involved various tasks such as architecture, program, data, and user interface design, programming, testing, and documentation. The development ended-up with a fully functional client-server multi-user software.

4.1.1 Collecting Software Requirements

Data Collection

To comprehend the requirements of data quality improvement software, researcher needs to identify:

1. Roles that will perform action and make contribution to data quality improvement;
2. Desired and effective high level business process for data quality improvement.

The latter can be best articulated by understanding the challenges and opportunities of data quality improvement in an organizational setting. For this purpose, interviews were used as it provide detailed information about personal experiences and opinions, and meaningful responses from further questions asked.

Interview Design

The qualitative process at this stage was conducted via semi-structured interviews,[132] which helps researcher to develop a keen understanding of topic interest. The questions and guide of semi-structured interview can be prepared in advance, which allow interviewees to get prepared and provide a more reliable and prepared answers. Furthermore, the open-ended questions in semi-structured interviews let interviewers follow relevant topic and meaningful interest of research, which help the interview to be consisted by structured and strong relevant questions.

This research has been reviewed for the protection of human subjects and approved by the institutional review board (IRB) of the University of Maryland, Baltimore County. The IRB protocol number is Y17GK12046. The IRB was obtained before any data collection. The combination of expert sampling[133] and convenience sampling[134] was adopted as a sampling strategy to get easiest accessible experts in this knowledge domain.

As there is a lack of observational evidence in data quality improvement in health organizations, candidates of interviews are in a position of data management in health organizations. Furthermore, it is not easy to get in touch with those professions in a short time; thus, participants of interview are suggested by MMIS professionals. The interview questions is in Appendix A. Questions were based on a number of well-established system analysis concepts and techniques, which are problem analysis, activity duration analysis,[135–137] activity based costing,[138–140] outcome analyis,[141–143] and technology analysis.

Systems analysis strategies[144] guided the research activities and investigation techniques to uncover valuable information that address the research questions mentioned above. Such strategies include: Interview questions and term definitions included in Appendix B were attached in the invitation emails, which was sent out to each participants individually. For those who give positive reply to participant in the study, more emails were exchanged to ensure an interview time. For those who do not reply, a further en-

couraged email was sent out, till they give a either positive or negative reply.

Each interview involved one participants. Participation in the study was voluntary and involved no incentives. The interviews were conducted over the phone, recorded, and transcribed verbatim to ensure accuracy. On average, each interview lasted for an hour. Participants were asked to figure out the status of data quality in MMIS, describe the data quality management activities they apply routinely, and talk about challenges in data quality improvement. Each participants answered questions related to their own experiences and give cases they have been encountered with.

This study followed the Framework[129–131] method, which has been used in many areas,[145–148] including health research. Data analysis was an ongoing and iterative process; this helped inform the researcher and better align the research activities. The framework method was used to analyze the evidence. Researchers[129] describe the framework method as "a matrix based method for analyzing qualitative data". It facilitates data management such that all the stages involved in the analytic hierarchy can be conducted. This method involved constructing a conceptual framework from which data was tagged accordingly. After tagging data elements with the appropriate tags or concepts, thematic charts were created to sort and synthesize the data. Categorization of the data elements followed where data is represented at a higher abstraction level. Typologies were created in which data elements or constructs fall into mutually exclusive sections. It is expected that as a result a central chart of data was generated and from which clusters and their associations may be unveiled leading to explanatory accounts and explanations.

The analysis adopted inductive approach[149] to explore perspectives of data quality improvements in the healthcare organization as there has been no study about it before. The Framework method summarize data during charting, which means, researchers from different expert domain can be engaged to provide their perspectives during analysis. What's more, summarized results are kept in a broader range of context, thus, enables a wider and deeper thinking of meaning and understanding. Besides, Framework methods

are flexible and easy to identify relevant concepts extracted from data. Finally, there is a clear flow in Framework method to convert original raw data to a well-structured and flexible final concept.

To ensure validity and relevance of the evidence, the following measures were incorporated into the research activities.[150] Respondent validation or member-checks was adopted better validation of data-collected and exclude researchers predispositions. Two researchers were analyzing the qualitative data to ensure mutual understanding from both the researcher and the participants.

4.1.2 Process Design and Tool Development

The design of process adopted custom development,[144] an appropriate strategy to respond to specific and unique requirements in novel systems, for this problem in a sustinct domain. Adopting process and supporting tools from other domains such as finance may not respond to the unique requirements in our research. Custom development provides a flexible development environment to create a system and solve problems, which maximized the requirement completion collected in previous phase. The custom development enabled a collaborative process for data quality improvement within the health organization, which facilitated the communication between different units and work efficiency. In addition, the custom development also would be more convenient to change the components and functions in the process of refinement.

The design included workflow design, program design, data storage design, architecture design, and user interface design. The workflow design is to identify the main functions and roles of usecases based on the requirement collection from interviews. Program design is the process moving from logical to physical implementation, accomplished by modeling the structure chart and exploring the programming specifications, and programming. The data storage design consists the design of data structure, determining the data flow within the system, and optimizing the data storage strategies. Architecture

design is the process to determine which architecture should be applied to the process, and apply the architecture development from programming design phase. Finally, the user interface design consists use scenario development, interface structure design, interface design prototyping, and interface evaluation. At the end, the process ended in a tool that improve the data quality in an efficient and effective manner that can be used by both command line and user interface.

This step resulted in a fully-functional software solution, DQT, which is a multi-user client-server software solution to support organizational process. In the development of DQT, special attention was devoted to detecting defects in large data sets in a fast manner by developing multi-threaded programs. The development involved the use of Tool Command Language (TCL), C programming to write TCL extensions when appropriate, and the SQLITE3 database. All of these tools are acknowledged to be useful tools in developing programs optimized for space and time complexity.

4.2 Results

4.2.1 Challenges and Opportunities of Data Quality Improvement

Analytical Hierarchy of Concept

The result identified from this stage was shown in table 4.1. There are five main concepts: demographics, problem attribute, current solution, challenge, and opportunity. Under those concepts subcategories are also listed.

Thematic Chart Summary

a. Demographics

Roles of participants include Health Policy Analyst, Medicare Program Specialist, Data Policy Supervisor, Claim Adjuster, and Compliance Policy Analyst. The **department** they

Table 4.1: Challenges and Opportunities of Data Quality Improvement in Healthcare Organization

Main Concept	Subconcept	Detailed Description
Demographics	Role	The role that participant acts in data quality improvement
	Department	The department that participant works in
	Work	The daily work that participant does for data quality improvement
Problem Attribute	Multiple defect types	Types of data quality problem that participant faced
	Data quality problem is severe	The severe degree of data quality problem that participant faced
	Disclosure	How did participant realized the problem of data quality
	Improve data quality is important	How does participant think the importance of data quality
	Bad consequence	The consequence occurred because of bad data quality in the organization
Current Solution	Manual check	Checking if record in system matches with paper based original record
	Adhoc approaches	Adopting adhoc approaches such as MS Access, Excel, or SQL
	Dictionary matrix	Checking if data violates the constraints in dictionary matrix
	Request form	Correct data by requirements listed on requests form from other group
Challenge	Difficulty of communication	The communications between different systems or groups are difficult
	Legacy system	MMIS is lack of advanced features to prevent and check defective entry
	Lack of documentation	Documentation of data and quality of data is not sufficient
	Different data standard	Difficult to unify different data standards among all multiple systems
	Weak data accessibility	Difficult to obtain data pieces other groups or systems
Opportunity	Data catalog	A data catalog to describe each element in data to facilitate understanding
	Accessible system	A system to make all data pieces accessible for all person in the organization

work for also varies from policy compliance division, behavioral health division, office of health services, and long term services and supports administration. Their daily **works** are checking claim payments and Provider enrollment, reporting analysis, checking data consistency within and across systems, long-term services management, research of potential frauds. As one participant noted, *"I work in the policy compliance division under health IT in the business intelligence data governance."*

b. Problem Attribute

Participants encountered with different **types of defects** in their dataset, it can be information mismatch, or typo in the system: *"For example, provider's name is typed wrong, or their billing address is wrong."* All of the types they described can fall in with one subcategory in the taxonomy of data defect.

As for **degree** of data quality problem they faces, most of participants agree that data quality problem they deal with is major because it requires communication with other people and other system: *"Sometimes it's pretty major and requires you know system update or review of the communication between MMIS and external system."* Only one of them think it is minor and can be corrected by simple fixes. The difference occurs because they adopt different approaches to improve quality of data as they face different problems.

The **disclosures** happened when they use data for reporting or claiming a payment: *"I realize there are problems when I actually need to use the data to answer some of questions I need to answer."*, look at one specific record, or interact with external people such as providers. All interviewees hold the opinion that data quality improvement is **very important** to their daily work as bad data quality will result in bad **consequences**, which can be problem in decision making about public health, fraudulent activities, wrong payment activities, sluggish in daily operations, or mis-communication. One interviewee stated: *"The quality of data is not there. And that will affect billing. It will affect claims. It will affect a mess."*

c. Current Solution

Some participants improve data quality by **adhoc approaches**: *"We are just deciding should*

we use excel, should we use access. Sometime MMIS system is the most like accessible." while others just **check manually** or process data according to a **request form**: "... what I do is just like people send me the request. And our team processes this request. And we check this request again with guidelines."

d. Challenge

Communication is the largest challenge for data quality improvement. A participant mentioned: *"You know I commonly ask for reports that I believe are clear but when I get the data it doesn't answer my question."* The next challenge is **documentation** of data pieces: *"If we don't have good documentation, we may end up denying someone falsely."* Some participants think the lack of **knowledge** is also the challenge: *"Usually the key error is lack of understanding or knowledge of why it's important, what fields are important depending upon which provider criteria."*

Different **data standards** cause difficulties for them to improve quality of data. One participant stated: *"... the federal and the state also have their own requirements which don't exactly match."* Some of them also complained the **accessibility of data** pieces that can facilitate data quality improvement is limited. One of them noted: *"... sometimes I just don't really capture the full picture of you know what's going on beyond this building, all you have is the data that you have on your side."* The **legacy system**, which is MMIS, they operates for data management, introduced more bad data and prevents data quality improvement efficiently: *"So if we can not completely upgrade the system, and the current system cannot take the improvement."*

e. Opportunity

A necessity of **data catalog** for data quality improvement raised from the interview: *"And I wish there would be a more prompt thing like how would this report be used and what do you want the end product to look like something like that, or what do you expect to the result to look like..."* The catalog should maintain information about the description and usage of each data element, such as a table or a column, with which user can understand the correlation

and meaning of data better.

Participants also mentioned an **accessible system** will also make data quality improvement more efficiently and effectively. One interviewee stated: *"... they receive updates from us once a week. Sometimes it will more beneficial if they can receive it everyday, but you know there are a lot of patients, sort of like that."* The accessibility to other pieces of data can enhance the verification of data defects and improve the accuracy of data defect corrections.

Typologies

Typologies were resulted from different type of work of data quality improvement people dealing with. People dealing with single data item focus more on details of cells in records, they manually check and correct data for quality improvement when there is a requirement of checking or correction. They claimed resources for them are limited and it prevents them to improve data quality efficiently.

People checking validity of a group of data adopt ad-hoc approaches with business rule matrix and data change request form to correct data, they encountered with problems of training people to understand these resources as the policy of data always changes, comprehending one material requires pre-acknowledgment of other related materials.

Healthcare professionals who are interacting with external systems care more about communication problem and data standard. Communication with external people outside the system is in low-efficiency and time consuming. Different data standards in different systems introduced more potential data defects in MMIS as well as increased the difficulty of data quality improvement.

4.2.2 Data Quality Toolkit Development

The development of DQT started with designing collaborative activities for the healthcare organization considering about the challenges and opportunities analyzed from in-

terviews. After that, DQT was developed with certain features to achieve operation of high level business process in a collaborative manner. The main features include business rule management, data defect detection, versioning control, collaborative communications, and auto updates. The fully functioning features supports DQT detects and manages data quality improvement initiatives in a efficiently and effectively, users are able to maintain a data quality with a better communication with it.

Collaborative Process Design

Processes were designed along the guidelines of data governance by identifying roles that will perform actions and processes that will help improve data quality issues and customize business rules while transmitting information between officials. Four operational roles were identified as part of data governance process for DQT in healthcare organizations namely health expert, data steward, data custodian and data operators. These roles and their processes unanimously lead to a collaborative contribution towards data quality improvement. The scenario of each role can be found in Appendix C. Figure 4.1 demonstrates both internal and external entities interact with DQT.

In DQT, a health expert can evaluate organizational policies and start creating business rules by creating a narrative as per requirement. Officials who have knowledge about organizational data implementation from business as well as more granular level point of view serve the role of data steward. A data steward can create structured business rules under the narratives that have been created, they can also import data and detect defects using DQT functionalities, prioritize defects for correction, assign tasks for data correction by creating tickets, compare versions of old and new data when corrected data is received from MMIS. Data custodians are the core technical personnels or programmers who assign queries or regular expressions to business rules in the form of creating formal specifications. These specifications add onto the business rule narrative and structured rule definitions and are used to scrutinize if the data abides by these organizational con-

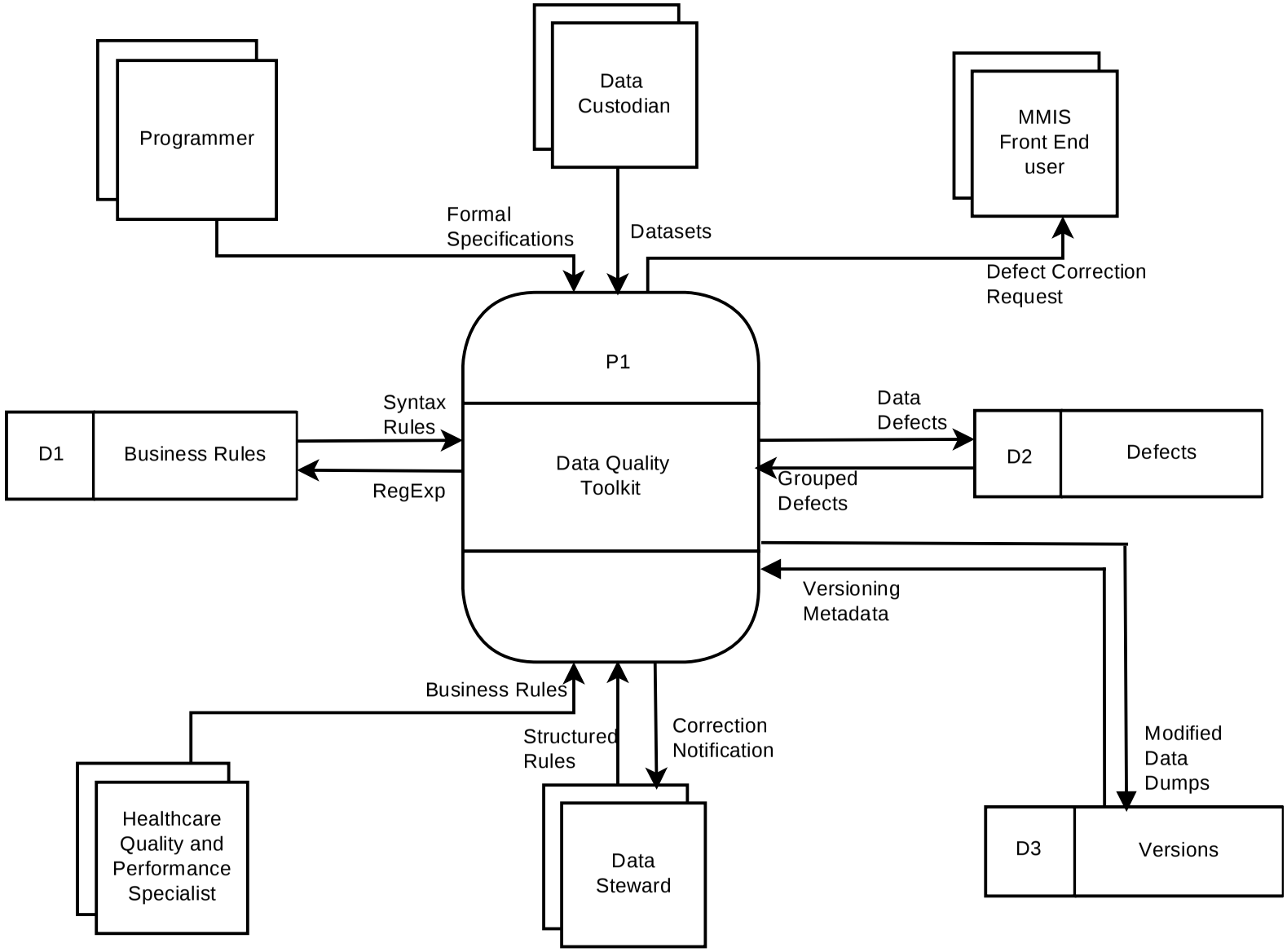

Figure 4.1: Context Diagram of DQT

straints. Data operators are MMIS front-end users who operates the system, the metadata about defects detected is sent to them for revision. Concepts like defect detection, business rules components such as narrative, structured rules, formal specifications will be explained in section **??**. The high level business processes are presented in figure 4.2, the cooperation and interactions between different roles are also shown.

Data Quality Toolkit Development

a. Business Rule Management

Business rule management is a major functionality of DQT. It enables users to edit the business rules associated to data fields. With the frequent changes in business policies, business rule management turned out to be immensely useful in changing business rules for the associated data fields. The business rule management functionality was designed in such a manner that, a user who is operating on DQT can be considered as unaware of technical or programming knowledge. The business rules in DQT are designed in three levels of management: business rule narratives, structured business rules and formal specifications.

The first level is narrative, a business rule initiative. Narratives are created as a first step for users to enter more textual descriptions about a business rule. Narratives are more friendly for officials who are more familiar with heathcare policies and have not much technical expertise.

The second level are structured rules, which are created by users who have more intricate knowledge about the rule's operations to supply information such as associated file and column names, column value, etc. The concept of structured rules was created to allow officials who have more knowledge about the implications and the data associated with it, in this manner, the narrative in the first level can be interpreted to a specific data table and data column.

Following the chain of narrative and structured rule, the third level of formal speci-

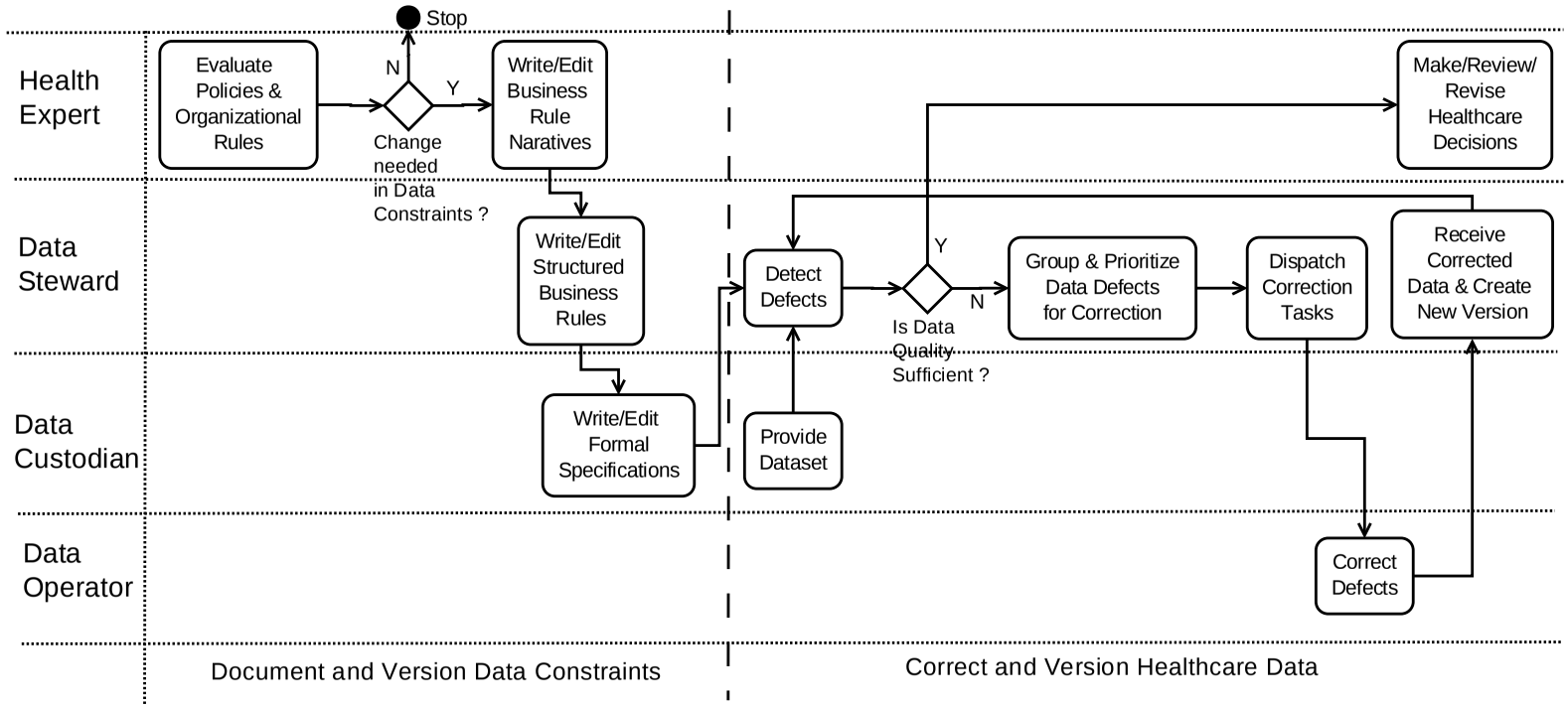

Figure 4.2: DQT Process Design

fications are created. It holds deeper technical information and specifications about the business rules such as SQL queries or regular expressions. Formal Specifications are entered by officials who have programming or technical expertise to add more definition to the business rule by tying queries or regular expressions to it. Formal specifications as third level of business rule establishes a link between the healthcare data and business rules created to govern the data. Information entered in the third level is directly used to check the compliance of the data to the organizational policies.

The user interface catering to business rules customization follows CRUD paradigm. CRUD in programming technology describes the elemental functions of create, read, update, or delete. The third programming level specification rule can be directly selected and utilized for data defect detection. With business rule management, a contextual description can be interpreted as a part of executable programming code and executed to detect data defects.

b. Data Defect Detection

The data governance processes enables users to improve the quality of the data used in healthcare organizations by importing data, performing processes to detect anomalies in the data, grouping detected data defects, and assigning them to other users to correct it. Data defect detection component of DQT automatically and efficiently identifies anomalies and discrepancies in data. An user is able to examine the quality of data files start with importing them into the application as part of a quality improvement initiative. Metadata about data defects such as location, value, and measurement is obtained by detection from selecting specific column from a table and applying a specific business rule to it. Defects detected are recorded and the required personnels are notified by using the ticket system. Finally the files are dispatched to the MMIS users for correction.

The data files imported into DQT can large, importing and performing defect detection processes on large data sets are time consuming. DQT hosts a splitting operation which breaks up the data file into smaller pages and performs parallel data processing

for defect detection. This function reduces the time required to obtain the files and operate on them. A grouping functionality in DQT categorizes data defects by four strategies of identified defects, allowing users to return to the system anytime and check all the defects that have been found by selecting the required strategy. DQT is efficient in detecting defects in large data sets promptly by developing multi-threaded programs, which is the mechanism of a central processing unit (CPU) to execute multiple processes or threads concurrently appropriately supported by the operating system. Multi-threaded programs lead to faster overall execution. By splitting data into pages and process data pages in threads, the performance can increase the speed by 47 times. DQT optimizes the users experience using these above features.

c. Collaborative Communication

The ticketing feature allows users to assign data correction task to officials. The person who detects the data defects, groups those defects and creates a ticket for the correction of those defects. Tickets can be assigned to certain assignees that users assigned, with an optional attached files. The status of a ticket can be new, in progress or completed. Once the assignees receive the email about the ticket, they can start working on data correction task and accordingly update the ticket status. Users also receives emails when the business rule they worked on get edited. All the users involved in a ticket receives an email whenever there is any change or update made to the ticket.

d. Version Control

DQT maintains a versioning system for both business rules and data which helps track their change over the time and maintain records of all the history versions. Any change in the business rules creates a new version of the rule which over the time helps track the changes that have occurred in the healthcare organizational policies. Uses are able to check drive any version of a business rule and check how did it changed over the period of time. The versioning control for business rule also enabled users who need to use different queries when they have different points of view, which resolve the problem of

conflict of use and interpretation of business rules.

In case of data versioning, DQT maintains only the current version of data as that reduces the demand for space. A diffing mechanism was adopted in DQT to compare the existing data with new version of data that is imported into the system. The new data dump does not replace the existing data if there is no change. After importing and only the latest data is kept, user can regenerate older versions of data and compare them between each other. Users can visualize the difference of any two versions of data by viewing two versions of data side-by-side, with differences marked by colors. Information about version of data is maintained in the database where the business rules version is also maintained.

4.3 Limitation

This research achieved its objectives by developing and enhancing DQT with collaborative activities of data quality improvement and business rule customization features. However, there are still several limitations in this research. For requirement collection, qualitative research is not immediate generalizable compared with quantitative research. But this research leveraged interviews for a deep investigation of data quality issues and solutions in real-life settings in health organization, thus alleviated the problem of qualitative research.

For DQT development, GUI needs to be improved and mechanism of business rule inputting can be simplified. Main functions are listed in menu bar on main working window, in which some of more important and more frequent used functions can be presented in a tool bar. However, the usages of each function needs to be collected from software deployment phase, and the development of tool bar should based on the usage in deployment. Management of business rules according to users based on their knowledge of health data policies, data structure, and programming language, which increased

the dependency of data quality checking on users.

4.4 Conclusion

It is a big concern that the update of information management systems falls behind of the changes of policies and lows at the state and federal levels. Data maintained in such system does not reach the expected level of quality. Efficient communication and documentation of data elements are demanded in healthcare organizations to facilitate functions in daily operation. Smooth automated high level business processes of data quality improvement in healthcare organizations is needed. The interview investigated the topic of data quality improvement in a in-practice fashion with collaboration with a state agency, which has not been done by in other data quality research. Health organizations will benefit from collaborative activities with fully functional multi-user client-server software developed for data quality improvement.

It is striking that, the lack of data quality is considerably affecting the daily operations such as dealing with rejected reimbursement claims detecting fraud, waste, and abuse, and making payments to vendors. The currently employed practices fall short in effectively and efficiently improving the quality of data. There is a clear need for systematic approaches and accompanying tools to address the challenges and opportunities identified in interview.

This research adopted qualitative methods and various software development activities to collect requirements and enhance DQT. First, the architecture of DQT was designed to define a collection of hardware and software components, and interfaces to implement the framework. Next, features such as operation design, business rule customization, data defect detection, ticketing system, and data version management are determined, constructed, and developed. Finally, the an advanced version of DQT was developed as a multi-user concurrent software with client-server mode. Users can work on and share a

global business rule repository for tracking of business rule change and detection of data defect.

Healthcare organizations will benefit from the collaborative processes developed for data quality improvement. Data quality checking, improving, and communication between multiple units can be tracked, managed, and maintained in an effective automated activities. Different roles in the healthcare organization will have a better understanding of each function in processes. A fully functional multi-user client-server software solution developed to support the associated activities will be a highly unique and novel contribution. Consequently, the data quality improvement initiatives will achieve higher degrees of effectiveness and efficiency.

Chapter 5

Aim 3: Implementation, Assessment, and Refinement

The refinement of the data quality improvement approach realized through its software prototype, DQT, was achieved by implementing DQT in an educational setting. In this context, implementation refers to installing DQT and putting it into use by data quality improvement teams.

5.1 Methods

Both qualitative and quantitative data were collected to assess whether DQT provides advantages; this assessment provided a basis for further refinements. At this stage, data quality reports, software usage logs, and detailed information of errors given by the software during its execution were collected during the implementation phase. In addition, the users reported their opinions and various issues they observed during the implementation in the form of tickets collected in a web-based repository made available to them. After that, focus groups were conducted to collect opinions and suggestions about the perceived usability and usefulness of the tool. Based on the results, DQT was refined by fixing its bugs and adding new features.

5.1.1 Assessment Settings

Due to its feasibility, the implementation of DQT took place in an educational setting. Student participants conducted month-long data-quality improvement projects within groups, each working on a different simulated homecare readmission dataset.

Participant Sampling

Convenience sampling was used to recruit 24 participants, who were among the graduate students of the Department of Information Systems. In terms of their general background and knowledge, these graduate students resemble the professionals who may be assigned to work on data quality improvement problems in the industrial settings. Therefore, their recruitment as participants in this study was appropriate.

Demographic data was collected from the participants through an online survey which aided the formation of homogeneous groups with the maximum within-group diversity according to a set of criteria that can affect their effectiveness and efficiency in evaluating DQT.[151] The criteria included background in statistics, programming language, databases, prior usability courses, GPA, and course load in the semester.

Data Simulation and Defect Injection

The student groups worked on the simulated datasets with defects injected in them. For simulation, based on the metadata of a real homecare re-admission dataset, 33 features were selected to be the columns. The column names and descriptions are provided in Appendix D. The researchers made assumptions about the distributional characteristics of the variables represented in the columns such as age. In addition, a correlation matrix with 33 dimensions was created to impose certain relationships between columns. These distributions and relationships were predetermined by the researchers based on the approximations that come from their experiences. This research does not include analyzing the data sets, therefore it is not affected by the assumptions made about the distributional

Table 5.1: Number of Records and Defects in Each Readmission Data Set

Group	Number of Records	Number of Unique Defects
A	99,539	8,201
B	99,615	7,980
C	99,546	9,865
D	99,592	6,849
Total	398,292	32,895

characteristics of the variables or the relationships among them. However, employing these predetermined assumptions made the data sets very similar to actual data sets. The simstudy package[152] in R was used for generating the simulated datasets based on the assumptions.

Next, data defects of different types were injected into dataset by randomly selecting cells. The defect types were missing data, dummy values, implausible values, invalid values, type mismatch, outside domain range, and duplicity. Overall, for each data sets, around 10 percent of the observations were injected with defects. There have been no published benchmarks or standards about the number of defects in real data sets. Therefore, 10 percent was used as reasonable assumption about defect density for this type of data. In our experience, as shown in Table 3.3, the defect density in actual healthcare administration data sets can actually be much higher. Table 5.1 represents the number of records created and number of unique defects injected foreach readmission data set.

5.1.2 Assessment of Perceivable Advantages

Perceived relative advantages of using DQT and its perceived complexity, the two factors that will play an important role in its diffusion,[153] were assessed throughout its implementation by conducting qualitative research to obtain rich contextual opinions, which would be infeasible to obtain by using a quantitative approach. Knowledge workers' intention to use new ideas and software solutions increase as they see higher perceived relative advantage and decrease as they see higher levels of complexity. For this purpose,

five focus groups were performed with each one of the focus groups which used DQT for data quality improvement for one month.

Adopting focus groups[154] was useful because it provides consolidated opinions and help better flash out answers compared with other qualitative research methods. It can help discover the differences of opinions among the individuals, and gather detailed information about both personal and group feelings and perceptions. Focus group could also provide a broader range of information from each participants while offering the opportunity to clarify the results. Conducting one-to-one interviews with the cohort of 24 available subjects would take longer, and the group conclusion would be missing. Focus groups allowed researchers to elicit the opinions which were analyzed and categorized as improvement requirements in the process of DQT refinement.

Focus Group Design

Each of the five focus groups lasted for approximately two and a half hours with the first half of discussing usefulness to understand the perceived relative advantages of DQT, and the second half discussing the usability issues to understand the perceived complexity of DQT. The facilitator conducted the discussion according to a facilitator guide, provided in Appendix E. Two group discussions were audio-recorded, and the main discussion topics were written on the board. The group facilitator also wrote down main concepts of each topic on a board. The audio recordings were transcribed and the transcripts were analyzed along with the notes. The focus group questions on perceived usefulness and usability of can be seen in Appendix F.

Data Analysis for Focus Group Data

The data analysis methods followed the the Framework method[129–131] which was discussed in Section 4.1.1, as it is a more flexible approach to identify the ideas and findings among context, and enables different perspectives and understanding to be merged to-

gether, which helps researcher make a in-depth investigation about the usability and usefulness about DQT. This method has been used by numerous researchers in the analysis of qualitative data collected via focus groups. The inductive approach guided new viewpoints of DQT to be discovered, which confirmed the comprehensiveness of the result and cognition of the usage of DQT.

Triangulation

To ensure the validity and relevance of the evidence, triangulation through two other source of data was also collected and analyzed, which are issues reported in ticketing repository and direct opinions.

An online ticketing system was made available to users for on going reporting of problems and tracking them to resolution. The ticketing system enables user to publish issues encountered while using DQT. Through this online ticketing system beyond DQT, researcher can get issue reports on time and give resolution in a timely manner. Messages were exchanged between user and researcher in the ticketing system, which promote the communication on the issues of DQT. By collecting data of issues posted in ticketing system, the researcher can evaluate whether DQT performs as expected.[155,156] In addition, the full documentation in ticketing system engaged DQT with more features that was not designed at the beginning, which fulfilled the inadequacy of functions in DQT.

Other than focus group and ticketing system, the participants also provided direct opinions in a group report. The contextual data from focus group, ticketing system, and direct opinions were leveraged to understand perceivable advantage of DQT.

5.1.3 Usage Characteristics

Usage characteristics of software systems need to be understood and improved in order to provide a better user experience.

Execution Profile

We implemented features to create a detailed runtime history by recording each operation performed by the users of DQT. As a result, the relative usage of the specific operations and functions was studied. By analyzing data from logs, the use frequency and use preferences of each function was identified which explained how DQT was utilized.[157–159] Those logs helped in prioritizing the refinement efforts. Functions used most frequently should be paid more attention for improvement and optimization.

Error Logs

The logs recorded detailed information of each error happened during execution, making it possible to perform further refinement and debugging more easily. Errors that occurred most frequently would be resolved first after implementation.

Defect Validation

Data quality reports were collected from participants. Data quality report includes location of each defective cell participants detected, which is also the unique row number of each defective cell in each column. There was one report collected from each group, as one report from one dataset. The result of report are consent from within the group. If there was any different points of view among group members, they would reach a consensus by discussions. As a result, comparing injected cells with detected cells, the usefulness of data quality improvement can be explored and investigated to evaluate the benefits of the approach.

5.2 Results and Discussions

5.2.1 Assessment of Perceivable Advantages

In total about ten hours of recordings were saved, 95 tickets were created in ticket repository, and direct opinions from 24 participants were assembled for the qualitative analysis, four themes with relative subthemes were emerged as the result of analysis.

Analytical Hierarchy of Concept

Hierarchy of concept is shown in table 5.2. There are four main concepts: usefulness, usability, issues, and suggestions. Under those concepts subthemes are also listed.

Thematic Chart Summary

Usefulness

For **rule management**, some user think it is is very useful and smooth because it provides the ability for non-technical users operate on DQT: *"The speration between narratives, rules and specifications makes working with the system easier as both non-technical and technical people can participate "*. What's more, business rule management make tracking operation history possible. But most of them stated structured rule and specification can be merged: *"I think rule and specification can be merged because they serve the same purpose"*. That is because all participants are from technical background, they know both data structure and programmings, so they hold the opinion that user can create a specification from a descriptive narrative. However, in the environment of healthcare organizations with less technicians, it is not the case.

Some participants noted **detection** provides accurate and consistent result about the location of each defect, it is easy to operate. The grouping function can merge similar type of defects together, which makes cleansing easier. One participant said *"It is very useful that the defects can be grouped according to taxonomy"*. But sometimes it shows empty result

Table 5.2: Perceivable Advantages and Required Improvements of DQT

Main Concept	Subconcept	Detailed Description
Usefulness	Rule is useful and friendly to non-technical users	Usefulness business rule management and three-layer concept of business rule
	Detection is useful	Usefulness of data defect detection feature
	Collaborative operation can be improved	Usefulness of the collaborative operation features
	Versioning is helpful for tracking history	Usefulness of version management of data and business rule
Usability	Good performance	Performance of DQT such as GUI operation data processing
	Accurate result	Accuracy of the result DQT provides
	Efficient	Efficiency of data quality improvement and usage of DQT
	Easy to use after learning	Ease of use of DQT
	Has advantages compared with other tools	Differences of DQT and other tools for data quality improvement
Issue	Complex navigation	Navigation of features and functionalities
	No result exporting	Exporting results to a file
	Layout needs redesign	Aesthetics and widget layout of DQT
	Terms are hard to understand	Terminologies in DQT is difficult for technical users
	Only ID shown in list	In list program only show ID of users or rules
Suggestion	Needs proper notification	Showing proper notification to user if there is any mistake in oepration
	Better tutorial	Providing an easy to learn tutorial
	GUI needs improvement	GUI needs improvement to make DQT not out-of-date
	Function for data dependency	Data dependency is required to be presented and managed
	Function for data cleansing	A feature of data cleansing in DQT can facilitate data quality improvement

window instead of notifying user "no defect detected".

Participants all agree that the **collaborative operation** make everyone see each others query and utilize that, which facilitate the communication and learning from each other *"One rule created for one project, can be reused for other projects also"*. However, some of them claimed the ticketing system should be improved to be more user friendly as they said *"The current ticketing system does not facilitate the tracking of task update"*. Some participants even stated ticketing system is useless so they adopted other tools for communication. They also have the same opinion that **versioning** of data and business rule is very useful to check mistake in business rule, track history of business rule, and check difference between two versions of data. One participant stated *"Versioning allows to work with same rules and allows to tarck previous versions"*

Usability

One participant said the threading improved the **performance** of data defect detection on large dataset. But some others said the performance is decreased because multi-task on DQT is not enabled: *"...no multi-tasking is possible, like one window needs to be closed before another can be opened"*. Some participants think the **accuracy** of DQT is good in terms of easy query was executed: *"Defect result is accurate when query is not complex"*, while others noted the results need to be double checked with Excel.[160]

Features such as business rule management and GUI operation improve the **efficiency**. And DQT is less time consuming compared with other tools. Participants state that DQT is difficult to use at first, but **easy in operation** after users learn it. A participant noted *"It was difficult at first, but easy to understand after you learn it"*. As there is no programming required, it is also friendly to non-technical users: *"...provide skeleton for less technical background"*.

Participants would like to use Python,[161] R, SQL,[162] SAS,[163] and Excel if there is no tool like DQT provided. **Comparing** with these tools, DQT provides a more intuitive GUI for user other than programming languages. DQT is more non-technical friendly, provides

more features to track history and progress of data quality improvement. Stated by one participant, *"...no way of version comparison in other tools... no collaboration operations among technical people and non-technical people"*. Although other tools provide more functionality than DQT, it requires user to coding and keep the record of codes by themselves.

Issue

Almost all participants claimed **navigation** of DQT needs to be improved as there are too many windows pop up for each feature: *"...taking too many steps to creating narrative, rule, and specification"*. And in result window, user was not able to **export result** into a file. They have to copy and paste results to a file created by themselves, when result is huge, it takes a lot of space on memory and the program will be suspended. Participants said *"When I tried to copy a large amount of defect detection result using ctrl+a, ctrl+c, and ctrl+v, I found that DQT halts. I also tried shift+click to copy some portion of the result but I couldn't select results in that way. So I think it would be convenient if DQT has a function that allows a user to export defect detection results in a csv file*. One participant also noted the **layout** of DQT is confusing users: *In some cases the transition are not smooth, and buttons are not placed at consistent places*.

Some participants also stated the **terminologies** in DQT is confusing for technical users. DQT is designed for both technical users and non-technical users, some technical terminologies are transferred to some non-technical words. However, it becomes confusing for technical users: *"The name of indicator is the primary key in database, it will be more understandable and more clear for me if it is shown as primary key"*. Participants also noted in some lists, user have to remember ID of an item such as business rule or a user instead of **showing name** to user. A participant claimed *"In ticketing system the user name should be there instead of id"*.

Suggestion

In DQT, if there is any error happen, a generalized message will pop up to tell user an error happened. There is not no detailed information provided further to user to tell them

which operation caused the error. Participants were noted that more **proper notification** should be provided to user to reduce anxiety: *"...there were too many unexpected errors happened, like log in issue, session timeout and other errors"*. The help manual was helpful for several participants, but most of them claimed it was not sufficient. They suggested to have a more understandable **tutorial** such as a video tutorial for user to make them learn DQT faster. One participant said *"The help manual in DQT is useful, but is a tutorial video will be more helpful to teach new users"*.

Participants noted **GUI improvement** is very necessary: *"The UI is vary basic and out-to-date, it need enhancement because it looks like a software from ancient times"*. One participant said a **data dependency** feature would also enhance DQT for business rule and data management: *"There should be an approval by administrator before business rules are actually published"*.

Some participants expect to have a **data cleansing** feature in DQT. One of them noted *"I have to rely on others tools for cleaning after detection using DQT. If there is a data cleaning option it will be better"*. However, considering about the confidentiality and security of healthcare data, data should only be modified in the original system. Thus, DQT will not include this feature to cause and confusing and insecurity.

5.2.2 Usage Characteristics

Execution Profile

Figure 5.1 shows the number of access to each functionality in the implementation. The most accessed function is rule management, the first step of defect detection in DQT, three-layer of business rule needs to be edited to accomplish an executable specification which can be utilized by detection. It is accessed most because participants needed to learn and give tries to each rule to make it valid for checking. It determines rule management is the most important functionality to obtain accurate result of data defect detection.

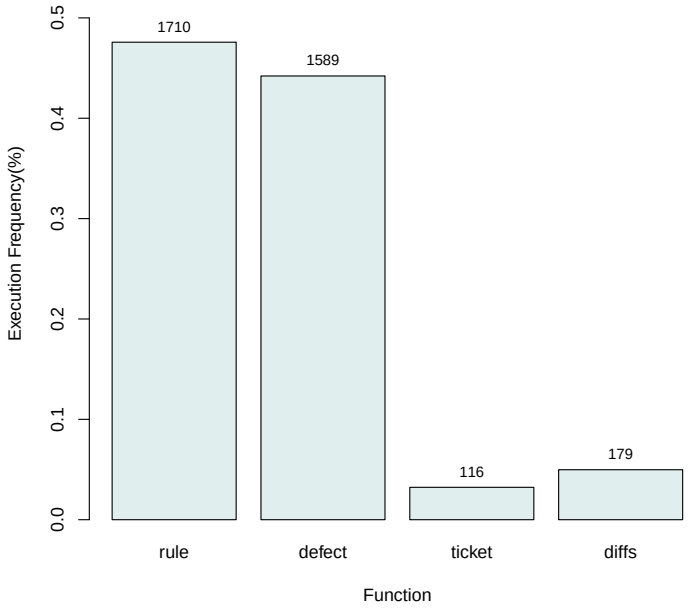

Figure 5.1: Execution Profile

Followed with rule management, data defect detection is the next function that been accessed most, only fall 7% behind rule management. Detection provides the core result in DQT for data quality improvement, which causes user access detection for more than 40% to detect defects and verify the validity of each specification. Rule management and data defect detection take 90% of usage of DQT, which reveals that these two function should be paid more attention to improve users satisfactory in terms of usability, performance, and accuracy.

Tickets and diffs are accessed less because features of these two are more simple, and users did not have too many requirements of cleaned dataset verification. As the conclusion of focus group, ticketing function provided a lower user experience, that also could be a reason of it is accessed least. Thus, user experience and more usability of diffs and ticketing should be provided and developed.

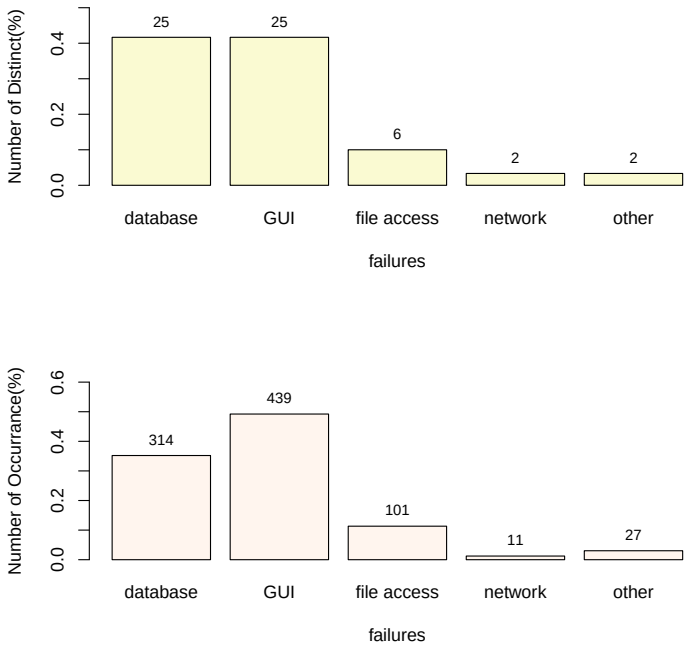

Figure 5.2: Failure Analysis

Error Logs

Figure 5.2 shows the number of distinct failures occurred and its number of occurrence during implementation. There are 60 different type of failures, database failure and GUI failure took 25 of each. Most types of database error are caused by empty entries, which can be prevented by bringing a friendly notification to user. Most of GUI failures are conflicts between windows that user opened. The six file access failures can also be prevented by notifying user about failure to access the file or missing of the file. Network failures and other failures can be caused by unexpected network problem and errors.

In terms of number of occurrence, GUI failure is more than database failures even though the number of failure types are same for them. The reason can be there are more GUI operation on DQT than database operation, and users will try to give more clicks on GUI to resolve the failure, however, in the contrast, it usually added more failures. Most occurred database failures is because user imported data files that are not accepted

Table 5.3: Percentage of Data Defect Detected by DQT

Group	Defective Cells Seeded	Defective Cells Detected	Detection Rate
A	8,201	7,311	89.15%
B	7,980	5,718	71.65%
C	9,865	8,631	87.49%
D	6,849	5,448	79.54%
Total	32,895	27,108	82.41%

by DQT, a notification can also be created for this type of failure to ask user change the format of imported data. There are 101 failures of file access, most of them are caused by wrong operation when user imported data. A more smooth process of data importing can be developed to improve.

Defect Validation

Table 5.3 shows the percentage of data defect detected by DQT. Group A, C, and D discovered and detected more than 79% of defects, while group B is below performance as they did not investigate enough business rules in their dataset. Since DQT is not built a predictive model, rather is built as a deterministic mechanism and rely on people making the right rules, the result highly depend on the decision and apprehending of defect of users. In total 75.83% of defects were detected by DQT, which indicates that DQT is performing in the desired way to provide accurate result for data defect detection efficiently.

5.2.3 Refinement

After implementation and analysis of data, bugs were fixed and four improvements were made on DQT.

Result Exporting

One exporting button was added on data defect result window, user can save results as a self-named csv file as tab separated dataset. Instead of copying and paste results to another file by taking memory space of computer, exporting results feature saves result in a better format as well as saving time by improving the performance and minimizing user efforts.

Data Dependency

Data dependency is a feature to show dependency of business rules and users who are operating on them. A tree structure is is showing to demonstrate the relations and hierarchy of business rules in terms of narrative, structured rule, and rule specification. Users who modified the rule is also shown to user to inform them about the dependency between users and rules. Besides, creation of each rule requires approval from data steward. Only rule approved by data steward can be utilized for data defect detection and visualized by other users. The data dependency make it possible for user to have an authenticate mechanism to manage business rule and prevent the repository of business rule get polluted.

Data Dictionary

Data dictionary is able to let user editing the description of each data element from domain, entity, and filed, which can also be interpreted as system, table, and column. By editing and viewing the description, constraint, and relationship of data elements, users can have a better understanding of data pieces and be more efficient on data quality improvement.

GUI Improvement

Based on suggestions and opinions collected, GUI of DQT was improved. Operation panel of business rule management was changed to a more easy to navigated format,

user can edit any layer of business rule on a panel by operating on a tree structure, which provides a visualization of relationship between business rules. Each function only pops up with one window, user switch by tabs for different features, instead of popping up with multiple windows. Widgets are settled on window in a more standard manner, which makes DQT more user friendly.

5.3 Limitations

The interview participants were all students from Department of Information Systems and Department of Computer Science, all of them know programming above average level, some of them are very skillful in regular expressions and SQL. It caused the problem of underestimating the usefulness of three-layer of business rule, some of them hold the opinion that directly writing program to check for data defects is better than operation on GUI, which will be a hard task for agencies in healthcare organization. However, writing programs to check defects is not able to track the history of business rule as policy of data changes.

Although in the simulated dataset defects were injected and kept in log, and description of each column are provided to participants at the beginning of implementation, participants usually misinterpret or overinterpret the relationship between columns. Some rules they created were should not be a constraint for data, and some constraints are undiscovered and defects were not found. It caused some bias in defects validation. But the result of defect validation still proves the DQT can serve the objective of improving quality of data efficiently.

5.4 Conclusion

Functions in DQT such as business rule management, data defect detection, collaborative work, and versioning are useful for efficient data quality improvement. The ticketing

system needs improvement with more simple and easy operations to facilitate communication between users. Overall DQT is easy to use after users learn how to use it. A better tutorial provided can improve user's motivation and reduce the time consuming in learning. DQT is especially friendly and easy to use for non-technical users. With GUI operation and contextual description of business rules, non-technical users are able work on DQT and contribute to data quality improvement. The GUI of DQT needs improvement to make it looks more professional other than only be professional in core functionality.

Compared with programming/scripting languages that can be used for data quality improvement such as Python, R, Excel, SQL, and SAS, DQT performs much faster due to its multicore custom programmed defect detection and data versioning algorithms. It allows users to create constraints in the form of business narratives, structured rules, and specifications to be applied on data. The versioning of data constraints can allow the users to track the history and changes of a business rule, while versioning of data can track the improvements on dataset. Although other tools provide more features in statistics, they requires programming, and the scripts cannot be tracked for previous version. On Excel, user can also operated on GUI to check defects, but the output is only the result, in which the process of detection operations are missing. DQT is also good at managing business rule and data quality in a collaborative manner, that other tools cannot provide.

Business rule management and defect detection are most accessed functions, more efforts should be put on these two functions to provide a better user experience and more accurate results. Over datasets in four groups, 72.12% of defects were detected by participants, therefore, DQT is testify to be a efficient tool to provide accurate data defect detection results.

Finally DQT was refined by bug fixing and four new features added. The data dependency feature enables users to check dependent relation between different level of business rules and users. It also make rule creation with user authentication, which pro-

tect business rule repository in a healthy environment. The data dictionary helps users to edit and comprehend the usage and restrict of each data element, largely expedite the understanding of data and improvement of data quality improvement. The result export function make it easier for user to save results and do second study on it as long as improve the performance of program. Improvement of GUI make navigation more logical and widget more coherent. A more professional look of DQT can motivate user to use the tool and conduct them process data intelligently.

Chapter 6

Conclusion

With the increasing importance and needs of data analysis in healthcare organizations, data quality is be basic requirement for useful and high-quality results. To invest the core problem of data quality, investigate the challenges and opportunities of data quality in healthcare organizations, and develop collaborate processed and DQT for data quality improvement, three studies were conducted in this research to provide a solution to improve data quality in healthcare organizations. The research was based and testified in real-life and educational settings, which confirmed that the processes and DQT developed are efficient and effective for data quality improvement in a collaborative environment.

By conducting documentation analysis, descriptive analysis, and literature review, a taxonomy of data defects with six main categories and seventeen subcategories was identified, which can help the understanding and categorization of root problems of data defects. Conditional required value missing and multiple attribute value duplicates in the taxonomy have not been mentioned and discovered in other study. Medicaid data in Provider and Procedure subsystems are detected with the taxonomy, in total more than two million data defects were detected, indicating that data quality in the MMIS system needs to be examined and improved. The most occurred defect type is syntax violation, which should be paid more attention to prevent in the future.

From the analysis of interviews, in healthcare organizations, lack of data quality is

troubling data managers with their daily work. All participants hold the opinion that data quality is very important and needs to be resolved in their organization as it can cause time and money lost. Poor data quality can caused by mis-communication, legacy system, lack of documentation, different data standards, and limit data accessibility. However, there is no systematic solution provided to them for collaborative data quality improvement. A data catalog and an accessible system to data would benefit the healthcare organization.

Based on the result of interview, DQT was developed to implement a efficient and effective tool for healthcare organizations for data quality improvement. Main features of DQT including business rule management, data defect detection, collaborative communication, and version control. The tool has GUI operation and process data in a fast manner by embedding threading functions.

Finally, DQT was implemented in an educational setting. Participants was selected to use DQT to improve quality of a dataset for one month. A focus group was conducted to investigate the perceivable advantages and objective assessment of DQT. Overall DQT serves the purpose of data quality improvement and provides accurate results. However, GUI and ticketing system needs to be improved for a better user experience and collaborative communication. Three features was developed to refine DQT: result exporting, data dependency, and data dictionary.

The overall research furnished in-depth studies of data quality improvement for healthcare organizations. With the data defect taxonomy and DQT, healthcare organizations are able to improve quality of data by identifying, checking, correction, and communication between different roles across the organizations. High quality of data obtained by using DQT will benefit healthcare organizations to prevent resource lost, ease and simplify the operation, improve the profits, and quality of healthcare results.

Appendix A

Interview Questions

1. A data defect refers to a discrepancy between the actual and expected correct values for a data item. To start with, could you give an example of a data defect from your context?

2. Defects detract from the quality of data. Higher number of data defects are associated with lower data quality. Through which mechanisms and how do you realize or know about data quality problems, i.e., high number of defects in your data?

3. Data quality improvement is identifying data defects and tracking them to their resolution to ensure that data is corrected. To what extent do you think data quality improvement is necessary in your opinion? Why? Would your organization need data quality improvement?

4. How do you currently perform data quality improvement? What are the approaches, steps, and tools?

5. What are the most important and recurring day-to-day problems associated with data quality improvement according to your experiences? What causes these problems?

6. What are the most time consuming tasks in data quality improvement? Why?

7. What are the most costly tasks in data quality improvement? Why?

8. What are the communication or collaboration difficulties in data quality improvement activities? Can you give examples?

9. In data quality improvement, what current activities can be possibly eliminated or possibly merged? Why would this be useful?

10. Are there other similar organizations you know with better data quality improvement mechanisms? If so, why are they better? What should your organization learn from them? How can such learning be facilitated?

Appendix B

Term Definition

1. **Data Defect:** Data defect refers to a discrepancy between the actual and expected correct values for a data item which requires a corrective change.

2. **Data Quality:** Defect detract from the quality of data. Higher number of data defects are associated with lower data quality.

3. **Data Quality Improvement:**Data quality improvement is identifying data defects and tracking them to their resolution to ensure that data is corrected.

4. **Collaborative Data Quality Improvement:** Collaborative data quality improvement is a practice of data quality improvement across functional and individual boundaries.

5. **Collaborative Data Quality management Toolkit:** A collaborative data quality management toolkit is a toolkit with full features of data quality management to improve quality of data.

6. **Business Rule:** Business rules are rules that define or constraint data and always resolve to either true or false, which are intended to assert data structure or value.

7. **Structured Business Rule:** Structured business rule is a rule that specify a certain constraint on data with format, structure, type or value.

8. **Business Rule Specifications:** Business rule specification is a command or section of command that specify and implement a business rule constraint on programming level.

9. **Data Versioning:** Data versioning indicates the action of saving new version and metadata of data when it is changed so it can be roll back and retrieve specific versions of data.

10. **Business Rule Versioning:** Business rule versioning means the action of saving new version and metadata of business rule when it is changed so it can be roll back and retrieve a specific version of business rule to be applied.

Appendix C

Scenarios of Roles

C.1 Health Quality and Performance Specialist

Your work involves carrying out policies and making decisions based on the analysis of healthcare data in this organization. One day, you receive a new policy document which defines a new series of codes for provider type. The change further divide "RX" which indicates to "pharmacy" into three new codes for provider types. To implement this policy change, you need to identify providers whose service type are pharmacy in the system, correct them with new codes, and generate a new disease distribution report. However, current codes have been applied for more than ten years, which produced millions of records that need to be corrected. This work will take huge effort because you could only search and check records one by one on the system with current features.

C.2 Data Steward

Your work involves making monthly payments to a Medicaid provider. One day, you discover overlaps among execution time of several procedures while looking over the monthly claims data submitted by the vendor. The problem spans back many months, and now, the leadership asks you to identify all suspicious claims for the past five years.

However, your investigation first reveals that the data accumulated is inconsistent and filled with invalid values. For example, the values in the column of start time for procedure are not only in different formats but many of them are plain wrong, some even include nonsensible values such as "headache". You do not know how to find all records with this specific inconsistency in the records, and even after you find it, you do not know how to correct them in a systematic way because a huge number of records exhibit in this problem.

C.3 Data Custodian

Your work involves describing standards for data in system, such as data value format of one value and affiliation relationships between two attributes. One day, you notice there is a bug in one of the standards, which should restrict provider base number in seven digits. But your description validate provider base number not less than seven digits. This description has been applied to system for several months but the constraint is wrong, which means provider base number with more than seven digits are also entered in system. You want to correct it but you are afraid that the modification would breakdown the system, and the system could not roll back to the current status neither.

C.4 Data Operator

Your work involves entering healthcare claims into an electronic health system. One day, your leadership who makes decision from analysis of data in this system tells you that he could not extract active provider after 01/01/2015 because provider service begin date is empty in most of records. You know you had never put service begin date and service end date in system because they were not used in the previous operations or data analysis. This practice has been in place for many years, and the other colleagues also follow the same practice. Now, you need to identify the records without this value and correct them.

Appendix D

Simulated Data Feature Descriptions

1. age
2. gender
    - male
    - female
3. race
    - African Americans (AFA)
    - European Americans (EUA)
    - Hispanic and Latino Americans (HLA)
    - Asian Americans (ASA)
    - Other (OTH)
4. marital status
    - single
    - married/partner
    - divorced/separated
    - widowed
    - other/unknown
5. area
    - rural
    - urban
6. income

7. education
    - high school
    - undergrad
    - graduated
    - other
8. mortality score
9. patient got better at walking or moving around
10. patient got better at getting in and out of bed
11. patient got better at bathing
12. patient had less pain when moving around
13. patient's breathing improved
14. patient's wounds improved or healed after an operation
15. patient got better at taking their drugs correctly by mouth
16. adverse events
    - adverse events (AEs)
    - adverse reactions (ARs)
    - serious adverse events (SAEs)
    - suspected serious adverse reactions (SSARs)
    - suspected unexpected serious adverse reactions (SUSARs)
17. drug users
18. alcohol users
19. the home health team began their patient's care in a timely manner
20. the home health team taught patient (or their family caregivers) about their drugs
21. the home health team checked patient's risk of falling
22. the home health team checked patient for depression
23. the home health team made sure that their patient has received a flu shot for the current flu season
24. the home health team made sure that their patient has received a pneumococcal vaccine (pneumonia shot)

25. for patient with diabetes, the home health team got doctor's orders, gave foot care, and taught patient about foot care
26. the home health team gave care in a professional way9. the home health team communicated with patient well
27. home health team discussed medicines, pain, and home safety with patient
28. home health patient had to be admitted to the hospital
29. patient receiving home health care needed any urgent, unplanned care in the hospital emergency room - without being admitted to the hospital
30. home health patient, who have had a recent hospital stay, had to be re-admitted to the hospital
31. home health patient, who have had a recent hospital stay, received care in the hospital emergency room without being re-admitted to the hospital
32. length of stay

Appendix E

Focus Group Facilitator Guide

Welcome

Hello to everyone, welcome and thank you again for agreeing to be part of our discussion today on Data Quality Toolkit (DQT).

Introductions

First of all, let me introduce our team here today: I'm (name of facilitator), the facilitator of this group.

We are interested in hearing about DQT. The focus will always be on usefulness and usability. We will divide the discussion into 2 sections.

You've been invited here today to give your opinion as a person who may have multiple roles in data quality improvement process within a group. We know you may have used and experienced all features of DQT in your group for this class, but today we would like your views as a person who works in a health care organization and may not have too much knowledge about technical or programming, rather than a master student from department of CS or IS who is skillful in SQL query and other programming language. We value your opinion and want you to know that we hope to use the information to learn more about needs of refinement in DQT regards to data quality improvement.

Ground rules

Before we begin, let me mention a few things about how we usually conduct these discussions:

1. I will be the facilitator for the group. My role is to ask the questions we have for the group, and to encourage everyone to participate. I won't be doing much talking, but may ask you to explain more or to give an example. Also, it's my job to see that everyone has a chance to voice their opinions, as well as to keep us moving along so that we have time to discuss all of the questions. So, at times, it might seem as though I am cutting you off, and this is not meant to be rude but rather to make sure that we have time to hear from everyone on each question. You'll also have the opportunity to continue this discussion online via a private forum.

   Since we only have until 7:00 PM today, we may not have time to hear many details of each person's perspective. We know that you have each been through your own experience and sharing your experience with others can be helpful. We hope you'll understand that for these next two and half hours we will ask you to focus on the questions asked. You can take extra time after the group is finished to talk more with each other if you wish.

   We want to thank each of you for being here, so please know that we value your ideas and comments

2. You have been divided into (X) groups for IS 777. We kindly ask that you remain with your group. You will be sharing your opinions and thoughts in your smaller groups and sharing summaries with the larger group.

   Each person will be handed a set of index cards. In each session a specific question

will ask you to write your comments and thoughts on these index cards first and then use them to guide your discussion in your respective groups. We will be collecting these cards at the end of each session, so please ensure that you're writing as legibly as possible. After the time is up, we'll ask a representative from each group to share the main points of their discussion with the larger group.

We have a volunteer student scribe with each group that will help take notes on the easel for you when you discuss in the group.

3. Each scribe has a short form we kindly ask everybody to fill it out. It asks for demographic information for description purposes. No personal information that will identify you or jeopardize your identity. We use this for research sample characteristics.

4. It's really important that everyone hear this: THERE ARE NO RIGHT OR WRONG ANSWERS, only differing points of view. Each person's experiences and opinions are valid, and we want to hear a wide range of opinions on the questions we'll be asking. You don't need to agree with others, but you must listen respectfully as others share their views. So, please speak up, whether you agree or disagree with what's being said, and let us know what you think. One person talks at a time for recording purposes.

5. Sometimes participants bring up private issues during these discussions, and we want to be sure that everyone agrees before we begin the group that anything of a private or personal nature that is mentioned in this room will NOT be repeated to others outside of this discussion group. Can I see a nod from everyone showing me that you agree with this confidentiality ground rule? *(If anyone is not willing to give their consent to confidentiality, they may be excused from the group.)*

6. Let me tell you about our recording process. As you can see, we have a recording

devices today using our smart phones and tablets. We usually record these focus groups because we want to get everything that all of you say, and we simply can't write fast enough to get it all down. We use first names only in the transcript, and when we put together the results from all the groups, we don't include any names.

It is VERY IMPORTANT that we speak ONE AT A TIME, so that we have a good quality recording. So, now that you know what our process is, is everyone OK with being recorded? Can I see a nod from everyone showing me that you agree.*(Start Recording)*

7. Let me mention before we start, that we plan to be finished with our discussion on this topic by 7:00PM.

Closing

Is there anything we've missed? Anything else we should know?

Thanks so much for being here today and for sharing your time and thoughts with us! We hope you continue this great discussion online via email and fossil repository.

Appendix F

Focus Group Questions

Question List

Usefulness

Usefulness is an individual's perception of using an IT system will enhance job performance.

1. What was your experience of using DQT?
2. What kind of problem will you face in your data if there is no DQT?
3. What approach would you like to use for data quality improvement if there is no DQT?
4. What are some ways that approaches mentioned above are different than DQT?
5. How much DQT can facilitate data quality improvement?
6. How do you feel about the results from DQT in terms of:
    - Accuracy
    - Completeness
    - Consistency

7. Are results obtained from DQT likely to be biased by any level of business rule?
8. What do you think about the usefulness of listed features:
    - Business rule management
    - Data defect detection
    - Collaborative communication and management
    - Data and business rule version control
9. Which feature in DQT is useless?
    - Is three-layer business rule structure useful in collaborative data quality improvement?
    - Is versioning control for data or business rule helpful in data governance?
    - Does ticketing system facilitate the communication within groups?
10. Which functionality would you like to add in DQT?

Usability

Usability is the ease of use and learnability of a human-made object such as a tool or device. In software engineering, usability is the degree to which a software can be used by specified consumers to achieve quantified objectives with effectiveness, efficiency, and satisfaction in a quantified context of use.

1. How easy is it to use DQT? (How much effort/time you need to pay to use DQT?)
2. How easy is it to learn DQT? (How much effort/time do you need to pay to learn DQT?)
3. How did you learn DQT?
4. What do you feel about the ease of learning by information provided in DQT?

5. How understandable of listed aspects in DQT?
    - Use of each widget
    - Concept of three-layer business rule
    - Purpose of each feature
6. How do you feel about design of DQT in terms of:
    - layout
    - content awareness
    - aesthetics
    - user experience
    - consistency
    - minimize user effort
    - navigation
7. How do you feel about listed features:
    - Business rule management
    - Data defect detection
    - Collaborative communication and management
    - Data and business rule version control
8. What are deficiencies of DQT in terms of:
    - layout
    - content awareness
    - aesthetics
    - user experience

- consistency
- minimize user effort
- navigation

9. What problem do you see in:
    - Business rule management
    - Data defect detection
    - Collaborative communication and management
    - Data and business rule version control
10. What feature would you like to add to improve DQT?

Bibliography

[1] Berwick DM, Nolan TW, Whittington J. The triple aim: care, health, and cost. Health affairs. 2008;27(3):759–769.

[2] Kumar S, Ghildayal NS, Shah RN. Examining quality and efficiency of the US healthcare system. International journal of health care quality assurance. 2011;24(5):366–388.

[3] Keehan SP, Sisko AM, Truffer CJ, Poisal JA, Cuckler GA, Madison AJ, et al. National health spending projections through 2020: economic recovery and reform drive faster spending growth. Health Affairs. 2011;p. 10–1377.

[4] Centers for Medicare & Medicaid Services, et al. National health expenditures 2012 highlights. Online verfügbar unter http://www cms gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/National-HealthExpendData/Downloads/highlights pdf. 2014;.

[5] Protection, Patient and Act, Affordable Care. Patient protection and affordable care act. Public law. 2010;111(148):1.

[6] Golay A. Pharmacoeconomic aspects of poor adherence: can better adherence reduce healthcare costs? Journal of medical economics. 2011;14(5):594–608.

[7] Srinivasan U, Arunasalam B. Leveraging big data analytics to reduce healthcare costs. IT professional. 2013;15(6):21–28.

[8] Egede LE, Gebregziabher M, Zhao Y, Dismuke CE, Walker RJ, Hunt KJ, et al. Impact of mental health visits on healthcare cost in patients with diabetes and comorbid mental health disorders. PLoS One. 2014;9(8):e103804.

[9] Abbass A, Kisely S, Rasic D, Town JM, Johansson R. Long-term healthcare cost reduction with intensive short-term dynamic psychotherapy in a tertiary psychiatric service. Journal of psychiatric research. 2015;64:114–120.

[10] Broderick RC, Fuchs HF, Harnsberger CR, Chang DC, Sandler BJ, Jacobsen GR, et al. Increasing the value of healthcare: improving mortality while reducing cost in bariatric surgery. Obesity surgery. 2015;25(12):2231–2238.

[11] Kaplan RS, Porter ME. How to solve the cost crisis in health care. Harv Bus Rev. 2011;89(9):46–52.

[12] Berwick DM, Hackbarth AD. Eliminating waste in US health care. Jama. 2012;307(14):1513–1516.

[13] Carayon P, Wetterneck TB, Rivera-Rodriguez AJ, Hundt AS, Hoonakker P, Holden R, et al. Human factors systems approach to healthcare quality and patient safety. Applied ergonomics. 2014;45(1):14–25.

[14] Melnyk BM, Gallagher-Ford L, Long LE, Fineout-Overholt E. The establishment of evidence-based practice competencies for practicing registered nurses and advanced practice nurses in real-world clinical settings: proficiencies to improve healthcare quality, reliability, patient outcomes, and costs. Worldviews on Evidence-Based Nursing. 2014;11(1):5–15.

[15] Margaret A, Edwards JR, Allen-Bridson K, Gross C, Malpiedi PJ, Peterson KD, et al. National Healthcare Safety Network (NHSN) report, data summary for 2013, device-associated module. American journal of infection control. 2015;43(3):206.

[16] Harvey AR, Basavaraju SV, Chung KW, Kuehnert MJ. Transfusion-related adverse reactions reported to the National Healthcare Safety Network Hemovigilance Module, United States, 2010 to 2012. Transfusion. 2015;55(4):709–718.

[17] Wu AW. Medical error: the second victim: the doctor who makes the mistake needs help too. BMJ: British Medical Journal. 2000;320(7237):726.

[18] World Health Organization. Global status report on alcohol and health, 2014. World Health Organization; 2014.

[19] Waring J, Allen D, Braithwaite J, Sandall J. Healthcare quality and safety: a review of policy, practice and research. Sociology of health & illness. 2016;38(2):198–215.

[20] McDermott AM, Steel DR, McKee L, Hamel L, Flood PC. Scotland 'Bold and Brave'? Conditions for Creating a Coherent National Healthcare Quality Strategy. In: Managing Change. Springer; 2015. p. 189–205.

[21] Vollaard H, van de Bovenkamp HM, Vrangbæk K. The emerging EU quality of care policy: From sharing information to enforcement. Health Policy. 2013;111(3):226–233.

[22] Batalden PB, Davidoff F. Batalden PB, Davidoff F, editors. What is "quality improvement" and how can it transform healthcare? BMJ Publishing Group Ltd; 2007.

[23] Koh HC, Tan G, et al. Data mining applications in healthcare. Journal of healthcare information management. 2011;19(2):65.

[24] Biafore S. Predictive solutions bring more power to decision makers. Health Management Technology. 1999;20(10):12–14.

[25] Christy T. Analytical tools help health firms fight fraud. Insurance & Technology. 1997;22(3):22–26.

[26] Kolar H. Caring for healthcare. Health management technology. 2001;22(4):46–47.

[27] Milley A. Healthcare and data mining. Health Management Technology. 2000;21(8):44–47.

[28] Dinov ID, Petrosyan P, Liu Z, Eggert P, Zamanyan A, Torri F, et al. The perfect neuroimaging-genetics-computation storm: collision of petabytes of data, millions of hardware devices and thousands of software tools. Brain imaging and behavior. 2014;8(2):311–322.

[29] Dinov ID. Volume and value of big healthcare data. Journal of medical statistics and informatics. 2016;4.

[30] Centers for Medicare & Medicaid Services (CMS) H, et al. Medicare and Medicaid programs; electronic health record incentive program. Final rule. Federal register. 2010;75(144):44313.

[31] Häyrinen K, Saranto K, Nykänen P. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. International journal of medical informatics. 2008;77(5):291–304.

[32] Churpek MM, Yuen TC, Park SY, Gibbons R, Edelson DP. Using electronic health record data to develop and validate a prediction model for adverse outcomes on the wards. Critical care medicine. 2014;42(4):841.

[33] King J, Patel V, Jamoom EW, Furukawa MF. Clinical benefits of electronic health record use: national findings. Health services research. 2014;49(1pt2):392–404.

[34] Hersh WR. Healthcare data analytics. Health informatics: practical guide for healthcare. 2014;.

[35] Donaldson MS, Corrigan JM, Kohn LT, et al. To err is human: building a safer health system. vol. 6. National Academies Press; 2000.

[36] Neame R. Design Principles in the Development of (Public) Health Information Infrastructures. Online journal of public health informatics. 2012;4(1).

[37] Crerand WJ, Lamb J, Rulon V, Karal B, Mardekian J. Building data quality into clinical trials. Journal of AHIMA. 2002;73(10):44–6.

[38] World Health Organization, et al. Improving data quality: a guide for developing countries. Manila: WHO Regional Office for the Western Pacific; 2003.

[39] AHIMA. Statement on Quality Healthcare Data and Information. The American Health Information Management Association; 2007. http://bok.ahima.org/doc?oid=101304#.WefHxWhSxPZ.

[40] Baum R, Bartram J, Hrudey S. The Flint Water Crisis Confirms That US Drinking Water Needs Improved Risk Management. Environmental science & technology. 2016;50(11):5436.

[41] Lofton KL. Report Finds Pill Mill Pharmacies Contributing to WV Drug Crisis. West Virginia Public Broadcasting; 2016. http://www.webcitation.org.

[42] Conrad C, Bradley HM, Broz D, Buddha S, Chapman EL, Galang RR, et al. Community outbreak of HIV infection linked to injection drug use of oxymorphone—Indiana, 2015. MMWR Morb Mortal Wkly Rep. 2015;64(16):443–444.

[43] Strong DM, Lee YW, Wang RY. Data quality in context. Communications of the ACM. 1997;40(5):103–110.

[44] Arts DG, De Keizer NF, Scheffer GJ. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. Journal of the American Medical Informatics Association. 2002;9(6):600–611.

[45] Goldhill D, Sumner A. APACHE II, data accuracy and outcome prediction. Anaesthesia. 1998;53(10):937–943.

[46] Lorenzoni L, Da Cas R, Aparo U. The quality of abstracting medical information from the medical record: the impact of training programmes. International Journal for Quality in Health Care. 1999;11(3):209–213.

[47] Seddon D, Williams E. Data quality in population-based cancer registration: an assessment of the Merseyside and Cheshire Cancer Registry. British Journal of Cancer. 1997;76(5):667.

[48] Federspiel CF, Ray WA, Schaffner W. Medicaid records as a valid data source: the Tennessee experience. Medical Care. 1976;14(2):166–172.

[49] Fowles JB, Lawthers AG, Weiner JP, Garnick DW, Petrie DS, Palmer RH. Agreement between physicians' office records and Medicare Part B claims data. Health care financing review. 1995;16(4):189.

[50] Botsis T, Hartvigsen G, Chen F, Weng C. Secondary use of EHR: data quality issues and informatics opportunities. Summit on Translational Bioinformatics. 2010;2010:1.

[51] Johnson CM, Johnson TR, Zhang J. A user-centered framework for redesigning health care interfaces. Journal of biomedical informatics. 2005;38(1):75–87.

[52] Fudge N, Wolfe CD, McKevitt C. Assessing the promise of user involvement in health service development: ethnographic study. Bmj. 2008;336(7639):313–317.

[53] Ash JS, Berg M, Coiera E. Some unintended consequences of information technology in health care: the nature of patient care information system-related errors. Journal of the American Medical Informatics Association. 2004;11(2):104–112.

[54] Weiner MG, Embi PJ. Toward reuse of clinical data for research and quality improvement: the end of the beginning? Annals of internal medicine. 2009;151(5):359–360.

[55] Burnum JF. The misinformation era: the fall of the medical record. Ann Intern Med. 1989;110(6):482–4.

[56] van der Lei J. Use and abuse of computer-stored medical records. Methods Archive. 1991;30:79–80.

[57] Yakout M, Elmagarmid AK, Neville J, Ouzzani M, Ilyas IF. Guided data repair. Proceedings of the VLDB Endowment. 2011;4(5):279–289.

[58] Houston L, Yu P, Martin A, Probst Y. Defining and Developing a Generic Framework for Monitoring Data Quality in Clinical Research. In: AMIA 2018 Annual Symposium Proceedings. San Fransisco, CA: Oxford University Press; 2018. p. 1300–1309.

[59] Lewis N. Fogarty S, editor. Poor Data Management Costs Healthcare Providers. InformationWeek Healthcare; 2012. Available from: http://www.informationweek.

[60] Rosenbaum S. Medicaid. Mass Medical Soc; 2002.

[61] Choma NN, Griffin MR, Huang RL, Mitchel EF, Kaltenbach LA, Gideon P, et al. An algorithm to identify incident myocardial infarction using Medicaid data. Pharmacoepidemiology and drug safety. 2009;18(11):1064–1071.

[62] Bright RA, Avorn J, Everitt DE. Medicaid data as a resource for epidemiologic studies: strengths and limitations. Journal of clinical epidemiology. 1989;42(10):937–945.

[63] Buescher PA, Whitmire JT, Brunssen S, Kluttz-Hile CE. Children who are medically fragile in North Carolina: using Medicaid data to estimate prevalence and medical care costs in 2004. Maternal and child health journal. 2006;10(5):461–466.

[64] Pedersen TB, Jensen CS, Dyreson CE. A foundation for capturing and querying complex multidimensional data. Information Systems. 2001;26(5):383–423.

[65] University of Oxford. Data Quality and Data Quality Assurance Policy, Planning and Resource Allocation. Planning & Resource Allocation; 2017. Available from: https://www.admin.ox.ac.uk.

[66] Schiefer J, List B, Bruckner R. A holistic approach for managing requirements of data warehouse systems. AMCIS 2002 Proceedings. 2002;p. 13.

[67] Rahm E, Do HH. Data cleaning: Problems and current approaches. IEEE Data Eng Bull. 2000;23(4):3–13.

[68] Hasan S, Padman R. Young SN, editor. Analyzing the Effect of Data Quality on the Accuracy of Clinical Decision Support Systems: A Computer Simulation Approach. American Medical Informatics Association; 2006. Available from: https://www.ncbi.nlm.nih.gov.

[69] Wand Y, Wang RY. Anchoring data quality dimensions in ontological foundations. Communications of the ACM. 1996;39(11):86–95.

[70] Weidema BP, Wesnaes MS. Data quality management for life cycle inventories—an example of using data quality indicators. Journal of cleaner production. 1996;4(3-4):167–174.

[71] Pipino LL, Lee YW, Wang RY. Data quality assessment. Communications of the ACM. 2002;45(4):211–218.

[72] Redman TC. The impact of poor data quality on the typical enterprise. Communications of the ACM. 1998;41(2):79–82.

[73] Fisher CW, Chengalur-Smith I, Ballou DP. The impact of experience and time on the use of data quality information in decision making. Information Systems Research. 2003;14(2):170–188.

[74] Haug A, Zachariassen F, Van Liempd D. The costs of poor data quality. Journal of Industrial Engineering and Management. 2011;4(2):168–193.

[75] Maletic JI, Marcus A. Data Cleansing: Beyond Integrity Analysis. In: IQ. Citeseer; 2000. p. 200–209.

[76] Sohn H, Farrar CR, Hunter NF, Worden K. Structural health monitoring using statistical pattern recognition techniques. Journal of dynamic systems, measurement, and control. 2001;123(4):706–711.

[77] Yang Y, Carbonell JG, Brown RD, Pierce T, Archibald BT, Liu X. Learning approaches for detecting and tracking news events. IEEE Intelligent Systems and Their Applications. 1999;14(4):32–43.

[78] Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis. vol. 344. John Wiley & Sons; 2009.

[79] Agrawal R, Imieliński T, Swami A. Mining Association Rules Between Sets of Items in Large Databases. SIGMOD Rec. 1993 Jun;22(2):207–216. Available from: http://doi.acm.org.

[80] Marcus A, Maletic JI. Utilizing association rules for the identification of errors in data. Technical Report CS-00-03. 2000;.

[81] Srikant R, Vu Q, Agrawal R. Mining association rules with item constraints. In: KDD. vol. 97; 1997. p. 67–73.

[82] Korn F, Labrinidis A, Kotidis Y, Faloutsos C. Ratio rules: A new paradigm for fast, quantifiable data mining. Research Showcase @ CMU. 1998;.

[83] Patel P. Perficient: Bad Data, Bad Decisions. Perficient; 2013. Available from: http://www.webcitation.org.

[84] Hasan S, Padman R. Analyzing the effect of data quality on the accuracy of clinical decision support systems: a computer simulation approach. In: AMIA annual symposium proceedings. vol. 2006. American Medical Informatics Association; 2006. p. 324.

[85] Wachter RM. Wachter RM, editor. Why Health Care Tech still So Bad. The New York Times; 2015. http://www.webcitation.org.

[86] Lin B, Chan HG. Managing data quality in the health care industry: Some critical issues. Journal of International Information Management. 2000;9(1):4.

[87] Dealy G. The Importance of Business Rules in Actionable Healthcare Data. Health IT Outcomes; 2016. http://www.webcitation.org.

[88] Davoudi S, Dooling JA, Glondys B, Jones TD, Kadlec L, Overgaard SM, et al. Data Quality Management Model (2015 Update). Journal of AHIMA. 2015;86(10).

[89] Walton J. The Role of Business Data Steward In Data Governance. Computer Task Group; 2015. http://www.webcitation.org.

[90] Rosenbaum S. Data governance and stewardship: designing data stewardship entities and advancing data access. Health services research. 2010;45(5p2):1442–1455.

[91] Belfils SE, Hillion S, Kormann T, Mathey C. Business Rule Management System. Google Patents; 2015. US Patent App. 14/481,392.

[92] Waller LA, Gotway CA. Applied spatial statistics for public health data. vol. 368. John Wiley & Sons; 2004.

[93] Newcombe CR, Waas FW. System and method for versioning data in a distributed data store. Google Patents; 2012. US Patent 8,266,122.

[94] Liu XL, Wang HZ, Li JZ, Gao H. EntityManager: Managing Dirty Data Based on Entity Resolution. Journal of Computer Science and Technology. 2017;32(3):644–662.

[95] Kwak SK, Kim JH. Statistical data preparation: management of missing values and outliers. Korean journal of anesthesiology. 2017;70(4):407–411.

[96] Cong G, Fan W, Geerts F, Jia X, Ma S. Improving data quality: Consistency and accuracy. In: Proceedings of the 33rd international conference on Very large data bases. VLDB Endowment; 2007. p. 315–326.

[97] Winkler WE. Methods for evaluating and creating data quality. Information Systems. 2004;29(7):531–550.

[98] Tian Y, Michiardi P, Vukolic M. Bleach: A Distributed Stream Data Cleaning System. arXiv preprint arXiv:160905113. 2016;.

[99] Zhao Z, Ng W. A model-based approach for rfid data stream cleansing. In: Proceedings of the 21st ACM international conference on Information and knowledge management. ACM; 2012. p. 862–871.

[100] Dallachiesa M, Ebaid A, Eldawy A, Elmagarmid A, Ilyas IF, Ouzzani M, et al. NADEEF: a commodity data cleaning system. In: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. ACM; 2013. p. 541–552.

[101] Fan W, Li J, Tang N, et al. Incremental detection of inconsistencies in distributed data. IEEE Transactions on Knowledge and Data Engineering. 2014;26(6):1367–1383.

[102] Khayyat Z, Ilyas IF, Jindal A, Madden S, Ouzzani M, Papotti P, et al. Bigdansing: A system for big data cleansing. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. ACM; 2015. p. 1215–1230.

[103] Andritsos P, Fuxman A, Miller RJ. Clean answers over dirty databases: A probabilistic approach. In: Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on. IEEE; 2006. p. 30–30.

[104] Taggart J, Liaw ST, Yu H. Structured data quality reports to improve EHR data quality. International journal of medical informatics. 2015;84(12):1094–1098.

[105] Arenas M, Bertossi L, Chomicki J. Consistent query answers in inconsistent databases. In: Proceedings of the eighteenth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM; 1999. p. 68–79.

[106] Team Rc, et al. R: A language and environment for statistical computing. R Foundation for Statistical Computing. 2013;.

[107] Maletic JI, Marcus A. Data Cleansing: Beyond Integrity Analysis. In: IQ. Citeseer; 2000. p. 200–209.

[108] Lee ML, Lu H, Ling TW, Ko YT. Cleansing data for mining and warehousing. In: International Conference on Database and Expert Systems Applications. Springer; 1999. p. 751–760.

[109] Hernández MA, Stolfo SJ. Real-world data is dirty: Data cleansing and the merge/purge problem. Data mining and knowledge discovery. 1998;2(1):9–37.

[110] Wei W, Zhang M, Zhang B, Tang X. A data cleaning method based on association rules. In: ISKE (International Conference on Intelligent Systems and Knowledge Engineering); 2007. p. 1–5.

[111] Răduţ C. The Quality of Data and Metadata in a data warehouse. RePEc: brc: journl. 2013;19:36–40.

[112] Yakout M, Elmagarmid AK, Neville J, Ouzzani M, Ilyas IF. Guided data repair. Proceedings of the VLDB Endowment. 2011;4(5):279–289.

[113] Franconi E, Palma AL, Leone N, Perri S, Scarcello F. Census data repair: a challenging application of disjunctive logic programming. In: International Conference on Logic for Programming Artificial Intelligence and Reasoning. Springer; 2001. p. 561–578.

[114] Demsky B, Rinard M. Automatic detection and repair of errors in data structures. In: Acm sigplan notices. vol. 38. ACM; 2003. p. 78–95.

[115] Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. Egems. 2016;4(1).

[116] Laranjeiro N, Soydemir SN, Bernardino J. A survey on data quality: classifying poor data. In: Dependable Computing (PRDC), 2015 IEEE 21st Pacific Rim International Symposium on. IEEE; 2015. p. 179–188.

[117] Wang RY, Strong DM. Beyond accuracy: What data quality means to data consumers. Journal of management information systems. 1996;12(4):5–33.

[118] Chen H, Ku WS, Wang H, Sun MT. Leveraging spatio-temporal redundancy for RFID data cleansing. In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of data. ACM; 2010. p. 51–62.

[119] Ousterhout JK, Jones K. Tcl and the Tk toolkit. Pearson Education; 2009.

[120] Owens M, Allen G. SQLite. Springer; 2010.

[121] Müller H, Freytag JC. Problems, methods, and challenges in comprehensive data cleansing. Professoren des Inst. Für Informatik; 2005.

[122] Oliveira P, Rodrigues F, Henriques PR. A Formal Definition of Data Quality Problems. In: IQ; 2005. p. 3–12.

[123] Barateiro J, Galhardas H. A Survey of Data Quality Tools. Datenbank-Spektrum. 2005;14(15-21):48.

[124] Gschwandtner T, Gärtner J, Aigner W, Miksch S. A taxonomy of dirty time-oriented data. In: International Conference on Availability, Reliability, and Security. Springer; 2012. p. 58–72.

[125] Kim W, Choi BJ, Hong EK, Kim SK, Lee D. A taxonomy of dirty data. Data mining and knowledge discovery. 2003;7(1):81–99.

[126] Li L, Peng T, Kennedy J. A rule based taxonomy of dirty data. GSTF Journal on Computing (JoC). 2018;1(2).

[127] Naumann F. Data profiling revisited. ACM SIGMOD Record. 2014;42(4):40–49.

[128] Kendall KE, Kendall JE, Kendall EJ, Kendall JA. Systems analysis and design. vol. 4. Prentice Hall New Jersey; 1992.

[129] Ritchie J, Lewis J, Nicholls CM, Ormston R. Qualitative research practice: A guide for social science students and researchers. Sage; 2013.

[130] Smith J, Firth J. Qualitative data analysis: the framework approach. Nurse researcher. 2011;18(2):52–62.

[131] Srivastava A, Thomson SB. Framework analysis: a qualitative methodology for applied policy research. Journal of Administration and Governance. 2009;.

[132] Louise Barriball K, While A. Collecting Data using a semi-structured interview: a discussion paper. Journal of advanced nursing. 1994;19(2):328–335.

[133] Ghosh S, Zafar MB, Bhattacharya P, Sharma N, Ganguly N, Gummadi K. On sampling the wisdom of crowds: Random vs. expert sampling of the twitter stream. In: Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. ACM; 2013. p. 1739–1744.

[134] Farrokhi F, Mahmoudi-Hamidabad A. Rethinking convenience sampling: Defining quality criteria. Theory and practice in language studies. 2012;2(4):784.

[135] Heckman JJ, Singer B. Econometric duration analysis. Journal of Econometrics. 1984;24(1-2):63–132.

[136] Grauman K, Betke M, Gips J, Bradski GR. Communication via eye blinks-detection and duration analysis in real time. In: Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on. vol. 1. IEEE; 2001. p. I–I.

[137] Shipan CR, Shannon ML. Delaying justice (s): A duration analysis of Supreme Court confirmations. American Journal of Political Science. 2003;47(4):654–668.

[138] Staubus GJ. Activity costing and input-output accounting. RD Irwin; 1971.

[139] Flanagan SR. Final report of the review of policing. Retrieved February. 2008;20:2009.

[140] Edwards S. Activity Based Costing. Topic Gateway Series No. 1. CIMA; 2014.

[141] Lee SJ, Earle CC, Weeks JC. Outcomes research in oncology: history, conceptual framework, and trends in the literature. Journal of the National Cancer Institute. 2000;92(3):195–204.

[142] Clancy CM, Eisenberg JM. Outcomes research: measuring the end results of health care. Science. 1998;282(5387):245–246.

[143] Nightingale F. Introductory Notes on Lying-in Institutions: Together with a Proposal for Organising an Institution for Training Midwives and Midwifery Nurses. Longmans, Green; 1871.

[144] Dennis A, Wixom BH, Roth RM. Systems analysis and design. John Wiley & Sons; 2008.

[145] Pettit PH. Equity and the Law of Trusts. Oxford University Press; 2012.

[146] Read S, Ashman M, Scott C, Savage J. Evaluation of the modern matron role in a sample of NHS trusts. Final Report to the Department of Health, The Royal College of Nursing Institute and The University of Sheffield School of Nursing and Midwifery, Sheffield and London, UK. 2004;.

[147] Gerrish K, Chau R, Sobowale A, Birks E. Bridging the language barrier: the use of interpreters in primary care nursing. Health & social care in the community. 2004;12(5):407–413.

[148] Gale NK, Heath G, Cameron E, Rashid S, Redwood S. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC medical research methodology. 2013;13(1):117.

[149] Thomas DR. A general inductive approach for analyzing qualitative evaluation data. American journal of evaluation. 2006;27(2):237–246.

[150] Mays N, Pope C. Qualitative research in health care: Assessing quality in qualitative research. BMJ: British Medical Journal. 2000;320(7226):50.

[151] Morgan DL. The focus group guidebook. vol. 1. Sage publications; 1997.

[152] Goldfeld K. CRAN - Package simstudy;. (Accessed on 10/29/2018). https://cran.r-project.org/web/packages/simstudy/index.html.

[153] Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS quarterly. 1989;p. 319–340.

[154] Kitzinger J. Qualitative research. Introducing focus groups. BMJ: British medical journal. 1995;311(7000):299.

[155] Boehm BW. A spiral model of software development and enhancement. Computer. 1988;21(5):61–72.

[156] Miller AA. System for automated problem detection, diagnosis, and resolution in a software driven system. Google Patents; 2004. US Patent 6,742,141.

[157] Jansen BJ. Search log analysis: What it is, what's been done, how to do it. Library & information science research. 2006;28(3):407–432.

[158] Kort J, de Poot H. Usage analysis: combining logging and qualitative methods. In: CHI'05 extended abstracts on Human factors in computing systems. ACM; 2005. p. 2121–2122.

[159] Froehlich J, Chen MY, Consolvo S, Harrison B, Landay JA. MyExperience: a system for in situ tracing and capturing of user feedback on mobile phones. In: Proceedings of the 5th international conference on Mobile systems, applications and services. ACM; 2007. p. 57–70.

[160] Levine DM, Berenson ML, Stephan D, Lysell D. Statistics for managers using Microsoft Excel. vol. 660. Prentice Hall Upper Saddle River, NJ; 1999.

[161] Van Rossum G, et al. Python Programming Language. In: USENIX Annual Technical Conference. vol. 41; 2007. p. 36.

[162] Elmasri R, Navathe S. Fundamentals of database systems. Addison-Wesley Publishing Company; 2010.

[163] DiMaggio C. Introduction. In: SAS for Epidemiologists. Springer; 2013. p. 1–5.