8-16-2016

DataGauge: A Model-Driven Framework for Systematically Assessing the Quality of Clinical Data for Secondary Use

Jose Franck DiazVasquez
University of Texas Health Science Center at Houston, jdiazvas@wakehealth.edu

DataGauge: A Model-Driven Framework for Systematically Assessing the Quality of Clinical Data for Secondary Use

A
Dissertation

Presented to the Faculty of
The University of Texas
Health Science Center at Houston
School of Biomedical Informatics
in Partial Fulfilment of the Requirements for the Degree of

Doctor of Philosophy

By

Jose Franck Diaz, MS

University of Texas Health Science Center at Houston

2016

Dissertation Committee:

Todd R. Johnson, PhD, PhD[1], Advisor
Elmer V. Bernstam, MD, MSE[1,2]
MinJae Lee, PhD[2]
Kevin Hwang, MD, MPH[2]

[1]The School of Biomedical Informatics
[2]McGovern Medical School

Dedication

To the people who have helped me turn chaos into order; for each and every one of you I am forever grateful.

Acknowledgements

It all begins with people, usually the closest ones.

So I would like a thank my wife Elaheh Rahbar, who inspired me to go beyond what I knew and understood. She has been my support during times of joy and sadness equally. Without her, this work would have never been. I would also like to thank my family for always providing earnest support, even when they did not understand what I was doing. In particular, I would like to thank my father, Dr. Pedro Pablo Diaz Vasquez, for leading by example and infusing me with the confidence of those who fight chaos every day. I also thank my mother, Josiane Garelli for her continuous support and prayer, as well as Nico, Caro, Alepo and Tia Edith as well as Alicia for their kind encouragement, love and respect. I am also forever grateful to my parents-in-law, Dr. Mohammad Hossein Rahbar and Afsaneh Zekri for their unfaltering support in so many aspects that it would be impractical to list here.

It all grows from interactions between people.

For which I thank my advisor, Dr. Todd R. Johnson. He has shown me patience and encouragement at all times. I am eternally grateful and will strive to pay this kindness forward. I'd also like to thank the members of my committee, Dr. Elmer V. Bernstam, Dr. MinJae Lee and Dr. Kevin Hwang, who have been generous with their time and comments to improve this work. For the same reason, I'd like to thank the faculty from

the CPRIT innovation fellowship Dr. Roberta Ness, Dr. Patricia D. Mullen and Dr. David S. Loose. These mentors have helped me accept that all is ultimately inaccurate, but that there is value in being progressively less wrong with a little help from my friends.

I'd also like to thank my CPRIT fellowship mates and fellow grad students for their camaraderie and diverse points of view. In particular, I'd like to thank Dr. Deevakar Rogith his diligent help and for sharing his grad school experience, Zhiguo (Stanley) Yu for his support as a lab mate and positive attitude, Dr. Swaroop Gantela for providing the most interesting and delightful distractions and Adriana Stanley, for always having time for a chat. I am sure I forget people but be kind, this is harder than it seems.

No analytics work is possible without the constant attention of people that create the infrastructures within which informaticians thrive. So I'd like to thank the UTH BIG team for their always excellent work and diligent support. In particular, Susan Guerrero, Chuck Bearden and Alejandro Araya have provided invaluable help throughout this adventure.

I'd also like to thank the SBMI faculty at large, for each and every one of you has always had time to provide guidance and advice any time I have needed of it. In particular, I thank Dr. Amy Franklin for her guidance during my early years at SBMI. She provided much of the ground work for my understanding of informatics, research and the importance of people in the endless game of data+meaning.

I'd also like to thank the whole SHARP-C team; It was a blast to work with you. You kept me going when the going got tough. I would like to acknowledge Dr. Krisanne

Graves, in particular, for her constant encouragement, her fairness, her wicked sense of humor.

Opportunities are provided by people.
So I would like to thank the two professors who made this achievement a possibility. First, I would like to express my gratitude to Dr. Jiajie Zhang who opened the doors for me to join SBMI. I am thankful he was able to see potential beyond my initial training gap. Then, I would like to thank Dr. James E. Moore Jr. who took a chance in giving me an opportunity before I even considered joining the research community. He opened the door to a brand new world of ideas that I never knew existed. I am very grateful for his support in the academic realm, but also for being a friend.

It all ends with people, because I have surely forgotten some that have inspired me to look further and keep on walking. This line acknowledges them.

Some things transcend words. So this line acknowledges all I fail to comprehend; in particular, what transcends language and thought.

⁂

This work was supported in part by a fellowship from the UTHealth Innovation in Cancer Prevention Research Training Program funded by the Cancer Prevention and Research Institute of Texas (CPRIT). Additional support was provided by the *Bridges Family Doctoral Fellowship in Informatics Innovation,* the *Doris L. Ross scholarship* and the *James T. Willerson Scholarship*.

Abstract

There is growing interest in the reuse of clinical data for research and clinical healthcare quality improvement. However, direct analysis of clinical data sets can yield misleading results. Data Cleaning is often employed as a means to detect and fix data issues during analysis but this approach lacks of systematicity. Data Quality (DQ) assessments are a more thorough way of spotting threats to the validity of analytical results stemming from data repurposing. This is because DQ assessments aim to evaluate 'fitness for purpose'. However, there is currently no systematic method to assess DQ for the secondary analysis of clinical data. In this dissertation I present DataGauge, a framework to address this gap in the state of the art.

I begin by introducing the problem and its general significance to the field of biomedical and clinical informatics (Chapter 1). I then present a literature review that surveys current methods for the DQ assessment of repurposed clinical data and derive the features required to advance the state of the art (Chapter 2). In chapter 3 I present DataGauge, a model-driven framework for systematically assessing the quality of repurposed clinical data, which addresses current limitations in the state of the art. Chapter 4 describes the development of a guidance framework to ensure the systematicity of DQ assessment design. I then evaluate DataGauge's ability to flag potential DQ issues in comparison to a systematic state of the art method. DataGauge was able to increase ten fold the number of

potential DQ issues found over the systematic state of the art method. It identified more specific issues that were a direct threat to fitness for purpose, but also provided broader coverage of the clinical data types and knowledge domains involved in secondary analyses.

DataGauge sets the groundwork for systematic and purpose-specific DQ assessments that fully integrate with secondary analysis workflows. It also promotes a team-based approach and the explicit definition of DQ requirements to support communication and transparent reporting of DQ results. Overall, this work provides tools that pave the way to a deeper understanding of repurposed clinical dataset limitations before analysis. It is also a first step towards the automation of purpose-specific DQ assessments for the secondary use of clinical data. Future work will consist of further development of these methods and validating them with research teams making secondary use of clinical data.

Vita

2007	B.S., Mechanical Engineering, Santo Domingo Institute of Technology (INTEC), Santo Domingo, Dominican Republic
2007-2010	M.S., B.S. as Ingénieur Diplomé (I.D.), Biomedical Engineering, ESIL Polytech' School of advanced studies in engineering, Luminy, Marseille, France
2012-2014	Graduate Research Assistant, The University of Texas Health Science Center at Houston (UTHealth) School of Biomedical Informatics, Houston, TX
2014 to present	Pre-doctoral Fellow, PhD Student, School of Public Health, School of Biomedical Informatics University of Texas Health Science Center at Houston, TX

Field of Study

Health Informatics

Table of Contents

Dedication ii

Acknowledgements iii

Abstract vii

Vita ix

Table of Contents xi

List of Tables xiii

List of Figures xiv

Chapter 1: Introduction 1

Chapter 2: Related Work 11

Chapter 3: DataGauge - A Model-Driven Process for the Systematic Assessment of Repurposed Clinical Data 21

Chapter 4: A Guidance Framework for the Development of DQ Requirements 47

Chapter 5: Evaluation of DataGauge 78

Chapter 6: Conclusions, Limitations, and Contributions 94

References 102

Appendix A: Definitions 118

Appendix B: DQ Requirement Development Guidance 121

Appendix C: Use Case Definitions 141

List of Tables

Table 1. DQ requirement development guidance table 36
Table 2. Data quality requirement examples 42
Table 3. Secondary analysis type coverage by case 59
Table 4. Clinical data type coverage by case 69
Table 5. Overview of the four dimensions of our framework for DQ requirement guidance 65
Table 6. Overview guidance checklist for DQ requirement development.....69
Table 7. Sample of DQ requirement development guidance questions.....70
Table 8. Overview guidance checklist for the research question " Is prednisone, a commonly-prescribed corticosteroid, is associated with weight gain?" 72
Table 9. Contextual guidance questions relevant to the clinical domain knowledge and the Vitals clinical data type.....73
Table 10. Negative Binomial regression results predicting the number of flags returned by each method based on the method and number of tests performed 91

List of Figures

Figure 1. Iterative analysis-specific DQ assessment method for the secondary use of clinical data 27
Figure 2. Evolution of the data needs model for the purpose of assessing a relationship between prednisone and weight gain using repurposed clinical data 40
Figure 3. Distribution of analysis types for dental data reuse analyses 54
Figure 4. Distribution of analysis types for clinical data request tickets 55
Figure 5. Distribution of clinical data types for dental data reuse analyses 56
Figure 6. Distribution of clinical data types for clinical data request tickets 57
Figure 7. Distribution DQ requirements by use case number.....62
Figure 8. Distribution DQ requirements by DQ dimensions and knowledge domain 67
Figure 9. Distribution DQ requirements by data granularity levels and knowledge domain 68
Figure 10. Distribution DQ requirements by clinical data type and knowledge domain 68
Figure 11. Data Needs Model for the research question " Is prednisone, a commonly-prescribed corticosteroid, is associated with weight gain?" 71
Figure 12. Number of flags returned by both methods.....85
Figure 13. Number of discarding and review flags returned by both methods.....86
Figure 14. Coverage of clinical data types by the control method and DataGauge...89
Figure 15. Coverage of knowledge domains by the comparison standard and DataGauge 89

Chapter 1: Introduction

There is growing interest in the reuse of clinical data for research and clinical healthcare quality improvement. However, direct analysis of clinical data sets can yield misleading results (Hersh et al., 2013; C. Safran, 2014). Notably, van der Lei formulated the first law of informatics (Van Der Lei, 1991): "Data shall be used only for the purpose for which they were collected." On the other hand, clinical data routinely serve multiple purposes including clinical, billing, administrative and legal. Data quality (DQ) flaws are often cited as one cause of these misleading results (Dentler et al., 2014; Dentler, ten Teije, de Keizer, & Cornet, 2013; Weiner & Embi, 2009). Thus, DQ assessment is generally recommended to prevent the hazards of data repurposing (Brown, Kahn, & Toh, 2013; Hersh, 2007; M. G. Kahn, Raebel, Glanz, Riedlinger, & Steiner, 2012).

Once data are acquired it is difficult to change their quality (Hogan & Wagner, 1997; Van Der Lei, 1991). However, verifying their accuracy and ability to satisfy the needs of their intended secondary uses has the potential to increase confidence by unlocking a deeper understanding of the dataset's strengths, weaknesses, flaws and limitations. The results of secondary analyses can be then be interpreted in the light of this information as an indication of their validity. This idea is analogous to the concept of statistical confidence interval (Brookmeyer & Crowley, 1982), which has enabled the understanding of many complex phenomena with a well-defined degree of certainty.

There is currently no generalized method or approach to carry out such evaluation systematically for the secondary use of clinical data. However, quality is routinely evaluated for products taking into account their intended purpose. To ensure systematicity, these evaluations apply quality control standards and methodologies (Dale, 2015; Evans & Lindsay, 1999; Juran, 1962; Taguchi, 1986; Walker & Gee, 2000) that require the explicit definition of quantitative requirements. Model-driven engineering (Schmidt, 2006) supports the definition and testing of these requirements for the systematic and purpose-driven evaluation of software products.

In this thesis I propose and evaluate DataGauge, a framework to systematically assess the quality of repurposed clinical datasets based on these model-driven software quality assessment methods. I define a general process for the assessment of repurposed clinical datasets and provide guidance for the development of DQ requirements specifically for the secondary use of clinical data. Finally, I evaluate the ability of this framework to catch more potential DQ issues than the current state of the art methods of systematic DQ assessment for the secondary use of clinical data.

1.1 - The Reuse of Clinical Data: Availability, Benefits and Limitations

Unprecedented amounts of data are created every day though the recoding of clinical care information in patient records. As a result, clinical enterprises hold large amounts of data. For instance, it has been reported that healthcare data storage needs at Beth Israel Deaconess Medical Center have increased approximately six orders of magnitude (i.e., from gigabytes to petabytes) over the past three decades (C. Safran, 2014). The HITECH

act of 2009 has also been a driving force for growth by providing incentives for healthcare institutions that successfully adopted interoperability-capable Electronic Health Records (EHR) (Blumenthal, 2010). These efforts have been fueled by the understanding that health IT can help lower costs and increase the quality of care in medical institutions (Jha AK, 2010). The Institute of Medicine's report defining learning healthcare systems (Institute of Medicine (US) Roundtable on Evidence-Based Medicine, 2007) has also stimulated the secondary use of clinical data for evidence-based medicine and healthcare quality improvement.

Benefits from the reuse of this wealth of data have been noted for several decades (Fries & McShane, 1979; C. Safran, 1991; Starmer, Rosati, & Fred McNeer, 1974). Providing new ways to approach evidence-based medicine, surveillance, clinical research and clinical care quality assurance are often cited as the main benefits (Guyatt G, Cairns J, Churchill D, & et al, 1992; Hersh, 2007; Charles Safran et al., 2007). Other applications include cohort analyses to determine readmission risk (Phillips, Safran, Cleary, & Delbanco, 1987); description of patient populations (Hansell, Hollowell, Nichols, McNiece, & Strachan, 1999; Herrmann & Safran, 1992); infection control and epidemiological monitoring (C, Kp, & W, 1994; Classen & Burke, 1995; Samore, Lichtenberg, Saubermann, Kawachi, & Carmeli, 1997); and discovery of pharmaco-epidemiological relationships (Brownstein, Sordo, Kohane, & Mandl, 2007; Chalasani, Aljadhey, Kesterson, Murray, & Hall, 2004; Herzig SJ, Howell MD, Ngo LH, & Marcantonio ER, 2009). An additional benefit is that, like other forms of retrospective research, the reuse of clinical data has the potential to yield valuable insights at very low

cost and extremely short time. Because they do not require patient recruitment or data collection, the research is reduced to the time of data extraction plus analysis.

However, repurposed clinical data have many limitations (Hersh et al., 2013). Issues such as inaccuracy (Hogan & Wagner, 1997), incompleteness (Nicole G. Weiskopf, Rusanov, & Weng, 2013), bias (George Hripcsak, Knirsch, Zhou, Wilcox, & Melton, 2011), coding standard inconsistencies (George Hripcsak, Knirsch, Zhou, Wilcox, & Melton, 2007), inaccessible data (e.g., clinical notes) (George Hripcsak et al., 1995), heterogeneity (De Lusignan et al., 2011) and clinical workflow influences on data recording (George Hripcsak, Albers, & Perotte, 2011) have been reported in the literature. Dentler et al. clearly showed the impact of these issues by attempting clinical quality indicators calculations directly from EHR data (Dentler et al., 2014). They found that only three out of eight quality indicators could be computed directly from repurposed clinical data due to an average record completeness of 50% and average correctness of 87%. These limitations have also been reported in the literature for decades. For example, a meta-analysis of data accuracy assessments on EHR data published in 1997 revealed highly variable results (Hogan & Wagner, 1997) (e.g., accuracy measurements varying from 44 to 100%). These results clearly reveal the limited reliability of repurposed EHR data and strongly caution against their direct reuse for purposes other than patient care. Such findings motivated van der Lei, in 1991, to formulate the first law of informatics (Van Der Lei, 1991): “Data shall be used only for the purpose for which they were collected.” By definition, secondary use violates this law, making it is necessary to assess whether clinical data are adequate for any intended purpose other than the original.

1.2 - Ensuring Analytical Validity Through Data Checking

Current practices suggest the use of data cleaning methods to ensure reliable results (Broeck & Fadnes, 2013). The data cleaning process happens as the analysis is carried out and aims to detect faulty data. Data cleaning happens in three steps: (1) Screen for anomalous data, (2) Diagnose possible issues and (3) Address the issues found. The screening step consists in detecting a lack or excess of data, screening for outliers, inconsistencies, "strange" patterns and suspect analytical results (Van den Broeck, Argeseanu Cunningham, Eeckels, & Herbst, 2005). The cleaning process is designed as a response to issues and discrepancies found during analysis rather than a preventative assessment procedure. Thorough checks are rarely done before the analysis in practice, which carries the risk of missing harmful issues. In this setup, only the issues detected by the analyst through dataset manipulation and analysis are addressed. This does not ensure that the assessment will cover all potential threats to the validity of analytical results in secondary use applications. For example, EHR-extracted clinical datasets may present issues such as missing and duplicate data appearing during the extraction process, which typically involves complex queries. Despite these potential problems, the data are rarely checked thoroughly for quality during the extraction process. Also, data cleaning is often driven by the purpose-independent application available testing tools (e.g., range checking, data validation and data format checking) rather than to detect potential threats to analytical results. This situation is much more alarming in the case of repurposed clinical data (Van Der Lei, 1991) because they are not specifically designed and recorded to satisfy secondary analytical needs. In such cases, issues may arise beyond the accuracy

and cleanliness of the data, such as not having the right variables to run the analysis or data that record implausible events. It is, therefore, imperative to conduct a systematic assessment of the data's suitability for a specific secondary use case *prior to the secondary analysis (e.g., statistical analysis, exploratory visualization, etc.)*. Thus, data assessments of repurposed clinical data must cover a broad spectrum of potential issues rather than only those detected by the analyst.

DQ assessment (Maydanchik, 2007a) is an alternate approach to data cleaning. Its goal is to help ensure valid analytical results through the evaluation of the dataset's ability to satisfy analytical needs (i.e., its fitness for purpose) (Holve, Kahn, Nahm, Ryan, & Weiskopf, 2013; Juran, 1962). This approach is usually carried out before performing the secondary analysis and the results serve as the basis to determine the dataset's strengths and weaknesses. It emphasizes the evaluation of the dataset for a specific application and is a more appropriate way of investigating a broader range of potential issues as compared to data cleaning. This approach is also very attractive for the secondary use of clinical data given that two purposes interact in such applications (Floridi, 2013): the primary purpose (i.e., recording the care of patients) and the secondary analytical purpose (e.g., prevalence estimation, clinical outcomes analysis, etc.). DQ assessments allow the user to evaluate whether the dataset will be good enough to provide reliable results for the secondary purpose, while still minding the primary purpose.

Kahn et al. (M. G. Kahn et al., 2012) recently proposed a generalized framework to support a comprehensive and systematic approach to assess DQ for the purpose of building EHR-based Clinical Data Warehouse (CDW). However, there is a second stage

where systematic approaches for the assessment of DQ are currently unavailable. These assessments aim to evaluate a subset of the CDW selected for a secondary analytical purpose (i.e., the *analytical dataset*). They typically focus on the independent and dependent variables directly related to the research question. A review of methods available for these secondary assessments (Nicole Gray Weiskopf & Weng, 2013) revealed that current methods are *not generalizable*, *not systematic* and *fail to take the secondary analytical purpose into account*. Moreover, it has been noted that one of the main barriers to the effective assessment and the transparent reporting of DQ results in secondary uses of clinical data are the ambiguity of DQ definitions (N. Weiskopf, Hripcsak, Swaminathan, & Weng, 2013) and the lack of a universally accepted set of DQ features to test for (M. Kahn et al., 2015). Making DQ requirements explicit would greatly support consistent DQ evaluations as well as clearer communication and reporting of DQ results.

To address the current limitations in these methods I developed a framework with the following characteristics:

- Supports purpose-specific DQ assessment
- Provides a DQ assessment process that is:
    - Generalizable to a wide range of secondary use cases
    - Systematic (i.e., executed according to a fixed sequence of steps)
- Makes DQ requirements *explicit* in order to:
    - Improve communication within the research team
    - Promote transparent reporting of DQ issues

1.3 - Dissertation Structure

To create such framework I developed a *purpose-specific DQ assessment process* (i.e., DataGauge) that possesses the features stated above, along with *a guidance framework for the development of comprehensive DQ assessments* of repurposed clinical datasets. Then, I e*valuated* DataGauge's ability to increase the *number of DQ potential issues* found before assessment. The dissertation is laid out in an analogous fashion. In Chapter 2, I present a review of the current state of the science in DQ theory, DQ assessments at large and DQ assessments for the secondary use of clinical data to show current needs and gaps in the literature. In chapter 3, I describe the DataGauge process and its development. In chapter 4, I detail the development of the DQ assessment guidance for the definition of DQ requirements within DataGauge. In Chapter 5, I evaluate DataGauge's ability to identify more DQ issues than the current systematic standard method of DQ assessment for the secondary use of clinical data. The dissertation concludes by describing its significance and contributions to the field of clinical research informatics and clinical data reuse.

1.4- Summary of Contributions

My work lays the practical foundation for the systematic DQ assessment of repurposed clinical data as an evaluation of fitness for purpose (Holve et al., 2013). This contributes to supporting the reliable secondary use of clinical data (Charles Safran et al., 2007) which is a critical step towards building learning healthcare systems (Institute of Medicine (US) Roundtable on Evidence-Based Medicine, 2007). DataGauge provides a

stronger foundation for research aiming to *learn from existing clinical data*, *generate novel yet data-driven research hypotheses* and *carry out cost-effective population-level analyses*. It also provides guidance to support a thorough definition of DQ requirements for the secondary use of clinical data, which is currently missing in the literature. Lastly, it provides a method to explicitly define and encode DQ requirements. This is a first step towards supporting communication within the analytical team, because the process provides a tangible set of documents to help organize the team member's diverse backgrounds and expertise with respect to the secondary use case.

This work innovates by breaking the current paradigm of data assessments for clinical data reuse, which is that ad-hoc, analyst-based, analysis-independent cleaning of data is sufficient to prevent misleading results in the reuse of clinical data. DataGauge proposes a new frame where repurposed datasets are to be assessed for their adequacy to answer a specific research question prior to analysis. DataGauge also stipulates that the assessments should be done by a team of experts in the domains of data science, statistics and medicine. This work supports cancer prevention by enabling a more trustworthy reuse of observational data from a broadly available yet still untapped data source: Electronic health records. EHRs are a very powerful source of knowledge that has the potential to enable the cost-effective analysis of population-level data. These databases will undoubtedly become an invaluable source of data for research fields such as cancer-prevention in the near future due to the increasing number of recorded variables and imminent inclusion genomic data.

Chapter 2: Related Work

Many tools are available for the secondary analysis of data but there are limited options to assess whether those data will yield valid results. To ensure that analytical valid results, one must ensure that the data accurately describe the observed objects and that they are likely to contain the necessary information (e.g., values, variables, sampling rate, etc.) to answer research question. Research shows that repurposed clinical data sources pose problems in both of these areas. These problems stem from measurement and recording processes (Aronsky & Haug, 2000) and inadequacies stemming from repurposing (Hersh et al., 2013). It is crucial to detect these issues before analysis in order to provide researchers with an understanding of data limitations and, in turn, the expected accuracy of analytical results. Such process is nothing more than the quality evaluation of a dataset, broadly known as a DQ assessment (Maydanchik, 2007a).

2.1 - Data Quality Definitions, Frameworks and Assessment Tools

DQ has been a subject of study for several decades in fields outside biomedical informatics (Dasu, 2013; Dentler et al., 2014, 2013; Madnick, Wang, Lee, & Zhu, 2009; Redman, 1998, 2013; Sadiq, 2013a; Trickey, 2012). Past research on DQ has produced definitions (Standardization, 1994), methods (Maydanchik, 2007a; Olson, 2003) and frameworks for DQ management (Fan, 2012; Sadiq, 2013b) and improvement (Batini & Scannapieca, 2006a, 2006c; Fan, Geerts, Ma, Tang, & Yu, 2013; Lee, Strong, Kahn, &

Wang, 2002; Redman, 2013; Wang & Strong, 1996). This wealth of research provides a foundation for assessment during the data production cycle and is designed to support database administration work. In fact, most of this knowledge is geared towards supporting the maintenance of enterprise databases through DQ management (Redman, 2013). This process entails the systematic assessment of whole databases according to user-defined rules (Maydanchik, 2007a) and flagging of common data issues such as duplicates and inconsistent input. Flagged data are then corrected using imputation methods or eliminated (Fan, 2012). The cycle is repeated to monitor and manage the quality of whole databases (Sadiq, 2013b). Though this research supports primary practical uses of data, it provides very little guidance for the assessment of repurposed clinical data.

Quality is defined as the ability to satisfy needs (Standardization, 1994). Those needs are defined by an intended purpose. Thus, the most widely accepted definition of DQ is 'fitness for purpose' (Holve et al., 2013; Juran, 1962). Three key features can be derived from this view of DQ. First, the quality of any dataset can only be measured with respect to a specific purpose. For example, a pre-Copernican, astronomical book would have very low DQ for its original purpose of understanding the laws governing the universe, but very high DQ for the secondary purpose of understanding the historical development of Ptolemaic astronomy (Floridi, 2013). Second, when repurposing data two purposes must be considered (Floridi, 2013): the original purpose for which the data were produced and the *secondary purpose, which generally entails a secondary analysis to answer a specific research question*. This means that the assessment for the secondary

purpose will necessarily have to take the initial purpose into account to define assumptions, expectations and evaluation parameters for the DQ assessment. Finally, systematically assessing quality in any industry or application requires a set of points of interest or *DQ criteria*. These criteria are represented by *DQ requirements*, which define specific conditions that the dataset must meet to be fit for purpose (Juran, 1962; Standardization, 1994). For example, the set of DQ requirements for assessing the quality of clinical data for treating individual diabetic patients is quite different from that needed when assessing the same data for the purpose of understanding the efficacy of a new diabetes treatment protocol. Routine clinical data from individual diabetic patients lacks randomization, controls, and systematic data collection. Thus, DQ requirements should aim to define the ideal dataset for the intended purpose as a way to provide a standard to evaluate whether the data possesses the necessary features, such as an evenly sampled population, to obtain valid and acceptably reliable results.

In practice, DQ assessments are carried out using a set of tools and techniques that aim to test for specific DQ issues. A large number of techniques are available in the literature to semi-automatically assess specific DQ issues (Borek, Woodall, Oberhofer, & Parlikad, 2011; Maydanchik, 2007b). We will call these techniques *DQ tests* and define them as a tool, algorithm, approach or strategy employed to test the adherence of a dataset to a specific DQ requirement. They serve as a means to gather evidence of a dataset's fitness for purpose in light to a specific DQ criterion. Since these tests are geared towards identifying discrete problems, they are only capable of detecting issues at the data level such as typos, erroneous formatting or outliers. This output is useful to flag low-level

problems but rarely provide enough information on their own to decide whether a dataset is fit for purpose. This is why a DQ assessment can be defined as a judiciously selected combination of DQ tests based on DQ requirements to assess a dataset's fitness for a specific analytical purpose based on domain knowledge, data science, research design and analytical tools (e.g., statistical methods, machine learning algorithms, visualizations, etc.).

Designing DQ assessments in an effective and reliable way is a challenging task because of the broadness of the question at hand: "Is this dataset good enough to yield valid results when analyzed to answer the research question?" To support this design process the literature provides frameworks that organize DQ knowledge and testing approaches (Batini & Scannapieca, 2006a, 2006b; M. G. Kahn et al., 2012; Lee et al., 2002; Madnick et al., 2009; Maydanchik, 2007a). One example is Wang & Strong's conceptual framework of DQ (Wang & Strong, 1996), which describe all aspects of DQ (i.e., DQ dimensions) that may be of interest to data consumers and classifies them into a taxonomy (see section 3.2.3 for full description). This work provides a list of aspects that should be considered to when evaluating DQ but the list is too abstract to be useful in domain-specific applications (Floridi, 2013; Nicole Gray Weiskopf & Weng, 2013).

Another example is Borek et al.'s classification of DQ assessment methods (Borek et al., 2011). This framework organizes DQ testing approaches according to target DQ problems and database mapping requirements (i.e., data model subsets) as a way to support systematic selection (see section 3.2.3 for full description). This framework is helpful in three ways. First, it gives a limited list of potentially testable DQ issues and

provides clear testing approaches. Second, it helps to break down the complexity of the dataset into more manageable pieces. Third, by linking DQ issues and data pieces to a finite number of testing methods, it limits the scope of the DQ assessment design. For example, if we were to do a single variable check we could only use range checking, data validation, lexical analysis or column analysis according to this classification (Borek et al., 2011). Based on this limited set of methods it is much easier to select the correct method. However, this framework does not interface with Wang & Strong's DQ dimensions and does not provide any domain-specific support. In general, the available guidance lumps domain-knowledge-specific DQ test definition into a 'business rule definition' task that provides no clear process to follow (Maydanchik, 2007a). This setup fails to support the systematic definition of DQ tests because it provides no specific structure execute the task, leaving domain experts are to handle this task ad hoc. Also, DQ testing approaches (i.e., DQ test tools classifications) that aim to evaluate purpose-specific issues are usually represented under the umbrella terms 'Domain Analysis' and 'Semantic Profiling' (Borek et al., 2011) but fail to define them in specific terms. This guidance ultimately does not ensure the systematic and thorough definition of DQ tests to support systematic and thorough definition of purpose-specific DQ assessments.

Thus, current DQ frameworks are still difficult to use for DQ assessment design and execution for three major reasons. First, there are large conceptual gaps between DQ theory and practice (Floridi, 2013). Second, major discrepancies and incompatibilities between nomenclature and DQ frameworks have been reported (Floridi, 2013; Nicole

Gray Weiskopf & Weng, 2013). Finally, the guidance tends to be abstract in nature and lack domain-specificity (Floridi, 2013).

2.2 - Data Quality Assessment for the Secondary Use of Clinical Data

It is well known that EHRs often contain inaccurate data and that accuracy varies between sites (Hogan & Wagner, 1997). It is also broadly accepted that clinical data are incomplete (N. Weiskopf et al., 2013). This has been partially attributed to variable execution in data entry workflows as well as limited integration between healthcare institutions (Finnell, Overhage, & Grannis, 2011; Parsons, McCullough, Wang, & Shih, 2012), so EHR data may not tell the patient's whole story. Also, some data contained in the record may not be easily accessible (e.g., clinical notes) (George Hripcsak et al., 1995). The data may not be recorded at regular intervals or at a satisfactory rate for the secondary analysis (N. Weiskopf et al., 2013). Since EHR data is not intended for research, coding may not be sufficiently complete or accurate for research purposes (Bernstam, Herskovic, Reeder, & Meric-Bernstam, 2010; George Hripcsak, Knirsch, et al., 2011). Ultimately, the root of the problem seems to be that clinical data are created as a byproduct of clinical practice and therefore, may not follow the same production quality standard as research staff would for a clinical research project (Hersh et al., 2013). All these issues are symptoms of poor DQ with respect to research. Therefore, it is crucial to carry out thorough assessments to detect DQ issues before analysis and to consider identified issues when interpreting the analysis results.

Current DQ assessment methods have several limitations in the field of biomedical informatics. A recent systematic literature review (Nicole Gray Weiskopf & Weng, 2013) found that they are not systematic or generalizable and fail to adopt the preferred 'fit-for-purpose' approach (Holve et al., 2013). They also fail to support the transparent reporting of DQ assessments results (M. Kahn et al., 2015). Even though general DQ testing approaches found in the literature can be applied to the secondary use of data (Borek et al., 2011; Maydanchik, 2007b), their disparate nature and data-level (as opposed to purpose-level) focus makes them inappropriate to support the evaluation of fitness for purpose (M. Kahn et al., 2015; M. G. Kahn et al., 2012; Wang & Strong, 1996; Nicole Gray Weiskopf & Weng, 2013). To address this limitation, Kahn et al. have developed a framework, based on existing strategies and DQ tests, (M. G. Kahn et al., 2012) that supports the assessment of DQ for repurposed clinical data. This framework provides a process to run evaluations paired to a list of DQ rule-types to be considered by database administrators and researchers to define DQ requirements combined into a standard to detect poor quality data. Quality is assessed in relationship to the defined DQ standard by flagging the data that infringes on the defined DQ requirements. This process and DQ requirement development guidance sets the foundation for a generalized DQ assessment for cross-site aggregation of clinical data. Though this initial framework increases the potential for systematization for clinical data reuse, it is heavily oriented towards the initial data production purpose (i.e., health record keeping) and does not consider an analytical reuse purpose (e.g., answering a research question). This means that this framework does not allow the user to assess fitness for purpose if the intended data use is

anything but data aggregation. For these secondary purposes, there are immensely large numbers of possible research questions to be considered, which complicates the definition of a general framework and systematic process to assess DQ.

In response to the incoherent definition of DQ testing approaches and discrepancies between frameworks, an ontology for the DQ assessment of repurposed clinical data has also been recently published (Johnson, Speedie, Simon, Kumar, & Westra, 2015). This work rigorously defines concepts to enable the automated computation of DQ measures (i.e., quantitative evidence of DQ requirement infringements within a given dataset for a specific purpose) through the application of DQ tests. It also defines relationships between DQ dimensions and 19 DQ measure types that aim to unambiguously define and catalog all possible DQ tests for the secondary use of clinical data. However, this work has several limitations in supporting the effective DQ assessment of repurposed clinical data. First, it does not provide a general process for the systematic execution and implementation of DQ assessments. Second, it fails to provide guidance as to how to develop DQ assessments that ensure a reliable evaluation of fitness for purpose. Third, it fails to bridge the gap between purpose, DQ theory and domain knowledge for the definition of the DQ requirements. This step is crucial to DQ assessments because it informs the calculation of the DQ measures through the selection and definition of specific DQ requirements that, in turn, define the DQ tests. Beyond DQ dimensions and theory, the definition of DQ requirements depends on two information sources: (1) purpose, which is difficult to model due to its great variability and (2) the domain knowledge held by the experts conducting the research, which is difficult to fully include

in an ontology. Though the ontology includes placeholders for these parameters, it fails to provide any guidance on how to integrate and use them in practice.

2.3 - Requirements for the Definition of a DQ Assessment Framework for the Secondary Use of Clinical Data

Effective DQ assessment depends on the purposeful combination of DQ tests (and, thus, the definition of DQ requirements) to assess fitness for purpose. In fields outside biomedical informatics, definition of DQ requirements has proven challenging and frameworks have been used to support the design of such assessments (Fan, 2012; Lee et al., 2002; Madnick et al., 2009; Stvilia, Gasser, Twidale, & Smith, 2007; Wang & Strong, 1996). A framework is currently available to support DQ assessments for the consolidation of multi-site clinical data into CDWs (M. G. Kahn et al., 2012). However, there is no framework to support DQ assessments for secondary analyses of clinical data. Given the current state of the science, such frameworks should at least (1) provide a generalizable and systematic method for assessing DQ consistently across datasets and purposes (Nicole Gray Weiskopf & Weng, 2013) and (2) support the purpose-specific assessment of repurposed data (Holve et al., 2013). In other words, the framework should encourage the definition of requirements and tests that evaluate the appropriateness of datasets for the research question at hand. Moreover, the necessity of a systematic approach to these assessments requires a unified yet general way to define and implement DQ assessments. Therefore, the framework must provide an unambiguous process to define and execute DQ assessments.

One additional limitation of current clinical data reuse practices is the tendency not to include detailed DQ assessment results in publications (N. Weiskopf et al., 2013; Nicole Gray Weiskopf & Weng, 2013). To address this, DQ assessment result reporting guidelines have been published (M. Kahn et al., 2015). These aim to promote transparency in published analyses through explicit reporting of employed DQ assessment methods and DQ results. The ultimate goal is to ensure a deeper, more precise understanding of repurposed clinical data limitations and, in turn, the analytical results. Defining all parameters and assumptions of the DQ assessment *explicitly* would facilitate the transparent reporting of DQ results in three ways. First, it would promote the unambiguous definition of DQ features to test. Second, it would promote the organized development of the DQ test lists, which would then facilitate the conversion into a publishable format. Lastly, using explicit DQ assessment documents would structure communication within the research team.

Thus, a framework that addresses the current limitations of DQ assessments for the secondary use of clinical data should have the following characteristics:

- Supports purpose-specific DQ assessment
- Provides a DQ assessment process that is:
    - Generalizable to a wide range of secondary use cases
    - Systematic (i.e., executed according to a fixed sequence of steps)
- Makes DQ requirements *explicit* in order to:
    - Improve communication within the research team
    - Promote transparent reporting of DQ issues

Chapter 3: DataGauge - A Model-Driven Process for the Systematic Assessment of Repurposed Clinical Data

In this chapter, I present DataGauge, a generalized and systematic procedure for the analysis-specific assessment of DQ for repurposed clinical data that addresses the limitations in the state of the science reviewed in Chapter 2. I also present an example showing its uses and advantages. Finally, I discuss DataGauge's strengths and limitations.

DataGauge consists of three stages: (1) Scope definition, (2) DQ specifications development and (3) Data processing according to these DQ specifications. DataGauge provides specific steps that can be applied consistently across analyses to promote the systematic assessment of DQ. It supports the purpose-specific assessment of DQ by generating DQ requirements and documentation specific to a particular research question. DataGauge relies on explicit standards and documentation, which promotes collaborative analysis by providing tools to structure communication within the analytics team and in published results. It also allows an iterative approach where the analytical scope and DQ standards are improved as the research progresses. Finally, DataGauge facilitates linking of DQ requirements to available assessment methods.

3.1 - Adapting a Model Driven Quality Assessment Process to Clinical Data

Quality assessment methods are widespread in many domains outside biomedical informatics and usually address one or more of the needs stated in Chapter 2. Basic quality assessments rely on qualitative evaluations (e.g., satisfaction surveys), that provide measures of perceived quality (Nelson & Niederberger, 1990). This type of assessment is generally *purpose-driven and developed based on a generalized set of guidelines* (Gómez, 2009) to ensure validity. However, such approaches have a tendency to produce ad-hoc evaluations rather than systematic assessments. To counteract this, standards organizations such as the ISO have defined quality control standards (Walker & Gee, 2000) and methodologies (Dale, 2015; Evans & Lindsay, 1999; Juran, 1962; Taguchi, 1986) that require the definition of *quantitative requirements and a systematic approach* to test them. These standards require *explicit design documentation* that defines quality requirements to be met by the evaluated product. One particularly interesting research field that resulted from the creation of these engineering standards is model-driven engineering (Schmidt, 2006). This field focuses on developing methods to support the explicit definition of formal requirements and their automatic evaluation. These model-driven methods enable the *systematic, generalizable and purpose-driven quality assessments of software products based on explicitly defined quality requirements*. Thus, this approach addresses similar challenges to those described in Chapter 2 for the quality assessment of software products. However, it has not yet been adapted to assess the DQ of repurposed clinical data, nor evaluated.

Model-driven software development and quality assessment is a well-developed branch of software engineering (Boytsov & Zaslavsky, 2013; Jordi Cabot, 2012; France & Rumpe, 2007; González & Cabot, 2014; Mayrand & Coallier, 1996; Whittle, Hutchinson, & Rouncefield, 2014). These similarities offer a unique opportunity to adapt these methods to the DQ assessment of repurposed clinical data. Beyond addressing the limitations of the current state of the science, translating these methods is advantageous because model-driven quality assessment methods follow the standards for systematic product quality control (Evans & Lindsay, 1999; Juran, 1962; Taguchi, 1986), which has proven useful in other fields. The adaptation of these methods is likely to be viable for two reasons: (1) Wang has shown that data can be evaluated for quality just like any other product (Wang, 1998) and (2) experimental model-driven approaches to data validity checking have been reported as successful in the context of structured data using finite state models (Mezzanzanica, Boselli, Cesarini, & Mercorio, 2011).

To define such a process I assessed the commonalities between model-driven software quality assessment methods. The methods shared three high levels stages: (1) Evaluation of needs and scope definition (France & Rumpe, 2007; Kan, 2002; Mayrand & Coallier, 1996), followed by (2) Explicit modelling of product specifications (i.e., the quality requirements) based on the needs (France & Rumpe, 2007; Kan, 2002; Mayrand & Coallier, 1996; Mezzanzanica et al., 2011; Whittle et al., 2014) and, finally, (3) Evaluation of the product based on the previously-defined requirements (Boselli, Cesarini, Mercorio, & Mezzanzanica, 2013; France & Rumpe, 2007; Kan, 2002; Mayrand & Coallier, 1996; Mezzanzanica et al., 2011). These three steps can be easily

adapted to the secondary use of clinical data as the following stages: (1) A data needs assessment that serves as a *definition of the data scope* and should include the analysis of the *research question* and a definition of the resulting data needs, (2) The specification development, which would include the *specification of the data needs in an explicit model* as well as the *definition of DQ requirements* and, finally, (3) the evaluation, which would entail the *assessment of the data* according to the DQ requirements.

At the heart of this process lies the definition of explicit models to represent DQ requirements. Multiple languages are used to describe such requirements in model-driven software quality assessment. For example, Universal Modelling Language (UML) and the Object Constraint Language (OCL) are routinely used to define software requirements. (J. Cabot, Clariso, & Riera, 2008; Jordi Cabot, 2012; Jordi Cabot & Gogolla, 2012; Demuth & Hussmann, 1999; Pinet et al., 2011; Selic, 2004; Zubcoff, Pardillo, & Trujillo, 2009). UML is also routinely used to describe databases and data models in practice through its entity-relationship diagrams (Selic, 2004). OCL is designed to fully integrate with UML and provides an additional layer of constraints on data models (Jordi Cabot & Gogolla, 2012). The combination of these two languages is a viable way of encoding DQ requirements in a standardized, unambiguous way, resulting in a unified DQ assessment specification model-based document.

To refine this initial process for the DQ assessment of repurposed clinical data I carried out three assessments of repurposed clinical data. The end result was a generalized process for the systematic DQ assessments based on UML entity-relationship diagrams

and OCL constraints that I have named DataGauge. The process is presented in the next section.

3.2 - The DataGauge Process

3.2.1 - Method overview

Figure 1 illustrates the three stages of the analysis-specific DQ assessment method called DataGauge. The stages are: (1) Data Need and Scope definition, (2) DQ specifications development, and (3) Data processing according to these specifications. These three stages are composed of five steps: (1) Define needs based on the research question and analytical study design, (2) Develop a data needs model (DNM) where we formalize the data needs, (3) Develop analysis-specific DQ requirements based on the analytical purpose, the DNM and the dimensions of DQ, (4) Extract data from the source dataset to fit the DNM, and (5) Evaluate the extract according to the DQ requirements where we flag all data that infringes on the DQ assessment standard.

DataGauge is a guide for DQ assessments that applies to most secondary uses of clinical data. It is designed to be carried out collaboratively by a team of domain experts (e.g., clinicians), data users (e.g., researchers, clinical personnel, etc.), informaticians, statisticians, and database administrators (Barlow, 2013). Such a team ensures input from all relevant perspectives. The team is expected to iterate over these steps several times to refine the specifications as the research progresses.

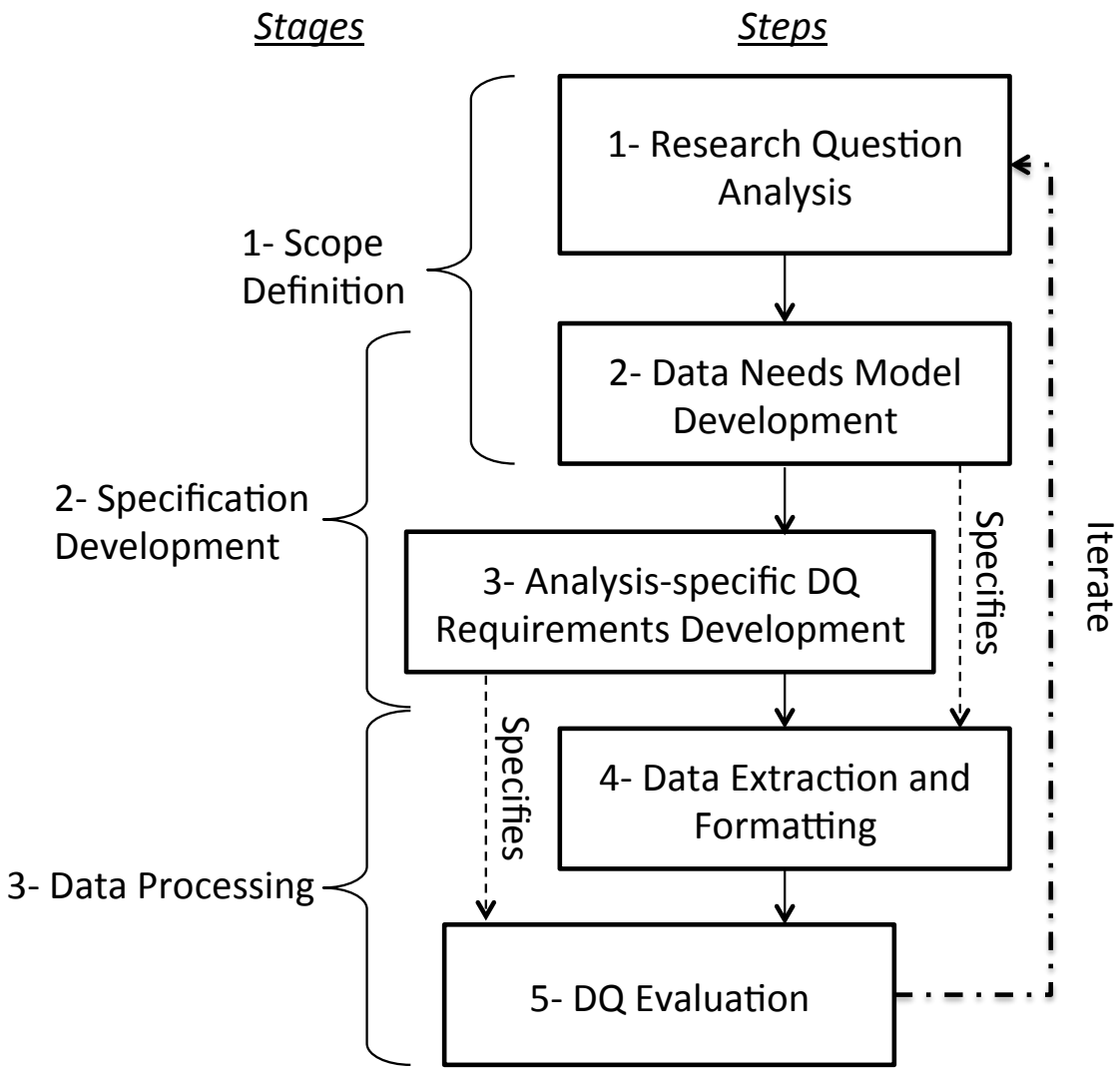


Figure 1 – Iterative analysis-specific DQ assessment method for the secondary use of clinical data. This process defines the general stages and steps for analysis-specific DQ assessment using data models and an analysis-specific DQ standard.

3.2.2 - Stage 1: Data Scope Definition

The initial stage defines the scope of research in terms of data. This is done in two steps.

Step 1: Research Question Analysis

First, the domain expert and statistician must define the research question and analytical design. This dictates the data needs. Objects of relevance to the analysis are identified (e.g., patients, prescriptions, diagnoses, etc.) along with independent and dependent variables based on the research question (Tabachnick & Fidell, 2001). I define data needs as all variables and metadata necessary to achieve the analytical goal (i.e., answering the research questions). For example, if we are studying the relationship between weight and age, the domain expert will define a person object that will contain the necessary patient demographic variables (i.e., weight measurements and age) plus all additional covariates to be selected based on their analytical needs and clinical domain knowledge.

Step 2: Data Needs Model Development

The second step is to build a Data Needs Model (DNM) that defines the ideal analytical dataset. We define the DNM as a fully-specified, explicit representation of all data needs for the analytical purpose, including the relationships between data elements. For example, if we are investigating relationships between weight and age, a simplistic DNM may be defined as a single table containing 'Subject ID', 'Weight' and 'Age' variables, where each line represents a weight observation. The model serves as a design specification document and an unambiguous means of communication among team members. This model defines the scope based on analytical requirements. This marks the end of stage 1 that ensures agreement on the scope of work, the necessary data and its

context for a specific analytical purpose. Step 2 continues into the first part of stage 2, described next.

3.2.3 - Stage 2: Specification Development

The second stage consists of iteratively refining the DNM along with DQ requirements that, if met, would ensure fitness for purpose. These two elements fully specify a DQ assessment for a clinical data source and a specific analytical purpose. Step 2 concludes when the analytics team has a fully defined and is satisfied with the DNM version. The qualities of a satisfactory DNM are difficult to define generically because they heavily depend on the purpose. The experts in the analytics team must, therefore, decide when the DNM is ready. It is important to note that it may be useful to run additional iterations of the DataGauge process after the initial evaluation. The DNM can then be revised and further refined.

This step is crucial for several reasons. First, it clearly defines the assessment scope. Second, it defines the object(s) of analysis fully and explicitly. Third, the DNM defines the variables and their relationships, but also assigns clinical meaning to them by grouping them into objects that make clinical sense such as patient, prescription and measurements. This provides a link between the source data, the analytical purpose and clinical domain knowledge. Revisiting the weight and age relationship example, the 'Patient ID' would be an identifier with no specific clinical meaning, but 'Weight' would be tied to an observation that is associated with an office visit in the medical record. The additional clinical workflow information allows the analytics team to relate the weight

variable to clinical domain and workflow knowledge. Defining the DNM also supports the generation of DQ requirements by outlining a finite DQ evaluation space. For example, if there are no numerical variables in the data model, the team will not have to set numerical thresholds for range checks. To ensure systematicity, the DNM design should also be in at least third normal form (Kent, 1983) or, equivalently, follow a tidy data format (Wickham, 2014). Although equivalent, we prefer the tidy data standard, because it defines the data format in terms that are easier for analysts and researchers to understand. These forms specify the observational units (modeled as different tables), the variables within a unit (columns of a table), and the observations of each unit (rows of a table), that allow the clear definition of data needs. An example of tidy-data-compliant DNM is provided in the next section. A clear specification of these three elements is required prior to beginning step 3.

Step 3: Analysis-specific DQ requirements Development

Once the analytics team has fully defined the DNM, the third step is the definition of an analysis-specific DQ standard composed of DQ requirements. I define the DQ standard as a document containing the full set of individual DQ requirements that fully describe a fit-for-purpose dataset for a specific research question and DNM. This document allows the analytics team to explicitly develop and agree on DQ for a particular case. The development of DQ requirements is task is complex because it requires the integration of multiple information sources (e.g., the DNM, the research question, DQ theory). To tease out this complexity, I first lay out the theoretical basis for the DQ requirement development in the next paragraphs as follows: (1) define DQ requirements by

differentiating them from inclusion/exclusion criteria, (2) explain how developing tidy-data-compliant models helps the analytics team to tease out the DNM's complexity and make the variables of interest more accessible, (3) I then explain how the levels of data granularity (Oliveira, Rodrigues, & Henriques, 2005) contribute to a thorough definition of DQ requirements and lastly (4) I explain the role of the DQ dimensions (Wang & Strong, 1996; Nicole Gray Weiskopf & Weng, 2013) and their integration into the DQ requirement generation task. The last paragraph of this section describes the DQ requirement development procedure.

I differentiate DQ requirements from inclusion/exclusion criteria by the object they define and their goal. Inclusion/Exclusion criteria focus on the main object of interest to the research question (e.g., patients, encounters, visits, prescriptions, etc...). They aim to define the features that qualify or make these objects unacceptable for the study. These criteria aim to define a cohort or population. For example, if patients are the objects of study, inclusion/exclusion criteria will be based on demographic and clinical considerations. The object of analysis will usually be a patient for clinical research question but may also be encounters, visits, clinical notes, etc. for quality improvement and other projects. On the other hand, DQ requirements define the minimum expectations of data to ensure that a specific dataset is valid and useful for a specific analytical purpose. They aim to define a fit-for-purpose dataset. For example, "patient is at least 18 years old" is an inclusion criterion whereas "patient date of birth is earlier than observation date" is a DQ requirement. Each DQ requirement corresponds to a DQ test that will be carried out. For example, if our secondary analytical purpose included

assessment of weight change over time, a requirement might be 'patients must have at least two weight measurements'. To test this, I would check that there are enough patients with at least two weight measurements in the dataset for the secondary analysis (i.e., the dataset has an adequate sample size). If the condition is not met, the dataset is not fit for this particular analysis.

DNMs often describe multiple aspects of a phenomenon, which can yield complex datasets. For example, a dataset for the secondary use of clinical data may contain cohort identification variables (i.e., variables that define patient characteristics and support their classification), outcome definition variables, exposure variables and relevant covariates (M. Kahn et al., 2015). To address this complexity we propose two strategies: the development of DNMs in the tidy data format (Wickham, 2014) and the decomposition of the DNM into levels of data granularity, described in the next paragraph (Oliveira et al., 2005). The requirements for a tidy dataset are "*each variable reside in a column*", "*each observation is presented in a row*" and "*each type of observational unit is a table*". In such form it is easy to isolate meaningful segments of the data model that correspond to variables and elements relevant to purpose. For example, let us build a DNM describing the necessary data to examine the relationship between patient demographics and patient weights over time. We may begin with a dataset that describes patients and their weight measurements and get the data in non-tidy-data-compliant formats; for instance, one line per weight measurements with all patient data attached to every line or one line per patient with multiple columns for weight measures. Neither of these forms is tidy-data-compliant because they combine two observational units in one table: patients

and weight measurements. Each observational unit must have a separate table. This means that we must necessarily have one table for patients and we must have a table for weight measurements. Also, *each observation must be recorded as an individual row*; this means that the database must contain one line per patient in the patient table and one line per weight measurement in the weights table. Finally, tidy-data standards require that *each variable must be recorded in a distinct column*. This means that the weights table must record the patient identifier, the weight value, the unit, a timestamp and any additional information as distinct columns. The same applies for the patient table where each column would correspond to a demographic variable. In the end we would have two tables: a patient table with all demographic variables (e.g., gender, date of birth, etc.) and a table with weight values and additional variables as described above. These two tables should be related to each other via a patient identifier, which plays the role of a primary key to the patient table and a foreign key to the measurements table. This data format makes patient demographics and weight measurement accessible for analysis and DQ assessment.

A model with accessible variables also allows the team to view the DNM as a combination of data elements along a scale of data granularity (Oliveira et al., 2005). I define data granularity as the different data levels at which the objects of interest (i.e., patients, outcomes, drug exposures, etc.) can be encoded into the DNM. Some examples of these levels are: single value, multiple values, observation, observational unit and dataset. The benefits of providing a way to break down the DNM into simpler data elements are twofold. First, it reduces the complexity of the elements to assess. Second, it

allows the use of DQ test selection guidance. Borek et al. (Borek et al., 2011; Oliveira et al., 2005) have mapped these levels of data granularity to DQ test methods (i.e., ways to check the data based on DQ requirements). This mapping reduces the number of possible tests to be applied by giving the DQ assessment designer a limited number of testing approaches to choose from. In essence, this work maps usual DQ problems (e.g., missing values, inconsistent data formats, incorrect values, etc.) to DQ testing approaches (e.g., range checking, data validation, lexical analysis, etc.), classifying their usefulness by data granularity level (e.g., single value, multiple values, observation, etc.). This is a good starting point for a systematic DQ assessment design strategy because it tasks the analytics team with reviewing a finite number of data elements, for a finite number DQ testing strategies. Its core weakness is that it does not map to DQ dimensions.

To define an assessment that evaluates fitness for purpose, it is necessary to account for all dimensions of DQ (Wang & Strong, 1996; Nicole Gray Weiskopf & Weng, 2013). The theoretical dimensions have been formally defined by Wang & Strong in their conceptual framework of DQ (Wang & Strong, 1996) based on aspects that may be important to data consumers. They fall into four distinct categories: (1) Intrinsic DQ refers to the qualities that the data should have regardless of their purpose, (2) Contextual DQ focuses on the qualities that the data should have, based on the purpose for which they will be used, (3) Representational DQ includes data modeling and information display, and (4) Accessibility DQ focuses on having data that are readily available to be processed and yet accessible by authorized users. Intrinsic DQ encompasses *accuracy, believability, objectivity and data source reputation* (Wang & Strong, 1996). These

dimensions relate closely to the initial purpose for which the data were produced rather than the secondary analytical purpose. On the other hand, contextual DQ relates to the purpose for which the data are to be used. Dimensions of *completeness, relevancy, timeliness, quantity of data and added value* in the context of the intended purpose constitute contextual DQ. These dimensions are particularly useful when assessing the fitness of a specific dataset for secondary analysis. DQ requirements should be built around these dimensions keeping the analytical purpose in mind. In practice, the assessment should also include data checks to ensure the *accuracy and believability* of the repurposed dataset in case preliminary DQ assessments of primary purpose failed to catch errors that may adversely impact the analysis. When combined with the DNM at different levels of data granularity and the analytical purpose, these dimensions are the key to a comprehensive analysis-specific DQ assessment. This will be illustrated in the following section.

Step 3 consists of reviewing the data model at all levels of data granularity to define the DQ requirements in terms of DQ dimensions and the analytical purpose. These requirements are assigned to specific DQ assessment methods (Borek et al., 2011) according to potential DQ issues and the level of data granularity (Oliveira et al., 2005). The team surveys the DNM in light of DQ dimensions to define the minimum standard of quality for a particular analysis. For example, if studying the relationship between patient weight and age, the domain expert should define criteria relating to the plausibility of the weight variable such as "weights are positive numbers", but also relating to the completeness of the dataset such as "each patient should have at least one weight

measurement" and so on. Each requirement will be mapped to a specific DQ checking method based on the assessed DQ dimension and data granularity level. We provide a table linking DQ dimensions, data granularity and DQ testing approaches to facilitate this (Table 1). We built this guidance table as a combination of Borek's et al.'s (Borek et al., 2011) classification of DQ testing approaches and the DQ dimensions relevant to the secondary use of clinical data defined by Weiskopf et al. (Nicole Gray Weiskopf & Weng, 2013) (i.e., correctness, completeness, concordance, plausibility and timeliness). I included an additional 'representation' dimension to assess the issues of data transformations and fit of the available data to the DNM specifications. This dimension accounts for database design considerations being respected in the dataset (e.g., primary and foreign key checks).

Table 1 - DQ requirement development guidance table. This table links DQ dimensions, levels of data granularity and DQ testing approaches as a way to provide an overview or 'menu' of the testing strategies at the disposal of the research team.

	Data Quality Dimensions				
Data Granularity Levels	***Correctness and Plausibility***	***Completeness***	***Concordance***	***Representation***	***Timeliness***
Cell/Value	Domain analysis, Data Validation, Lexical analysis	Domain Analysis, Lexical Analysis	Domain Analysis	Column Analysis, Lexical Analysis, Schema Matching	Domain Analysis
Column/Variable	Column Analysis, Data Validation, Semantic Profiling	Column Analysis, Domain Analysis	Column Analysis, Data Validation	Column Analysis, Schema Matching	Column Analysis, Domain Analysis
Line/Observation	Domain Analysis, Semantic Profiling	Domain Analysis, Semantic Profiling	Domain Analysis, Semantic Profiling	Domain Analysis, Schema Matching	Domain Analysis, Semantic Profiling
Table/Observational unit	Domain Analysis	Domain Analysis, Column Analysis	Column Analysis, Semantic Profiling	Schema Matching	Semantic Profiling, Domain Analysis
Multiple Tables/ Dataset	Semantic Profiling, PK/FK analysis, Column Analysis	Domain Analysis, Semantic Profiling	Domain Analysis, PK/FK Analysis, Semantic Profiling	Column analysis, PK/FK Analysis, Semantic Profiling, Schema Matching	Semantic Profiling, Domain Analysis
Multiple Databases/ Multiple Datasets	Semantic Profiling, Domain Analysis, Column Analysis	Domain Analysis, Semantic Profiling	Semantic Profiling, Domain Analysis	Column analysis, Schema Matching, Semantic Profiling	Semantic Profiling, Domain Analysis

3.2.4 - Stage 3: Data Processing

This last stage uses the DNM and analysis-specific DQ standards to extract, format and assess the data.

Step 4: Data Extraction and Fitting

The fourth step uses the DNM to guide the data extraction from the original database and to fit the data into the format defined in the specifications. The database administrator creates a schema with tables matching the DNM then loads the source clinical data into the tables. This schema should have all database rules such as variable type definitions, primary key rules, table relationship rules and other data validation triggers built in. Using this predefined schema to load the extracted data ensures that the values match the

agreed upon data model, variable types and database relationships. This step is an initial representational DQ test; if the data are not in the right format or variable types do not match, the database software should produce an error.

Step 5: DQ Evaluation

The fifth and last step consists of evaluating DQ based on the previously defined DQ requirements. Appropriate DQ test methods (Maydanchik, 2007b) are selected and implemented to test each DQ requirement. This process evaluates DQ standard compliance and flags discrepancies. These flags allow further analysis, data diagnosis and imputation (Van den Broeck et al., 2005). Several indicators (i.e., DQ measures) can be calculated from these flags as measures of DQ (e.g., compliance percentage for each variable or patients with no data flaws divided by the total number of patients). These results provide quantitative evidence of non-compliant data and can serve as a basis for experts to judge the fitness for purpose.

3.3 - Example

We used DataGauge to assess DQ for a repurposed clinical dataset and address the challenges of analysis-specific DQ assessment. The analytical purpose was to determine whether prednisone, a commonly-prescribed corticosteroid, is associated with weight gain. We chose this association because weight gain is a known and clinically-significant side effect of prednisone (PredniSONE Tablets [Package Insert], 2012) that is likely to be detectable through retrospective review of clinical data. Our data source was a CDW

containing routinely recorded clinical data from six academic outpatient clinics in a large metropolitan area in the southern United States.

We used a UML-based database modeling tool (MySQL workbench data modeler; Oracle Corp., Redwood Shores, CA) to develop the DNM. A team composed of a clinician, a statistician and an informatician (the author), who also played the role of database administrator, developed the final UML diagram for the research question (Figure 2c). A series of models were iteratively created and discussed according to their ability to satisfy the analytical purpose as well as data availability in the CDW. Note the changes that led to the final data model in Figure 2 and the variations in DQ requirements per iteration in Table 1. Figure 2 shows variables that were removed because they were unavailable in the CDW (e.g., no drug exposure variable was found or could be reliably calculated from our CDW data). Figure 2 shows the evolution of the DNM from (a) single-table format that is not tidy data-compliant into (b) a tidy data compliant model with four observational units (i.e., Patient, Visit, PrednisonePrescription and Weight). The final DNM (c) improves on the tidy-data compliant model by removing the Visit observational unit, which is not directly relevant to the research question, and adapted the model to the data available in the CDW (e.g., changes in the variables describing the prednisone prescription). Our final model conforms to the tidy data guidelines (Wickham, 2014) providing one observational type per table (i.e., patients are represented as a separate table from weights and prescriptions), each line represents an observation (i.e., each line on the patient table a different patient, and in the prescription table a different prescription) and each variable corresponds to a column (e.g., patientID, gender and DoB

are separate variables in the patient table); this data format also corresponds to the third normal form (Kent, 1983). The final DNM served as a data specification document to guide the data extraction. Database tables were created to match the DNM and the raw data were extracted from the source database into the DNM schema using standard SQL queries. DQ requirements were defined in the form of Boolean expressions and Object Constraint Language constraints (Jordi Cabot & Gogolla, 2012). We chose OCL due to its integration with the UML diagrams previously used for the data models (Demuth & Hussmann, 1999; J. Cabot, Clarisó, & Riera, 2014; Seiter, Wille, Soeken, & Drechsler, 2013).

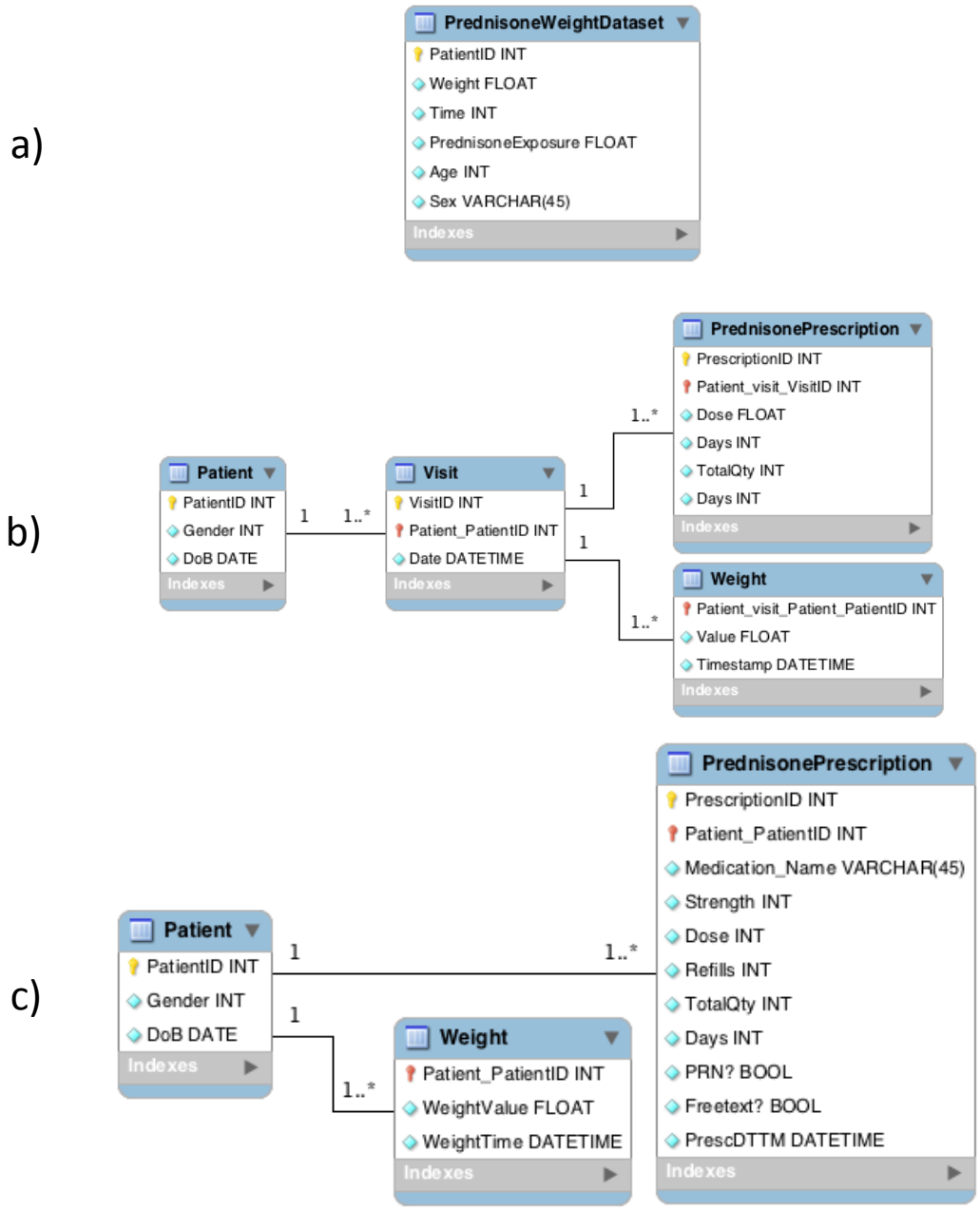


Figure 2 - Evolution of the data needs model for the purpose of assessing a relationship between prednisone and weight gain using repurposed clinical data. This data model defines the data needs for the evaluation of an association between prednisone and weight gain. a), b) and c) show the three versions of the DNM for the each iteration.

We generated DQ requirements iteratively and collaboratively. Each DQ dimension was surveyed at different levels of data granularity (e.g., single value, multiple values, observation, observational unit, dataset, etc. (Oliveira et al., 2005)) running through all variables of the DNM and using the overview provided in Table 1 to guide the process. For example, when we combined the accuracy dimension with the single value level for the final DNM we came up with requirements such as "Dose must be positive" or "Refills must be positive or 0"; both of these requirements were mapped to a range checking method. The concordance dimension at the observation level yielded criteria such as "the prescription date should be later than the patient's date of birth" which was mapped to the semantic profiling DQ check method. At the observational unit or table level we assessed the timeliness of the data with the "Patient has a second weight measurement within 4 months of the first prescription" requirement. This requirement was also mapped to the semantic profiling check method. DQ requirements were generated until the analytics team was satisfied with the DQ standard. We used the DQ requirements to evaluate the quality of the extracted data based on the third version of the DQ standard. We covered analysis-specific DQ requirements as well as generic requirements to test accuracy and believability of the data. Of 52 requirements, 17 were analysis-specific. Analysis-specific requirements tended to be more complex and concern a larger number of variables. Table 2 shows how the requirements evolved over iterations; note the increasing precision and analysis-specificity (e.g., "2 values per patientID" in iteration 2 followed by "50% patients with 2 weight measures within 4 months of first prescription"). Each new DNM represented a specific data model designed to satisfy the same analytical purpose; each

iteration for the DQ requirement created an increasingly complete, refined and analysis-specific set of requirements.

Table 2 - Data quality requirement examples. The table shows DQ requirement examples as they were generated. The requirements became more specific and analysis-specific as the development progressed.

Iteration	*DQ Dimension*	*Variable Granularity*	*Variable(s)*	*Analysis Specific?*	*Requirement*	*DQ assessment method*	*DQ Result (% compliance or Pass/Fail)*
1	Accuracy	Value	Gender	No	In {'M','F','U'}	Data Validation	99.99
	Accuracy	Value	WeightValue	No	>0	Range Checking	92.65
	Believability	Value	WeightValue	No	<400	Range Checking	99.95
	Accuracy	Value	Strength	No	>0	Range Checking	97.37
	Believability	Value	Strength	No	<2*[Max dose]	Domain Analysis	100
	Accuracy	Value	Dose	No	>0	Range Checking	51.68
	Believability	Value	Dose	No	<2*[Max pills at min strength]	Domain Analysis	100
	Accuracy	Value	Refills	No	>=0	Range Checking	100
2	Accuracy	Value	WeightTime	No	>[System Installation Date]	Data Validation	100
	Accuracy	Column	PatientID	No	Unique	Column Analysis	100
	Concordance	Line	WeightTime, DoB	No	Timestamp>DoB	Domain Analysis	100
	Concordance	Line	PrescDTTM, DoB	No	PrescDTTM>DoB	Domain Analysis	100
	Concordance	Table	PatientID, WeightTime, WeightValue	Yes	Patient weights on prescription date are less than 2% apart	Domain Analysis	92.45
	Completeness	Table	PatientID, WeightValue	Yes	2 weight measurements per patient	Domain Analysis	85.92
	Completeness	Line	PatientID, WeightTime	Yes	Patient has weight measurement on prescription date	Domain Analysis	97.54
	Timeliness	Table	PatientID, WeightTime	Yes	Patient has second weight measure within 4 months of prescription	Domain Analysis	48.62
3	Amount of data	Table	Strength, Dose, Days, Refills	Yes	Can calculate total milligrams prescribed for 50% of prescriptions	Domain Analysis	Failed
	Amount of data	Table	Patient, PRN	Yes	Less than 25% PRN prescriptions	Domain Analysis	Passed
	Amount of data	Dataset	PatienID, WeightTime	Yes	50% patients 2 weight measures within 4 months of first prescription	Domain Analysis	Failed
	Completeness	Dataset	PatientID, WeightValue, WeightTime, PrescriptionTable	Yes	Patients with at least 2 unflawed weights after an unflawed prescription	Domain Analysis	13.1
	All	Dataset	All Variables	No	Patient records with no general DQ flaw	Domain Analysis	2.93

The DQ tests revealed several DQ flaws. We were able to identify specific DQ issues such as inaccuracies (e.g., 84 weight values were above 400kg), inconsistencies (e.g., 56 instances where weight changed more than 20% over 2 days) and incompleteness (e.g., 43,135 patients with less than two weight measurements within 3 months of the prescription). This showed the approach's effectiveness at catching DQ issues and screening data at the basic data level. We also excluded 14.1% of the patient records as

they contained a single weight measurement and weight gain can only be calculated with two or more. We flagged all data items that violated DQ criteria and then calculated the number of patients with no flagged data in their records, having at least two weight measurements after their first prednisone prescription. Thus, only 2,379 patients out of 80,990 (13.1%) could be confidently used for analysis. The massive censoring of patient records is likely to bias in the final dataset used for analysis and potentially render subsequent analytical results unreliable. Also, in this particular case we are looking at a commonly prescribed drug and a broad population. However, large clinical datasets are often used to investigate features of rare diseases and specific cases for which few patients records may qualify. This high level of censoring could drastically reduce the sample size to levels inadequate for secondary statistical analyses.

This example illustrates how DataGauge can advance current practices in DQ assessment for the secondary use of clinical data. First, DataGauge provides an analysis-specific DQ assessment method. We showed a way to define DQ requirements based on the DNM, which is dictated by the analytical purpose, but also to define them based on the intended purpose. For example, "2 values per patientID" is a DQ requirement that depends on the DNM, but also relevant to the analytical purpose because two weight measurements are necessary to detect change over time.

DataGauge is general because it can be applied across cases by generating new analysis-specific DNMs and DQ standards. Second, DataGauge provides guidance for DQ standards and tool selection by allowing a DNM to be decomposed into multiple pieces. Paired to potential DQ issues, these pieces can be mapped to specific DQ test tools and

standards (Borek et al., 2011; M. Kahn et al., 2015). In surveying the DNM at multiple levels of granularity we were able to identify the DQ requirements. The levels of data granularity informed method selection by explicitly listing and reducing number DQ tests that could be applied. For example, if we were to do a single variable check we could only select from range checking, data validation, lexical analysis or column analysis (Borek et al., 2011). Based on this limited set of methods and the actual requirement it is much easier to select the correct method.

The DNM in the tidy data format allows the research team to clearly identify variable types (e.g., cohort definition, exposure, covariates, etc.) to organize their requirements and results to match current DQ reporting standards (M. Kahn et al., 2015). Finally, the application of DataGauge provides explicit documentation (i.e., a DNM and a DQ standard) that can structure the communication among team members. These documents can be evaluated, discussed and improved as the work progresses. The team can focus on analytical needs and research goals to develop a model that is then refined by the constraints of data availability.

3.4 - Conclusion

I have presented DataGauge, an iterative team-based method to carry out analysis-specific DQ assessments for the secondary use of clinical data. DataGauge requires five steps: (1) Define needs based on the research question and analytical study design, (2) Develop a DNM where we formalize the data needs, (3) Develop analysis-specific DQ assessment requirements based on the analytical purpose, the DNM and the dimensions

of DQ, (4) Extract data from the source dataset to fit the DNM, and finally (5) Evaluate the extract according to the DQ requirements. DataGauge addresses limitations in the state of the science of analysis-specific DQ assessment for the secondary use of clinical data by providing a systematic and analysis-specific approach. DataGauge is designed to be a general DQ assessment process, as its steps can be applied to any dataset and analytical purpose. It is purpose-specific because the first two stages are dedicated to capturing the complexities of the research question at hand and developing assessment documents to fully specify a purpose-specific DQ assessment. These documents describe the assessment's assumptions and parameters explicitly. Finally, DataGauge supports systematicity because it allows the consistent definition and implementation of DQ assessments. Variables of interest and DNMs are defined in a systematic way; however, a systematic definition of DQ requirements would require further support due to the its complexity (i.e., combining research question, DMN, DQ theory and domain knowledge to define requirements).

We have provided preliminary guidance for the development of a well-defined DNM and analysis-specific DQ requirements. This guidance combines knowledge from the literature about the link between DQ dimensions relevant to the secondary use of clinical data, levels of data granularity and DQ testing approaches. It provides an overview of the general DQ dimensions to consider, data granularity levels that support the decomposition of the DNM into more manageable pieces and provides DQ testing approaches for each DQ dimension-data granularity level combination. This guidance is useful because it provides a finite set of aspects to consider when developing DQ

requirements for a specific purpose and maps them to specific testing approaches. However, *it was difficult to determine whether the definition of DQ requirements had covered all relevant aspects. Ensuring this is critical because the efficacy of the DQ assessment at catching DQ issues is dependent on the comprehensiveness of the DQ standard*. We have provided four preliminary guides to support the development and testing of DQ requirements based on previous work: 1) a tidy data format organize data models in such a way to make variables accessible (Wickham, 2014), 2) a scale in levels of data granularity to break down the DNM's complexity (Oliveira et al., 2005), 3) a mapping between levels of granularity and available DQ check methods (Borek et al., 2011) and 4) the dimensions of DQ that dictate the aspects to consider when testing fitness for purpose (Wang & Strong, 1996). DataGauge integrates prior work into a single procedure that guides the user to explicitly define their data needs and DQ requirements. However, there are still challenges in defining DQ requirements comprehensively. We attribute this to three reasons: (1) The DNM, the data granularity levels and the DQ dimensions don't fully define the problem space to define all DQ requirements, (2) the DQ dimensions are vague (e.g., there are multiple definitions of completeness (N. Weiskopf et al., 2013)) and (3) the DQ dimensions are too far removed from clinical domain to be useful in secondary use applications. Thus, further work is needed to develop domain-specific guidance for the definition of analysis-specific DQ standards. Such work would ensure a thorough and systematic development of DQ requirements and, in turn, more reliable DQ assessments. In the next chapter, we describe the development of such guidance.

Chapter 4: A Guidance Framework for the Development of DQ Requirements

DataGauge is designed to enable systematic, purpose-specific DQ assessments by providing a general process to implement DQ evaluations across repurposed datasets and research questions. However, the systematicity of a DQ assessment can only be fully ensured if the analytics team is provided with a way to generate the DQ requirements comprehensively and consistently. In fact, DQ assessments can be defined as a judiciously selected combination of DQ tests based on DQ requirements. This means that the definition of DQ requirements is the key component of the assessment's design. Generating such requirements can be a daunting task because it requires integrating multiple information sources: the research question, the DNM, DQ theory and the domain knowledge necessary to interpret the other three elements. Though the current version of DataGauge provides some support to address this (e.g., DQ dimensions (Wang & Strong, 1996) and a classification of DQ test approaches (Borek et al., 2011)), the generation of DQ requirements remains a complex task. This complexity threatens the systematic definition of DQ requirements. Thus, further guidance to support this task is needed.

There is very little literature available to guide the development of purpose-specific requirements and the available guidance presents three key limitations: (1) the DQ requirement generation process is not described in detail (Maydanchik, 2007a), (2) the guidance provides no specific structure to address the task systematically, implying that

experts should execute this task ad hoc (Borek et al., 2011; Lee et al., 2002; Wang, 1998) and (3) purpose-specific DQ testing approaches are usually lumped under the umbrella terms 'Domain Analysis' and 'Semantic Profiling' (Borek et al., 2011), which ultimately don't provide transparent guidance.

A similar situation exists in biomedical informatics, where very little guidance exists to support fitness-for-purpose assessments (Holve et al., 2013). The literature provides three pieces of guidance. First, DQ dimensions have been adapted to the secondary use of clinical data that provide some support (Nicole Gray Weiskopf & Weng, 2013) but remain vague, ambiguous and removed from the clinical domain (N. Weiskopf et al., 2013). Second, an assessment framework that provides guidance as to what types of general checks can be performed on clinical data has been published (M. G. Kahn et al., 2012). This guidance fails to include considerations about the purpose or the domains of expertise that come into play when making secondary use of clinical data (Barlow, 2013). Finally, the DQ ontology for the secondary use of clinical data (Johnson et al., 2015) provides some insight as to which types of tests that can be used to assess fitness for purpose. Even though, providing such list of possibilities is useful, the interplay of complex information sources (i.e., dataset and domain knowledge) as well as the inherent vagueness of the research purpose and DQ dimensions still make this work unlikely to support the systematic generation of DQ requirements for a specific research question and a specific dataset.

Three key challenges result from this state of the science; they are also relevant the current guidance provided by DataGauge. First, the original guidance posed problems

because it is abstract and too far removed from the knowledge domain of application. In other words, the DQ dimensions (Wang & Strong, 1996; Nicole Gray Weiskopf & Weng, 2013), data granularity types (Borek et al., 2011) and DQ testing approaches (Maydanchik, 2007a) do not provide any guidance about the kinds of threats that affect data based on clinical domain or EHR data knowledge. For example, if we were assessing concordance in weight values within a day, and detected a 30% increase, this would clearly be a problem based on clinical knowledge (i.e., patient weights don't fluctuate that much over the course of a day) whereas this would not be a problem if we were assessing the systolic blood pressure measurement values; systolic blood pressure varies considerably over a day. Current literature provides no guidance on this type of distinction and lacks explicit structure to facilitate the detection of such issues through the integration of expert knowledge. Second, the current guidance does not provide a clear overview or list the aspects to be reviewed to ensure a comprehensive and systematic generation of DQ requirements. Though DataGauge provides a limited list of items to assess when combined with the DNM, it fails to include important dimensions such as the expertise involved in defining DQ requirements. An overview would help the analytics team track the coverage of DQ assessment aspects and potential DQ threats to be considered. Though the list of DQ criteria, data types, variables in the DNM and DQ testing approaches provided a finite list to review, it does not identify specific DQ threats in repurposed clinical data. Finally, the DQ dimensions that aim to guide the requirement generation are vague. For example, completeness could be interpreted in multiple ways (N. Weiskopf et al., 2013). If completeness was thought of as overall completeness (i.e.,

the overall number of recorded observations), DQ requirements such as "patients must have at least X measurements" could be defined. In contrast, completeness is sometimes thought of in statistical terms, a corresponding requirement could be defined as "the dataset must contain enough observations to provide adequate statistical power".

Based on these limitations, this framework should possess three features. First, it should provide a finite list of potential aspects to consider while developing DQ requirements for a specific dataset and research question. This list should serve as an overview to the DQ requirement generation process. Second, it should provide a way to bridge domain knowledge and the DQ requirement generation task. Third, it should list issues and threats that the requirements should check for in a specific way.

To develop such guidance, I did a preliminary review to define the most frequent analysis types and clinical data types employed in clinical data reuse projects. I based my work on a literature review article and a set of clinical data request tickets. I then asked a clinical expert to define research questions based on their secondary use activities. These served as use cases. The six final use cases covered 90% of the most common analyses and clinical data types revealed by the literature review and clinical data request tickets. I then applied DataGauge to these six use cases in order to answer the following question: What criteria should be considered when assessing DQ for repurposed clinical data? The resulting DQ requirements represented a broad range of issues to test for when making secondary use of clinical data. Based on the generated DQ requirements, I defined an overview to guide DQ requirement generation in the form of a checklist. This overview partitions the problem space by knowledge domains and aspects of clinical data to bridge

the gap between task and domain knowledge. I also provide a list of specific questions to guide the generation of DQ requirements within each task of the overview checklist. The questions provide requirement-level guidance to ensure the coverage of specific DQ issues that may threaten fitness for purpose.

In this chapter I describe the methodology and results of my DQ requirement development guidance efforts. First, I describe the process used to define six use cases that cover most applications and data types used in clinical data reuse projects. Then I describe the methodology that I used to develop the guidance framework based on the application of DataGauge to these use cases; a presentation of the results follows. I then describe the general guidance framework and I give an example of its use. Finally, I discuss the contributions, strengths and limitations of this work as well as future development directions.

4.1 - Defining Secondary Use Coverage

Clinical data are routinely repurposed for a vast number of applications, so to address our research question I defined a specific scope. Because covering all possible secondary uses of clinical data is a monumental endeavor, I limited our scope to uses where clinical data are reused to answer a specific research question. The set of all possible research questions that can be answered reusing clinical data is extremely large; therefore, it is necessary to sample the most representative cases. I chose to cover the most common analysis types and clinical data types. This will ensure the coverage of the most common modalities of research and the most frequently used sections of repurposed EHR data

(i.e., the clinical data types). The research question was: "What are the most common analyses and clinical data types used for secondary use of clinical data?"

4.1.1 - Methods

I used two data sources to establish a distribution of secondary analysis types and clinical data types used in clinical data reuse research. First, a systematic review by Song et al. (Song, Liu, Abromitis, & Schleyer, 2013) that summarized the secondary uses of dental care data in the literature. Dental EHR records present much overlap with general EHR so I was able to use these results as a starting point. Variables relevant to the dental healthcare were not included (e.g., caries activity and periodontics procedure data). I classified analyses by type and clinical variables used to describe the distribution. The article defined the clinical variable types. The analysis type categories were defined using a qualitative grounded theory approach. They were derived from the research questions reported for every analysis in the literature review. The resulting categories were outcomes and distribution descriptions, association detection, prevalence estimation, effect size (e.g., treatment effectiveness) and other. This gave me a preliminary idea of the most frequently used data and their intended use. Second, I reviewed data requests submitted to a CDW team. The requests were recoded as service tickets in a tracking system database for the informatics core of a major healthcare research institution and served to bridge the gap between dental data and EHR. I reviewed each request and extracted the required clinical data types as well as the intended secondary analysis type, when available. I used the same clinical data type categories as for the previous dataset adding clinical notes, which were often used in this dataset but missing in the review

article. The considered analysis types were prevalence estimation, distribution, method validation and other. A large portion of the requests dealt with electronic patient recruitment (i.e., search of patients corresponding to a specific profile in the EHR database for potential enrolment in a randomized controlled trial enrolment or clinical chart review). This category was added as a secondary use type. The frequency of each analysis and data type was then computed. To confirm the validity of the secondary analysis categories, a second biomedical informatician independently classified the dental data analyses from the systematic review and the data request tickets. I used this second classification to calculate an unweighted Cohen's Kappa measure of inter-observer reliability.

4.1.2 - Results

The review of secondary uses of dental data (Song et al., 2013) presented 60 publications. All were included in this review. Then, I reviewed 238 clinical data request tickets spanning 5 years. 106 requests were excluded because they were misclassified technical requests (68 items), lacked information on requested variables (22 items), were requests of data other than clinical data (13 items) or were duplicate requests (3 items).

The distribution of analysis types was similar for both data sources (Figures 3 and 4). We found that Association studies, Distribution & Outcomes, Prevalence Estimation and Electronic Patient Recruitment analysis types covered 91.67% of the dental data research and 96.69% of ticket cases. The inter-observer reliability (i.e. Cohen's Kappa) for the purpose classification was respectively 0.869 and 0.968 for the dental secondary analyses

and the data request tickets respectively, which denotes a high level of agreement between observers and, in turn, a reliable classification. In terms of clinical data types, I also found similar proportions between the review paper and the data requests (Figures 5 and 6). I found that Demographics, Diagnoses, Appointments, Medications, Labs, Vitals covered 90.14% of the dental secondary analyses and 96.18% of the clinical data requests tickets.

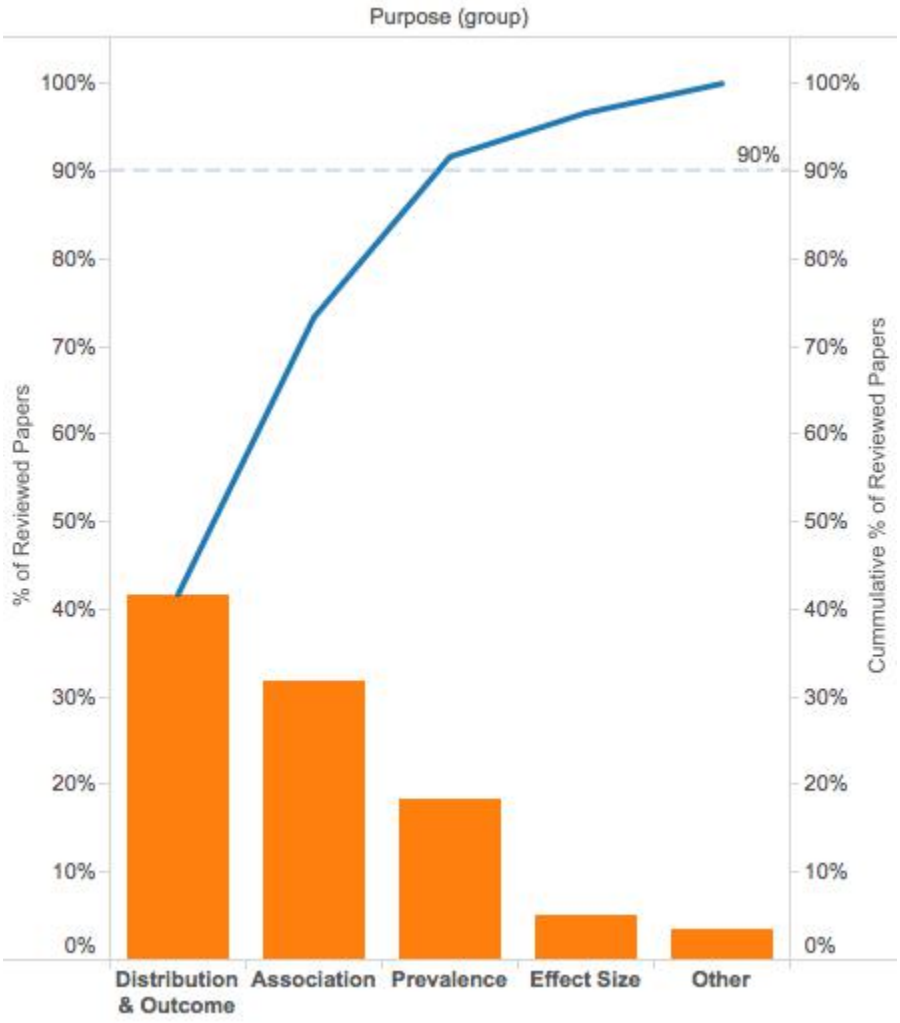


Figure 3 - Distribution of analysis types for dental data reuse analyses. Over 90% of the analyses are Distribution/Outcome, Association or Prevalence estimation analyses.

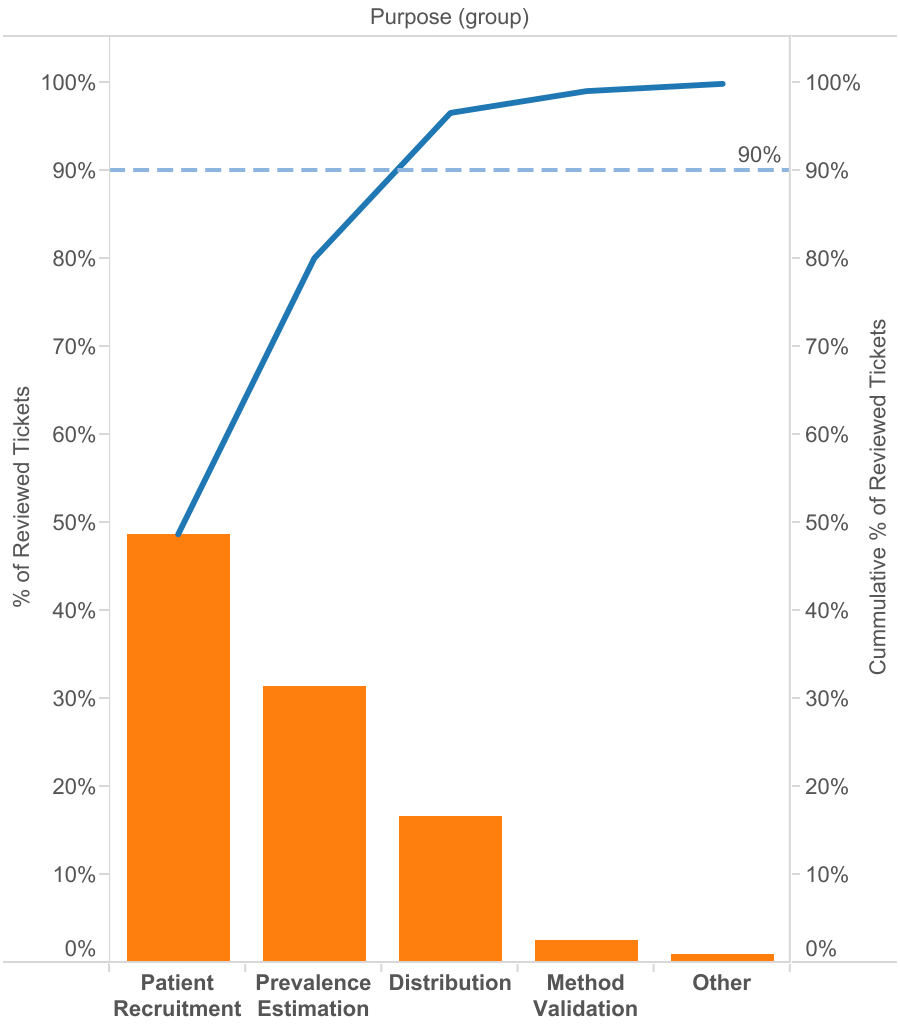


Figure 4 - Distribution of analysis types for clinical data request tickets. Over 90% of the analyses are Electronic Patient Recruitment, Prevalence Estimation or Distribution analyses.

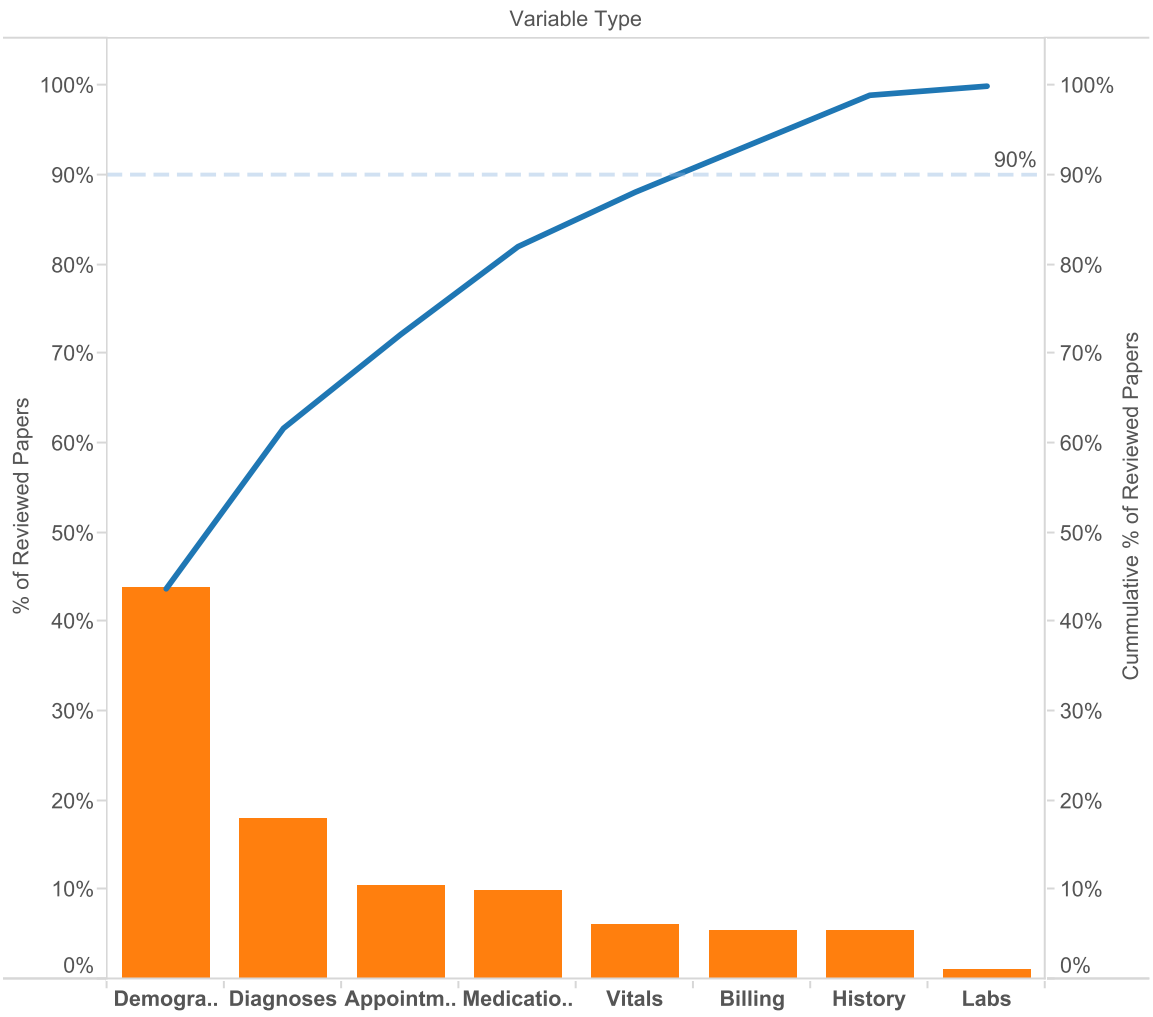


Figure 5 - Distribution of clinical data types for dental data reuse analyses. Over 90% of the analyses use Demographics, Diagnoses, Appointments, Medications, Vitals and Labs.

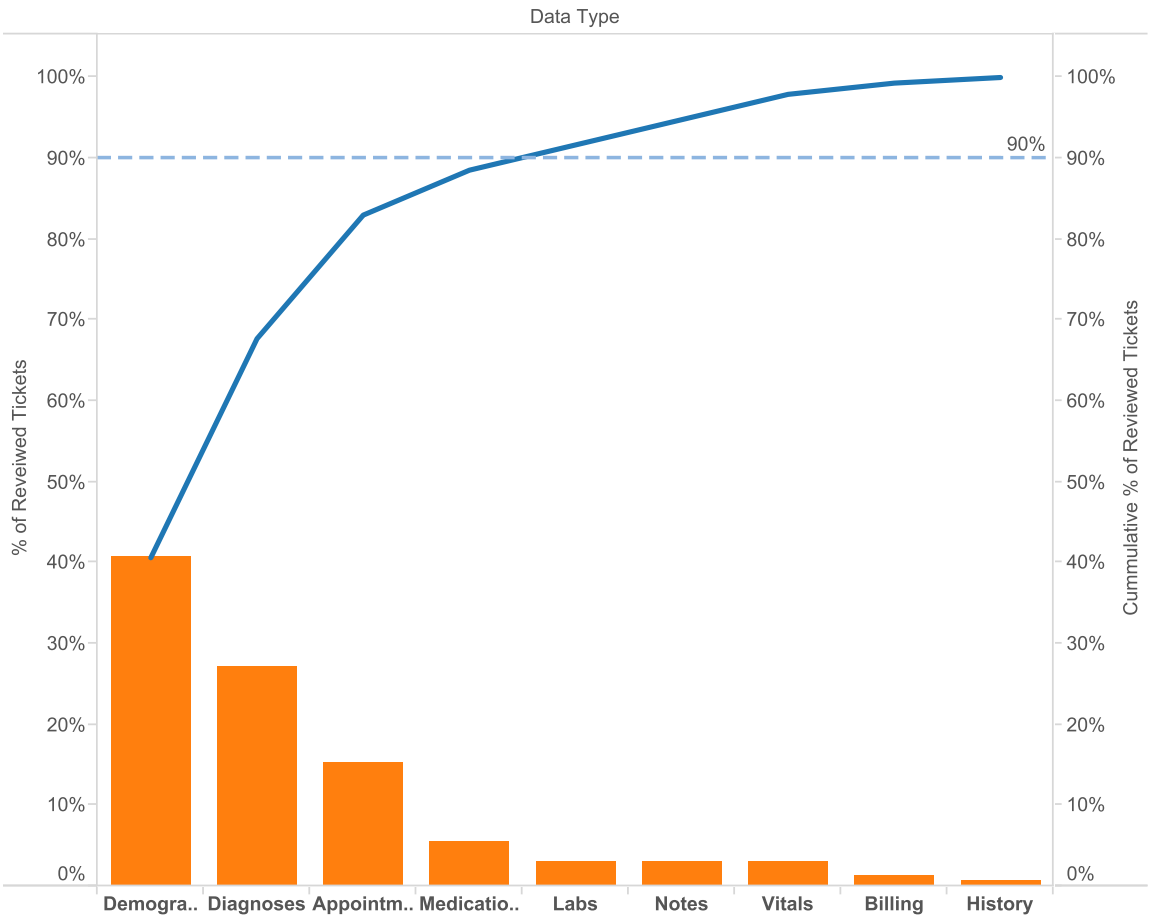


Figure 6 - Distribution of clinical data types for clinical data request tickets. Over 90% of the analyses use Demographics, Diagnoses, Appointment, Medication, Labs and Vitals data.

4.1.3 - Discussion

I was able to define a limited number of analysis types and clinical data types that would cover over 90% of secondary analyses of clinical data. However, this approach has three main limitations. First, I reviewed only two data sources. The first one was a literature review on dental EHRs. I was able to use this review as a viable initial analysis because dental EHRs have similar data structures to clinical EHRs. The second data source was a group of unstructured data requests that required qualitative analysis but provided a broader overview of the types of analyses done for a CDW beyond published analyses. Thus, the combination of these two data sources provides a reasonable basis to define the main categories of analysis types and clinical data types used in secondary uses of EHR data. Second, I did not cover all possible features that define secondary use application. However, clinical data types and analysis types are two critical aspects that define secondary use projects and can be generalized across applications. Third, the analysis types were generated using qualitative grounded theory methods from reviewing research questions. These categories may not be exhaustive but the data shows that they cover most cases, leaving only 3% of the analyses in the 'Other' category.

4.1.4 Use-Case Development

After identifying analysis and clinical data types that cover over 90% of secondary use cases, I developed a set of specific secondary use cases to cover all combinations of these types. I asked a clinical expert to define research questions serially as they came up in his research activities. I then developed those research questions into full-fledged use cases

by including all additional information on assumptions and research design. A short list of these research questions is presented below; a full description of the use cases can be found in appendix C. Tables 3 and 4 show the coverage of these use cases across secondary analysis types and clinical data types.

1. Is the second BP measure statistically lower than the first BP measure taken within a visit?
2. Are dual BP measurements provider-dependent?
3. How do patient weight measurements vary over time?
4. Is prednisone exposure correlated with weight gain?
5. Are HbA1C lab values correlated with BMI?
6. Can we find patients with BMI>25 and age>21?

Table 3 - Secondary analysis type coverage by case. Columns show each case and their relevant analysis types.

Case	1	2	3	4	5	6
Association		✓		✓	✓	
Distribution	✓		✓	✓		
Prevalence	✓					✓
Recruitment						✓

Table 4 - Clinical data type coverage by case. Columns show each case and their relevant clinical data types.

	1	*2*	*3*	*4*	*5*	*6*
Demographics	✓	✓	✓	✓	✓	✓
Diagnoses	✓		✓			
Appointments	✓	✓		✓		
Meds			✓	✓		
Labs			✓		✓	
Vitals	✓	✓	✓	✓	✓	✓

4.2 - Method

A data analytics team conducted DQ assessments using DataGauge for the six previously described use cases. We developed the DQ requirements using the preliminary guidance (see Table 1) that took into account the DQ dimensions, data granularity levels and DQ testing approaches. The team was composed of three domain experts: a clinician, a statistician and a data scientist who also served as a database administrator. We first generated a data needs model iteratively for each case. The model was discussed collectively. At least two iterations were carried out for every case. Once the DNM was established for every research question, we led one-hour interviews with the clinician and statistician; the team's data scientist led the interviews. The requirements were encoded from notes taken during each interview sessions. An approximate total of 20 hours was spent with each expert to generate the final list of requirements. The final DQ standards contained a list of DQ requirements for each case. The six final standards represented the list of minimum requirements for each dataset to be fit for purpose.

Once all requirements were available for each case, I selected dimensions to define an overview. These dimensions were selected based on their ability to cover the DQ requirement generation problem space (i.e. the DQ theory, domain knowledge, assessed data, etc.). This overview would serve as a checklist to ensure that the team generating the DQ requirements would cover all relevant aspects. The dimensions were chosen to improve DataGauge's current guidance (see section 3.2.3 for details) by (1) providing an overview of aspects to cover, (2) bridging the gap between the guidance and clinical

domain knowledge, while (3) supplying the analytics team with concrete questions to develop all relevant DQ requirements.

For each sub-section of this overview we listed all relevant requirements and we generated questions to guide the generation of DQ requirements. These questions aim to help users to consider all potential flaws relevant to that specific section of the overview. Providing concrete issues to consider aims to support the generation of DQ requirements by focusing the expert's attention to known potential threats to fitness for purpose. Phrasing the guidance in the form of questions forces the analytics team to respond to specific queries and structure their search around specific interest points rather than searching a vast problem space. The overview helps the team have a clear idea of the extent to which potential issues have been covered. An example of the guidance is described in section 4.4, a full version of the guidance framework (i.e., overview checklist and question lists) is available in Appendix B.

4.3 - Results

4.3.1 - DQ Requirements Dataset Overview and Descriptive Statistics

Our experts generated 389 requirements across the six cases. Figure 7 shows the distribution number of requirements by case. These requirements describe features a dataset must have to be fit for purpose.

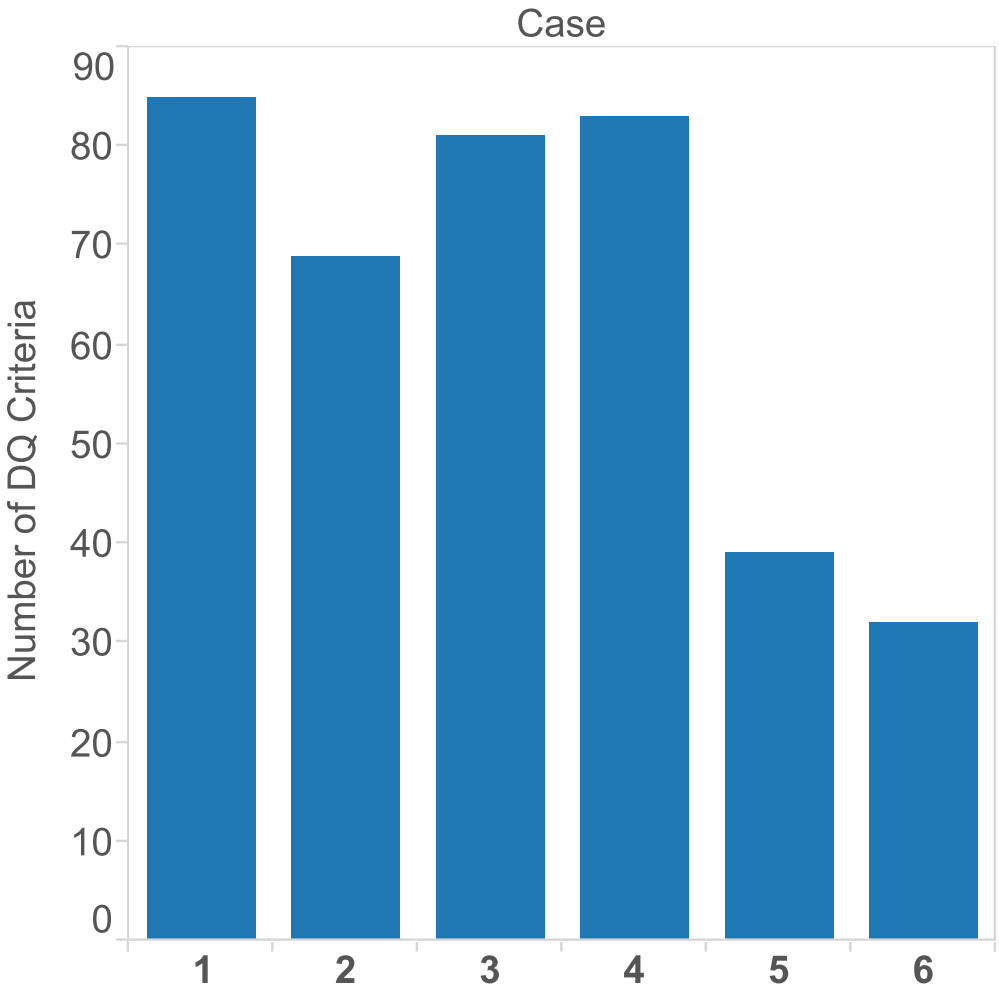


Figure 7 - Distribution of DQ requirements by use case number.

4.3.2 - Guidance Dimensions

The final guidance included four dimensions (Table 5). The two dimensions from the original DataGauge guidance were preserved. (1) The DQ dimensions as described in chapter 3 (i.e., correctness, plausibility, completeness, concordance, representation and timeliness) remained the link to DQ theory. (2) The levels of data granularity (Borek et al., 2011) (i.e., single value, multiple values, observation, observational unit and dataset)

remained the primary link in the DNM and reduce complexity. I used these two initial dimensions to develop the DQ requirements for the six cases. Two new dimensions were added as the result of the team-based development work. (3) To address the lack of integration with domain knowledge, I defined a list of knowledge domains. These directly represent the different areas of expertise necessary to run a secondary use analysis project and could be mapped to the three expert roles by Barlow (Barlow, 2013). The clinical expert held knowledge about the *clinical/medical science domain* and the *clinical workflow*. The statistician held the knowledge about the *analytical tools* and the *research design*. The data scientist held the knowledge about *data representation* and *data manipulation*. I also defined an additional *research goal knowledge domain* common to all three experts that represents the research question and purpose-specific considerations. These knowledge domains and their limitations surfaced continuously during the expert interviews, revealing knowledge gaps between experts as well as the need for integrated expertise and teamwork. The knowledge domains dimension is useful to separate out the types of background knowledge to be thought of individually when generating DQ requirements and allowing to communicate with experts. This also grounds the DQ generation task by providing some structure. The knowledge domains dimension contributes to the separation of concerns (Painter, 2006) and contributes to a more orderly development of DQ requirements. (4) To further ground the guidance to make it less abstract, while bridging the gap between the evaluated datasets and the clinical knowledge domain, I defined the clinical data types dimension. I used the same

clinical data types from those defined in section 4.1 (i.e., demographics, appointments, diagnoses, prescriptions, lab results and vitals).

Table 5 - Overview of the four dimensions of our framework for DQ requirement guidance.

	DQ Dimensions	**Levels of Data Granularity**	**Knowledge Domains**	**Clinical Data Types**
Aspects of DQ Assessments Represented	DQ theory	DNM integration and dissection	Overarching expertise needs, Domain knowledge, Multi-disciplinary approach	Data model of origin, Data production considerations, Clinical domain knowledge
Elements	Correctness, Plausibility, Completeness, Concordance, Representation, Timeliness	Single value, Variable, Observation, Observational Unit, Multiple Observational Units, Dataset	Analytical Tool, Clinical, Data Manipulation, Representation, Research Design, Research Goal, Workflow	Demographics, Appointments, Diagnoses, Vitals, Prescriptions, Lab results
Guidance Need Addressed	-Connection to DQ theory	-Provides a way to break down the complexity of datasets -Provides a way to examine the dataset as a series of subsets	-Provides an overview of relevant areas of expertise -Creates a link to domain knowledge -Separates concerns (Painter, 2006) when developing DQ requirements	-Reduces vagueness and abstraction -Provides a link to the assessed data and the clinical domain knowledge
Source Work	DQ Dimensions (Wang & Strong, 1996) adapted the secondary use (Nicole Gray Weiskopf & Weng, 2013), modified in Chapter 3	Classification of DQ testing approaches (Borek et al., 2011; Oliveira et al., 2005), adapted in Chapter 3	Expert interviews (see Section 4.2) as an extension of the areas expertise needed to carry out secondary analyses of data (Barlow, 2013)	Clinical data reuse case definition (see Section 4.1), partially based on (Song et al., 2013)

4.3.3 - Overview Dimensions and DQ Requirement Generation Contexts

Providing an overview of DQ requirement generation activities is one goals of this new guidance framework. The utility of an overview is to provide the research team with a list of aspects to consider when generating the DQ requirements. The knowledge domain dimension is most fit to serve as a base to build the the framework for four reasons: (1) It represents the expertise held by the team as a whole, (2) It can be used to allocate responsibility and separate concerns during the DQ requirement generation (Painter, 2006), (3) Domain knowledge is necessary to carry out every step of the DQ assessment process and (4) It is the most encompassing and comprehensive dimension.

To identify another dimension that could be included in the overview, I checked whether the knowledge domain dimension could be paired with any of the other three dimensions. The goal was to identify a combination that would create separation of concerns (Painter, 2006), while giving a comprehensive overview of the problem space. Ideally, the combined dimensions would allow DQ requirement generation for every dimension element combinations (e.g., clinical knowledge and the completeness DQ dimension). I examined this by creating heat maps the number of DQ requirements generated from the six use cases for combinations of the knowledge domains and other three dimensions (Figures 8-9). I found that the generated DQ requirements fully covered the combination of Knowledge Domains and Clinical Data Types (Figure 9). This means that for every combination of knowledge domain and clinical data type elements it is possible to

generate DQ requirements. I refer to the combination of a knowledge domain element (e.g., clinical domain) and a clinical data type (e.g., vitals) as a *DQ requirement generation context*. To emphasize the need to cover each DQ requirement generation context I converted Figure 9 into a checklist (Table 6). This checklist gives the analytics team a clear list of all possible DQ generation contexts; each context corresponds to a checkbox. This overview should help the analytics team to clearly keep track of the contexts covered during the design of the DQ assessment.

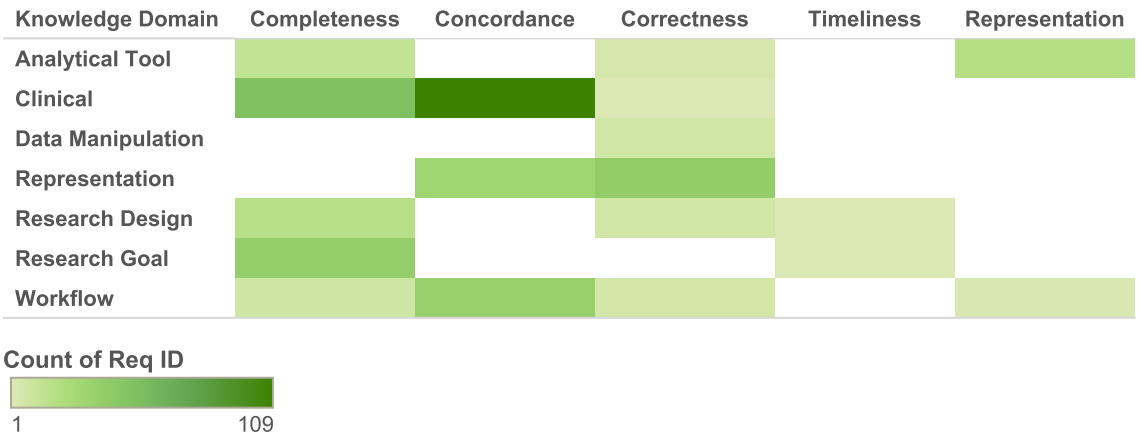


Figure 8 - Distribution of DQ requirements by DQ dimensions and knowledge domain. The generated requirements fail to cover all sub-sections of this two-dimensional space.

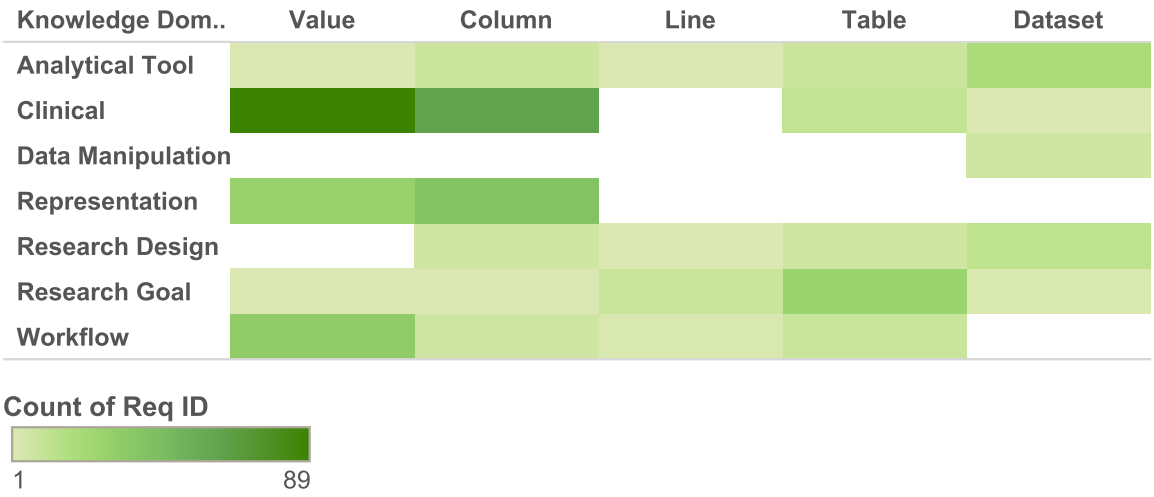


Figure 9 - Distribution of DQ requirements by data granularity levels and knowledge domain. The generated requirements fail to cover all sub-sections of this two-dimensional space.

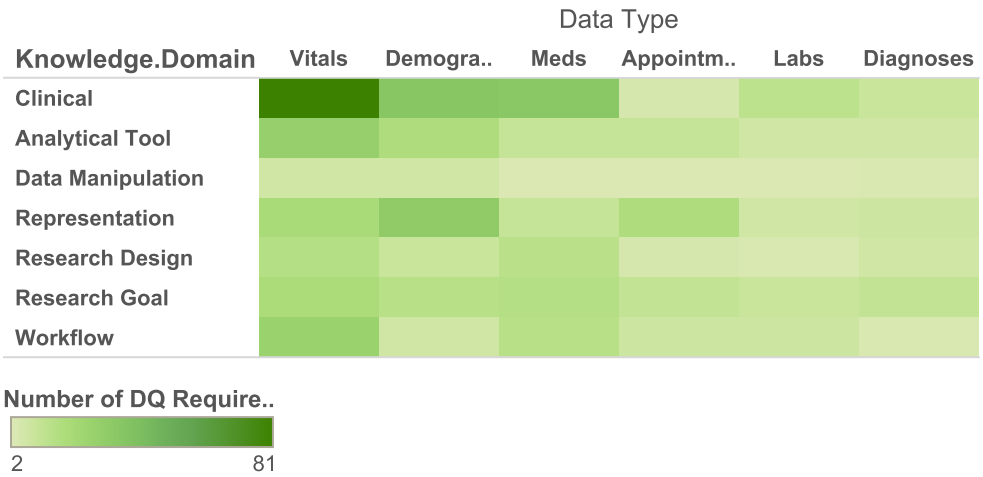


Figure 10 - Distribution of DQ requirements by clinical data type and knowledge domain. The generated requirements cover each sub-section of this two-dimensional space. This makes these two dimensions useful to define a comprehensive problem-space.

Table 6 - Overview guidance checklist for DQ requirement development. This checklist partitions the problem-space of DQ requirement development into sub-sections or DQ generation contexts defined by the knowledge domain (left column) and the clinical data type (top row). This partitioning provides the analytics team with a clear idea of the aspects to cover and breaks the complexity of DQ requirement generation into manageable sub-tasks.

	Vitals	Demographics	Meds	Appointment	Labs	Diagnoses
Clinical	☑	☐	☐	☐	☐	☐
Analytical Tool	☐	☐	☐	☐	☐	☐
Data Manipulation	☐	☐	☐	☐	☐	☐
Representation	☐	☐	☐	☐	☐	☐
Research Design	☐	☐	☐	☐	☐	☐
Research Goal	☐	☐	☐	☐	☐	☐
Workflow	☐	☐	☐	☐	☐	☐

4.3.4 - Contextual DQ Requirement Generation

The two remaining dimensions (i.e., DQ dimensions and levels of data granularity) were allocated to provide guidance within each DQ requirement generation context (i.e., the combination of a knowledge domain and a clinical data type). I provide two sources of guidance: (1) the original guidance that combines the DQ dimensions and levels of data granularity as well as (2) a series of context-specific questions generated from the DQ requirements available from the six use cases. The original guidance provided by the DataGauge framework (see section 3.2.3) is used to break down the DNM into simpler data structure at every data granularity level.

To provide more specific and domain-appropriate guidance, I formulated guidance questions from the DQ requirements generated for the six use cases for each specific DQ requirement development context. The role of these questions is to direct the analytics team's attention to specific threats to fitness for purpose. I classified the questions according to the evaluated DQ dimensions. Table 7 shows a sample of these of questions for the clinical knowledge domain and the "Vitals" clinical data type. All levels of granularity are relevant to all questions if they are represented in the DNM. The data granularity levels are involved in the process by aiding the DNM decomposition into a list of pieces for which the domain expert will answer the questions. In the next section I present an example of the use of this guidance for the development of DQ requirements.

Table 7 - Sample of DQ requirement development guidance questions. These questions aim to guide the analytics team to address general DQ issues that may be a threat to fitness for purpose. The team will define DQ requirements based on these questions taking into account the assessed data and the research question. The resulting requirements define the features of a dataset that would be fit for purpose. An example of their use is shown in the following section.

DQ Dimension	Guidance Question
Completeness	Are there enough values to carry out the analysis?
Completeness	Are all expected vital values present?
Completeness	Are there any missing values for each vital record?
Correctness	Are the values within the clinically expected range?
Plausibility	Are all measurements taken after the patient's date of birth?

4.4 - DQ Requirement Definition Guidance Example

Within DataGauge, DQ requirements are generated after the research question and the DNM are defined (see Chapter 3). The DNM defines the relevant clinical data types; all knowledge domains should be considered for every research question. For this example, I use the same research question used in Chapter 3: " Is prednisone, a commonly-prescribed corticosteroid, associated with weight gain?" The associated DNM model is shown in Figure 11. The relevant data types for this specific DNM are Demographics, represented by the Patient table, Medications, represented by the PrednisonePrescription table and Vitals, represented by the Weights table. As stated earlier, all knowledge domains are relevant. The relevant guidance checklist and overview is shown in Table 8.

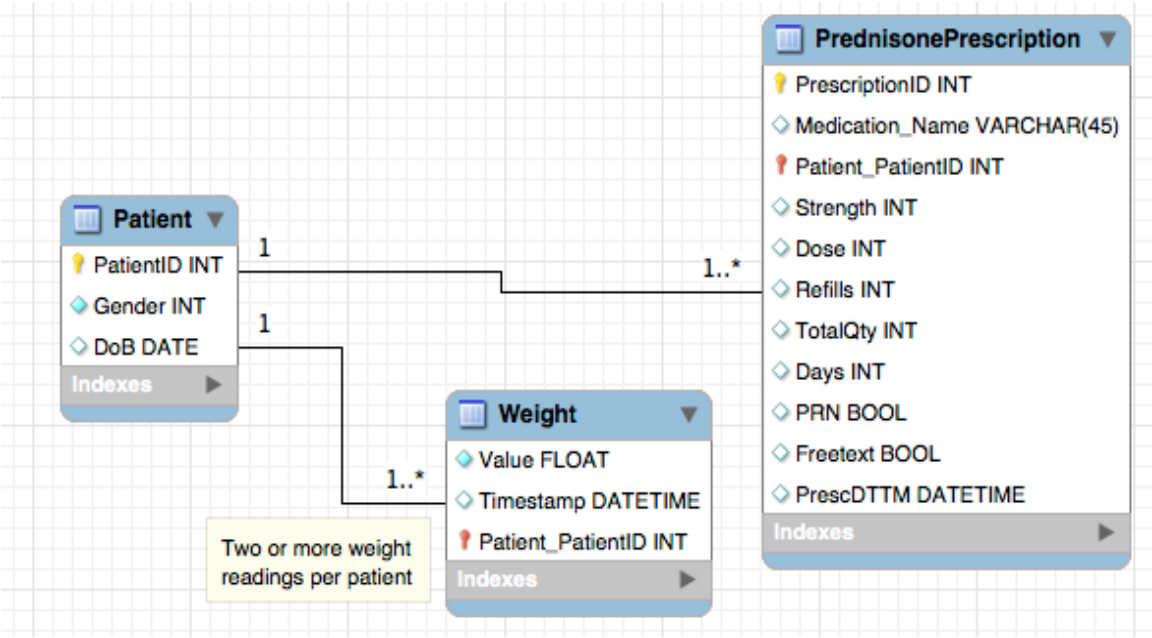


Figure 11 - Data Needs Model for the research question " Is prednisone, a commonly-prescribed corticosteroid, is associated with weight gain?"

Table 8 - Overview guidance checklist for the research question " Is prednisone, a commonly-prescribed corticosteroid, is associated with weight gain?" Clinical data types are eliminated based on their relevance. In this case, Appointments, Labs and Diagnoses were not relevant.

	Vitals	Demographics	Meds
Clinical	☐	☐	☐
Analytical Tool	☐	☐	☐
Data Manipulation	☐	☐	☐
Representation	☐	☐	☐
Research Design	☐	☐	☐
Research Goal	☐	☐	☐
Workflow	☐	☐	☐

For a relevant data type, the research team selects a specific knowledge domain and the questions relevant to that context (i.e., the combination of a clinical data type and knowledge domain) are retrieved. For our example, we first select the Vitals clinical data type and the clinical knowledge domain (i.e., the top left checkbox on Table 8). The team then selects portions of the DNM relevant to the assessed clinical data type. For our example, the relevant data elements are: the values within the Timestamp, Value and PatientID variables, these three variables themselves, each observation containing a combination of Value, Timestamp and PatientID, the Weight table, the combination of the patient and Weight tables and the whole dataset containing information about vitals, patients and prescriptions. With this information available, we can then proceed to

answer the contextual guidance questions for each element of this list. The list of relevant questions for this requirement generation context is shown on Table 9.

Table 9 - Contextual guidance questions relevant to the clinical domain knowledge and the Vitals clinical data type. This is the full list of questions available for this context (i.e., top left check box on Table 7). The full list of questions for the other contexts is available in Appendix B.

DQ Dimension Addressed	DQ Generation Guidance Question
Completeness	Are all expected vital values present?
Completeness	Are there any missing values for each vital record?
Completeness	Do related values have a match? (e.g., systolic + diastolic blood pressure)
Completeness	Is the numeric distribution of values and value frequencies as expected?
Correctness	Are all values within plausible limits for the population?
Completeness	Are there an adequate number of measurements within the desired observation window?
Correctness	Are the measurements taken after the patient's date of birth?
Completeness	Are there sufficient values over time for the analysis?
Completeness	Is the number of vital measurements as expected? (e.g., at least one weight measurement per visit)
Correctness	Is the difference in consecutive values in an encounter within acceptable range?
Correctness	Is the difference between variable values between two time points in acceptable proportion?
Plausibility	Are time stamps in the expected order? (e.g., order before admin)
Plausibility	Are time stamps within the expected time frame? (e.g., BPs measured within the encounter window vs. outside).
Timeliness	Is the time between events of the expected range?
Concordance	Are there any sudden changes over time? Are they valid?
Concordance	Are the values coherent or vary as expected within a visit?
Concordance	Is the overall vital measure variability as expected?

The combination of a question and the data granularity level DNM pieces (e.g., a weight value, the weight variable or a weight observation which combines a weight value, a timestamp and patient ID) may generate one or multiple DQ requirements. For each question, the expert must consider the research question to inform the definition of requirements for a fit-for-purpose dataset. For example, when the clinical expert considers the question "Are all expected vital values present?", which addresses the completeness DQ dimension, and we are evaluating the weight values, which is a considered part of the “Vitals” data type, the team may generate requirements such as "There are adult patients with weights between 50 and 100kg". This is purpose-specific because the research question (i.e., "Is prednisone, a commonly-prescribed corticosteroid, is associated with weight gain?") is likely to be answered using adult patient data, given that children weights increase over time and the prednisone effect may be masked. At the observation level (i.e., the combination of a weight value, a timestamp and a patientID), the expert may come up with "All weight values have a timestamp value", which is necessary to calculate the rate of weight gain to answer the research question.

When considering the guidance question "Are there sufficient values over time for the analysis?” which addresses the DQ dimension of timeliness, for the timestamp variable within this same context, the expert may come up with requirements such as "Patients have at least one weight per encounter" to ensure there is no censoring, "Patients have at least two weights overall" to be able to calculate the rate of weight change, "Patient has at least one weight per year between the first and last visit" to ensure reasonable temporal resolution. When considering the whole dataset, the expert may come up with a

requirement such as "Patient has at least two measurements within a year of the first prednisone prescription" to ensure that there is enough data in the timeframe where the effect is expected.

When considering the question "Are there any sudden changes over time?", which considers the dimension of concordance, the experts may come up with requirements such as "Weight within a day should not vary more than X%", "Weights may not vary more than X% per month" and "Weights may not vary more than X% overall", which would reveal potentially inaccurate values that may distort relationship value found in the analysis.

The team will run through the whole list of questions for that specific checkbox and then move on to the next knowledge domain. When all knowledge domains are covered, the team moves on to the next relevant clinical data type and repeats the procedure. Once all clinical data types are reviewed, all checkboxes should be marked. All generated requirements are aggregated into the DQ standard, which is specific to the research question and the DNM in question. These requirements define the DQ tests that will yield the DQ measures to evaluate fitness for purpose within the DataGauge framework.

4.4 - Conclusion

In this chapter, I presented a framework to support the development of DQ requirements for a specific research question and DNM within DataGauge. The framework addresses the shortcomings of current guidance available in the literature and the DataGauge process to enable the systematic generation of DQ requirements. The framework provides

(1) A clear overview of the issues to be considered while developing DQ requirements for a specific dataset and research question, (2) A bridge between domain knowledge and the DQ requirement generation task and (3) A list of specific issues and threats to be considered in the form of questions.

This framework provides support to the user by providing an overview of the main points to be considered when assessing DQ. This provides an overarching frame to pose questions about the appropriateness of a given dataset for a specific purpose. Because of its dimensions, it supports the separation of concerns (Painter, 2006), which is likely to provide a more organized and thorough DQ requirement generation. This framework also addresses the gap between the evaluated data, the domain knowledge and the DQ theory stated in the literature (Floridi, 2013) and provides an overarching map to ensure systematic coverage of all concerns. Finally, this framework supports the systematic definition of DQ requirements. This step is critical because a DQ assessment is the combination of DQ tests that are directly defined by the DQ requirements. Thus, the thoroughness of the DQ assessment is directly dependent on the thorough definition of DQ requirements.

This framework provides an unambiguous checklist and overview to check for issues as a way to ensure coverage and systematicity. It also provides specific questions as a way to focus the generation of DQ requirements on specific threats to fitness for purpose. However, this is an initial framework that does not cover all possible clinical data types and was developed based on a limited number of cases. Nevertheless, our cases cover the most broadly used data types and secondary use applications (see section 4.2). A single

analytics team was involved in the framework's development because we aimed to develop this initial framework that will be further expanded as other research groups adopt DataGauge. The guidance checklist provides a development base that is applicable to other projects as a launching point for future development. Finally, the current version of the guidance questions may not be phrased in the ideal way. Thus, validation and further refinement is still necessary and will be part of our future research. This work will be carried out after the initial DataGauge evaluation described in the following chapter.

Chapter 5: Evaluation of DataGauge

This chapter presents an evaluation of DataGauge's ability to improve upon the current state of the art methods of systematic DQ assessment. My *hypothesis* is that DataGauge will increase the number of DQ issues flagged for a specific secondary use of clinical data over the current systematic state of the art method. This evaluation will test DataGauge's potential usefulness, inform future research directions and identify further development routes. One particularly interesting aspect in this evaluation is that DataGauge is one of the first systematic DQ assessment methods to allow a systematic, fitness-for-purpose approach to the assessment of repurposed clinical data. This sort of approach has been advocated as the preferred method for DQ assessment in the field of clinical informatics (Holve et al., 2013), yet a viable candidate is still missing. Comparing the performance of DataGauge against a general DQ assessment method would also provide evidence to evaluate the advantage, if any, of a fit for purpose approach versus a generic one.

In this chapter I describe the process carried out to evaluate DataGauge against a state of the art DQ assessment method (i.e., the evaluation standard). First, I describe the methods used for the evaluation, including the rationale for the evaluation standard selection. Then I present and compare the performance results for DataGauge and the evaluation standard

over the six cases defined in section 4.2. Finally, I discuss these results and their implications.

5.1 - Methods

5.1.1 - Defining a DQ Assessment Standard

To evaluate the performance of our assessment method it is necessary to select a control or baseline method. In most analyses carried out for research purposes some kind of data cleaning procedure (Broeck & Fadnes, 2013) is carried out by the analyst. This consists of detecting, diagnosing and rectifying anomalies in the dataset as the analysis progresses. The main pitfall of this approach is that, being analyst-dependent; there is a risk of missing errors that threaten the assessment's thoroughness. This is especially problematic when the data in question are repurposed because an additional set of complex issues come into play (Hersh et al., 2013; Van Der Lei, 1991). In fact, data cleaning is geared towards cleaning datasets produced and used for the same purpose such as randomized controlled trial datasets (Van den Broeck et al., 2005) rather than repurposed data. An additional pitfall is that, because cleaning happens in parallel to the analysis, the analyst may find "satisfactory" results from flawed data on first approach and never question their accuracy. Thus, data cleaning is unlikely to be a reliable comparison standard.

Though many individual tools are available to test discrete aspects of DQ (Borek et al., 2011), very few tools are available to carry out whole DQ assessments. However, an automated tool to assess the DQ in the initial phase of integrating EHR data into a CDW

is currently available. The Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems (ACHILLES) (G Hripcsak et al., 2015) is a platform, which enables the characterization, quality assessment and visualization of observational health databases. ACHILLES was released by the Observational Health Data Sciences and Informatics (OHDSI) collaborative and provides a module called ACHILLES Heel (G Hripcsak et al., 2015) that allows the user to automatically detect database integrity issues and DQ flaws in clinical datasets modelled using the OMOP common data model (Overhage, Ryan, Reich, Hartzema, & Stang, 2012; Reisinger et al., 2010). ACHILLES is a state-of-the-art tool to automatically assess DQ in CDWs. It is capable of detecting issues such as predefined value compliance, limited numeric value range checks, temporal inconsistencies and data model integrity checks (e.g., primary key and foreign key check).

Because ACHILLES is an automated, code-based tool, it is capable of consistently testing all relevant DQ aspects that an expert user with knowledge of the database would run. This sort of consistency is desirable for a comparison standard or baseline because the end output would be the same no matter how many tests are run as long as the tested database remains unchanged. Another virtue of ACHILLES is that, by design, it does not consider the research question or analytical purpose. One of the limitations of ACHILLES is that, in its current version, it does not check for statistical outliers. Statisticians and data analysts routinely use the following rule for outlier detection: values outside three standard deviations to each side of the mean are considered probable outliers in normally-distributed data; the range grows to five standard deviations for any

other kind of distribution (Amidan, Ferryman, & Cooley, 2005; Rahbar et al., 2013). To address this shortcoming, I supplemented ACHILLES with this outlier detection function. The final standard of comparison selected for evaluation was the *ACHILLES module supplemented by the statistical outlier detection* rule described above.

5.1.2 - Testing Protocol

My test dataset consisted of a teaching outpatient clinic's EHR's database from a major metropolitan area. This dataset contained 10 years of data for 954,891 patients. The EHR database was restructured to comply with the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) Version 5 (G Hripcsak et al., 2015; Overhage et al., 2012). This enabled me to run the ACHILLES module automatically.

For each test case, I created a new database containing only the data relevant to the research question. This allowed a complete, yet focused ACHILLES assessment of the whole database where only issues relevant to the appropriate data would be flagged. ACHILLES ran automatically on the filtered datasets. To run DataGauge, I coded SQL-based query tests for each requirement generated by experts in the previous experiment (see Section 4.2 for detailed methods). Each query, I counted the number of violations in the dataset for the requirement it tested. Each violation flag represented a potential DQ issue. These counts are the DQ assessment results or DQ measures that quantify the dataset's adherence to the DQ requirements defined using DataGauge.

Both methods flag requirement violations that could ultimately be a threat to the validity of analytical results. On one hand, the comparison standard focuses on identifying

general integrity issues in the dataset. The issues found can be of two kinds: (1) they could affect the intended use and become a significant data problem or (2) they may not threaten the purpose directly. These indirect threats are issues that do not jeopardize answering the research question but may distort the answer if they are too prevalent. On the other hand, the requirements generated by experts using DataGauge requirements describe a fit-for-purpose dataset. Therefore, each violation is a potential threat to the secondary analysis because they represent a shortcoming of the assessed dataset in relation to the ideal dataset. However, not all requirements represent the same level of threat to fitness for purpose. To quantify the level of threat, we asked the analytics team generating the requirements to assign a value to the issue in one of two categories: (1) Discarding Requirements, which reveal a direct threats to fitness for purpose (e.g., Patient records that infringe on the requirement "Patient has at least two weight measurements" for the calculation of weight change over time); (2) Review Requirements, which indicate a potential flaw that may be rectified via imputation methods or are not a direct threat to the dataset's fitness for secondary use (e.g., Breaches of "Weight values don't change more than 2% within a day" should be reviewed and may be potentially corrected using imputation methods if there are enough values to provide a stable baseline). Each DQ test from DataGauge had such value assigned during the initial generation. The same classification process was done for the evaluation standard results for each case.

5.1.3 - Major Differences in DQ Assessment Methods

There are three major differences between the DQ assessment tools that must be taken into account when interpreting the results. First, the knowledge available when

generating the test code for each method is different. The evaluation standard (i.e., ACHILLES + Outlier detection) is an automated method designed to be broadly applicable and only taps into knowledge about the data structure. On the other hand, this standard is easier to use, requiring minimal human effort and involvement because it is based on executable code. This makes its results reproducible, as they are not based on human expertise. In contrast, DataGauge benefits from a team of experts and their knowledge about the data source, the purpose and domain knowledge. This means that the DQ tests performed by DataGauge should be much more specific and purposeful than those done by ACHILLES. Second, DataGauge is designed to be purpose-specific, whereas the standard does not take the purpose into account by design. Therefore, the issues flagged by the standard may not be a threat to the purpose. To address this, the issues found by both methods were assigned a level of threat value: discarding requirements (i.e., direct threats to fitness for purpose) or review requirements (i.e., potential flaws that may be rectified via imputation methods or are not a direct threat to fitness for secondary use). Finally, it is important to note that our comparison standard is static executable code, whereas DataGauge is based on a set of requirements generated by a team of experts. This means that the issues flagged by DataGauge are extremely likely to be relevant threats to the fitness for purpose of the dataset by design (given that each assessment is tailored to each research question and dataset's needs), whereas the comparison standard flags issues that are likely to be data flaws without taking the purpose into account. Therefore, it is necessary to review the issues found by the standard to confirm their relevancy to assessing fitness for purpose.

5.2 - Results

I compared the results in three ways. First, I compared the number of issues found by each method. I show overall performance in flagging potential DQ issues for each method, followed by a breakdown by DQ requirement severity based on the requirement's level of threat to fitness for purpose (i.e., discarding vs. review requirement). I then reviewed the most common DQ issues flagged by both methods to confirm that the issues found were true DQ issues. Second, I compared the coverage of each method based on the previously-defined overview dimensions (i.e., clinical data types and the knowledge domains -see section 4.1) derived from EHR database design features and types of expertise required for a data reuse project (Barlow, 2013). This shows the comprehensiveness of each method. Finally, I built a regression model to predict the number of flags found by each method. I counted the total number of tests carried out by each method and the total number of tests flagged as positive. I used these two variables for each method to build a negative binomial to predict the number of issues found based on the selected methods and the number of test instances. Each case was considered a separate experiment. All analyses were conducted using R Version 3.2.2 (R Core Team, 2013) assuming statistical significance at $p<0{\cdot}05$.

5.2.1 - Number of Flags

Data Gauge flagged roughly ten times more issues than the comparison standard. These issues represent individual data elements that infringed on a DQ requirement. Overall, the standard returned 1.4 million flags whereas DataGauge returned 19 million (Figure 12).

Review flag counts were 1 million flags from the comparison standard versus 9.3 million for DataGauge (Figure 13). Discarding flag counts showed, 0.4 million for the standard versus 9.6 million for DataGauge (Figure 13).

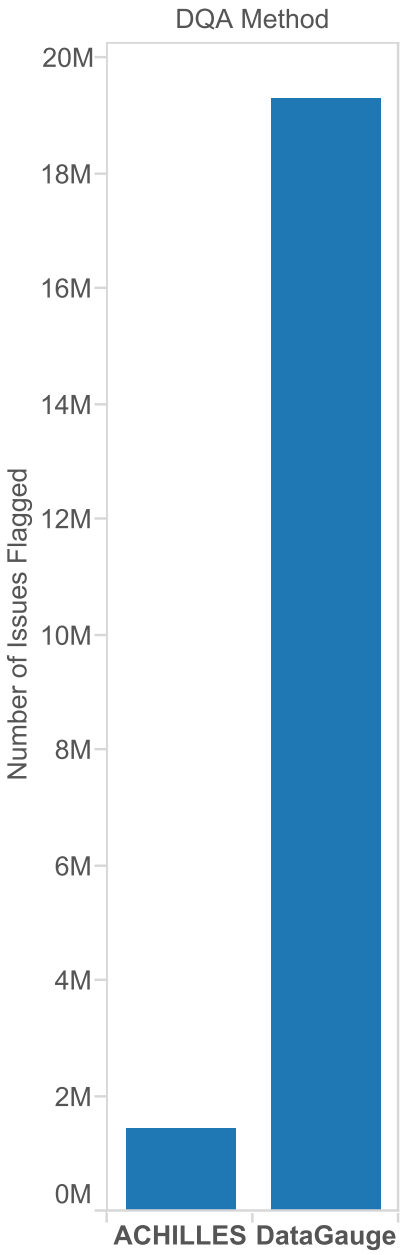


Figure 12 - Number of flags returned by both methods. DataGauge returned close to ten times more flags than the comparison standard.

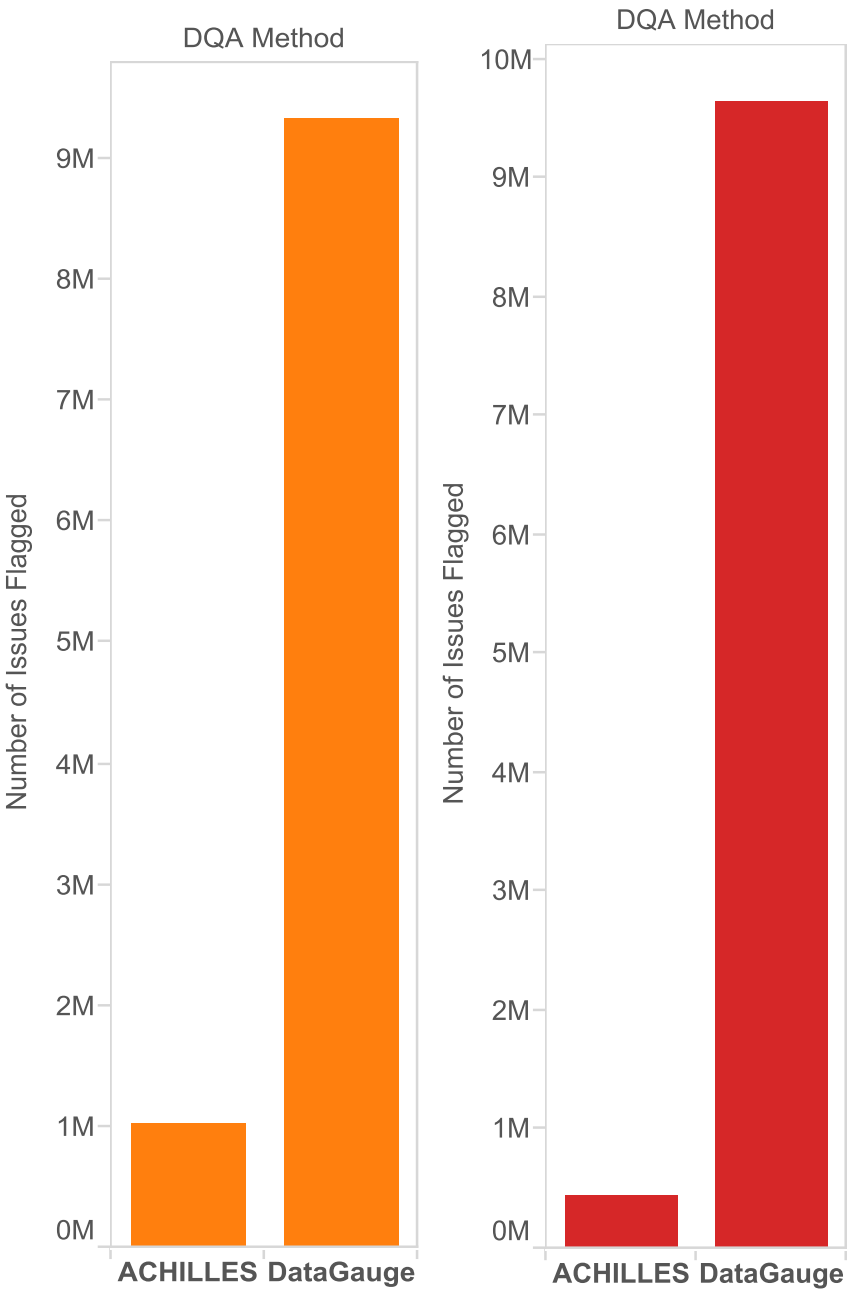


Figure 13 - Number of discarding and review flags returned by both methods. DataGauge returned close to ten times more review flags than the comparison standard (left) and close to twenty times more discarding flags (right).

Reviewing the most common flags revealed multiple overlaps, yet DataGauge included all flags found by the standard plus more specific issues. Dataset integrity issues were found by both methods. For example, the control method flagged person, provider and care site IDs with no correspondence in the table that contained the variables that defined them (e.g., the demographics table for patients); DataGauge identified the same issues as primary key-foreign key relationship breaches. Data values were checked by both methods but DataGauge did this in a more purposeful way. More specifically, the control method checked for statistical outliers, which provides a knowledge-independent and broadly applicable way of screening for outliers. These outliers are extremely likely to be true outliers but the results provide no insight to determine the nature of the DQ issues. On the other hand, DataGauge flagged values for specific reasons such as vital values out of range (e.g., Weight measurements over 400kg that are possible but not likely), abrupt changes over time (e.g., weight changes larger than 10% in the same day) and impossible timelines (e.g., measurement taken before the patient's date of birth). Beyond these two categories, DataGauge identified issues that were specific threats to the fitness for purpose that the control method was unable to flag. For example, patients with overlapping or co-occurring prescriptions for a research question dealing with drug exposure and weight variation; the threat to the secondary analysis is that we are unable to confidently calculate the effective drug exposure because the uncertainty in the drug exposure timeframe.

Examining the most common issues for review and discarding requirements confirmed the drastic difference in test specificity. This confirmed the value of purpose-specific tests in providing data to evaluate fitness for purpose; review flags clearly showed this difference. For example, the control method found unidentifiable providers and care sites, which were marginally threatening to the purpose in most cases, whereas DataGauge found direct threats to the purpose such as missing covariate data and unexpectedly high value changes over time. Discarding requirements showed the same differences. The control method flagged statistical outliers and person IDs that could not be mapped to the demographics table rendering the record in question unusable. In contrast, DataGauge found patients without the necessary data to conduct the primary analysis (excluding covariates), data outside the plausible timeframe, prescriptions with missing values that prevented the calculation of an effective dose, amongst other issues.

5.2.2 - Coverage of Clinical Data Types and Knowledge Domains

Dimension coverage revealed great differences between methods. DataGauge found issues in all Clinical Data Types (Figure 14) and Knowledge domains (Figure 15). On the other hand, the comparison standard failed to identify any issues in two out of six clinical data types and three out of six knowledge domains. The overview also shows a larger number of flags on the DataGauge side for every category where both methods returned flags.

	DQA Method	
Data Type	ACHILLES	DataGauge
Apointment	14	1,706,739
Demographics	3,516	451,144
Diagnoses	0	7,313
Labs	0	740,668
Meds	7,071	761,706
Vitals	5,527	1,437,716

Figure 14 - Coverage of clinical data types by the control method and DataGauge. DataGauge covers all clinical data types, whereas the comparison standard fails to flag any issues for diagnoses and labs. DataGauge also flags more issues for in all data types covered by both methods.

	DQA Method	
Knowledge.Domain	ACHILLES	DataGauge
Analytical Tool	9,070	10,485
Clinical	0	1,092,118
Representation	4	1,602,378
Research Design	0	9,082
Research Goal	0	1,657,841
Workflow	7,054	742,866

Figure 15 - Coverage of knowledge domains by the comparison standard and DataGauge. DataGauge covers all knowledge domains, whereas the comparison fails to

flag any issues in three domains. DataGauge also flags more issues in all domains covered by both methods.

5.2.3 - Statistical Comparison

Statistical comparison of both methods revealed a statistically significant difference between the two methods when comparing values for only six cases. I compare the number of issues flagged between DataGauge and the evaluation standard using negative binomial regression model that accounts for the over-dispersed count data, using the total number of tests done by each method as an offset variable (Table 10). To account for potential confounding effects, I compare matched DQ assessment results for a series of cases (i.e., six matching research question and dataset pairs), where only the applied method varies. I also explored covariates such as data granularity, analytical unit and DQ dimension but none of them significantly impacted the estimates of interest . The expected number of flags returned by DataGauge was significantly higher than the number returned by standard (Rate Ratio=7.00; 95% CI=[3.22,15.2]; $p<0.0001$).

Table 10 – Comparison of the number of flags returned by each method based on multivariable Negative Binomial regression model . The expected number of flags returned by DataGauge is almost seven times greater than the number returned by the control method.

	Rate Ratio (95% Confidence Interval)	**p-value**
Method	7.00 (3.22,15.2)	<0.0001

5.3 Discussion

DataGauge was able to flag more potential issues than the state of the art method. It identified more direct threats to the analytical purpose, but also more DQ issues that could be potentially corrected via review or imputation. This difference was confirmed to be statistically significant. The issues identified by DataGauge were also more specific, which allowed for a deeper understanding the dataset's limitations. Finally, DataGauge showed broader coverage of Knowledge Domains and Clinical Data Types.

This analysis provides evidence of DataGauge's contribution to the current state of the art but has three main limitations. First, we only used a single clinical database containing outpatient data. Though this is only one type of clinical data, but is one of the most broadly available for secondary use. Second, we only assessed DQ for six use cases. However, we justified their coverage of secondary uses and clinical data types in Chapter

4. Moreover, these six cases provided adequate statistical power to uncover a statistically significant difference between DataGauge and the comparison standard. Third, only one research team was employed in this evaluation, thus the results may not generalize.

Based on these results, I can conclude that DataGauge improves the current state of the art by improving the detection of DQ issues in two ways. First, DataGauge increases the number of opportunities to detect DQ issues by providing a broader coverage of the assessed data and relevant knowledge domains. Second, DataGauge focuses the detection of specific issues that are directly relevant to the intended secondary use. DataGauge is the first systematic method to assess DQ in truly fitness-for-purpose-oriented way (Holve et al., 2013; Juran, 1962). Even though the method is human intensive, the great threats posed by data repurposing (Hersh et al., 2013; Van Der Lei, 1991) demand such thorough evaluation. Also, it is currently very difficult to truly assess fitness for purpose in an automated way given that it requires so much domain knowledge that is not readily structured. Nevertheless, DataGauge provides an initial systematic process that can support future automation efforts. Also, when applied thoroughly by a team of experts, DataGauge is more likely that the current systematic methods to provide a useful body of evidence to justify or reject the use of a data source for a specific secondary use.

This evaluation provides evidence to confirm that DataGauge improves the current state of the art. DataGauge builds upon Kahn's framework for pragmatic DQ assessment framework (M. G. Kahn et al., 2012). It provides practical implementation support for a previously unavailable fitness for purpose DQ assessment of repurposed clinical datasets (Holve et al., 2013). It also provides methodological grounding for the implementation of

the recently published the ontology for the secondary use of clinical data (Johnson et al., 2015).

These results also encourage the use of fitness for purpose approaches for DQ assessment of repurposed clinical datasets (Holve et al., 2013) because of their ability to find specific issues that will encourage or discourage the use of a dataset for a specific purpose. Though these results are encouraging, a deeper evaluation of the method's usefulness and utility to real analytics teams is necessary. In this preliminary evaluation I employed automated methods as a starting point of comparison. However, research teams making secondary use of data usually carry out additional ad hoc tests and custom data cleaning. This limitation will be addressed in a future work.

Chapter 6: Conclusions, Limitations and Contributions

In this dissertation, I have described and evaluated DataGauge, the first systematic yet purpose-specific procedure for the DQ assessment of repurposed clinical data. DataGauge addresses current limitations in the state of the art of DQ assessments. It provides a generalized and systematic way of assessing repurposed clinical data taking into account the research question for the secondary analysis. Additionally, it makes DQ assessment parameters explicit, supporting communication within the analytics team and the transparent reporting of clinical data. I also presented a framework for the development of DQ requirements within DataGauge, which lists the major concerns to be checked when repurposing clinical data to promote thorough assessments. This framework aims to promote comprehensiveness and separation of concerns (Painter, 2006) when designing DQ assessments, which contributes to a thorough and orderly development of DQ requirements when applying DataGauge. Finally, I evaluated the DataGauge process comparing it to a current state of the art systematic DQ assessment method. DataGauge flagged more issues than the evaluation standard, uncovering bigger threats to fitness for purpose and covering more aspects of the assessed data and knowledge domains.

DataGauge improves the state of the art by supporting the systematic design and implementation of DQ assessments taking into account the secondary use purpose or

research question. Previous DQ assessment methods for the reuse of clinical data have focused on the application of multiple DQ tools to uncover discrete problems (e.g., typos, missing values, outliers) (Batini & Scannapieca, 2006b; Maydanchik, 2007a) and have failed to take the analytical purpose into account. Faulconer and de Lusignan (Faulconer & de Lusignan, 2004), for example, proposed a statistical procedure to identify DQ problems generically using mathematical tools rather than assessing potential threats to achieving the analytical purpose. Hogan and Wagner (Hogan & Wagner, 1997) also suggested statistical probes to identify specific data issues, but only addressed correctness and completeness. Kahn et al. (M. G. Kahn et al., 2012) proposed a purpose-based DQ assessment framework for the reuse of clinical data. However, this framework focuses on detecting DQ flaws based the data's primary clinical purpose rather than taking analysis-specific considerations into account. Johnson et al. (Johnson et al., 2015) have proposed an ontology to make DQ measure calculations systematic, but do not provide a way to select these measures systematically to assess fitness for purpose. Thus, DataGauge is the first method to address the problem of analysis-specific DQ assessments.

DataGauge is designed to be general enough to support a team of experts to systematically design and carry out their DQ assessment for any data source and analytical purpose. However, the development of DQ requirements across analytics teams is a potential source of inconsistencies. To address this limitation, we have provided a framework to support the systematic development of DQ standards. We achieve this by providing a list of potential DQ issues to consider when repurposing clinical data. This guidance framework is a first step towards supporting the systematic definition of DQ

requirements. Further development and testing of this framework will be part of my future work.

The definition of this method and evaluation results have multiple implications. For *clinicians and researchers making secondary use of clinical data*, DataGauge assessments allows for a more trustworthy repurposing of clinical data, by supporting thorough and purpose-specific assessments *before* the secondary analysis takes place. DataGauge also supports the reporting of DQ results by making the DQ assessment assumptions explicit, and therefore easier to report. DataGauge also promotes a more thorough assessment of repurposed clinical data because it accounts for the research question (i.e., the use purpose). This should lead to a better understanding of the limitations of repurposed clinical data and, in turn, the analytical results. *Informaticians* will benefit from this work by having a systematic way of assessing the quality of clinical data that is fully integrated with the data extraction process. DataGauge offers a streamlined way of integrating multiple workflows into a single process. This means that the work of the analytics team making secondary use of clinical data, the CDW team running the data extraction and the DQ assessment work are harmonized into a single workflow. The process is also supported by a set of documents that make assumptions explicit and may improve communication within and beyond the analytics team. DataGauge may also support the development of design tools and interfaces for secondary use and data extraction. Such tools are currently available but provide little to no support for DQ assessment.

6.1 - Limitations

While DataGauge was shown to flag more potential DQ issues than the current state of the art method, there are a number of limitations to this work. The DataGauge process presents four main limitations. First, we do not provide guidance in terms of the number of iterations to reach satisfactory specifications as this depends on multiple factors such as analysis type, data needs and research goals. This must, therefore, be left to expert judgment. Second, DataGauge is tailored to support analysis-specific DQ assessments and therefore, assumes that the input dataset has been a pre-cleaned to meet the DQ standards expected from a CDW. Third, DataGauge is human-intensive because the exhaustive definition of DQ requirements and their testing require considerable effort. Also, in its current state, DataGauge requires custom coding for the testing of every DQ requirement, which is much more time consuming than the fully automated evaluation standard. However, the great threats that arise from data repurposing (Hersh et al., 2013; Van Der Lei, 1991) demand thorough evaluations that are often not detectable using automated data checks. Finally, DataGauge only focuses on assessing defined by the DNM and data contained in the analytical dataset rather than all potentially relevant data in the CDW. Nevertheless, there are currently no other ways of defining analysis-specific data needs and quality requirements in a consistent way, which makes DataGauge a valuable tool for the clinical research informatics community.

The guidance framework to support DataGauge presents four main limitations. First, it may not include all relevant dimensions. Generating all relevant DQ requirements depends on many aspects of the data, their intended use and their meaning, yet the

dimensions included are the ones found to be most significant from our research experience. Second, the comprehensive nature of the framework may lead to the generation of redundant DQ requirements. Third, *the framework, in its current state, may tax users with a large number of questions for DQ requirement generation. Therefore, it may not be practical for day-to-day DQ assessment projects.* Refinement and testing of this guidance framework will be done in future work. Finally, though it interfaces with other DQ frameworks (Johnson et al., 2015; M. Kahn et al., 2015; M. G. Kahn et al., 2012) , it may be useful to integrate them in future work as well.

The evaluation of DataGauge presents four limitations. First, the evaluation was based on a single clinical database containing outpatient data. Though this is only one type of clinical data it is one of the most broadly available for secondary use and one database is a reasonable starting point for an initial evaluation. Second, a limited number of cases were used for the development of the guidance framework and evaluation. Still, these cases were justified to cover over 90% of secondary uses and clinical data types. Third, only two experts were interviewed to develop the DQ requirements that served as the basis to develop the guidance framework the evaluation. This is acceptable because the method is still in its early stages of development and this evaluation was preliminary in nature. Finally, the ideal comparison standard for the evaluation of DataGauge would have been a naive research team using their habitual cleaning methods. However, this option was impractical in our current setting for three reasons: (1) finding research teams willing to adopt a new, untested technique can be challenging, (2) including a human element in the evaluation process would introduce much variability in the results and

would, therefore, require a much larger sample size and (3) the funding needed to run such study was not currently available. DataGauge is in its early stages of its development and, thus, a proof of concept evaluation against the current state of the art tools is prudent and will inform future research. Thus, we chose to use a standardized baseline method that would minimize the involvement of humans and maximize the feasibility and repeatability of the evaluation.

6.2 - Contributions

DataGauge contributes to the current state of the science in three ways. First, it advances applications by supporting more through checks of repurposed clinical data and enabling their systematic yet purpose-specific DQ assessment. Such checks are necessary because repurposed clinical data may not be appropriate for their intended secondary purpose (Van Der Lei, 1991). This work lays the practical foundation for the systematic DQ assessment of repurposed clinical data as an evaluation of fitness for purpose (Holve et al., 2013). This contributes to support the reliable secondary use of clinical data (Charles Safran et al., 2007), which is a critical step towards building learning healthcare systems (Institute of Medicine (US) Roundtable on Evidence-Based Medicine, 2007). Second, DataGauge has been shown to improve upon the current state-of-the-art systematic DQ assessment method by providing a systematic yet purpose-specific approach. This new assessment methodology flags more potential issues than an alternative automated approach. It also improves current methods and practices by promoting the explicit definition of DQ assessment requirements (M. Kahn et al., 2015; M. G. Kahn et al., 2012) and data extraction parameters. These explicit definitions also show promise in enabling

a smoother data extraction process through the definition of explicit requirements. In our experience, data extraction is the source of much inefficiency and 'back and forth' between the analytics teams and database administrators this is partly due to the ambiguity of requests, usually submitted as plain text descriptions. Providing a standard for the explicit definition of data extraction requirements, such as the DNM, can eliminate much inefficiency by improving communication. The explicit definition of DQ requirements is also a contribution to current practices given that transparent reporting of DQ results is a current route of development in the field (M. Kahn et al., 2015), yet no applied methodological support to such practices is found in the literature. Also, explicit DQ requirements have the potential to set the groundwork for automated DQ assessment in the future. Lastly, my work contributes to the *field of biomedical informatics* by providing a preliminary inventory of concerns and potential issues to be checked while making secondary use of clinical data. The guidance framework provides a list of knowledge domains that ensure all expertise needed for the secondary use of data is accounted for. The guidance framework and, more specifically, the overview checklist supports the separation of concerns (Painter, 2006) in DQ requirement generation tasks. The principle of separation of concerns is responsible the orderly and modular design of current computer programs. I anticipate that this framework and overview based on this principle will be a first step towards an orderly and modular generation of DQ requirements. This is particularly important because of the complexity involved in the generation of DQ requirements, which stems from the multiple sources of information to be taken into account for a specific dataset and research question. This guidance also

promotes a team-based (Barlow, 2013) definition of requirements that is desirable, yet poorly supported by current practices (Broeck & Fadnes, 2013).

6.3 - Future Work

Future work will consist of further testing and development of the DataGauge and the guidance framework. DataGauge will be tested with analytics teams making secondary use of clinical data. An evaluation where a control group uses their native methods to clean their datasets and an experimental group applies the DataGauge procedure will be carried out. The DQ results for each group will be compared. Also, I will further expand the guidance framework by exploring more secondary use cases. Finally, the DataGauge process will be iteratively improved and streamlined based on end-user feedback. Because DataGauge is intentionally designed to require considerable human input, I plan to develop and evaluate tools to support analytical teams as they work through the process, much as software engineers and programmers now use a number of different tools to capture requirements, track changes, and track issues.

References

Amidan, B. G., Ferryman, T. A., & Cooley, S. K. (2005). Data outlier detection using the Chebyshev theorem. In *2005 IEEE Aerospace Conference* (pp. 3814–3819). http://doi.org/10.1109/AERO.2005.1559688

Aronsky, D., & Haug, P. J. (2000). Assessing the Quality of Clinical Data in a Computer-based Record for Calculating the Pneumonia Severity Index. *Journal of the American Medical Informatics Association*, *7*(1), 55–65. http://doi.org/10.1136/jamia.2000.0070055

Barlow, M. (2013). *The Culture of Big Data*. Retrieved from http://www.oreilly.com/data/free/culture-of-big-data.csp

Batini, C., & Scannapieca, M. (2006a). *Data Quality : Concepts, Methodologies and Techniques*. Berlin: Springer.

Batini, C., & Scannapieca, M. (2006b). Methodologies for Data Quality Measurement and Improvement. In *Data Quality: Concepts, Methodologies and Techniques* (pp. 161–200). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/3-540-33173-5_7

Batini, C., & Scannapieca, M. (2006c). Models for Data Quality. In *Data Quality: Concepts, Methodologies and Techniques* (pp. 51–68).

Bernstam, E. V., Herskovic, J. R., Reeder, P., & Meric-Bernstam, F. (2010). Oncology research using electronic medical record data. *ASCO Meeting Abstracts*, *28*(15_suppl), e16501.

Blumenthal, D. (2010). Launching HITECH. *New England Journal of Medicine*, *362*(5), 382–385. http://doi.org/10.1056/NEJMp0912825

Borek, A., Woodall, P., Oberhofer, M., & Parlikad, A. (2011). A classification of data quality assessment methods. In *ICIQ 2011-Proceedings of the 16th International Conference on Information Quality* (pp. 189–203).

Boselli, R., Cesarini, M., Mercorio, F., & Mezzanzanica, M. (2013). Inconsistency Knowledge Discovery for Longitudinal Data Management: A Model-Based Approach. In A. Holzinger & G. Pasi (Eds.), *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data* (pp. 183–194). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-39146-0_17

Botsis, T., Hartvigsen, G., Chen, F., & Weng, C. (2010). Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. *AMIA Summits on Translational Science Proceedings*, *2010*, 1.

Boytsov, A., & Zaslavsky, A. (2013). Formal verification of context and situation models in pervasive computing. *Pervasive and Mobile Computing*, *9*(1), 98–117. http://doi.org/10.1016/j.pmcj.2012.03.001

Broeck, J. V. den, & Fadnes, L. T. (2013). Data Cleaning. In J. V. den Broeck & J. R. Brestoff (Eds.), *Epidemiology: Principles and Practical Guidelines* (pp. 389–

399). Springer Netherlands. Retrieved from http://link.springer.com/chapter/10.1007/978-94-007-5989-3_20

Brookmeyer, R., & Crowley, J. (1982). A Confidence Interval for the Median Survival Time. *Biometrics*, *38*(1), 29–41. http://doi.org/10.2307/2530286

Brown, J. S., Kahn, M., & Toh, D. (2013). Data Quality Assessment for Comparative Effectiveness Research in Distributed Data Networks: *Medical Care*, *51*, S22–S29. http://doi.org/10.1097/MLR.0b013e31829b1e2c

Brownstein, J. S., Sordo, M., Kohane, I. S., & Mandl, K. D. (2007). The Tell-Tale Heart: Population-Based Surveillance Reveals an Association of Rofecoxib and Celecoxib with Myocardial Infarction. *PLoS ONE*, *2*(9), e840. http://doi.org/10.1371/journal.pone.0000840

Cabot, J. (2012, September 10). *MDE 2.0 : Pragmatical formal model verification and other challenges*. Universitat Politécnica de Catalunya. Retrieved from http://tel.archives-ouvertes.fr/tel-00915282

Cabot, J., Clariso, R., & Riera, D. (2008). Verification of UML/OCL Class Diagrams using Constraint Programming. In *IEEE International Conference on Software Testing Verification and Validation Workshop, 2008. ICSTW '08* (pp. 73–80). http://doi.org/10.1109/ICSTW.2008.54

Cabot, J., Clarisó, R., & Riera, D. (2014). On the verification of UML/OCL class diagrams using constraint programming. *Journal of Systems and Software*, *93*, 1–23. http://doi.org/10.1016/j.jss.2014.03.023

Cabot, J., & Gogolla, M. (2012). Object Constraint Language (OCL): A Definitive Guide. In M. Bernardo, V. Cortellessa, & A. Pierantonio (Eds.), *Formal Methods for Model-Driven Engineering* (pp. 58–90). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-30982-3_3

C, C.-B., Kp, A., & W, K. (1994). MONI: an intelligent database and monitoring system for surveillance of nosocomial infections. *Medinfo. MEDINFO*, *8 Pt 2*, 1684–1684.

Chalasani, N., Aljadhey, H., Kesterson, J., Murray, M. D., & Hall, S. D. (2004). Patients with elevated liver enzymes are not at higher risk for statin hepatotoxicity. *Gastroenterology*, *126*(5), 1287–1292. http://doi.org/10.1053/j.gastro.2004.02.015

Classen, D. C., & Burke, J. P. (1995). The computer-based patient record: the role of the hospital epidemiologist. *Infection Control and Hospital Epidemiology: The Official Journal of the Society of Hospital Epidemiologists of America*, *16*(12), 729–736.

Dale, B. (2015). Total Quality Management. In *Wiley Encyclopedia of Management*. John Wiley & Sons, Ltd. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9781118785317.weom100042/abstract

Dasu, T. (2013). Data Glitches: Monsters in Your Data. In S. Sadiq (Ed.), *Handbook of Data Quality* (pp. 163–178). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-36257-6_8

De Lusignan, S., Liaw, S., Krause, P., Curcin, V., Vicente, M., Michalakidis, G., … Mendis, K. (2011). Key concepts to assess the readiness of data for International

research: Data quality, lineage and provenance, extraction and processing errors, traceability, and curation. *Yearbook of Medical Informatics*, *6*(1), 112–20.

de Lusignan, S., & Mimnagh, C. (2006). Breaking the first law of informatics: the Quality and Outcomes Framework (QOF) in the dock. *Informatics in Primary Care*, *14*(3), 153–156.

Demuth, B., & Hussmann, H. (1999). Using UML/OCL Constraints for Relational Database Design. In R. France & B. Rumpe (Eds.), *«UML»'99 — The Unified Modeling Language* (pp. 598–613). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/3-540-46852-8_42

Dentler, K., Cornet, R., Teije, A. ten, Tanis, P., Klinkenbijl, J., Tytgat, K., & Keizer, N. de. (2014). Influence of data quality on computed Dutch hospital quality indicators: a case study in colorectal cancer surgery. *BMC Medical Informatics and Decision Making*, *14*(1), 32. http://doi.org/10.1186/1472-6947-14-32

Dentler, K., ten Teije, A., de Keizer, N., & Cornet, R. (2013). Barriers to the reuse of routinely recorded clinical data: a field report. *Studies in Health Technology and Informatics*, *192*, 313–317.

Evans, J. R., & Lindsay, W. M. (1999). The management and control of quality.

Fan, W. (2012). Data Quality: Theory and Practice. In H. Gao, L. Lim, W. Wang, C. Li, & L. Chen (Eds.), *Web-Age Information Management* (pp. 1–16). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-32281-5_1

Fan, W., Geerts, F., Ma, S., Tang, N., & Yu, W. (2013). Data Quality Problems beyond Consistency and Deduplication. In V. Tannen, L. Wong, L. Libkin, W. Fan, W.-C. Tan, & M. Fourman (Eds.), *In Search of Elegance in the Theory and Practice of Computation* (pp. 237–249). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-41660-6_12

Faulconer, E. R., & de Lusignan, S. (2004). An eight-step method for assessing diagnostic data quality in practice: chronic obstructive pulmonary disease as an exemplar. *Informatics in Primary Care*, *12*(4), 243–254.

Finnell, J. T., Overhage, J. M., & Grannis, S. (2011). All health care is not local: an evaluation of the distribution of Emergency Department care delivered in Indiana. *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium*, *2011*, 409–416.

Floridi, L. (2012). Big Data and Their Epistemological Challenge. *Philosophy & Technology*, *25*(4), 435–437. http://doi.org/10.1007/s13347-012-0093-4

Floridi, L. (2013). Information quality. *Philosopy and Technology*, *26*.

France, R., & Rumpe, B. (2007). Model-driven Development of Complex Software: A Research Roadmap. In *Future of Software Engineering, 2007. FOSE '07* (pp. 37–54). http://doi.org/10.1109/FOSE.2007.14

Fries, J. F., & McShane, D. (1979). ARAMIS: A National Chronic Disease Data Bank System. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 798–801.

Gómez, C. C. (2009). Assessing the Quality of Qualitative Health Research: Criteria, Process and Writing. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *10*(2). Retrieved from http://www.qualitative-research.net/index.php/fqs/article/view/1294

González, C. A., & Cabot, J. (2014). Formal verification of static software models in MDE: A systematic. *Information and Software Technology*, *56*, 821–838.

Guyatt G, Cairns J, Churchill D, & et al. (1992). Evidence-based medicine: A new approach to teaching the practice of medicine. *JAMA*, *268*(17), 2420–2425. http://doi.org/10.1001/jama.1992.03490170092032

Hansell, A., Hollowell, J., Nichols, T., McNiece, R., & Strachan, D. (1999). Use of the General Practice Research Database (GPRD) for respiratory epidemiology: a comparison with the 4th Morbidity Survey in General Practice (MSGP4). *Thorax*, *54*(5), 413–419. http://doi.org/10.1136/thx.54.5.413

Herrmann, F., & Safran, C. (1992). Real time exploration of routinely collected data: An analysis of admissions for AIDS in a teaching hospital, 878–882.

Hersh, W. R. (2007). Adding value to the electronic health record through secondary use of data for quality assurance, research, and surveillance. *Clin Pharmacol Ther*, *81*, 126–128.

Hersh, W. R., Weiner, M. G., Embi, P. J., Logan, J. R., Payne, P. R. O., Bernstam, E. V., … Saltz, J. H. (2013). Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research. *Medical Care*, *51*(8 0 3), S30–S37. http://doi.org/10.1097/MLR.0b013e31829b1dbd

Herzig SJ, Howell MD, Ngo LH, & Marcantonio ER. (2009). ACid-suppressive medication use and the risk for hospital-acquired pneumonia. *JAMA*, *301*(20), 2120–2128. http://doi.org/10.1001/jama.2009.722

Hogan, W. R., & Wagner, M. M. (1997). Accuracy of Data in Computer-based Patient Records. *Journal of the American Medical Informatics Association*, *4*(5), 342–355. http://doi.org/10.1136/jamia.1997.0040342

Holve, E., Kahn, M., Nahm, M., Ryan, P., & Weiskopf, N. (2013). A comprehensive framework for data quality assessment in CER. *AMIA Summits on Translational Science Proceedings*, *2013*, 86–88.

Hripcsak, G., Albers, D. J., & Perotte, A. (2011). Exploiting time in electronic health record correlations. *Journal of the American Medical Informatics Association*, amiajnl–2011–000463. http://doi.org/10.1136/amiajnl-2011-000463

Hripcsak, G., Duke, J., Shah, N., Reich, C., Huser, V., Schuemie, M., … others. (2015). Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *MEDINFO*, *15*.

Hripcsak, G., Friedman, C., Alderson, P. O., DuMouchel, W., Johnson, S. B., & Clayton, P. D. (1995). Unlocking Clinical Data from Narrative Reports: A Study of Natural Language Processing. *Annals of Internal Medicine*, *122*(9), 681–688. http://doi.org/10.7326/0003-4819-122-9-199505010-00007

Hripcsak, G., Knirsch, C., Zhou, L., Wilcox, A., & Melton, G. B. (2007). Using discordance to improve classification in narrative clinical databases: An

application to community-acquired pneumonia. *Computers in Biology and Medicine*, *37*(3), 296–304. http://doi.org/10.1016/j.compbiomed.2006.02.001

Hripcsak, G., Knirsch, C., Zhou, L., Wilcox, A., & Melton, G. B. (2011). Bias Associated with Mining Electronic Health Records. *Journal of Biomedical Discovery and Collaboration*, *6*, 48–52.

Institute of Medicine (US) Roundtable on Evidence-Based Medicine. (2007). *The Learning Healthcare System: Workshop Summary*. (L. Olsen, D. Aisner, & J. M. McGinnis, Eds.). Washington (DC): National Academies Press (US). Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK53494/

Jha AK. (2010). Meaningful use of electronic health records: The road ahead. *JAMA*, *304*(15), 1709–1710. http://doi.org/10.1001/jama.2010.1497

Johnson, S. G., Speedie, S., Simon, G., Kumar, V., & Westra, B. L. (2015). A Data Quality Ontology for the Secondary Use of EHR Data. Presented at the AMIA Symposium, San Francisco, CA: AMIA. Retrieved from https://knowledge.amia.org/59310-amia-1.2741865/t007-1.2744224/f007-1.2744225/2246427-1.2744284/2246427-1.2744285?timeStamp=1456356254401

Juran, J. M. (1962). Quality control handbook. In *Quality control handbook*. McGraw-Hill.

Kahn, M., Brown, J., Chun, A., Davidson, B., Meeker, D., Ryan, P., … Zozus, M. (2015). Transparent Reporting of Data Quality in Distributed Data Networks. *eGEMs (Generating Evidence & Methods to Improve Patient Outcomes)*, *3*(1). http://doi.org/10.13063/2327-9214.1052

Kahn, M. G., Raebel, M. A., Glanz, J. M., Riedlinger, K., & Steiner, J. F. (2012). A Pragmatic Framework for Single-site and Multisite Data Quality Assessment in Electronic Health Record-based Clinical Research. *Medical Care*, *50*(0). http://doi.org/10.1097/MLR.0b013e318257dd67

Kan, S. H. (2002). *Metrics and Models in Software Quality Engineering* (2nd ed.). Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Kent, W. (1983). A Simple Guide to Five Normal Forms in Relational Database Theory. *Commun. ACM*, *26*(2), 120–125. http://doi.org/10.1145/358024.358054

Lee, Y. W., Strong, D. M., Kahn, B. K., & Wang, R. Y. (2002). AIMQ: a methodology for information quality assessment. *Information & Management*, *40*(2), 133–146. http://doi.org/10.1016/S0378-7206(02)00043-5

Madnick, S. E., Wang, R. Y., Lee, Y. W., & Zhu, H. (2009). Overview and Framework for Data and Information Quality Research. *J. Data and Information Quality*, *1*(1), 2:1–2:22. http://doi.org/10.1145/1515693.1516680

Maydanchik, A. (2007a). *Data Quality Assessment*. Technics Publications.

Maydanchik, A. (2007b). On Hunting Mammoths and Measuring Data Quality.

Mayrand, J., & Coallier, F. (1996). System acquisition based on software product assessment. In *, Proceedings of the 18th International Conference on Software Engineering, 1996* (pp. 210–219). http://doi.org/10.1109/ICSE.1996.493417

Mezzanzanica, M., Boselli, R., Cesarini, M., & Mercorio, F. (2011). Data Quality through Model Checking Techniques. In J. Gama, E. Bradley, & J. Hollmén (Eds.), *Advances in Intelligent Data Analysis X* (pp. 270–281). Springer Berlin

Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-24800-9_26

Morton, R. B. (1999). *Methods and Models: A Guide to the Empirical Analysis of Formal Models in Political Science*. Cambridge England ; New York: Cambridge University Press.

Nelson, C. W., & Niederberger, J. (1990). Patient satisfaction surveys: an opportunity for total quality improvement. *Hospital & Health Services Administration*, *35*(3), 409–428.

Oliveira, P., Rodrigues, F., & Henriques, P. R. (2005). A Formal Definition of Data Quality Problems. In *IQ*.

Olson, J. E. (2003). *Data Quality: The Accuracy Dimension*. Morgan Kaufmann.

Ossher, H., & Tarr, P. (2001). Using multidimensional separation of concerns to (re) shape evolving software. *Communications of the ACM*, *44*(10), 43–50.

Overhage, J. M., Ryan, P. B., Reich, C. G., Hartzema, A. G., & Stang, P. E. (2012). Validation of a common data model for active safety surveillance research. *Journal of the American Medical Informatics Association*, *19*(1), 54–60. http://doi.org/10.1136/amiajnl-2011-000376

Painter, R. R. (2006). *Software Plans: Multi-Dimensional Fine-Grained Separation of Concerns*. Citeseer.

Parsons, A., McCullough, C., Wang, J., & Shih, S. (2012). Validity of electronic health record-derived quality measurement for performance monitoring. *Journal of the*

American Medical Informatics Association, amiajnl–2011–000557. http://doi.org/10.1136/amiajnl-2011-000557

Phillips, D. R. S., Safran, C., Cleary, P. D., & Delbanco, T. L. (1987). Predicting emergency readmissions for patients discharged from the medical service of a teaching hospital. *Journal of General Internal Medicine*, *2*(6), 400–405. http://doi.org/10.1007/BF02596366

Pinet, F., Kang, M.-A., Boulil, K., Bimonte, S., De Sousa, G., Roussey, C., & Schneider, M. (2011). Using OCL to Model Constraints in Data Warehouses: *International Journal of Technology Diffusion*, *2*(3), 36–46. http://doi.org/10.4018/jtd.2011070104

PredniSONE Tablets [Package Insert]. (2012). Ridgefield, CT : Boehringer-Ingelheim Inc.

Rahbar, M. H., Gonzales, N. R., Ardjomand-Hessabi, M., Tahanan, A., Sline, M. R., Peng, H., … Grotta, J. C. (2013). The University of Texas Houston Stroke Registry (UTHSR): implementation of enhanced data quality assurance procedures improves data quality. *BMC Neurology*, *13*(1), 61. http://doi.org/10.1186/1471-2377-13-61

R Core Team. (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Redman, T. C. (1998). The Impact of Poor Data Quality on the Typical Enterprise. *Commun. ACM*, *41*(2), 79–82. http://doi.org/10.1145/269012.269025

Redman, T. C. (2013). Data Quality Management Past, Present, and Future: Towards a Management System for Data. In S. Sadiq (Ed.), *Handbook of Data Quality* (pp. 15–40). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-36257-6_2

Reisinger, S. J., Ryan, P. B., O'Hara, D. J., Powell, G. E., Painter, J. L., Pattishall, E. N., & Morris, J. A. (2010). Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases. *Journal of the American Medical Informatics Association*, *17*(6), 652–662. http://doi.org/10.1136/jamia.2009.002477

Sadiq, S. (2013a). *Handbook of Data Quality: Research and Practice*. Springer Publishing Company, Incorporated.

Sadiq, S. (2013b). Prologue: Research and Practice in Data Quality Management. In S. Sadiq (Ed.), *Handbook of Data Quality* (pp. 1–11). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-36257-6_1

Safran, C. (1991). Using routinely collected data for clinical research. *Statistics in Medicine*, *10*(4), 559–564.

Safran, C. (2014). Reuse Of Clinical Data. *IMIA Yearbook*, *9*(1), 52–54. http://doi.org/10.15265/IY-2014-0013

Safran, C., Bloomrosen, M., Hammond, W. E., Labkoff, S., Markel-Fox, S., Tang, P. C., & Detmer, D. E. (2007). Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper. *Journal*

of the American Medical Informatics Association, *14*(1), 1–9. http://doi.org/10.1197/jamia.M2273

Samore, M., Lichtenberg, D., Saubermann, L., Kawachi, C., & Carmeli, Y. (1997). A clinical data repository enhances hospital infection control. *Proceedings of the AMIA Annual Fall Symposium*, 56–60.

Schmidt, D. C. (2006). Model-driven engineering. *COMPUTER-IEEE COMPUTER SOCIETY-*, *39*(2), 25.

Seiter, J., Wille, R., Soeken, M., & Drechsler, R. (2013). Determining relevant model elements for the verification of UML/OCL specifications. In *Design, Automation Test in Europe Conference Exhibition (DATE), 2013* (pp. 1189–1192). http://doi.org/10.7873/DATE.2013.247

Selic, B. (2004). Tutorial: An overview of UML 2.0. In *Proceedings of the 26th International Conference on Software Engineering* (pp. 741–742). IEEE Computer Society.

Song, M., Liu, K., Abromitis, R., & Schleyer, T. L. (2013). Reusing electronic patient data for dental clinical research: A review of current status. *Journal of Dentistry*, *41*(12), 1148–1163. http://doi.org/10.1016/j.jdent.2013.04.006

Standardization, I. O. for. (1994). *ISO 8402: 1994: Quality Management and Quality Assurance-Vocabulary*. International Organization for Standardization.

Starmer, C. F., Rosati, R. A., & Fred McNeer, J. (1974). Data bank use in management of chronic disease. *Computers and Biomedical Research*, *7*(2), 111–116.

Stvilia, B., Gasser, L., Twidale, M. B., & Smith, L. C. (2007). A framework for information quality assessment. *Journal of the American Society for Information Science and Technology*, *58*(12), 1720–1733. http://doi.org/10.1002/asi.20652

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics*.

Taguchi, G. (1986). *Introduction to quality engineering: designing quality into products and processes*.

The Merriam-Webster Dictionary. (n.d.).

The Philosophy of Information. (2013) (Reprint edition). Oxford: Oxford University Press.

Trickey, A. W. (2012). *Data quality in trauma transfusion studies and the impact of missing data on predicting massive transfusion*. THE UNIVERSITY OF TEXAS SCHOOL OF PUBLIC HEALTH. Retrieved from http://gradworks.umi.com.ezproxyhost.library.tmc.edu/35/50/3550588.html

Van den Broeck, J., Argeseanu Cunningham, S., Eeckels, R., & Herbst, K. (2005). Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities. *PLoS Med*, *2*(10), e267. http://doi.org/10.1371/journal.pmed.0020267

Van Der Lei, J. (1991). Use and abuse of computer-stored medical records. *Methods of Information in Medicine*, *30*(2), 79–80.

Walker, A., & Gee, C. (2000). {ISO 9001 model support for software process assessment}. *Logistics Information Management*, *13*(1).

Wang, R. Y. (1998). A Product Perspective on Total Data Quality Management. *Commun. ACM*, *41*(2), 58–65. http://doi.org/10.1145/269012.269022

Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, *12*(4), 5–33.

Weiner, M. G., & Embi, P. J. (2009). Toward Reuse of Clinical Data for Research and Quality Improvement: The End of the Beginning? *Annals of Internal Medicine*, *151*(5), 359–360. http://doi.org/10.7326/0003-4819-151-5-200909010-00141

Weiskopf, N., George Hripcsak, Swaminathan, S., & Weng, C. (2013). Defining and measuring completeness of electronic health records for secondary use. *Journal of Biomedical Informatics*, *46*(5), 830–836. http://doi.org/10.1016/j.jbi.2013.06.010

Weiskopf, N. G., Rusanov, A., & Weng, C. (2013). Sick Patients Have More Data: The Non-Random Completeness of Electronic Health Records. *AMIA Annual Symposium Proceedings*, *2013*, 1472–1477.

Weiskopf, N. G., & Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, *20*(1), 144–151. http://doi.org/10.1136/amiajnl-2011-000681

Whittle, J., Hutchinson, J., & Rouncefield, M. (2014). The State of Practice in Model-Driven Engineering. *IEEE Software*, *31*(3), 79–85. http://doi.org/10.1109/MS.2013.65

Wickham, H. (2014). Tidy data. *Under Review*.

Zubcoff, J., Pardillo, J., & Trujillo, J. (2009). A UML profile for the conceptual modelling of data-mining with time-series in data warehouses. *Information and Software Technology*, *51*(6), 977–992. http://doi.org/10.1016/j.infsof.2008.09.006

Appendix A: Definitions

Data: "A datum is a putative fact regarding some difference or lack of uniformity within some context." (*The Philosophy of Information*, 2013) In other words, data are discrete, atomic answers to specific questions about an object of interest.

Clinical Data: Discrete, atomic answers to specific questions about a patient's health status and healthcare procedures. In this dissertation, I refer to clinical data as all data recorded in an electronic health record.

Quality: "The totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs" (Standardization, 1994)

Data Quality: A dataset's ability of satisfying the needs for a specific purpose (Holve et al., 2013) (i.e., fitness for purpose (Juran, 1962))

Purpose: "Something set up as an object or end to be attained" (*The Merriam-Webster Dictionary*, n.d.). We refer to purpose in this dissertation as the use of data to answer a specific research question.

Initial purpose: The purpose for which the data are first produced; also known as the production purpose (Floridi, 2012, 2013).

Data Repurposing: Using data for any purpose other than for which they were produced(de Lusignan & Mimnagh, 2006; Van Der Lei, 1991).

Secondary Purpose or Secondary use: Any use of data other than the use for which they were collected; also known as the analytical or secondary purpose(Botsis, Hartvigsen, Chen, & Weng, 2010; Floridi, 2012, 2013; C. Safran, 2014).

Criterion: "A standard on which a judgment or decision may be based or a characterizing mark or trait" (*The Merriam-Webster Dictionary*, n.d.). For this dissertation, I define a criterion is a axis of interest upon which a value of adequacy can be assigned, an evaluation can be carried out or a decision can be made.

DQ criterion: A point of interest upon which a dataset's fitness for a specific use can be evaluated. The evaluation will provide a degree of adequacy for the use in relation to a specific criterion. Each criterion defines a sub-feature of the overall fitness for purpose.

DQ requirement: A condition describing a specific feature that must be respected by a dataset in order to be fit for a specific use or purpose. Each requirement defines a specific sub-feature of the overall fitness for purpose. Requirements describe the desirable aspect for a dataset in the light of a criterion, which is defined by the intended use. DQ requirements define the minimum expectation on data values to ensure that a specific dataset is valid and useful for a specific analytical purpose. They aim to define a "fit-for-purpose" dataset in a specific case.

DQ Standard: In this dissertation, I refer to DQ standard as the combination of all DQ requirements for a specific dataset and intended use. DQ standards are not interchangeable and must be generated in the light of domain knowledge, the available data and the intended purpose or research task.

Inclusion/Exclusion Criteria: We differentiate DQ requirements from Inclusion/Exclusion criteria by the object they define and their goal. Inclusion/Exclusion criteria define the subjects that qualify for the study based on demographic and clinical considerations; they aim to define a patient population. For example, "patient is at least 18 years old" is an inclusion criterion whereas "patient date of birth is earlier than observation date" is a DQ requirement.

DQ rule: "Data quality rules are constraints that validate data relationships and can be checked using computer programs" (Maydanchik, 2007a). In this dissertation I refer to DQ rules as an explicit, unambiguous limits that are easily encoded into machine-executable code to automatically flag infringing data within the assessed datasets. We differentiate them from DQ requirements by their unambiguous and formal (Morton, 1999) nature.

DQ test: A practical tool, algorithm, approach or strategy employed to test the adherence of a dataset to a specific DQ requirement or the breach of a DQ rule. They serve as a means to gather evidence of a dataset's fitness for purpose for a specific DQ criterion.

DQ Assessment: A judiciously selected combination of DQ tests based on DQ requirements to assess a dataset's fitness for a specific analytical purpose based on clinical, data science and analytical tool knowledge.
DQ assessments vs. DQ tests: A DQ assessment is a combination of DQ tests designed to evaluate whether a dataset is fit for a specific purpose. DQ requirements and DQ rules define the parameters necessary to run the DQ tests.

Data Needs Model: An explicit, unambiguous external representation of the minimal data needed to achieve a specific research task or goal. In this dissertation, data needs

model refers to a tidy-data-compliant (Wickham, 2014) UML model (Demuth & Hussmann, 1999) of the minimal clinical data required to answer a specific research question.

Research Question Analysis: To define all objects of interest in light of a specific research question. In this dissertation, I refer to research question analysis as the steps taken to determine the clinical objects (e.g., patients, prescriptions, labs, etc.), which must necessarily be known to answer the research questions. From these objects are derived the variables that are likely to contain the necessary information to answer the research question via secondary analysis. For example, if our research question is "Is Prednisone exposure related to weight gain?" we will need information about the patients, their prednisone prescription history and their weight measurement values. Within these three objects we will then define the variables to describe then. For example, a patient may be described by a patient identifier, a date of birth and a gender variable.

Data extraction: Selection of all data that may be relevant to a specific research question that is transformed to fit a specific data needs model and subsequently used for DQ evaluation and analysis.

DQ evaluation: The implementation of all DQ tests (based on the DQ standard) into machine-executable code to detect violations of the predefined DQ requirements. Their goal is to provide evidence to support the analytics team's decision on the dataset's fitness for purpose. Each DQ evaluation is specific to a research question, a data needs model, a DQ standard and an extracted dataset.

DQ results: They are the quantitative evidence of infringement of DQ requirements within a given dataset for a specific purpose. They usually take the form of counts, percentages or true/false flags and serve to support the expert's decision about a dataset being fit or unfit for a specific purpose. They can be assimilated to DQ measures as described by Johnson et al. (Johnson et al., 2015). In this dissertation, I represented all DQ results as infringement counts or flag counts.

Separation of concerns: "The ability to identify, encapsulate and manipulate only those parts of software that are relevant to a particular concept, goal, or purpose" (Ossher & Tarr, 2001). In other words, it is the ability of teasing out the different pieces of the puzzle interacting in a task. In the task of generating DQ requirement, multiple sources of information are at play as well as multiple sources of knowledge (i.e., domain experts). In this case, the separation of concerns is the definition of an information structure or framework that delineates the role and interactions of each information and knowledge source. The guidance framework described in Chapter 4 provides such structure.

Appendix B: DQ Requirement Development Guidance

In this Appendix I present the details of the guidance framework introduced in Chapter 4. The framework is composed of two parts: (1) the Data-Knowledge checklist and (2) the guidance questionnaires. The Data-Knowledge checklist serves as an overview of all possible contexts (i.e., the combination of a clinical data type and a knowledge domain) that must be assessed to ensure a thorough DQ assessment design. For each DQ requirement generation context (i.e., the combination of a knowledge domain and a clinical data type instance), the framework assigns a guidance questionnaire that presents all relevant concerns in the form of questions to assess DQ dimension. Each question is used as a focus to generate DQ requirements around that specific point of interest taking as inputs a specific Data Needs Model piece at a specific data granularity level (e.g., a value, an observation, a variable, etc.) and a specific Research Question. The DQ requirements should aim to define the ideal dataset for the Research Question or Research Goal. Once the analytics team has addressed all pertinent questionnaires and the checklist is full, it can be assumed that there is reasonable coverage of potential DQ issues in repurposed clinical data.

Data-Knowledge Requirements Development Checklist

The Data-Knowledge checklist serves as an overview of the different contexts to be covered for a thorough DQ assessment design (see table below). Two dimensions define it: (1) the Clinical Data Types, which are sub-sections of the electronic medical record and (2) the Knowledge Domain, which represent the different types of knowledge necessary to carry out a secondary analysis of clinical data. Thus, we can refer to contexts as the combination of specific elements from these to dimensions. For example, Appointment data type and Analytical Tool would be the first context in the table below. Addressing all combinations between all elements between these two dimensions ensure systematicity through thorough coverage of all relevant issues. It is recommended that research teams work in groups with all experts present in the room. The Research Question and the Data Needs Model should be visible to all and consulted as part of the process. It is suggested that the team work their way through the checklist by breaking the Data Needs Model down for a Clinical Data Type at all relevant Data Granularity Levels, then running thought the questionnaires each knowledge domain and finally move on to the next Clinical Data Type. This will facilitate cross-domain thought and dialog between the experts, while keeping the evaluated target (i.e., the DNM pieces) stable during DQ requirement generation.

	Clinical Data Types:	Appointment	Demographics	Diagnoses	Labs	Meds	Vitals
Knowledge Domain	**Analytical Tool**	☑	☐	☐	☐	☐	☐
	Clinical	☐	☐	☐	☐	☐	☐
	Data Manipulation	☐	☐	☐	☐	☐	☐
	Representation	☐	☐	☐	☐	☐	☐
	Research Design	☐	☐	☐	☐	☐	☐
	Research Goal	☐	☐	☐	☐	☐	☐
	Workflow	☐	☐	☐	☐	☐	☐

Question-Driven DQ Requirement Definition

The guidance questionnaires correspond to specific contexts; that is, a specific combination of Knowledge Domain and Clinical Data Type (e.g., Analytical Tool and Appointment). For each context, the analytics team will identify the relevant subset of the Data Needs Model and break them down in as many possible pieces for every Data Granularity Level (i.e., value, variable, observation, observational unit, multiple observational units and dataset). They will then address each question for a specific DNM piece in the frame of a specific Knowledge Domain. These features are the DQ requirements. Once the team has answered every question, they may move on to the next context and answer its questionnaire. It is important that the team keep in mind the research question throughout this work in order to generate requirements that define a fit for purpose dataset. A full example of the use of this guidance is described in section 4.4.

Contextual Questionnaires In Alphabetical Order

1. Appointment

1.1. Analytical Tool

DQ Dimension	Question
Tool Limitations	Are all statistical assumptions met by the data?
Tool Limitations	Can you assume independence between observations?
Completeness	Are there enough observations within the timeframe of interest to run the analysis?
Completeness	Are there enough patients with all relevant variables within the timeframe of interest to run the analysis?
Completeness	Are there enough observations to provide appropriate statistical power?
Completeness	Is the censorship rate acceptable?
Completeness	Are the measurements recorded at reasonably regular intervals?
Completeness	Are there gaps in recording within the timeframe of interest?
Correctness	Are there duplicate observations?
Completeness	Are there enough data points by appointments?
Completeness	Are there enough data points by patient and/or care provider?

1.2. Clinical

DQ Dimension	Guidance Question
Plausibility	Is the appointment date within a plausible timeframe?
Timeliness	Is the time between appointments in the expected range?

1.3. Data Manipulation

This Knowledge Domain is beyond DQ dimensions and requirements. Good data management practices should prevent any DQ issues in this context, yet it is recommended to verify that no issues arose from extraction. At this stage, the data scientist designs tests and fail-safes to avoid corrupting the dataset during the data manipulations and extract-transform-load (ETL) procedures and/or identify them once the data has been extracted. One example such issues is the creation of duplicates when using joins. To avoid such issues, the data base administrator should verify the counts after every join to ensure no duplication has taken place. These checks are usually done before the dataset is extracted and assessed for DQ. DQ requirements are not necessary beyond the tests defined by the data scientist but could still be defined for thoroughness purposes.

1.4. Representation

DQ Dimension	Guidance Question
Representation	Are all Primary Keys unique?
Representation	Do all foreign keys correspond to one primary key?
Representation	Are all values formatted appropriately or as expected?
Representation	Are all values of the right sort (e.g., character, numeric, etc.)?
Representation	Are all predefined values (i.e., concepts) found in the dictionary? (e.g., ICD-9)
Representation	Are all defined concepts of interest found in the dataset? (e.g., if the concept 'male' is defined to be used in the gender variable, the variable should contain 'male' values)

1.5. Research Design

DQ Dimension	Guidance Question
Completeness	Are any of variables needed for the analysis missing?
Completeness	Are there missing values?
Completeness	Are there values missing not at random?
Completeness	Are there enough values within the desired time range?
Completeness	Is the analytical observational unit (i.e., the main outcome variable, dependent variable and covariates in one table line) complete?
Completeness	Is the censorship level for each variable of the observational unit acceptable?
Correctness	Does the observational unit follow the ideal timeline of collection? (e.g., vitals are recorded on visit days)
Completeness	Is the rate of censorship acceptable? Are there enough encounters per patient for the designated research design?
Completeness	Are the missing values spread randomly over time?
Correctness	Does the distribution of demographics correspond to population estimates at large? (i.e., is the sampling random?)
Completeness	Are the data accessible for analysis? (e.g., are the data not locked in the clinical notes)
Completeness	Is the ratio of censorship acceptable (i.e., # missing data/# available data)?

1.6. Research Goal

DQ Dimension	Guidance Question
Completeness	Are the appointment records complete?
Plausibility	Is the age of patient plausible at the appointment time?
Correctness	Does the appointment record map to a single patient?
Correctness	Is there at least one appointment per patient?

1.7. Workflow

DQ Dimension	Guidance Question
Concordance	Are all providers found and defined in the providers table?
Concordance	Are all locations found and defined in the care sites table?
Concordance	Are all provider occupations as expected? (e.g., no medical assistants prescribing drugs)
Correctness	Are time stamps in the expected order? (e.g., order before admin).
Timeliness	Are time stamps within the expected time frame? (e.g., BPs measured within the encounter window vs. outside)
Correctness	Is the time between events of the expected range?

2. Demographics

2.1. Analytical Tool

DQ Dimension	Guidance Question
Tool Limitations	Are all statistical assumptions met by the data?
Tool Limitations	Can you assume independence between observations?
Completeness	Are there enough observations within the timeframe of interest?
Completeness	Are there enough patients with all relevant variables within the timeframe of interest?
Completeness	Are there enough observations to provide appropriate statistical power?
Completeness	Is censorship at an acceptable rate?
Completeness	Are the measurements recorded at reasonably regular intervals?
Completeness	Are there gaps in recording within the timeframe of interest?
Concordance	Are there duplicate observations?
Completeness	Are there enough data points by encounter?
Completeness	Are there enough data points by patient and/or care provider?

2.2. Clinical

DQ Dimension	Guidance Question
Correctness	Is gender Male, Female or Unknown?
Completeness	Does race contain all expected values? In the expected proportions?
Correctness	Are ages within inclusion criteria, non-negative and below 130?
Plausibility	Are all dates of birth later than Jan 1st 1900?

2.3. Data Manipulation

This Knowledge Domain is beyond DQ dimensions and requirements. Good data management practices should prevent any DQ issues in this context, yet it is recommended to verify that no issues arose from extraction. At this stage, the data scientist designs tests and fail-safes to avoid corrupting the dataset during the data manipulations and extract-transform-load (ETL) procedures and/or identify them once the data has been extracted. One example such issues is the creation of duplicates when using joins. To avoid such issues, the data base administrator should verify the counts after every join to ensure no duplication has taken place. These checks are usually done before the dataset is extracted and assessed for DQ. DQ requirements are not necessary beyond the tests defined by the data scientist but could still be defined for thoroughness purposes.

2.4. Representation

DQ Dimension	Guidance Question
Representation	Are all Primary Keys unique?
Representation	Do all foreign keys correspond to one primary key?
Representation	Are all values formatted appropriately or as expected?
Representation	Are all values of the right sort (e.g., character, numeric, etc.)?
Representation	Are all predefined values (i.e., concepts) found in the dictionary? (e.g., ICD-9)
Representation	Are all defined concepts of interest found in the dataset? (e.g., if the concept 'male' is defined to be used in the gender variable, the variable should contain 'male' values)

2.5. Research Design

DQ Dimension	Guidance Question
Completeness	Are there missing variables?
Completeness	Are there missing values?
Completeness	Are the values missing not at random?
Completeness	Are there enough values within the desired time range?
Completeness	Is the observational unit defined for the analysis complete? (i.e., are all necessary variables present)
Completeness	Is the censorship level for the analytical unit observational unit acceptable? (i.e., are there enough values for the analysis overall)
Correctness	Does the observational unit follow the ideal timeline of collection?
Completeness	Are the missing values spread randomly over time?
Correctness	Does the distribution of demographics correspond to population estimates at large? (i.e., is the sampling random?)
Completeness	Are the data accessible for analysis? (e.g., the data are not locked in clinical notes)
Completeness	Is the ratio of censorship acceptable (i.e., # missing data/# available data)?

2.6. Research Goal

DQ Dimension	Guidance Question
Completeness	Are the demographics records complete? (e.g., patients have gender, date of birth and race values)
Completeness	Does the demographic data have null values?
Plausibility	Is the age of the patient plausible? (e.g., not over 150 years and not negative)
Correctness	Do values map to a single patient?
Concordance	Are the patient demographics coherent with issue of interest? (e.g., no pregnant males)

2.7. Workflow

DQ Dimension	Guidance Question
Completeness	Are the demographics records entered within the expected observation timeframe? (e.g., data entered during a visit)
Completeness	Are there missing values in the demographics data of interest?
Plausibility	Is the age of the patient plausible?
Completeness	Does the patient have the minimum data available to fully describe each visit? (e.g., weight and BP must be taken for every visit as good practice)

3. Diagnoses

3.1. Analytical Tool

DQ Dimension	Guidance Question
Tool Limitations	Are all statistical assumptions met by the data?
Tool Limitations	Can you assume independence between observations?
Completeness	Are there enough observations within the timeframe of interest?
Completeness	Are there enough patients with all relevant variables within the timeframe of interest?
Completeness	Are there enough observations to provide appropriate statistical power?
Completeness	Is censorship at an acceptable rate?
Completeness	Are the measurements recorded at reasonably regular intervals?
Completeness	Are there recording gaps within the timeframe of interest?
Correctness	Are there duplicate observations?
Completeness	Are there enough data points by encounter?
Completeness	Are there enough data points by patient and/or care provider?

3.2. Clinical

DQ Dimension	Guidance Question
Completeness	Diagnoses of interest appear at least once?
Plausibility	Are all diagnoses recorded after the patient's date of birth?
Timeliness	Are the diagnoses recorded during a visit that falls within the expected observation timeframe or before the first visit (i.e., medical history)?
Completeness	Are diagnoses recorded at regular intervals in the timeframe of interest?

3.3. Data Manipulation

This Knowledge Domain is beyond DQ dimensions and requirements. Good data management practices should prevent any DQ issues in this context, yet it is recommended to verify that no issues arose from extraction. At this stage, the data scientist designs tests and fail-safes to avoid corrupting the dataset during the data manipulations and extract-transform-load (ETL) procedures and/or identify them once the data has been extracted. One example such issues is the creation of duplicates when using joins. To avoid such issues, the data base administrator should verify the counts after every join to ensure no duplication has taken place. These checks are usually done before the dataset is extracted and assessed for DQ. DQ requirements are not necessary beyond the tests defined by the data scientist but could still be defined for thoroughness purposes.

3.4. Representation

DQ Dimension	Guidance Question
Representation	Are all Primary Keys unique?
Representation	Do all foreign keys correspond to one primary key?
Representation	Are all values formatted appropriately or as expected?
Representation	Are all values of the right sort (e.g., character, numeric, etc.)?
Representation	Are all predefined values (i.e., concepts) found in the dictionary? (e.g., ICD-9)
Representation	Are all defined concepts of interest found in the dataset? (e.g., if the concept 'male' is defined to be used in the gender variable, the variable should contain 'male' values)

3.5. Research Design

DQ Dimension	Guidance Question
Completeness	Are there missing variables?
Completeness	Are there missing values?
Completeness	Are the values missing not at random?
Completeness	Are there enough values within the timeframe of interest?
Completeness	Is the analytical observational unit complete?
Completeness	Is the censorship level for the analytical observational unit acceptable?
Correctness	Does the observational unit follow the ideal timeline of collection?
Plausibility	Is the censorship rate acceptable?
Correctness	Are missing values spread randomly over time?
Correctness	Does the distribution of demographics correspond to population estimates at large (i.e., is the sampling random)?
Completeness	Are the data accessible for analysis? (i.e., are the data not locked in clinical notes?)

3.6. Research Goal

DQ Dimension	Guidance Question
Concordance	Are there diagnoses that conflict or interact with the phenomenon in questions?
Timeliness	Are there comorbidities at the time of the visit of interest?
Completeness	Are the diagnoses of interest present?
Concordance	Do the diagnoses evolve as expected?
Completeness	Is the ratio of censorship acceptable (i.e., # missing data/# available data)?

3.7. Workflow

DQ Dimension	**Guidance Question**
Timeliness	Are there conflicting diagnoses overlapping in time?
Concordance	Are there duplicate diagnoses?
Plausibility	Are all diagnoses recorded after patient's data of birth?
Correctness	Are the diagnoses within the expected observational timeframe?
Completeness	Are there diagnoses of interest within the desired timeframe?

4. Labs

4.1. Analytical Tool

DQ Dimension	Guidance Question
Tool Limitations	Are all statistical assumptions met by the data?
Tool Limitations	Can you assume independence between observations?
Completeness	Are there enough observations within the timeframe of interest?
Completeness	Are there enough patients with all relevant variables within the timeframe of interest?
Completeness	Are there enough observations to provide appropriate statistical power?
Completeness	Is censorship at an acceptable rate?
Completeness	Are the measurements recorded at reasonably regular intervals?
Completeness	Are there recording gaps within the timeframe of interest?
Correctness	Are there duplicate observations?
Completeness	Are there enough data points by encounter?
Completeness	Are there enough data points by patient and/or care provider?

4.2. Clinical

DQ Dimension	Guidance Question
Completeness	Are all expected lab values present?
Completeness	Are there any missing values for each lab record? (e.g., value with timestamp)
Completeness	Do matching values have a match? (e.g., lipid panel)
Completeness	Is the distribution of values distribution as expected?
Completeness	Are the values within limits for the population?
Completeness	Are there enough measurements within the expected observation window?
Plausibility	Are the labs dated after the patient's date of birth?
Completeness	Are there sufficient values over time for the analysis?
Correctness	Is the frequency of lab values as expected?
Plausibility	Is the difference in consecutive measurements in an encounter within acceptable range?
Plausibility	Is the difference in consecutive measurements within an acceptable range of than the average difference for the individual?
Plausibility	Is the difference between variable values between two time points in acceptable proportion?
Plausibility	Are there any sudden changes over time? Are they valid?
Concordance	Are the values coherent or vary as expected within a visit?
Concordance	Is the overall vital measure variability as expected?
Concordance	Are there statistical outliers?
Completeness	Are there timestamps at regular intervals within expected observation timeframe?
Completeness	Lab results contain positive, negative and numeric values?
Correctness	Is the temporal frequency of labs as expected for a patient?
Completness	Is the temporal density as expected for the population?

4.3. Data Manipulation

This Knowledge Domain is beyond DQ dimensions and requirements. Good data management practices should prevent any DQ issues in this context, yet it is recommended to verify that no issues arose from extraction. At this stage, the data scientist designs tests and fail-safes to avoid corrupting the dataset during the data manipulations and extract-transform-load (ETL) procedures and/or identify them once the data has been extracted. One example such issues is the creation of duplicates when using joins. To avoid such issues, the data base administrator should verify the counts after every join to ensure no duplication has taken place. These checks are usually done before the dataset is extracted and assessed for DQ. DQ requirements are not necessary

beyond the tests defined by the data scientist but could still be defined for thoroughness purposes.

4.4. Representation

DQ Dimension	Guidance Question
Representation	Are all Primary Keys unique?
Representation	Do all foreign keys correspond to one primary key?
Representation	Are all values formatted appropriately or as expected?
Representation	Are all values of the right sort (e.g., character, numeric, etc.)?
Representation	Are all predefined values (i.e., concepts) found in the dictionary? (e.g., ICD-9)
Representation	Are all defined concepts of interest found in the dataset? (e.g., if the concept 'male' is defined to be used in the gender variable, the variable should contain 'male' values)

4.5. Research Design

DQ Dimension	Guidance Question
Completeness	Are there missing variables?
Completeness	Are there missing values?
Completeness	Are the values missing not at random?
Completeness	Are there enough values within the timeframe of interest?
Completeness	Is the analytical observational unit complete?
Completeness	Is the censorship level for the observational unit acceptable?
Correctness	Does the observational unit follow the ideal timeline of collection?
Correctness	Are missing values spread randomly over time?
Correctness	Does the distribution of demographics correspond to population estimates at large? (Is the sampling random?
Completeness	Are the data accessible for analysis? (e.g., are the data not locked in clinical notes?)
Completeness	Is the ratio of censorship acceptable (i.e., # missing data/# available data)?

4.6. Research Goal

DQ Dimension	Guidance Question
Completeness	Are the lab results complete and with adequately recorded values?
Completeness	Does the lab result contain all expected values?
Correctness	Are the lab values in the expected format and unit?
Completeness	Does the data have null values?
Correctness	Is the age of patients plausible at the time of the lab?
Correctness	Do labs correspond to a single patient?
Timeliness	Are there enough values within the desired timeframe?
Completeness	Does the patient have enough data once non-compliant values have been eliminated?
Concordance	Are there sudden changes in values over time?
Concordance	Are the values of interest coherent with the patient's history? Is it a potential outlier?

4.7. Workflow

DQ Dimension	Guidance Question
Concordance	Are there overlapping labs? Do they present disparate values?
Concordance	Are there duplicate labs?
Plausibility	Are the labs recorded after patient's date of birth?
Completeness	Are there enough labs within the desired timeframe?

5. Meds

5.1. Analytical Tool

DQ Dimensions	Guidance Question
Tool Limitations	Are all statistical assumptions met by the data?
Tool Limitations	Can you assume independence between observations?
Completeness	Are there enough observations within the timeframe of interest?
Completeness	Are there enough patients with all relevant variables within the timeframe of interest?
Completeness	Are there enough observations to provide appropriate statistical power?
Completeness	Is censorship at an acceptable rate?
Completeness	Are the measurements recorded at reasonably regular intervals?
Completeness	Are there gaps in recording within the timeframe of interest?
Correctness	Are there duplicate observations?
Completeness	Are there enough data points by encounter?
Completeness	Are there enough data points by patient and/or care provider?

5.2. Clinical

DQ Dimension	Guidance Question
Completeness	Does the dataset contain prescriptions of the drug of interest?
Completeness	Are numeric values within the expected range?
Completeness	Is the sig complete for all prescriptions?
Concordance	Are there overlaps between prescriptions?
Concordance	Are there multiple prescriptions of the same drug at the same time?
Completeness	Is a drug exposure variable calculable from the available data for every prescription?
Correctness	Is the dose of medications prescribed in multiples of the commercially available strength?
Correctness	Is the daily dose within an acceptable range?
Plausibility	Is the number of refills within an acceptable range?
Plausibility	Does the total quantity dispensed match the duration and dose prescribed?
Correctness	Is the total quantity dispensed within an acceptable dose range for the medication?
Correctness	Is the number of days prescribed within an acceptable range?
Correctness	Is days>0 and <200?
Correctness	Are all strength values>0 and <[max commercial strength]?
Correctness	Are doses>0 and <2*[max daily dose]?
Correctness	Are refills>=0 and <10?
Plausibility	Is total quality=days*dose?
Plausibility	Is total quantity>0 and <600?

5.3. Data Manipulation

This Knowledge Domain is beyond DQ dimensions and requirements. Good data management practices should prevent any DQ issues in this context, yet it is recommended to verify that no issues arose from extraction. At this stage, the data scientist designs tests and fail-safes to avoid corrupting the dataset during the data manipulations and extract-transform-load (ETL) procedures and/or identify them once the data has been extracted. One example such issues is the creation of duplicates when using joins. To avoid such issues, the data base administrator should verify the counts after every join to ensure no duplication has taken place. These checks are usually done before the dataset is extracted and assessed for DQ. DQ requirements are not necessary beyond the tests defined by the data scientist but could still be defined for thoroughness purposes.

5.4. Representation

DQ Dimension	Guidance Question
Representation	Are all Primary Keys unique?
Representation	Do all foreign keys correspond to one primary key?
Representation	Are all values formatted appropriately or as expected?
Representation	Are all values of the right sort (e.g., character, numeric, etc.)?
Representation	Are all predefined values (i.e., concepts) found in the dictionary? (e.g., ICD-9)
Representation	Are all defined concepts of interest found in the dataset? (e.g., if the concept 'male' is defined to be used in the gender variable, the variable should contain 'male' values)

5.5. Research Design

DQ Dimension	Guidance Question
Completeness	Are there missing variables?
Completeness	Are there missing values?
Completeness	Are the values missing not at random?
Completeness	Are there enough values within the timeframe of interest?
Completeness	Is the analytical observational unit complete?
Completeness	Is the censorship level for the observational unit acceptable?
Correctness	Does the observational unit follow the ideal timeline of collection?
Plausibility	Is the censorship rate acceptable?
Correctness	Are the missing values spread randomly over time?
Correctness	Does the distribution of demographics correspond to population estimates at large? (i.e., is the sampling random)
Completeness	Are the data accessible for analysis? (e.g., are the data locked in the notes)?
Completeness	Is the ratio of censorship acceptable (i.e., # missing data/# available data)?

5.6. Research Goal

DQ Dimensions	Guidance Question
Completeness	Does the prescription contain the medication of interest?
Correctness	Does the prescription correspond to a patient?
Concordance	Are there medications that interact with the drug of interest?
Concordance	Are there simultaneous prescriptions? Of the same drug?
Concordance	Are there overlapping prescriptions?
Concordance	Are there prescriptions that interact with the phenomenon of interest?
Completeness	Is it possible to calculate an effective dose?
Concordance	Are the prescriptions renewed? Stopped?
Concordance	Are the prescriptions coherent with the diagnoses?
Completeness	Are there diagnoses, vitals or other clinical data available at the time of prescription?
Completeness	Is the sig complete?
Timeliness	Are there enough patients with prescriptions in the timeframe of interest?

5.7. Workflow

DQ Dimension	Guidance Question
Concordance	Are there overlapping prescriptions?
Completeness	Are the prescriptions renewed?
Plausibility	Are the prescriptions recorded after patient's date of birth?
Timeliness	Are the prescriptions within the expected observational time?
Completeness	Are there prescriptions within the desired timeframe?
Concordance	Are the prescriptions recorded during a visit or encounter?

6. Vitals

6.1. Analytical Tool

DQ Dimension	Guidance Question
Tool Limitations	Are all statistical assumptions met by the data?
Tool Limitations	Can you assume independence between observations?
Completeness	Are there enough observations within the timeframe of interest?
Completeness	Are there enough patients with all relevant variables within the timeframe of interest?
Completeness	Are there enough observations to provide appropriate statistical power?
Completeness	Is censorship at an acceptable rate?
Completeness	Are the measurements recorded at reasonably regular intervals?
Completeness	Are there gaps in recording within the timeframe of interest?
Correctness	Are there duplicate observations?
Completeness	Are there enough data points by encounter?
Completeness	Are there enough data points by patient and/or care provider?

6.2. Clinical

DQ Dimension	Guidance Question
Completeness/ Correctness	Is the numeric distribution of values and value frequencies as expected?
Completeness/ Correctness	Is the number of vital measurements as expected? (e.g., at least one weight measurement per visit)
Completeness/ Correctness	Do related values have a match? (e.g., systolic + Diastolic)
Completeness/ Correctness	Are there any missing values for each vital record?
Completeness/ Correctness	Are there an adequate number of measurements within the desired observation window?
Concordance	Is the overall vital measure variability as expected?
Concordance	Are there any sudden changes over time? Are they valid?
Concordance	Are the values coherent or vary as expected within a visit?
Correctness	Is the difference in consecutive values in an encounter within acceptable range?
Correctness	Is the difference between variable values between two time points in acceptable proportion?
Correctness	Are the measurements taken after the patient's date of birth?
Correctness	Are all values within plausible limits for the population?
Completeness	Is the time between events of the expected range?
Correctness	Are time stamps within the expected time frame (e.g., BPs measured within the encounter window vs. outside).
Timeliness	Are time stamps in the expected order? (e.g., order before admin).
Timeliness	Are there sufficient values over time for the analysis?

6.3. Data Manipulation

This Knowledge Domain is beyond DQ dimensions and requirements. Good data management practices should prevent any DQ issues in this context, yet it is recommended to verify that no issues arose from extraction. At this stage, the data scientist designs tests and fail-safes to avoid corrupting the dataset during the data manipulations and extract-transform-load (ETL) procedures and/or identify them once the data has been extracted. One example such issues is the creation of duplicates when using joins. To avoid such issues, the data base administrator should verify the counts after every join to ensure no duplication has taken place. These checks are usually done before the dataset is extracted and assessed for DQ. DQ requirements are not necessary beyond the tests defined by the data scientist but could still be defined for thoroughness purposes.

.

6.4. Representation

DQ Dimension	Guidance Question
Representation	Are all Primary Keys unique?
Representation	Do all foreign keys correspond to one primary key?
Representation	Are all values formatted appropriately or as expected?
Representation	Are all values of the right sort (e.g., character, numeric, etc.)?
Representation	Are all predefined values (i.e., concepts) found in the dictionary? (e.g., ICD-9)
Representation	Are all defined concepts of interest found in the dataset? (e.g., if the concept 'male' is defined to be used in the gender variable, the variable should contain 'male' values)

6.5. Research Design

DQ Dimensions	Guidance Question
Completeness	Are there missing variables?
Completeness	Are there missing values?
Completeness	Are the values missing not at random?
Completeness	Are there enough values within the desired time range?
Completeness	Is the observational unit defined for the analysis complete?
Completeness	Is the censorship level for the observational unit acceptable?
Correctness	Does the observational unit follow the ideal collection timeline?
Completeness	Is the ratio of censorship acceptable (i.e., # missing data/# available data)?
Completeness	Are the missing values spread randomly over time?
Correctness	Does the distribution of demographics correspond to population estimates at large? (i.e., is the sampling random?)
Completeness	Are the data accessible for analysis? (e.g., is the data locked in clinical notes)

6.6. Research Goal

DQ Dimensions	Guidance Question
Completeness	Are the vital records complete?
Completeness	Does each observation contain all specific measurements required for analysis?
Completeness	Does the data set have null values for ideal analytical observation unit?
Plausibility	Is the age of patient plausible at the time of measurement?
Correctness	Do vital observations map to a single patient?
Timeliness	Are there enough values within the desired timeframe?
Completeness	Does the patient have enough data once the record is cleaned?
Concordance	Are there sudden changes in values over time?
Concordance	Is the value of interest coherent with the patient's history? Is it a potential outlier?

6.7. Workflow

DQ Dimension	Guidance Question
Completeness	Are there vitals for every observation year?
Plausibility	Are all measures after the patient's date of birth?
Correctness	Are the measures within the expected observation range?
Plausibility	Are all expected vital measures within each encounter? (e.g., routinely recorded values such as weight, BP, etc.)
Concordance	Is the provider the same for all measures within the encounter?
Concordance	Are all readings for the same visit entered within an hour of each other for outpatient data?
Completeness	Is the measurement rate as expected for repeated measures?
Concordance	Do matched measurements have the same caregiver? (e.g., systolic and diastolic)
Correctness	Do patients have more measurements than expected?

Appendix C: Use Case Definitions

Case 1- Duplicate BP measurements

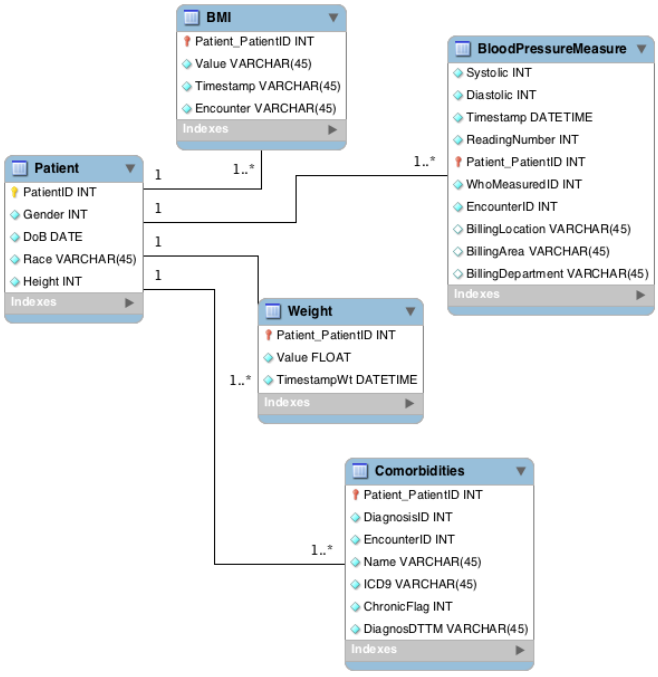
Goal: Use of a CDW to assess whether there is a difference between blood pressure (BP) measurements when a repeated measures protocol is in place.

Research Question: Is the second BP measure statistically lower than the first BP measure taken within a visit?

Hypothesis: Subsequent BP measurements (second, third, etc.) are lower than the first BP measurement taken during that visit.

Possible Analyses:
Linear Regression model- Predict 2nd BP value based on 1 BP value (used for validation)
Linear Regression model 2- Outcome Variable = Difference between BP measurements
GEE or mixed effect model (repeated measures)- for multiple data points

Final Data Needs Model:



Case 2- Caregiver relationship to BP measurements

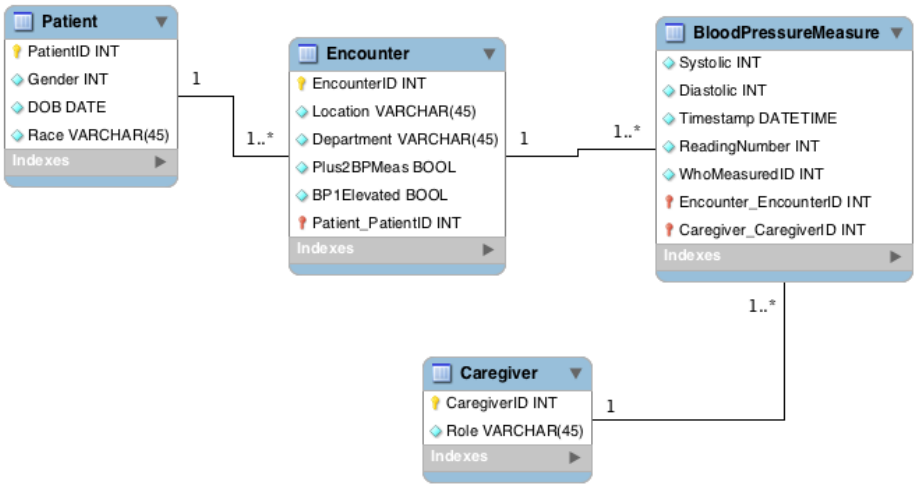
Goal: Describe the number of blood pressure recordings by caregivers along visit units to assess the relationship, if any, between the caregiver and the implementation of the double measurement blood pressure protocol.

Research Question: Are dual BP measurements provider-dependent?

Hypothesis: There is a relationship between the number of BP measurements recorded and the caregiver's background, training and occupation.

Analysis:
Logistic GEE or mixed model - Accounts for patient-level correlation and covariates.
Main outcome variable= 2 BP measurements per visit? Binary Variable (yes/no)

Final Data Needs Model:



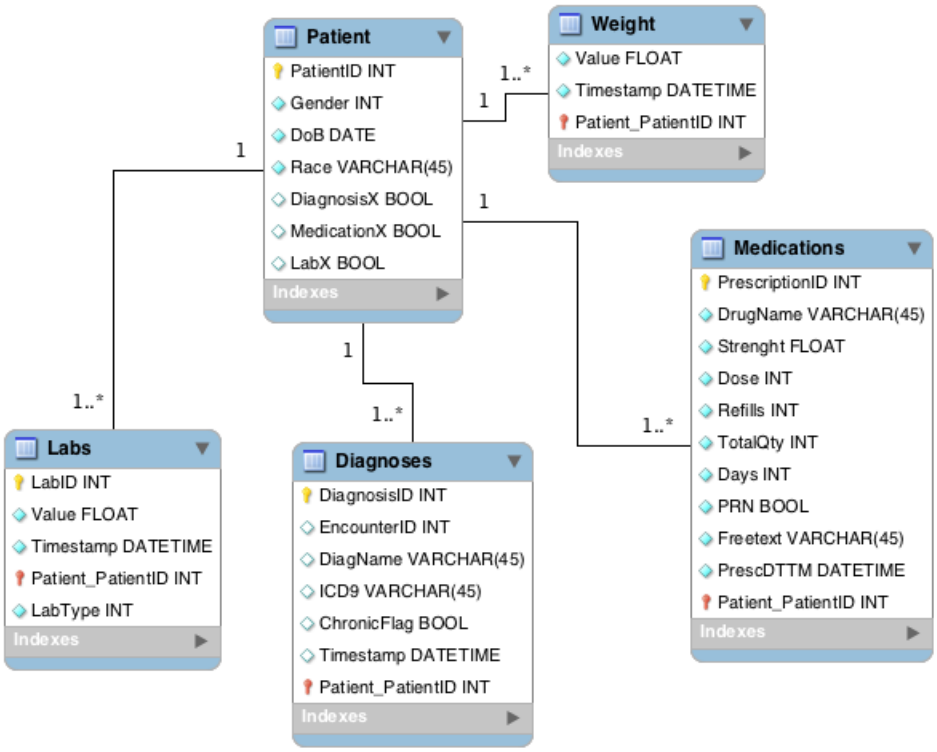
Case 3- Observational study of weight trends in population
Goal: Identify and describe the types trends in the weight changes of patients from the point of view of an outpatient clinic.

Research Question: What are the modes of variation of weight values over time?

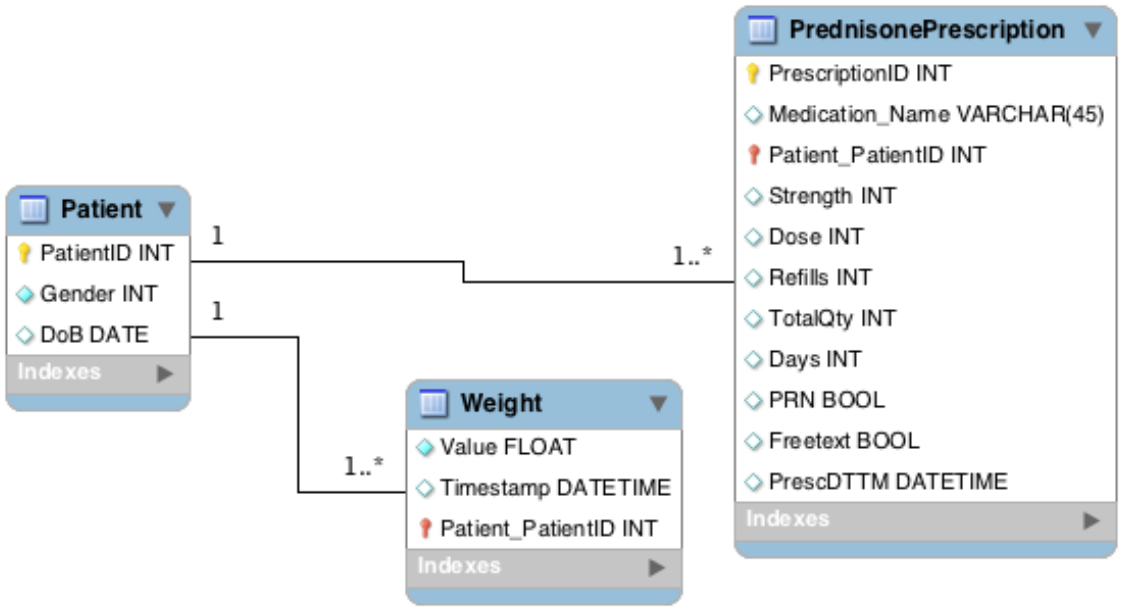
Hypothesis: There are district patterns of weight evolution over time that depend on the patient's demographic and health information.

Possible Analyses:
Sparkline, trend line and line plot visualization techniques.
Linear regression for adjusted mean weight.

Final Data Needs Model:



Case 4- Relationship between drug intake and side effect
Goal: Identify the relationship between prednisone and weight gain in clinical care data.
Research Question: 4. Is prednisone exposure correlated with weight gain?
Hypothesis: There is a positive relationship between prednisone exposure and weight change over time.
Possible Analysis:
GEE regression model predicting weight change over time after first prescription based on prednisone exposure levels.



Case 5-Relationship between BMI and HbA1c lab values

Goal: Identify and describe the relationship between BMI and HbA1C and glucose lab test results.

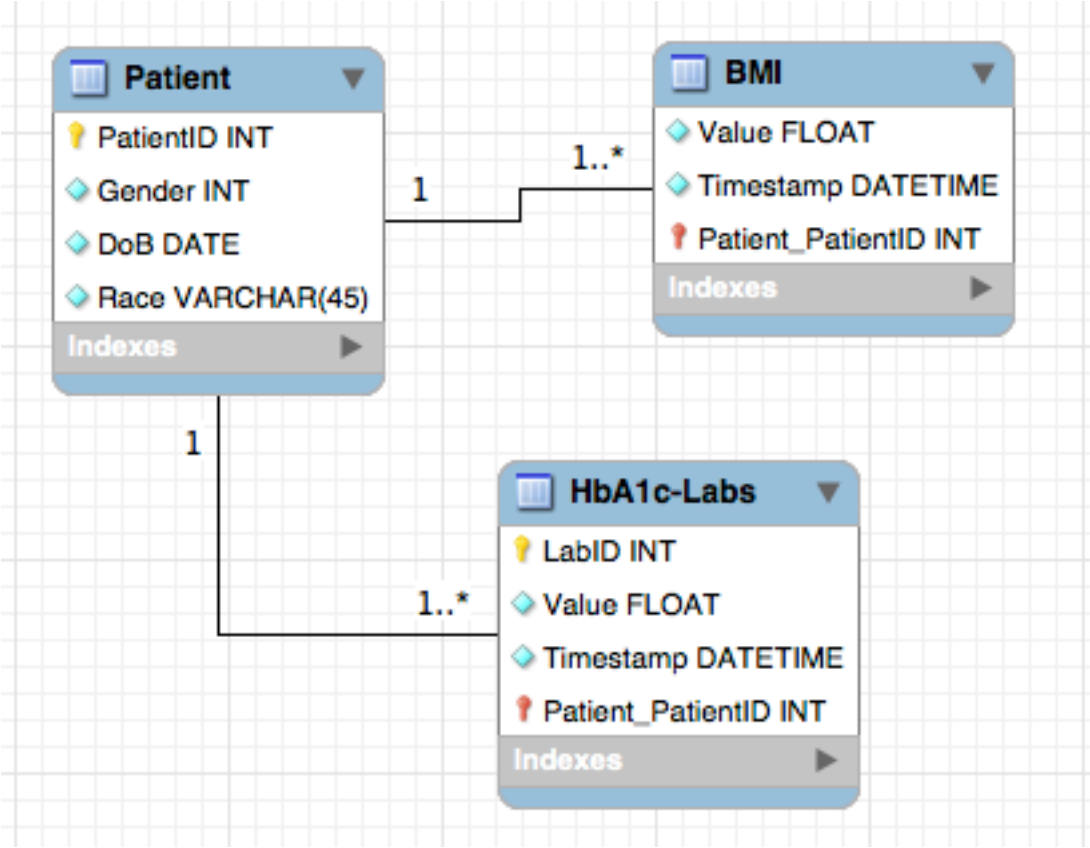
Research Question: Are HbA1C lab values correlated with BMI?

Hypothesis: Patients with higher BMI have higher glucose and HbA1C readings

Analysis:

Linear Regression predicting HbA1c values based on BMI values

Final Data Needs Model:



Case 6-Recruiting patients with BMI>25 and age>21
Goal: Identify patients with the following characteristics: BMI>25 and age>21

Research Question: Can we find patients BMI>25 and age>21?

Hypothesis: N.A.

Analysis: N.A.

Final Data Needs Model: