Secondary Use of Electronic Clinical Data: Barriers, Facilitators and a Proposed Solution

Daniel Capurro Nario

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

Doctor of Philosophy

University of Washington

2012

Reading Committee:

Peter Tarczy-Hornoch, Chair

Ira Kalet

James Tufano

Program Authorized to Offer Degree:
Biomedical and Health Informatics

University of Washington

Abstract

Secondary Use of Electronic Clinical Data: Barriers, Facilitators and a Proposed Solution

Daniel Capurro Nario

Chair of the Supervisory Committee:
Professor Peter Tarczy-Hornoch
Biomedical Informatics and Medical Education

The increasing adoption of electronic medical records is producing a massive accumulation of routinelly collected electronic clinical data (ECD). This data can be used not only for direct patient care but for secondary purposes such as clinical research, quality improvement, and public health. However, using clinical data collected for one purpose does not necessarily render it usable for secondary purposes. This dissertation seeks to explore (1) researchers' data needs and whether ECD is fit for use for research purposes, (2) the barriers and facilitators to secondary use faced by clinical researchers, and (3) to propose a solution to help address one of the barriers identified. To do this, this dissertation is composed of three different but interrelated studies. The first one consists of a Delphi process to develop a tool to systematically assess the fitness for use of ECD for research and its subsequent application to a set of clinical data requests. The second study is a qualitative inquiry into the barriers and facilitators to secondary use of clinical data experienced by researchers at the University of Washington, Group Health Research Institute and the Veterans Affairs' Northwest Center for Outcomes Research in Older Adults. The third study describes the development of a system to query clinical relational databases based on temporal abstractions and patterns, which could enable researchers to identify high-level con-

cepts from clinical databases. The results of this dissertation make contributions that should allow us to improve the reutilization of ECD for research purposes.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS AND ABBREVIATIONS

AAHC	Association of Academic Health Centers.
AAMC	American Association of Medical Colleges.
ADE	adverse drug events.
AHA/ACC	American Heart Association/American College of Cardiology.
AHIC PHC	American Health Information Community's Personalized Health Care Workgroup.
AHRQ	Agency for Healthcare Research and Quality.
APHA	American Public Health Association.
ARDS	acute respiratory distress syndrome.
ARRA	American Recovery and Reinvestment Act.
ASCO	American Society of Clinical Oncology.
CDC	Center for Disease Control and Prevention.
CDR	clinical data repository.
CDR-CAT	Clinical Data Request Complexity Assessment Tool.
CDSS	clinical decision support system.
CER	comparative effectiveness research.
CICTR	Cross-Institutional Clinical Translational Research.

CMS	Center for Medicare and Medicaid Services.
CTSA	Clinical and Translational Science Awards.
CVC	central venous catheter.
DEDUCE	Duke Enterprise Data Unified Content Explorer.
DHHS	US Department of Health and Human Services.
EBM	Evidence-Based Medicine.
ECD	electronic clinical data.
ESSENCE	Electronic Surveillance System for the Early Notification of Community-based Epidemics.
ET	event time.
GHRI	Group Health Research Institute.
HIE	health information exchanges.
HIPAA	Health Insurance Portability and Accountability Act.
HITECH	Health Information and Technology for Economic and Clinical Health Act.
HMO	health maintenance organization.
HMORN	HMO Research Network.

i2b2	Integrating Biology and the Bedside.
ILINet	Influenza-like Illness Surveillance Network.
IOM	Institute of Medicine.
IRB	institutional review board.
KBTA	Knowledge Based Temporal Abstraction.
NCAB	National Cancer Advisory Board.
NER	named entity recognition.
NHANES	National Health and Nutrition Examination Survey.
NICE	National Institute for Health and Clinical Excellence.
NIH	National Institutes of Health.
NLP	natural language processing.
OECD	Organization for Economic Cooperation and Development..
ONC	Office of the National Coordinator for Health Information Technology.
PROTEMPA	Process-Oriented Temporal Analysis.

QI	quality improvement.
RCT	randomized Controlled Trial.
RHIO	regional health information organizations.
SCIP	Surgical Care Improvement Project.
SCOAP	Surgical Care and Outcomes Assessment Program.
TT	transaction time.
UDT	user defined time.
UW	University of Washington.
VA	US Department of Veterans Affairs.
VT	valid time.
WHO	World Health Organization.
WSD	word sense disambiguation.

ACKNOWLEDGMENTS

Many people were were instrumental in this dissertation. I would like to thank every member of my dissertation committee. To Dr. Peter Tarczy-Hornoch, my advisor and committee chair, your knowledge of this field and your continuous and untiring effort to enable a learning health care system are an inspiration. To Dr. Ira Kalet, member of my reading committee, I am deeply grateful to you for trusting me more than I trusted myself. Your patience to let me discover, your guidance, and your willingness to take risks exploring problems outside the comfort zone are invaluable qualities. To Dr. James Tufano, member of my reading committee, thank you for all your guidance and, especially, for your effort during the last stages of this dissertation. To Tony Black, your deep experience with secondary uses of clinical data was invaluable for the success of this dissertation. To Dr. Meliha Yetisgen-Yildiz, thank you for sharing what you know and enduring my limited programming skills.

I would also like to thank all the friends I made during my stay in Seattle. To my classmates Wynona Black, Denny Bromley, Melissa Clarckson, Dr. Walter Curioso, Rupa Patel, and Steve Rysavy: our endless discussions were essential in shaping my view of this field. To Alicia Guidry: thank you for all your patient support during the dissertation writing stage. To Tony Black, Dr. Meliha Yetisgen-Yilditz, and Dr. Marcelo Lopetegui for all their help during this project. To the members of the TransPHorM project, Dr. Anne Turner, Megumu Brownstein, Kate Cole and Shomir Chaudhuri: I signed up for the project because I needed a team and that is precisely what I found. To all the friends that became our family during our four years in

Seattle, the list is too long and my appreciation is infinite.

Finally, I would like to thank all those who were part of this process from the beginning. To Dr. Juan Pablo García Huidobro: you showed me that a career in research in Chile is not only possible but full of joy. To my father, Carlos Capurro: your passion and vision has always been an inspiration, from you I learned that science is everywhere, that I cannot advance without learning. To my mother, Ethel Nario: your no-questions-asked trust in me and my projects, and your invisible but strong guidance throughout my life gave me a firm ground to stand on; I hope I can do the same for my children. To my wife, Dr. Javiera Martinez, your infinite generosity, world vision, and faith in others are true lights in our path.

Lastly, I would like to thank all the institutions that made this work possible: the School of Medicine of the Pontificia Universidad Católica de Chile for thinking that biomedical informatics was a good idea and the Fulbright-MECESUP Faculty Development Scholarship for funding me.

"Knowing is not enough; we must apply. Willing is not enough; we must do."

—Goethe

DEDICATION

This dissertation is dedicated to my family:

To my parents who supported and encouraged me along the way.
To my wife, Javiera, the happiest person I know and the one who instills balance into my life.
To my children, Olivia, Mateo, and the little one on its way, who have taught me how to be a father and a better person.

Chapter 1

EXECUTIVE SUMMARY

1.1 Overview

Health care systems around the world are plagued with inefficiencies. Effective treatments that are never prescribed, preventive measures insufficiently ordered, and interventions of unknown or unproven effectiveness are frequently performed. Electronic medical records are being increasingly adopted worldwide with the expectation that their use can bring significant benefits to patients through improving the effectiveness, efficiency and safety of direct patient care. The adoption of electronic medical records for routine clinical care is generating massive amounts of electronic clinical data (ECD)[1]. This data has the potential to, not only improve direct patient care, but also to change the way the whole health care industry functions—including clinical research, quality improvement, and public health. This use of routinely collected electronic patient data for purposes other than direct patient care is usually referred to as *secondary use of clinical data*.

1.1.1 Potential Benefits of Secondary Use of Routinely Collected Clinical Data

Secondary use of electronic patient data has the potential to benefit almost all aspects of health care. This dissertation will focus primarily on the benefits of and barriers to secondary use of clinical data by researchers.

---

[1] *Data* is the plural form of *datum*, however, the use of data as a mass noun—used in a similar fashion as the word *information*—is increasingly accepted as correct and will be the form used throughout this dissertation

Clinical research and other forms of research that utilize patient data are complex and expensive to conduct. Every step of the research process—from study design to patient recruitment, monitoring, and outcomes assessment—is essential to obtain valid results, but are exposed to significant hurdles. Secondary uses of large patient databases generated via the use of electronic medical records systems (EMRs) can facilitate patient recruitment, monitoring, and outcomes assessment. Large patient databases can also enable new ways to conduct clinical research.

When translated into clinical practice, evidence generated through comparative effectiveness research (CER)—research that seeks to compare the relative effectiveness of different diagnostic or therapeutic options—can serve as a key contributor to the efficiency of health care systems (see section 2.5.2). Analyzing large patient databases could make possible the continuous monitoring of the effects of diagnostic approaches and therapeutic strategies for large cohorts of patients, in the real world. This also means that adverse effects of interventions—which are infrequently reported in conventional clinical trials—can be effectively measured and incorporated into the clinical decision-making process. This should ultimately provide a better understanding of the risks and benefits associated with every critical decision.

Similarly, other types of research that utilize patient data can be supported or catalyzed by the availability of high quality, accessible, electronic patient data. Genomic research frequently involves correlating gene expression patterns with concrete manifestations of diseases, or phenotype. Clinical databases, which can also be considered a mere representation of patients' phenotype, are a rich source of phenotypic information. Using them could lead to an explosion in the research revealing the role of the environment and genes in shaping patients' phenotypes. With the advent of -omics data (genomics, proteomics, metabolomics, and so on), it is highly likely that electronic clinical data will play a huge role in giving meaning to the mountains of

data being generated by biomedical sciences, data that we we still don't fully understand (see section 2.5.3).

Besides supporting research, the use of routinely collected clinical data for purposes other than direct patient care has the potential to improve the way the quality of health care delivery is monitored and improved, and how public health is practiced. Readily available electronic clinical data can be used to monitor patient outcomes and to provide feedback to clinicians and policy-makers, almost in real time (see section 2.5.4). Similarly, this type of data could be used to improve surveillance of conditions of public health importance or to continuously measure disease burden (see section 2.5.5).

Altogether, an efficient use of routinely collected clinical data that can better support clinical decision making, facilitate clinical research, quality improvement and public health is the foundation for a *learning health care system.* A learning health care system is one in which active learning happens during routine patient care. Active learning entails the active and continuous assessment and questioning of what works, what doesn't, and how can we improve the way we deliver care. The Institute of Medicine (IOM) defines the learning health care system as one in which "*knowledge generation is so embedded into the core of the practice of medicine that it is a natural outgrowth and product of the health care delivery process and leads to continual improvement in care.*" [1]. Electronic clinical data is necessary, but not sufficient, to build a learning health care system (see section 2.4).

1.1.2 Known Barriers to Secondary Use of Routinely Collected Clinical Data

In spite of the huge potential that lies within secondary uses of clinical data, significant barriers exist today that still limit that potential from becoming true. Broadly, and based on the literature review presented in chapter 3 we can divide the barriers to

secondary use of clinical data into *data-related barriers* and *societal and organizational barriers*. Data-related barriers are those that pertain to characteristics of the data itself or to how the data is represented, stored, transmitted and analyzed. Examples of data-related barriers include: data captured using inadequate formats, lack of data standards to ensure adequate semantic and syntactic interoperability, and inadequate data quality.

Similarly, there are several societal and organizational barriers to secondary use of clinical data. Some of them include: health care systems that work under a payment or reimbursement model which discourages secondary use of clinical data to improve efficiency, concerns about patient privacy, and insufficient adoption of adequate tools to support secondary use.

This dissertation seeks to characterize researchers' needs when using electronic clinical data for research, explore the barriers and facilitators faced by them and to develop a system that addresses some of the barriers identified.

1.2 Motivation for this Dissertation and Initial Work

Motivation for this dissertation emerged from my initial work related to the implementation of a clinical decision support system (CDSS) to aid in the recommendation of colon cancer screening [2]. The research project involved the development of a decision engine that, using routinely collected clinical data, could recommend colon cancer screening strategies to health practitioners based on the US Preventive Services Task Force recommendations published in 2008 [3]. The system performed correctly with test data, but when it was time to assess real data from an electronic medical record—past medical and surgical history, family history, previously performed screening tests, results of previous tests—it was clear that the system could not be implemented at our academic medical center. The information was not easily acces-

sible. This led me to state a preliminary and exploratory research question: How complex is it to extract clinical data from an electronic medical record for purposes other than direct patient care?

To start addressing this question I performed an initial analysis of a set of clinical data requests submitted by researchers at the University of Washington to the local clinical data repository. In addition to this analysis, I conducted informal interviews with key stakeholders (database administrators, clinical and health service researchers), and elaborated a list attributes that made the extraction of clinical data for research a complex task. The following list summarizes the initial findings:

- Long iterative process of building complex database queries to meet researchers expectations, data availability and the knowledge to extract it.
- Low data quality, especially for diagnosis coding.
- Lack of an up-to-date medication list.
- Excessive reliance on manual chart abstraction.
- High frequency of relative temporal queries.

This initial assessment then led to hypothesize that there are significant barriers that impede the full utilization of electronic clinical data for purposes other than direct patient care, that these barriers needed further exploration, as well as innovative solutions to overcome them.

1.3 Research Questions

Within the context of my exploratory research question, this dissertation's overarching question is: **How can we improve researchers' abilities to use electronic clinical data for clinical and epidemiological research?** Furthermore, this dissertation will seek to answer the following sub-questions:

- What are researchers' clinical data needs and is the data fit for this particular use?
- What are the barriers and facilitators faced by researchers when using clinical data for secondary purposes?
- Is it possible to build a clinical database query system with the potential to overcome some of the barriers frequently faced by researchers?

1.4 Outline of This Dissertation

1.4.1 Chapter 2. Secondary Uses of Clinical Data: Overview

This chapter is a literature review that provides an introduction to the topic of secondary use of clinical data as well as an overview of the prevailing conditions that make necessary the exploitation of routinely collected clinical data to improve health care delivery and research. The different uses for electronic patient data, ranging from helping in the development of personalized medicine to improving clinical and translational research, as well as public health are discussed. The challenges involved in using clinical data for secondary purposes are discussed in chapter 3.

1.4.2 Chapter 3. Known Barriers to Secondary use of Clinical Data

This chapter delivers a systematic description of the challenges involved in the secondary use of clinical data described in the literature. Barriers to secondary use are organized into data-related barriers and societal and organizational barriers. The impact that these barriers have in the ability to make use of routinely collected clinical data is also discussed. This chapter provides the foundation for chapters 4 through 6 that explore the needs and barriers experienced by researchers engaged in secondary uses of clinical data.

1.4.3 Chapter 4. Researchers' Data Needs: A Systematic Characterization of Researchers' Clinical Data Requests (Aim 1)

This chapter addresses the following research question: **What are researchers' clinical data needs and is the data fit for this particular use?** Due to the lack of instruments to systematically characterize researcher's data requests, this chapter describes the development of such an instrument using a Delphi process, with experts from research institutions in the USA and United Kingdom (section 4.2.1). This instrument was later used to systematically and prospectively characterize investigator initiated clinical data requests (sections 4.3.5 and 4.3.7).

1.4.4 Chapter 5. Understanding Researchers' Barriers and Facilitators to Secondary Use of Clinical Data (Aim 2)

This chapter addresses the following research question: **What are the barriers and facilitators faced by researchers when using clinical data for secondary purposes?** This was done by taking a broader approach—using qualitative research methods—to identify the barriers and facilitators that arise, not only from the data and the informatics infrastructure themselves, but also from organizations and the health care system as a whole. This study included researchers from the University of Washington, the Group Health Research Institute and the VA Northwest Center for Outcomes Research in Older Adults (section 5.6.4).

1.4.5 Chapter 6. Querying Temporal Patterns in Clinical Databases (Aim 3)

This chapter addresses the following research question: **Is it possible to build a clinical database query system that has the potential to overcome some of the frequent barriers faced by researchers?**. One of the barriers identified during the early stages of this dissertation, and further confirmed on aims 1 and 2, was researchers' need to query clinical databases using high-level concepts. One of

the identified concepts was the presence of temporal relations. This describes the development and initial testing of temporal abstraction and query system.

1.4.6 Chapter 7. Conclusions

Finally, this chapter provides a summary of the findings presented in all three dissertation aims. It also discusses the implications of such findings and lays down possible avenues for future research that can continue to advance this field and, ultimately, bring closer to reality the idea of a learning health care system.

1.5 Contributions

This dissertation makes several contributions to the fields of biomedical informatics and to clinical and epidemiological research.

Aim 1 delivers a new tool to systematically assess researchers' data needs. It also provides a new understanding of the fitness for use of electronic clinical data for research. In the particular case studied, a majority of it was not fit for use because it was not available, not easily accessible, or the intrinsic data quality was inadequate or unknown. These difficulties significantly reduce options of using clinical data for purposes other than direct patient care and the chances of establishing a learning health care system.

Aim 2 presents a rich description of the organizational factors influencing researchers' abilities to use electronic patient data for research. Identified barriers and facilitators common to the three organizations studied were related to knowledge management ,the organizational structure, and organizational support. Addressing the identified barriers could facilitate secondary uses of clinical data.

Aim 3 presents a new tool to query time stamped clinical databases based on temporal patterns. In the case presented, this tool was able to query a clinical database for temporal patterns with greater precision when compared to routine human annotation. This could eventually be used to address some of the difficulties to secondary use identified in aims 1 and 2.

Overall, these contributions should be valuable additions to our current knowledge and can help advance the implementation of a learning health care system.

Chapter 2

SECONDARY USES OF CLINICAL DATA: OVERVIEW

2.1 Introduction

The expansion of the amount of clinical and genetic knowledge available, along with the continuous rise in health care costs are forces driving the need for an efficient utilization of clinical data for purposes other than direct patient care. This dissertation aims to address the need to improve the re-utilization of data generated during patient care activities (electronic clinical data) for purposes other than direct patient care, in particular for clinical research. The overarching research question is: **How can we improve researchers' abilities to use clinical data for clinical and epidemiological research?**.

In this chapter I present the results of a review conducted to identify the body of literature describing the need for, and potential benefits of, using clinical data for secondary purposes. Landmark publications and proceedings of the main conferences in the biomedical informatics domain were retrieved and their citations were reviewed to further identify relevant publications in a snowball sampling fashion. In addition, to identify additional articles, I conducted a broad search in PubMed and EMBASE using terms such as "Medical Records Systems, Computerized"[Mesh], "electronic medical record"[EMBASE subject heading], "clinical data repository", "clinical data warehouse".

2.2 Motivating Forces for "Secondary" Uses of Clinical Data: Ineffective, Inefficient and Costly Healthcare

Effective treatments that are never prescribed, preventive measures insufficiently ordered, frequently performed interventions of unknown or unproven effectiveness plague most health care systems regardless of their overall design— public, private or mixed— . For example, a study by Bundgard in 2000 showed that only between 15 and 44% of all patients with an atrial fibrillation and no contraindication for anticoagulant therapy were prescribed warfarin [4], a treatment that has been consistently shown to reduce the risk of stroke and death [5, 6]. Similarly, but in the opposite direction, the United States has seen a significant surge in the number of patients with a diagnosis of prostate cancer operated using minimally invasive approaches—including robotic surgery—with the goal of reducing surgical complications [7], however there is insufficient evidence to support this claim and there are studies that suggest that these less-invasive approaches might increase long-term morbidity [8]. These evidence-practice gaps also translate into wide geographical variations in the services provided and the cost of providing health care, with higher spending not necessarily being translated into better health related outcomes [9].

Medical errors also add significant costs to patient care—between US$6 and US$39 billions worldwide, annually, according the World Health Organization (WHO) [10]— and reduces the resources available for effective and efficient care. A cornerstone publication in this domain, by the US Institute of Medicine in 2000 [11], concluded that approximately 98,000 preventable deaths occur each year as a consequence of preventable medical errors. This high incidence of medical errors is not unique to the US, it has been found in other countries as well [12, 13]. The same WHO report estimates that in developed countries, 1 in 10 patients is harmed while receiving hospital care. The same report estimates that this number can be even higher in developing

countries

The increasing cost of health care is a persistent concern worldwide [14]. It is a phenomenon that is most frequently associated with technologically advanced countries, but it is equally important in developing countries. For example, the average annual growth in per capita health expenditure between 1993 and 2008 for the countries belonging to the Organization for Economic Cooperation and Development. (OECD)—an organization of mostly rich countries—was 3.9%, but for middle income countries like Chile and Turkey it was 5% and 8.3% respectively [15]. There are multiple explanations for this phenomenon. Among them, the higher burden of chronic diseases [16], the population's increasing age [17] and the incorporation of high-cost technological innovations [18]. However, some of the most disturbing contributors to the high costs of providing health care are related to inefficiencies in the system.

2.3 Conventional Solutions and Their Shortcomings: Evidence-Based Medicine, Personalized Medicine and Learning Health Care Systems

Improvements in effectiveness and the efficiency of health care delivery can be achieved using some approaches that involve re-utilization of routinely collected clinical data. Among them we can find (1) measures to increase the generation of and adherence to recommended best practices— including the adoption of Evidence-Based Medicine principles and the development and adoption of clinical guidelines—and (2) measures to personalize health care.

2.3.1 Evidence-Based Medicine and Randomized Controlled Trials

Evidence-Based Medicine (EBM) has been defined as "*... the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of in-*

dividual patients" [19] and seeks to incorporate the best available clinical evidence—along with patients' preferences and clinical experience—to clinical decision-making. Deciding between one intervention or another ideally involves assessing their relative effectiveness through a head-to-head comparison in a randomized Controlled Trial (RCT). However, most clinical questions cannot be answered in this way because there is no RCT available comparing the two options of interest—too many RCTs compare an intervention against placebo—or there is simply no RCT available. Although multiple reasons can explain this lack of availability of clinically relevant RCTs, the enormous cost of conducting a high quality RCT—both in terms time and resources—is a major contributor. Although the cost of conducting a randomized controlled trial can vary widely depending on the number of patients included, the type of intervention and the length of the follow-up period, large clinical trials for common conditions such as acute coronary syndromes and health failure can cost well over 100 million dollars [20].

In addition to exorbitantly elevated costs, frequently the results of a RCT do not necessarily translate into routine clinical practice. This translation, the applicability of a study's results to real-world situations, is called the external validity [21]. Multiple reasons help explain this less-than-desired external validity. First, RCTs frequently test the efficacy of an intervention against placebo when, in reality, decision-makers need to decide between two viable options, usually two alternative treatments. Second, and in order to reduce variability between subjects, RCTs usually have strict inclusion and exclusion criteria when enrolling patients; decision makers need interventions that can be applied to a broad and diverse patient base. Finally, RCTs can only measure some outcomes and routinely need to leave other ones out. Among the outcomes less frequently measured are longer-term outcomes—because of the higher costs of longer follow-ups–, and adverse effects [22], both of which can significantly affect our understanding of the real effectiveness of a new intervention being

considered. These constraints—high costs and limited external validity—help explain why there are so many areas where there is insufficient evidence to support commonly performed interventions [23, 24]. This situation has led to the search of more innovative ways to understand the real-world effectiveness of relevant health interventions.

2.3.2 Pragmatic Clinical Trials

One such approach the de conduction of pragmatic or practical clinical trials [25]. Pragmatic clinical trials can be defined as "*... trials for which the hypothesis and study design are formulated based on information needed to make a decision* [26]" and are designed to test an intervention as they would happen in a real world clinical setting, including its risks, benefits and costs. These trials are useful when comparing two pharmacological alternatives when there is no incentive for the manufacturer of the drugs to compare their products head-to-head or when there is the need to compare a drug against a non-pharmacological intervention. Although these trials solve the issue of answering a clinically relevant question, they still need significant resources given the number of patients that need to be included and the length of the follow-up period. For example, the ALLHAT study, a trial to compare different drugs for hypertensive patients that included over 42,000 individuals and followed them for more than 7 years [27], had a cost of up to US$130 million [26].

2.3.3 Systematic Reviews

A second approach is the development of systematic reviews. Systematic reviews are a category of studies that "*... aim to identify, evaluate and summarize the findings of all relevant individual studies, thereby making the available evidence more accessible to decision- makers*" [28]. To conduct a systematic review, authors define a clinically relevant question, systematically search for studies answering the clinical question,

assess the methodological quality of all included studies and summarize the findings in a way that is meaningful for decision-makers [29]. Systematic reviews are considered a higher level of evidence because they make systematic efforts to reduce bias by including published and unpublished studies and by assessing the quality of the included studies. However, the fact that they rely on previously conducted clinical trials addressing the question of interest means that they will not be available in areas where no good-quality clinical trials have been conducted.

2.3.4 Analysis of Routinely Collected Clinical Data

A third approach to generate evidence about "what works" is to analyze routinely collected clinical data and make inferences about the relative effectiveness of different health services. Although this approach presents problems related to observational studies—such as the lack of control over exposure variables and unmeasured confounders [30]—they offer the ability to test hypotheses that cannot be tested using a RCT, to measure outcomes that have not been measured in previously conducted RCTs, and to include a large number of subjects without the added costs of RCTs. Even with the advantages of these type of studies, the main limitations derive from the need for accessible, high-quality health care data and, as we will later discuss, this condition is not easily met.

2.3.5 Personalized Medicine

Clinical trials report on the magnitude of the effect of a specific intervention for the average patient included in that trial. However, it is clear that patients are heterogeneous even within a well-designed study. This means that some patients might benefit more than others, or might be harmed more than others. That is the reason why investigators perform sub-group analyses; to identify the group of patients that

benefits the most from a specific intervention. This approach, extended to the individual level, is what characterizes personalized medicine.

Personalized medicine is health care tailored to the individual. It takes into account the individual characteristics of the patient, or the individual characteristics of the disease that is affecting him or her [31] to design the most effective preventive, diagnostic and therapeutic strategies. It uses the wealth of new biological and clinical information available to adjust interventions to the individual. Although not yet widespread, several instances of personalized medicine are being used today; examples of them are anticoagulation and lung cancer therapy.

Anticoagulation is a widely used therapy to prevent multiple thromboembolic diseases such as pulmonary thromboembolic disease, and systemic emboli due to the presence of an atrial fibrillation [32]. Along with the benefits of reducing embolic events, anticoagulant drugs increase the risk of major bleeding episodes, which might increase morbidity and mortality. Warfarin, the most widely used anticoagulant and has a very narrow therapeutic range—the range at which the benefits are greater than the risks—,which means that over and under dosing is associated with worse outcomes. It is known that patients with certain genotypic variations metabolize the drug differently and need different dosing schemes. Using a personalized approach, Epstein et al. [33] conducted a quasi-experimental study that showed that patients receiving individually tailored warfarin therapy had a lower risk of being hospitalized over a 6-month period when compared to historical controls. Similarly for lung cancer—a frequent and lethal condition—Rosell et al. [34] showed that patients whose tumors had specific Epidermal Growth Factor Receptor mutations had better outcomes when treated with Erlotinib, a specific inhibitor of such receptor.

These two examples help illustrate the fact that personalized medicine, through

the reduction of hospitalizations or adverse effects, or through better selection of patients that will benefit from a costly or risky therapy, can help improve the efficiency of health care. In the case of the Warfarin study described, authors used a clinical database to extract patients' phenotype and genotype to draw their conclusions. In a broader sense, the American Health Information Community's Personalized Health Care Workgroup (AHIC PHC) [35] proposed a dataset required to implement personalized health care. This dataset includes demographic information, personal health information—such as history of specific disorders, laboratory and pathology data, previous treatments and environmental exposures—, family history, and genetic/genomic information. This, again, stresses the need to access high-quality clinical data which can, in this case, advance the field of personal medicine.

2.4 The Learning Health Care System

In the previous sections I have discussed some of the potential benefits of learning from the data generated during routine patient care. This is what has been called a *learning health care system*. A learning health care system is one in which active learning can happen in the context of routine patient care. Active learning means the continuous, meaningful and timely questioning of: what works in health care, what are the costs, what are the risks, and what are the benefits. According to the Institute of Medicine (IOM), a learning health system is the one in which "*knowledge generation is so embedded into the core of the practice of medicine that it is a natural outgrowth and product of the health care delivery process and leads to continual improvement in care.*" [1]

Answering those questions meaningfully requires the ability to effectively access patient data being collected during patient encounters and stored in patient records. Although epidemiologists have been conducting observational studies using clinical data from medical records for a long time, the increasing adoption of electronic med-

ical records in the US is leading to an increased availability of clinical data in electronic form, which has the potential to significantly increase this kind of studies. The adoption of electronic medical records by office-based physicians has been steadily increasing. According to the US Department of Health and Human Services, in 2001 only 18.2% of them were using any kind of electronic medical records. In 2010, that percentage was estimated to be 50.7%. This adoption has been fueled by the provisions included in the Health Information and Technology for Economic and Clinical Health Act (HITECH) [36], part of the American Recovery and Reinvestment Act (ARRA), which allocated more than 19 billion US dollars to stimulate the adoption of electronic medical records by hospitals, clinics and individual providers. Furthermore, the bill introduces the concept of *meaningful use* of electronic medical records [37], that is the use of electronic medical records to achieve improvements in the quality of health care. Examples of the criteria that define meaningful use are the ability to report clinical quality measures to the Center for Medicare and Medicaid or State health agencies, submit information to immunization registries, send syndromic surveillance data to public health agencies and laboratory data on notifiable conditions—all of these are examples of non-clinical uses of clinical data—. Users of electronic medical records must demonstrate this kind of meaningful use in order to qualify for the reimbursements defined in the bill. A main goal of the HITECH act is to create the information technology infrastructure that will enable a learning health care system.

Although the infrastructure for a national learning health care system is far from complete—given the still insufficient adoption of electronic medical records—there are several examples of successful studies conducted using large clinical databases. Graham et al. [38] conducted a nested case-control study using Kaiser Permanente's patient database, in which they demonstrated an association between the use of a ciclooxigenase 2 selective inhibitor, Rofecoxib, and increased risk of serious coronary

heart disease. This finding is comparable to what was found in a randomized controlled trial initially designed to test the effects of using Rofecoxib to prevent colon cancer [39]. Both publications were part of the key studies that led to the withdrawal of the drug from the marked by its manufacturer. This example highlights the potential impact of having high-quality patient information for post-market pharmacologic vigilance. Similarly, but in the area of quality of care, Kwon at al. [40] conducted a study using the Surgical Care and Outcomes Assessment Program (SCOAP) database, one that contains information from surgical procedures performed in multiple hospitals in Washington State. They concluded that only 66% of patients taking beta-blockers before surgery continued the therapy during the postoperative period for non-cardiac surgery. In addition, failure to continue with β-blockers was associated with a 2-fold increase of adverse events during the 90-days following surgery. These are just a few examples of the possible uses of routinely collected clinical data for purposes other than direct patient care.

2.5 Secondary Uses of Clinical Data

As defined in the previous section, secondary uses of clinical data involve the utilization of routinely collected clinical data for purposes other than direct patient care. The fundamental characteristic of secondary use of clinical data is that patient data is queried, viewed, or analyzed in an aggregate fashion, as opposed to looking at individual patient records. A second characteristic is that, at least ideally, and since datasets usually involve numbers in the order of thousands of patients, data should be in a format suitable for computer processing. These two characteristics, along with the fact that it is data describing patients' clinical attributes, are the ones that will define most of the complexities around secondary use of clinical data.

There are multiple types of secondary uses. I have grouped them into the most well known categories which, although with some degree of overlap, comprehensively

describe the range of possible uses.

2.5.1 Clinical Epidemiology

Clinical epidemiology is a branch of medical sciences that studies the impact of different aspects of clinical practice and health care. According to Haynes et al. [41], the core questions it tries to solve are:

- How to screen for diseases
- How to prevent, treat, ameliorate, or rehabilitate health problems
- How to predict the course of disease
- How to measure the burden of illness, quality of life and the effects of health services innovations
- How to increase the quality of health care and improve its outcomes
- How to systematically summarize evidence from research

There is a broad variety of study designs that can answer clinical questions in the areas mentioned. However, we can—very broadly—classify them into a few categories according to the presence or not of an experimental intervention, and the position in time that the observer takes. In terms of the presence or absence of an experimental interventions, we can define them as experimental designs—in which the researcher designs and evaluates the effect of an intervention—and non-experimental designs—in which there is no explicit intervention or when the intervention being studied was part of a natural experiment. In terms of the position in time that the researcher assumes, we can divide them into prospective, retrospective and cross-sectional designs. Independent of the study design, there are tasks that common to all of them. A researcher must identify the possible participants, screen them for inclusion and obtain patients demographic and clinical data for analysis. Each step of these common tasks could be expedited utilizing clinical data routinely collected during patient care.

One example of secondary use of clinical data in the context of clinical epidemiology is for patient recruitment for clinical trials. Patient recruitment is a critical component in clinical trials. It is the number-one reason explaining the failure of clinical trials and it is estimated that 86% of clinical trials are delayed because of slow patient enrollment; 13% of clinical trials are delayed six months or more [42]. To improve this situation, researchers are relying on querying electronic medical records databases to accelerate patient identification and recruitment. Li et al. [43] showed that using Natural Language Processing of clinical notes, in addition to ICD-9 codes, improved the identification of potential participants in a clinical trial.

A second example of secondary use of clinical data in the context of clinical epidemiology is the utilization of electronic medical data to assess patient outcomes. Asche et al. conducted a study to identify the frequency of adverse events in patients with type 2 diabetes who were utilizing oral antidiabetic drugs. In this case, researchers used a combination of chief complaints, ICD-9 codes, and laboratory values to assess adverse drug events (ADE) [44] such as diarrhea, abdominal pain and lactic acidosis.

2.5.2 Comparative Effectiveness Research

According to the IOM, comparative effectiveness research (CER) is defined as "... *the generation and synthesis of evidence that compares the benefits and harms of alternative methods to prevent, diagnose, treat and monitor a clinical condition, or to improve the delivery of care. The purpose of CER is to assist consumers, clinicians, purchasers, and policy makers to make informed decisions that will improve health care at both the individual and population levels.*" CER is not a new kind of research, but has received increased attention during the last decade—especially in the US—given the increasing costs of health care and the need to select the most effective interventions to high prevalence conditions.

Several countries—all of them with a centralized payer systems—have a history of conducting CER to decide on which drugs or devices are paid for by health insurance. Australia has been using such information to make coverage decisions since the 1950s [45]. Similarly, in the United Kingdom, the National Institute for Health and Clinical Excellence (NICE) reviews medical technologies to decide whether they will be covered. Similar agencies exist in Canada and Germany [46].

In the US, the Agency for Healthcare Research and Quality (AHRQ) is the public agency in charge of conducting health services research, including CER. This mandate has been further incentivized after the 2009 American Recovery and Reinvestment Act [47], which allocated $1.1 billion US dollars to support CER.

Multiple methods can be used to conduct CER studies. Randomized controlled trials and systematic reviews of clinical trials are considered key components of CER, but, as mentioned in section 2.3.1, randomized controlled trials are expensive to conduct and frequently have characteristics that limit their external validity. Systematic reviews rely on the existence of relevant clinical trials. As a consequence, one of the approaches that can overcome these drawbacks is to compare the effectiveness of alternative interventions through the conduction of observational studies using large clinical databases.

The IOM recommended that the "... *CER Program should help to develop large-scale, clinical and administrative data networks to facilitate better use of data and more efficient ways to collect new data to yield CER findings*" [48]. This secondary use of clinical data is an integral component of the learning health care system discussed in section 2.4.

The possibility of using clinical data for CER is real, but it faces several challenges such as integrating data from disparate sources—which requires significant efforts to standardize and share data –, preserving patient privacy and data security, and data quality, among many others. These challenges will be discussed in the next chapter.

2.5.3 Translational Research

The astonishing speed at which new biological knowledge is being produced, coupled with the increasing complexity of conducting clinical research, is creating a significant gap—in terms of knowledge generation—between basic biological and clinical research. This chasm impedes, or at least delays, the development and adoption of new and efficacious treatments based on the latest biological discoveries. To bridge the gap, the scientific community has focused on removing the existing barriers between basic biological and clinical research through the conduction of what is now called *translational research*.

Translational research can be described as the endeavor of conveying knowledge from the "bench to the bedside", which seeks to streamline the translation of basic biological knowledge into clinical applications [49]. This is not necessarily a new type of research but a new approach to biomedical research that focuses on integrating clinical and biological knowledge, removing the barriers for adequate translation, to accelerate the development of new treatments and interventions that will ultimately lead to the improvement of patients' lives.

One of the key components of translational research is to integrate new biological knowledge with clinical knowledge. This, most often than not, involves the identification of cohorts of patients and their clinical data, using computational methods. For example, in a study conducted by Chen et al. [50] to identify biomarkers of acute

rejection of solid organs, the authors queried a public repository of microarray data to find information on genes that were differentially expressed in damaged tissues and, at the same time, detectable in serum or urine. These potential biomarkers for acute rejection were then validated, in terms of their diagnostic precision, using data from biopsies of patients with and without evidence of acute organ rejection. This study, as many others in the domain of translational research, exemplifies the need of fluidly integrating novel biological knowledge—the repository of microarray data—and clinical data—biopsy reports of patients with suspected organ rejections–. Again, access to high quality clinical data is vital for translational research.

In the US, the National Institutes of Health (NIH), through the establishment of the Clinical and Translational Science Awards (CTSA), has been instrumental in the recent explosion of translational research. The NIH has funded 60 centers across the nation to incentivize this kind of research, reaching more than half a billion dollars per year in funding [51]. This same approach has been adopted in the United Kingdom and in other European countries [52]. Very recently (July 2012) the US National Institutes of Health released a new vision for the CTSAs that expands the objectives for CTSAs [53]. These changes are intended to foster the adoption of national research practices and tools to transform the research environment. These practices and tools certainly include the ones that allow the use of routinely collected clinical data for translational research.

Knowledge translation "from bench to bedside" is not the only definition of translational research available. Considering the slow speed at which effective interventions are widely adopted as routine clinical practice, experts have identified additional gaps that need to be bridged. The Institute of Medicine also identified the translation of clinical knowledge into actual improvement of patients' health as the second roadblock that needs to be addressed by translational research [54]. This second roadblock can

be addressed with the demonstration of the clinical effectiveness of different interventions through CER as discussed in section 2.5.2. As previously discussed, secondary use of clinical data is a key component of CER.

In addition to these two roadblocks, some authors advocate for the recognition of a third one that prevents the translation of biomedical research into improved health. This roadblock exists between the actual demonstration of the effectiveness of clinical interventions and the widespread adoption of them in the health care system [55]. Considering that the widespread adoption of beta-blockers as routine therapy for patients with an acute myocardial infarction took more than 25 years [56], this seems like a major roadblock. Research related to the measurement of the quality of care and the implementation and evaluation of health care systems re-design are fundamental ways to tackle this third roadblock.

2.5.4 Quality Improvement

Inadequate quality of health care—in the form of underuse, misuse and overuse of health care services—reduces the effectiveness of any health care system while adding significant costs. According to the Institute of Medicine "*quality problems occur typically not because of a failure of goodwill, knowledge, effort, or resources devoted to health care, but because of fundamental shortcomings in the ways care is organized*" [57]. As a response to this situation, health care organizations have undertaken initiatives to continuously improve the quality of the health care they provide. This has lead the industry to move from a quality assurance model—one that seeks to find responsibilities when something does not go as planned—to a quality improvement model, where quality is understood as a property of the system as a whole and not of its individual components.

Quality improvement (QI) is a formal process to analyze and systematically im-

prove the performance of productive systems. In the case of health care, it relates to the productive process of providing health care and all its associated components. This approach is not exclusive to health care. It originated in non-health industries that were aiming at standardizing procedures in order to decrease variability and, since the 1970s, has been progressively incorporated as a method to assess every productive process [58]. Although different methods of quality improvement exist, they share some basic common elements. With variations, quality improvement methods consist of the definition of a problem, the assessment or analysis of the current state, development and execution of a plan and an evaluation of the success [59].

Health care organizations have also started to adopt quality improvement principles from other industries into their daily practices. One successful example is the case of Seattle Children's Hospital, which implemented the Continuous Performance Improvement methods, based on practices developed by Toyota for the auto industry and was able to improve the quality of care while reducing overall cost [60]. A similar experience has been pioneered at the Virginia Mason Medical Center, also in Seattle [61]. In this case, they report savings between 12 and 15 million US dollars during the initial six years of the implementation of the quality improvement system. In addition to improving the costs of providing health care, this hospital has been able to consistently achieve patient health outcomes above national averages [62].

In the US, as well as in many other parts of the world, there are several public and private organizations that have established quality measures for health care delivery organizations. In 2001, the US Department of Health and Human Services (DHHS), through the Center for Medicare and Medicaid Services (CMS) announced The Quality Initiative, a program to improve health care quality for all Americans [63]. This initiative consisted of several programs that aimed at the collection of performance data from a diverse set of health care providers. In parallel, the Joint Commission,

an independent health care certification and accreditation agency, worked on the definition of core quality measures for hospitals. Since 2003 both agencies have worked together to align their core quality measures. Examples of these core quality measures are for example the proportion of patients with an acute myocardial infarction that received aspirin within 24 hours of admission, the proportion of patients that received a dose of prophylactic antibiotics within 60 minutes prior to a surgical procedure, or the proportion of hospitalized patients that were screened for tobacco use in the past 30 days [64]. The type of data needed to calculate these quality measures includes demographics, diagnostic codes, and times when health services were provided, among others.

Traditionally, the information source for quality measure reporting has been health insurance claims [65]. However, insurance claims do not provide detailed clinical information, such as blood pressure or physical examination findings, which can be key elements in deciding whether the care being provided is adequate [66]. In addition, the validity of diagnostic and therapeutic codes contained in insurance claims has been questioned repeatedly [67]. This is why quality reporting still involves significant human abstraction of medical records. Nowadays, the increasing availability of electronic clinical data is seen as a chance to improve the validity and ease of quality measures reporting.

2.5.5 Public Health

So far I have discussed current and potential uses of patient data so high quality information and knowledge is available to make better and safer individual patient decisions. But the ability to access and analyze individual patient data at regional levels has the potential to impact the way public health conducts its mission.

Public health was defined in 1920 by Winslow [68] as "*... the science and the art of preventing disease, prolonging life, and promoting physical health and efficiency through organized community efforts for the sanitation of the environment, the control of community infections, the education of the individual in principles of personal hygiene, the organization of medical and nursing service for the early diagnosis and preventive treatment of disease, and the development of the social machinery which will ensure to every individual in the community a standard of living adequate for the maintenance of health.*" This definition stands true up to this day.

Today, in an attempt to more specifically define the activities performed by public health agencies, the American Public Health Association (APHA) has defined the 10 essential public health services [69]. Those public health services are:

- Monitor health status to identify community health problems.
- Diagnose and investigate health problems and health hazards in the community.
- Inform, educate, and empower people about health issues.
- Mobilize community partnerships to identify and solve health problems.
- Develop policies and plans that support individual and community health efforts.
- Enforce laws and regulations that protect health and ensure safety.
- Link people to needed personal health services and assure the provision of health care when otherwise unavailable.
- Assure a competent public health and personal health care workforce.
- Evaluate effectiveness, accessibility, and quality of personal and population-based health services.
- Research for new insights and innovative solutions to health problems.

It is clear that several of these services can be improved or facilitated by having access to population health data. Monitoring the population's health status is

conducted through surveillance. Disease surveillance can be classified as active or passive. Active surveillance involves the direct actions taken by public health agencies to collect information. An example of active surveillance is National Health and Nutrition Examination Survey (NHANES), the periodical survey conducted by the National Center for Health Statistics to assess the "*the health and nutritional status of adults and children in the United States*" [70]. The survey has been conducted since 1971 and consists of interviews, physical examinations and laboratory tests of about 5,000 individuals in the country every year. The survey is currently used to estimate the prevalence of various diseases and risk factors as well as for health services and epidemiological research. Although the data collected is of superb quality, the cost involved in collecting such information is significant. Similar surveys are conducted yearly in other countries as well [71, 72]. It is easy to envision a scenario in which the data collected during routine health care could be aggregated at a national level and used for the same purposes at a fraction of the cost and, given the increased coverage, with greater precision.

Passive surveillance involves the collection and analysis of data submitted by non-public health entities—such as hospitals, clinics, laboratories, individual providers or the public—to continuously assess the population's health status. One example of passive surveillance is what is conducted by states through the collection of information sent by providers regarding notifiable conditions. For example, Washington State mandates health care providers the notification of multiple infectious and non-infectious conditions such as cases of hepatitis A infections or occupational asthma [73]. The amount of data required to report a single case is significant. In the case of Hepatitis A, a clinician must fill a 2-page long paper form that includes more than 50 fields [74]. This leads to significant underreporting of these conditions. A study conducted in Ireland demonstrated that, overall, 18% of notifiable conditions remained unreported; the most extreme case of underreporting was found to be related to cases

of acute viral meningitis, for which 86% of the cases were not reported [75]. There is significant room for improvement through the automated collection of electronic data.

Several initiatives have been undertaken to collect clinical data to improve public health surveillance. The Center for Disease Control and Prevention (CDC) has established a national reporting system named Influenza-like Illness Surveillance Network (ILINet), which collects weekly counts of patients complaining of fever and cough. The information is currently collected from approximately 3,000 care providers across the nation and is sent through fax or the Internet [76]. Similarly, the US Department of Veterans Affairs (VA) implemented Electronic Surveillance System for the Early Notification of Community-based Epidemics (ESSENCE), a nation-wide system that analyzes ICD-9 codes associated with emergency room visits and—comparing to historical data—is able to detect events with greater than expected frequencies [77]. This system was able to accurately detect the 2009 H1N1 Influenza pandemic. This successful example speaks about the possibilities for public health assessment when electronic clinical data is available. Moreover, this system succeeded despite the fact that it can only access ICD-9 codes for emergency care visits, and cannot analyze data related to hospitalizations or outpatient visits. The potential benefits would be enormous if it could access richer clinical datasets.

2.5.6 Meaningful Use of Electronic Health Records

The concept of *meaningful use* introduced in section 2.4—the set of uses that an adopter of electronic medical records needs to demonstrate in order to receive federal incentives—also ties together the ideas presented here, around using clinical data for quality improvement and public health. The incentives for providers to adopt electronic health records are dependent on them meeting several meaningful use criteria. These criteria will be required in a three-stage process and, so far, only criteria for stages I and II have been released . A large proportion of the meaningful use cri-

teria include the use of routinely collected clinical data for secondary purposes [78]. To receive the incentives providers must demonstrate that they are able to use their electronic medical records to:

- identify and report cancer cases to a State cancer registry,
- submit electronic data to immunization registries,
- submit electronic syndromic surveillance data,
- calculate clinical quality measures,
- submit electronic reportable laboratory results, and
- use information to identify patients who should receive reminders for preventative care,

All these uses require that most or all the known and unknown barriers to secondary use be properly addressed.

2.6 Summary

In this chapter I have presented the current environment in which almost every health care system is embedded in: elevated and increasing costs and system-wide inefficiencies, along with a fast-paced generation of new biomedical knowledge that is not always translated adequately to clinical care. A health care system that is able to learn from its routine practices and foster research is one of the most attractive avenues to address these issues. This necessarily requires the extensive use of high-quality clinical information.

In the next chapter I will present the results of a literature review assessing the current knowledge regarding barriers to the use of routinely collected clinical data for secondary purposes, which will help highlight the current knowledge gaps and lay out the background for this dissertation.

Chapter 3

KNOWN BARRIERS TO SECONDARY USE OF CLINICAL DATA

3.1 Introduction

In the previous chapter I discussed the current need to leverage the enormous data that is being collected during routine clinical care, and use it for purposes other than direct patient care. These secondary uses could help us improve health care efficiency, quality, and streamline research and the practice of public health. The overarching question addressed in chapters 4 to 6 is **how can we improve researchers' abilities to use electronic clinical data for clinical and epidemiological research?**. In this chapter I will report on a literature review—using the same methodology I described in chapter 3—that describes the current landscape regarding barriers to secondary uses of clinical data. This description of the known barriers for secondary use of clinical data will highlight existing gaps, some of which this dissertation addresses. Chapter 4 focuses on characterizing researchers' needs, chapter 5 focuses specifically in identifying barriers and facilitators to secondary use of clinical data faced by researchers and chapter 6 focuses on developing a system to facilitate and are presented in chapters 4 through 6.

3.2 Organization of this Chapter

Secondary use of clinical data requires, on one hand, access to high quality clinical data and, on the other, the ability to query and analyze such data. The barriers that can impede or hinder the secondary use of clinical data can be divided into *data-*

related barriers and *societal and organizational barriers* and I will discuss them in sections 3.3 and 3.4. Finally, in section 3.5, I will introduce the approach adopted throughout this dissertation to address some of the identified barriers.

3.3 Data-related Barriers

The most obvious set of barriers to secondary use of clinical data are related to the data itself. Issues such as its accuracy, the way the data is captured, and the consequences of patient information being dispersed across multiple actors in the health care system arise as significant barriers.

3.3.1 Data Quality

One of the most studied barriers for secondary use is the issue of inadequate data quality. Data quality has been defined as *"a set of characteristics that data should own... in order to provide services in both public and private contexts"* [79]. Frequently, data quality is understood as the degree to which an information system reflects the status of the real world [80]. When the distance between the real world and its corresponding representation in an information system is large, then the quality is considered sub-optimal and can lead to erroneous judgment, analysis and decision-making.

The concept of data quality has evolved over time and multiple ways to assess it coexist today.

3.3.1.1 Intrinsic Data Quality

The traditional way of measuring data quality—mainly represented in the information management literature—focuses on assessing different attributes of the data it-

self. This is defined as the data's *intrinsic data quality*. Among the attributes that constitute intrinsic data quality we can find data correctness, completeness and consistency [81]. *Data correctness* is the concept describing the degree of accuracy of a data element—such as a blood pressure measurement—when compared to a gold standard. *Data completeness* can be assessed from a database point of view (does the database schema contain all the needed data elements?) or from the data point of view (is there missing data?). *Data consistency* is related to the consistent representation of the same data element within a database, for example if there are multiple instances in which a database stores a patient's gender, it should be consistently represented across all instances.

This approach—measuring intrinsic data quality—has been applied extensively when assessing health information data quality. For example Stein et al. explored the concordance (data consistency, using the previously delivered definition) between structured data elements and the same information stored as free text [82]. They found that 4 to 13% of the cases assessed contained contradictions between the structured data and free text fields. A similar study was conducted by Weaver et al. in which the authors studied data consistency on information related to vaccinations [83]. They found significant inconsistencies in the information about vaccinations for a single patient when it was documented in different systems within the same organization. Inadequate intrinsic data quality can significantly affect the reutilization of clinical data for purposes other than direct patient care.

The value of these intrinsic data quality assessments is unquestionable, but they are insufficient to fully characterize the attributes that make a clinical data source useful for secondary use and a broader concept of data quality is needed.

3.3.1.2 A Broader View of Data Quality

Wand and Wang proposed in 1996 the utilization of a broader perspective to understand data quality [84]. This broader perspective includes not only the intrinsic data quality—how well it represents the real world—but also the context in which the data exists, which ultimately influences the ability of the end-users to effectively use the data. This is well aligned with the concept of "fit for use" [85], which means that the data (as well as the information system it is embedded in) is useful for its purpose.

To develop their data quality framework, Wand and Wang first clarify the difference between the internal and external views of an information system. In their words, the internal view *"... addresses the construction and operation necessary to attain the required functionality, given a set of requirements..."* and the external view *"... is concerned with the use and effect of an information system. It addresses the purpose and justification of the system and its deployment in the organization ..."*. Intrinsic data quality is mostly related to the internal view of information systems; a broader approach also considers the external view of an information system.

As a consequence, the framework proposed by them includes other data quality categories to be assessed:

- Contextual data quality: this category contains attributes such as relevancy and timeliness, which depend on the context in which the data is being used.
- Representational data quality: includes attributes such as concise representation and ease of understanding.
- Accessibility data quality: includes attributes such as accessibility and access security.

This more comprehensive view of data quality has been extensively adopted in many domains. The latest review of existing frameworks to assess data quality re-

vealed the existence of 13 different instruments to measure data quality [86] and most of them now include the concepts introduced by Wand and Wang. However, this view has been scarcely applied in health care and has not been applied in the domain of secondary use of clinical data.

This need to apply a broader view of data quality has two direct implications: the first one is that data quality is relative [85]. It is relative because it depends on who the user is and what is it being used for. In the case of secondary uses of clinical data, users are numerous and diverse, and so are the uses. Although studies assessing data quality of clinical data for research have been published [87], they usually focus on internal data quality and consider researchers as a single type of user with homogeneous needs.

The second implication is the importance of feedback in assuring data quality. If clinical data is produced and stored inside the database and there is no measure of whether the data collected is "fit for use", it is impossible to devise and implement measures to improve the quality of the data.

These two consequences also highlight current gaps regarding data quality for secondary use. First, we need a specific understanding of secondary users and their needs. Second, tools are needed to make sure that the "fitness for use" of clinical data for secondary uses is captured and communicated back to database administrators and to individuals involved in the original data collection.

Despite the lack of a systematic assessment of this broad concept of data quality for secondary use, there is literature describing many individual data quality issues that impose barriers to secondary use. Among them we can mention the capture of data in non-computable formats as well as data fragmentation and its consequences.

3.3.2 Data Capture and Format

Electronic medical records, as information systems, have been originally designed and implemented to assist in the care of individual patients, their primary use [88]. This primary use implies that users of electronic medical records will usually be individual clinicians entering and retrieving information on individual patients. Since individuals are very efficient at generating and decoding free text, a significant proportion of clinical information is captured in that fashion. According to a study conducted at the University of Washington Institute of Translational Health Sciences, approximately 50% of all data required to calculate surgical quality of care measures using an electronic medical record were stored as free text. Sources of free text included clinical notes, radiology reports and discharge summaries [Black et al. unpublished data]. This abundance of data stored in this unstructured format is a significant barrier to secondary uses of clinical data. The inability to accurately compute large amounts of unstructured clinical data has led to the continued reliance on manual patient chart abstraction, even in large organizations with a long history of use of electronic medical records [89]. Multiple initiatives have been undertaken to address the problem of free text and increase the access to structured clinical data for secondary purposes.

3.3.2.1 Natural Language Processing

natural language processing (NLP) has been extensively studied as a method to overcome this barrier [90]. Natural Language Processing, part of the broader concept of text-mining, is a domain at the intersection of linguistics, artificial intelligence and computer science that seeks to develop methods to extract meaning from free-text documents. It started in the 1950's outside biomedical disciplines [91]. Within the biomedical field, the earliest published report accessible through the National Library

of Medicine's PubMed database dates from 1978. Since then, almost 2,000 articles have been published regarding the extraction of concepts from free-text clinical documents.

NLP can be used to solve a diverse set of high-level problems present when extracting content from free text documents in any knowledge domain, but these pose specific challenges in health care. Examples of those problems are named entity recognition (NER), word sense disambiguation (WSD), identification of negations and uncertainties, extracting relationships between terms, performing temporal inferences and, ultimately, information extraction [92]. Although researchers have been continuously making progress and NLP engines have increased accuracy, it is still not ideal. A 2011 study conducted by Harkema et al. [93] researchers built a NLP engine to identify quality measures in colonoscopy reports. The authors report that the system's accuracy was sufficient for less than half of the quality measures studied. In addition, the time required to develop a single NLP classifier and their limited transferability to different settings [94] has limited its widespread adoption and it is still considered a technology in development. However it is expected that it will be progressively incorporated into commercial clinical information systems in the future.

3.3.2.2 Increasing the Amount of Structured Data

A second approach to increase the amount of structured clinical data for secondary use revolves around improving structured data capture when clinicians are producing data, a process called clinical documentation. Clinical documentation is a critical component of clinical care since it allows transferring essential information between providers. It is frequently performed using free-text since that allows rich and flexible descriptions of clinical situations. Given the enormous variability seen in the presentation and evolution of clinical conditions, free-form data entry methods such as free

text or dictations are frequently preferred by clinicians . However, as we mentioned above, this hampers the re-use of data captured in such a way. This has led to the development of systems to allow structured clinical documentation. The implementation of such systems is not innocuous as they can significantly interfere with clinical workflow [95]. As Rosenbloom et al. [96] report, the *"... continuous tension between structure and expressiveness"* of clinical notes remains to be solved.

3.3.2.3 Clinical Registries

A third approach is the establishment of clinical registries. A clinical registry have been defined as [97]:

> *"... an organized system that uses observational study methods to collect uniform data (clinical and other) to evaluate specified outcomes for a population defined by a particular disease, condition, or exposure, and that serves one or more predetermined scientific, clinical, or policy purposes"*

Clinical registries usually contain a greater proportion of structured clinical data since they have been designed with the specific goal of secondary analysis of clinical data. Registries are frequently populated using a combination of routinely collected clinical data and data specifically collected for the purposes of the registry. This reliance on additional data to complement data collected during clinical care make registries into a valuable source of structured information, but the additional effort and cost required to collect that data puts them into a different category, distanced from clinical databases populated exclusively during clinical care.

Data quality issues and the capture of clinical data in non-computable formats are significant barriers at all levels, whether inside a single institution of across multiple

ones. The following barriers that I will discuss—data fragmentation, data interoperability and data provenance— affect the re-use of clinical data at the level of single institutions too, but their effect is maximal when secondary use requires using data from multiple organizations.

3.3.3 Data Fragmentation

Along with data being captured in non-structured formats, which renders it of little use for secondary analysis, the fact that data usually resides in different information systems and organizations produces a significant amount of data fragmentation.

The US health care system is extremely fragmented. According to a 2006 Harvard Business Review article, more than 50% of physicians work in practices of three or less clinicians, 25% of all community hospitals and 50% of nursing homes are independent organizations. Public health organizations share the same pattern. There are approximately 2700 independent local public health agencies in the US, all working independently without an overarching central public health agency [98].

A fragmented health care system leads to fragmented health information. The multitude of individual entities involved in health care provision, each one generating and storing clinical data in its own repository, using multiple ways to represent such clinical data, and with few incentives to share it, are the ideal setting to facilitate the development of multiple health information silos. Although the federal government has allocated funds to incentivize the establishment of Health Information Exchanges [47]—regional partnerships between health-related entities that agree to share health information—the existence of these silos is still the rule. For example, a study by Bourgeois et al. [99] conducted in the state of Massachusetts concluded that 31% of all patients visited 2 or more hospitals for acute care and 1% visited 5 or more hospitals in the same period. Needless to say, if those hospitals do not share health

information about their patients, the information collected during one episode is not available to clinicians seeing the patient during a different episode. Moreover, the transmission of key health information would depend entirely on the patient recalling and communicating it.

The impact of the fragmentation of health information has also been well documented. One study conducted in Canada, looking at clinical information gaps produced when older adults were transferred from a nursing home to a hospital to receive emergency care, found that information gaps were present in 85% of all cases [100]. This means that previously known information was not available for physicians treating the patients at the hospital. Such information gaps included critical information needed to make decisions such as baseline cognitive function (36.5%) or advanced directives (46.4%). In the domain of secondary use of clinical data, Wei et al. [101] conducted a study to determine the effect of information fragmentation of the accuracy of cohort discovery from electronic medical records. Using data from two hospitals increased sensitivity for patients with type 2 Diabetes Mellitus from 67.1% to 100%.

The problem of data fragmentation is obvious when realizing that information from a single patient can reside in multiple organizations that do not share data. However, this is also an issue within single organizations. Frequently, large health care provider organizations build their information infrastructure using multiple components such as hospital information systems, radiology information systems, laboratory information systems, surgical and anesthesiology information systems, and so on. For instance, the University of Washington's clinical data warehouse integrates patient data from 14 different information systems [102]. The task of integrating information from each one of them into a single repository is not a trivial task.

Information fragmentation has been addressed—at least at the multi-institution

level—using several methods. Among them, the most relevant are information integration strategies for clinical care and research, and by means of government incentives.

3.3.3.1 Integrated Heath Care Systems

Although the integration of health information was probably not its main objective, the most obvious information integration strategy, which allows having data from a large number of individuals, are integrated health care systems. In the US, this is the case of Kaiser Permanente and Group Health, among others. Both organizations are vertically integrated health care payers and providers within which, although not always the case, patients receive all their preventive and therapeutic care. This lets these organizations have complete or near-complete data about the health services received by their members, thus allowing them to conduct operational and clinical research with greater ease.

3.3.3.2 Health Information Exchanges

A second strategy to integrate clinical information is the development of organizations to integrate disparate health care organizations. This is the case of regional health information organizations (RHIO) or health information exchanges (HIE). They are third-party organizations that facilitate the exchange of health information between participant providers. These organizations have been funded and incentivized since the early 2000' but have frequently faced issues of non-sustainable business models and unclear return on investments for their members, threatening long term sustainability [103].

3.3.3.3 Research Information Exchange

In the case of information integration for research, where the incentive is the conduction of relevant clinical research, there are several successful initiatives. One example is the HMO Research Network (HMORN) [104]. The HMORN is a *"consortium of 19 health care delivery organizations with both defined patient populations and formal, recognized research capabilities"* across the US and Israel. The network's fundamental asset, in terms of health information, is a federation of clinical databases [105]. Each member organization transfers a subset of routinely collected clinical data from their daily operations' clinical information systems into a separate, locally hosted, database. This database has a data model shared across all HMORN organizations, thus allowing data interoperability. Since this is a database federation, each member organization retains ownership and custody of their own clinical information. Data use agreements govern the access, analysis and interpretation of HMORN clinical data, as well as the publication of findings [106]. The main limitation of this dataset is that it only includes a subset of a patient's medical record, excluding all unstructured clinical data such as progress notes, radiology and pathology reports, discharge summaries, etc.

A second example of a federation of clinical databases for research is the one established between the University of Washington and the University of California at Davis and San Francisco denominated Cross-Institutional Clinical Translational Research (CICTR) [107]. This project established a de-identified federation of clinical databases through the selection of a common technical platform, establishing a secure network, and mapping clinical terms for semantic and syntactic interoperability. The CICTR, at its full implementation, contained clinical data from more than 5 million patients. The main limitation of the cases presented is the fact that researchers need to agree in a common data model, which is usually the minimum common denomi-

nator of available data, which limits the dataset's richness.

3.3.3.4 Incentives to Exchange Information

In terms of incentives, since the early 2000' the US government has provided funds to establish systems to the exchange of health information [108]. The latest round of such incentives started in 2011 through the implementation of the first stages of Meaningful Use of health information technology included in the HITECH act [36]. The results of such incentive programs, and its impact on secondary uses of clinical data, are still unknown.

3.3.4 Data Interoperability

A consequence of data fragmentation is that health information originates and is stored in diverse information systems. This is true for data within a single institution—where health information for a single patient can originate from different hospital, laboratory and laboratory information systems—and, of course, for data residing in multiple institutions. To make that information useful, exploiting all of its potential, different information sources and systems need to interoperable. Interoperability has been defined as *"the ability of two or more systems or components to exchange information and to use the information that has been exchanged..."* [109] and in the context of clinical data, the Institute of Medicine has defined interoperability as the *"efficient exchange (and aggregation) of sufficient high quality, meaningful information upon which trustworthy decisions can be leveraged to improve health care for all of us, as patients"* [110]. The US Department of Health and Human Services' Office of the National Coordinator for Health Information Technology (ONC) has defined data interoperability as a key element of the national health information technology

infrastructure [111].

To make health information systems interoperable, two conditions should to be met. *Syntactic interoperability*, in which two systems are able to communicate and exchange data, and *semantic interoperability*, in which the exchanged information can be adequately interpreted [112]. Full interoperability requires the adoption of robust information exchange standards that completely describe the information's syntax and semantic. Despite the existence of multiple clinical information standards, they are still not robust enough or have been insufficiently adopted, thus creating significant barriers to the exchange and aggregation of clinical data across organizations [107].

Interoperability of health information systems can be achieved through the adoption of standards of health information. There has been extensive work conducted in this area. Briefly, standard terminologies and ontologies can be used to establish semantic and syntactic interoperability of health information. Terminologies provide a standard way of naming clinical concepts, with limited representation of relations between terms. On the other hand, ontologies provide, in addition to a standard naming convention, rich representations of the relations between clinical terms [113]. A detailed description of terminologies and ontologies is beyond the scope of this dissertation.

The lack of full semantic and syntactic interoperability in current health information systems, and adequate standards to achieve it [114], is a significant barrier to the secondary use of clinical data.

3.3.5 Data Provenance

Even when disperse data is integrated, barriers to adequately interpret that data persist. Information about the information—what we call meta-data—is essential.

Data provenance is the information that *"summarizes the history of ownership of items and actions performed on them"* [115] or maintaining information about "where it came from, how it was constructed, what processes it has undergone since" [110].

In the case of clinical information, data provenance is critical. For example, a simple arterial blood pressure reading can have different interpretations depending on the context in which the reading was obtained. For example, a single systolic blood pressure reading of 180 mmHg can mean that a patient is highly likely to have hypertension, if the reading was taken with the patient resting and without any blood pressure altering drugs. On the other hand, the same reading, in the context of hundreds of readings obtained through continuous and invasive blood pressure measurement in a hypotensive patient, might only be a measurement error. Even information regarding the specific site where the blood pressure reading was obtained can have major effects on interpretation [116]. When electronic clinical data is transferred from one organization to another with the purpose of aggregating and analyzing data from multiple sites—or between information systems within a single organizations—the issue of data provenance becomes highly significant.

Although clinical data provenance is recognized as an issue in the setting of health care information integration, it has been infrequently reported in the literature. Most reports on clinical data provenance are related to biobanks [117, 118] and neuroimaging [119]. Clinical data provenance is starting to be discussed in the context of defining new measures of health information systems Meaningful Use [120, 121] and should

become a significant issue as more clinical information systems begin sharing data.

3.4 Societal and Organizational Barriers

Probably more significant than the technical barriers, the way the society has decided to organize its health care system and the way health care organizations work can also be a source of barriers for secondary use.

3.4.1 Barriers Related to the Health Care System

As we discussed in section 3.3.3 the US the health care system is composed by multiple independent organizations. These organizations are embedded in a system that relies mostly on private initiative to provide health services to the population. In addition to this, most health care providing organizations—either private or publicly owned—receive reimbursements for the services provided in a fee-for-service fashion. This means that providers receive payments directly related to the number of services provided and, even when they may be non-for-profit organizations, they compete with each other for market share [122].

If one of the benefits of sharing health care information and using it for secondary purposes is the improvement of the quality of care [123] and, ultimately, the population's health, then, under the current market incentives, the whole idea of exchanging health information is counterproductive since it could eventually lead to the provision of fewer services [124]. And even if the market share remained unchanged, there is not a compelling reason to allocate resources to a project with few or uncertain economic benefits [103, 125].

A study conducted by Vest et al in 2010 [126] showed that, among other factors,

private, for-profit and non-networked hospitals were less likely to adopt health information exchange programs. Also, hospitals that were subject to higher marked competition were less likely to implement such a program. The series of incentives included in the HITECH act mentioned in section 3.3.3, aim at better aligning healthcare provider's interests with interoperability and health information exchange, with the ultimate goal of creating the conditions for a learning healthcare system [36].

3.4.2 Patient Privacy

A second societal barrier to secondary use relates to the responsibility to make sure that patient information is not misused. Patients, through the act of seeking care, release personal health information with the assurance it will be used to ensure they receive proper care. Consequently, it is expected that people involved in their care will treat that information accordingly and make sure that it is kept confidential. However, there are circumstances in which disclosing a patient's personal health information can be seen as beneficial for the same patient and the society as a whole.

At the individual level, an adequate flow of patient health information is crucial to ensure that, no matter where a patient is seeking care, his or her information will be available to be used by clinicians to make the best decisions possible. In addition, a fair amount of health information exchange is expected to occur to ensure the correct processing of medical bills by insurance companies. At the societal level, as we have discussed in previous sections, personal health information is critical for the adequate conduction of biomedical research and for the practice of public health. These uses might not benefit a particular individual but benefit society as a whole. As a consequence, there is a need to balance between ensuring patient privacy and allowing disclosure of personal health information for non-clinical uses.

This need for balance is what motivated the provisions included in the Health Insurance Portability and Accountability Act (HIPAA) that regulate the exchange of personal health information [127]. The US Congress passed HIPAA in 1996 as legislation to improve the delivery of health care and to increase the number of insured individuals. This act contained several provisions not related to health information, but it included provisions to achieve "administrative simplification". This component of HIPAA included directions for the DHHS to elaborate regulations concerning electronic transmission of health information. The transmission of health data for research was also included in these regulations.

The main component of HIPAA that regulates the transmission of personal health information is the Standards for Privacy of Individually Identifiable Health Information rule, better known as the HIPAA Privacy Rule [128]. This rule defines the entities that are covered by the rule, the type of information that is protected, the authorized uses and the penalties for noncompliance. Although this rule was not elaborated specifically to regulate the use of clinical data for biomedical research, public health or quality improvement, the rule specifies the conditions under which personal health information might be disclosed without the patient's consent.

In the case of public health, for example, personal health information might be disclosed to public health authorities "... authorized by law to collect or receive such information for preventing or controlling disease, injury, or disability..." [128]. For research, personal health information might be disclosed without patient consent when the research project has been approved by an institutional review board (IRB) or Privacy Board, when the information is solely used in preparation for research and it will not be removed from the covered entity, or when research is conducted using personal health information of deceased patients. For quality improvement, the HIPAA Privacy Rule allows the utilization of protected health information for health care

operations such as quality assessment and quality improvement.

Despite its impact on biomedical research, researchers did not have a significant role in the elaboration of the HIPAA Privacy Rule [129]. When the rule was first implemented in 2003, researchers began to raise concerns that the rule did not adequately protect patient privacy and did have a significant effect in limiting epidemiological research. For example, in a report elaborated by the Joint Policy Committee of the Society of Epidemiology [130], researchers manifested that the implementation of the rule had added burden without adding protections to the existing regulations included in the Common Rule that regulates human subjects research [131]. As one researcher pointed out "An already cumbersome patient consent form now has an additional page-and-a-half explaining HIPAA restrictions. This detracts from the informed consent process pertaining to the more critical issue: the actual medical risks and benefits of participating."

Despite limitations inherent to the method itself, several researcher surveys agree that the HIPAA Privacy Rule has set additional barriers to clinical and epidemiological research. A report by the Institute of Medicine on the effects of the Privacy Rule on clinical and epidemiological research [129] summarizing the results of surveys conducted by the American Association of Medical Colleges (AAMC), the National Cancer Advisory Board (NCAB), the AHRQ, the HMORN, the American Heart Association/American College of Cardiology (AHA/ACC), the Association of Academic Health Centers (AAHC), and the American Society of Clinical Oncology (ASCO) found the following perceived negative effects://

- Increased cost and time required in the conduction of a research project.
- Recruitment of research participants more difficult and increased the likelihood of selection bias.

- Increased participants' confusion.
- Standards for de-identification have not created an effective way for researchers to collect data.
- Led researchers to abandon some studies.
- Created new barriers to use patients' specimens collected during clinical trials.
- Different interpretations of the rule have made conducting research more difficult.

To address the issues around patient privacy regarding secondary use of clinical data, researchers have adopted two main approaches: informed consent and de-identification of records.

3.4.2.1 Informed Consent

Informed consent involves requesting approval from patients to use their clinical data for research. This can be done either prospectively=–requesting approval before the data is collected—or retrospective—requesting approval after the data is collected. Obtaining consent from every patient is the ideal situation in terms of respecting a patient's preferences, but the task is frequently impossible. In contrast with clinical trials, in which researchers have direct contact with study participants, when clinical data is being used for purposes other than direct patient care, that contact does not necessarily exist, posing major difficulties to obtain it. Barriers such as the large number of patients, the proportion of patients that are unreachable, or the need to contact patients at less than ideal circumstances can turn this task into something unfeasible [132]. To solve this regulations such as HIPAA usually allow for the use of previously collected clinical data for research when there is approval by the corresponding IRB [127]. Similar procedures exist in other countries as well [97, 133].

3.4.2.2 De-identification of Medical Records

The second approach is to use de-identified datasets. De-identified datasets are those to which personal identifiers have been completely removed. HIPAA legislation defines 18 different patient identifiers and when they are completely removed the dataset is considered de-identified and can be used for research without patient consent. Furthermore, studies using de-identified data are not considered human subjects research and do not require IRB approval. The task of de-identifying a dataset is not trivial though. Personal identifiers can be easily eliminated or obfuscated when they are stored as structured database fields, however, when they are embedded in clinical narratives such as admission notes, daily progress notes or discharge summaries, the task is complex.

A recent systematic review by Meystre et al. [134] assessed the published literature on automatic de-identification of narrative texts. They identified studies describing 18 different systems to automatically remove personal identifiers from clinical text. Systems used either pattern matching or machine learning, or a combination of both. The systems identified usually performed well, with precision and recall around or above 90%. The authors, however, highlight the fact that there was significant effort involved in developing each one of these systems and that they offer little transferability to other clinical systems or settings. It is also worth noting that a precision and recall of 90% is, given the current regulations, not enough to ensure patient privacy. In addition to not being completely accurate, there is growing concern that data can still be re-identified. El Emmam et al. conducted a systematic review on re-identification attacks on health data. On average, 34% of de-identified health records could be re-identified. Although, when looking deeper into the records that were re-identifiable, the researchers found that a significant proportion of them were not really de-identified according to current standards. Records that were adequately

de-identified, had a re-identification rate of 0.013%.

In summary, ensuring patient privacy is a key component of secondary use of clinical data but ensuring it, through informed consent or data de-identification, is not trivial. Automated systems are not completely accurate and manual de-identification consumes enormous resources. On the other hand, if complete de-identification is achieved, the probability of re-identification is minimal.

3.4.3 Barriers Related to Trust

An additional, less structural but equally important barrier, is trust. It is especially significant when secondary use of clinical data involves the transmission of patient information between organizations. Transmitting health information to and using information from an external organization requires trust. Trust in that the information sent will be handled with the care—this is intimately related to the issue of patient privacy discussed above—and trust in the accuracy of the information received, which will be ultimately used for health related decision-making and research purposes. Concordantly, lack of trust has been shown to be negatively associated with the adoption of health information exchanges [135], which, as I have discussed, are important for secondary use.

The issue of trust is intimately related to data governance. Data governance is *"a system of decision rights and accountabilities for information-related processes, executed according to agreed-upon models which describe who can take what actions with what information, and when, under what circumstances, using what methods"* [136]. This means that the depositaries of such responsibility need to ensure that an organization's information assets are used according to policies, regulations and in alignment to the organization's mission and, in this case, have to assess whether data

sharing its data with other organizations fits with all those constraints. These are necessary precautions; nonetheless, they still impose additional barriers to secondary use of clinical data.

3.4.4 Barriers to Adopt a Clinical Data Warehouse

Once an organization has decided that it will use its clinical data for secondary uses, it needs to invest in the necessary infrastructure. A central component of this infrastructure is the availability of a clinical data repository (CDR). CDRs are databases that contain clinical data, usually integrating data from disparate sources. These systems are different from transaction-based clinical care systems, which are databases used for daily patient care. CDR are separate entities used for multiple purposes such as supporting clinical operations, quality improvement, clinical research, among others [137].

In spite of the variety of potential uses for a CDR, their adoption is not widespread. A survey conducted among all CTSA centers in the US, premier research institutions with specific funding allocated to encourage translational research and secondary uses of clinical data, showed that 76% of them had some form of CDR, up from 64% in 2008 [137]. A literature search did not identify a study reporting adoption rates in non-academic, non-research institutions, but it is reasonable to expect that adoption rates would be significantly lower. One of the factors that might influence this less-than-ideal adoption of CDRs is the high rate of failure of data warehouse initiatives. Although not studied in health organizations, failure rates of data warehouse adoption initiatives are thought to be close to 50% [138].

The diffusion of innovations theory [139] provides an interesting framework to understand the factors influencing the adoption of an innovation such as a CDR. Among

these we can find the perceived benefit of the innovation, financial factors, compatibility with beliefs, compatibility with previous ideas, and incentives to adopt. In the context of information technologies, Klein et al. [140] described that, from an organizational point of view, resource availability, management support, an adequate implementation climate, formal implementation policies and practices, and perceived benefits can influence the adoption of a new computer-based technological advances.

The factors influencing the adoption of CDRs have not been systematically explored. However, a study conducted by Ramamurthy et al. [138] can cast some light on the factors involved in adopting such technology. In that study, using Roger's diffusion of innovations theory as a framework [139], the authors explored the factors associated with successful implementation of a enterprise wide data warehouse. The authors identified five attributes associated with adoption of a data warehouse. From the organization's point of view, its size, its absorptive capacity (*"the ability to create and nurture an environment to absorb and transfer the skill base to exploit the nuances of an innovation"*), and its commitment to adopt the innovation (such as senior management support and stakeholder's buy-in) were all positively associated with successful adoption. From the innovation's point of view, the relative advantage, or perceived usefulness, was positively associated with adoption; the complexity of the tool was negatively associated with successful adoption.

In the health care arena, two studies have been conducted that allow us to understand some barriers to adopt a clinical data repository. The first one is a study conducted by Schubart et al. in 1999 [141]. In this study, the authors surveyed users of the University of Virginia clinical data repository and found that previous knowledge of clinical coding systems and computer software training, as well as time to test the system were positively associated with the use of the CDR. Privacy concerns and complexity of the tool were described as issues that might impede the use of the

tool. However, the authors defined users as individuals with privileges to submit a query and adoption as actually submitting a database query. These results might not necessarily represent actual users of clinical data for secondary purposes, such as researchers or quality managers, but programmers in charge of extracting data.

The second study is the survey of CTSA centers mentioned above [137]. The survey asked about the biggest obstacles faced while getting a CDR project approved and the biggest challenges for ongoing CDR projects. The former set of obstacles included obtaining sponsorship or buy-in, adequate funding, issues around data ownership and adequate staffing. For ongoing CDR projects, the obstacles also included staffing and funding, but also patient privacy/HIPAA and data quality issues.

Several measures have been adopted to increase the adoption of CDRs by researchers and other secondary users of clinical data. One approach has been to improve the user interface to increase the adoption by less experienced database users. For example, the Informatics for Integrating Biology and the Bedside (i2b2) [142], a project funded by the National Institutes of Health to develop infrastructure to accelerate translational research, developed a set of tools that include a clinical data repository with a user-friendly interface to allow non-experts to execute simple queries using drag-and-drop functionalities. Similarly, the implementation of the Duke Enterprise Data Unified Content Explorer (DEDUCE) included a Guided Query component to allow the construction of database queries without "knowledge of the underlying clinical database structure" [143]. In addition to improving the user's experience, user training has been a centerpiece component of initiatives to improve the adoption of CDRs [144].

The extent to which data quality—understood as a broad concept that encompasses not only the quality of the data itself but also the ease of access, analysis

an interpretation—influences the use or adoption of clinical data repositories by researchers has not yet been explored.

3.5 Summary and Implications

The overarching research question for this dissertation is: How can we improve researchers' abilities to use electronic clinical data for clinical and epidemiological research? Up to this point, in chapter 2 I have presented evidence about the potential benefits of using routinely collected electronic clinical data for secondary purposes, which if fully implemented could lead to establishing a learning health care system. However, as the literature review presented in this chapter shows, several barriers exist that limit our ability to use electronic patient data for secondary purposes.

Barriers arise from characteristics of the data itself, as well as from data residing within information systems and them, in turn, residing inside organizations and societies (see figure 3.1). As a consequence, they can be grouped, although with significant overlap, into two broad categories: Data-related barriers and societal and organizational barriers.

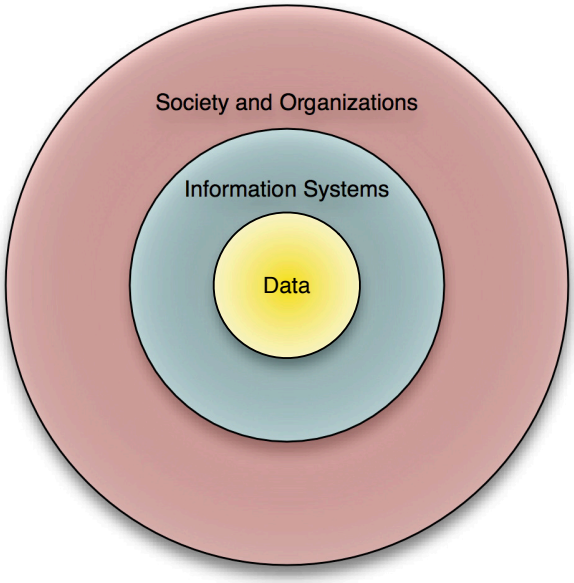


Figure 3.1: Levels at which barriers to secondary use of clinical data exist.

Data-related barriers include inadequate data quality, the availability of non-computable formats, and consequences of data fragmentation. The vision presented here highlights the fact that data quality is not an attribute that depends exclusively from the data itself but also depends on the characteristics of the information systems it is embedded in and who the users are. Users of clinical data for secondary purposes are diverse and so are their needs. So far, and especially for researchers, they have not been fully characterized and, thus, the assessment of data quality for their purposes remains incomplete.

Societal and organizational barriers include barriers that arise from how the health care system is designed and operated, how patient privacy issues are handled and how organizations solve their needs to use clinical data for secondary purposes. The extent to which an organization facilitates or not the a researcher's ability to use clinical data is so far unknown.

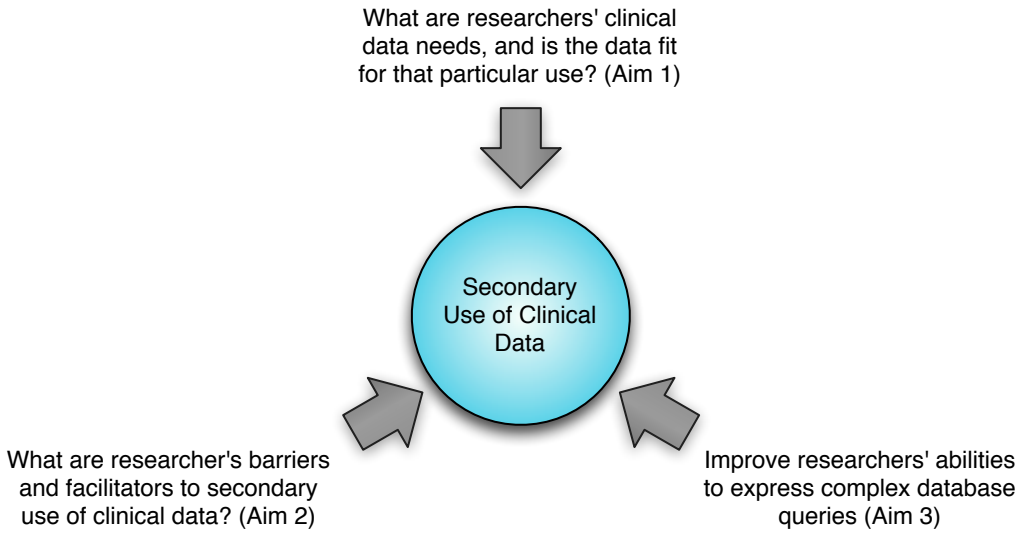


Figure 3.2: Three aspects of secondary use of clinical data that this dissertation will address.

Some barriers identified in the preliminary inquiry leading to this dissertation have not been systematically studied or addressed in the literature so far. These barriers include (see section 1.2):

- Long iterative processes of building complex database queries to meet researchers expectations, data availability and the knowledge to extract it.
- Excessive reliance on manual chart abstraction.
- High frequency of relative temporal queries.

Given these current areas of uncertainty, I decided to address this issue from three different but interrelated fronts (see figure 3.2):

- Develop, refine and apply a tool to systematically assess the complexity underlying clinical data requests.

- Explore the barriers and facilitators to secondary use of clinical data experienced by researchers.
- Develop and test a tool to allow researchers to build complex database queries.

As a result, this dissertation is composed by the following aims:

- **Aim 1**: Develop and apply a Clinical Data Request Complexity Assessment Tool (CDR- CAT) which will be used to systematically evaluate the complexity of clinical data requests sent to a clinical data repository.

- **Aim 2**: Conduct a qualitative study to identify the barriers and facilitators perceived by researchers conducting outcomes research using clinical data as their primary data source.

- **Aim 3**: Develop and implement a framework to describe and query clinical data based on temporal intervals.

The next three chapters describe the three proposed aims.

Chapter 4

AIM 1: A SYSTEMATIC CHARACTERIZATION OF RESEARCHERS' CLINICAL DATA REQUESTS AND OF ELECTRONIC CLINCAL DATA'S FITNESS FOR USE

4.1 Introduction

To answer the question: How can we improve researchers' abilities to use electronic clinical data for clinical and epidemiological research?, the literature reviews presented in chapters 2 and 3 described the potential benefits of using routinely collected clinical data for secondary purposes—one of which is clinical research—and the known barriers to secondary use of clinical data. Despite all the existing literature on current barriers, little is known about the kind of clinical data researchers need or whether electronic clinical data is fit for use. This chapter addresses those questions.

To apply the concept of 'fitness for use' to clinical data, which I introduced in chapter 3, requires that we understand users' needs as well as the resources available to meet those needs. Therefore, to assess the quality of electronic clinical data for this kind of secondary use it is necessary to characterize the type of data researchers request from clinical data repositories. As a first step to improve the use of clinical data for secondary purposes, this chapter will answer the following questions (see figure 4.1): **What kind of data do researchers need to extract from clinical databases? Is the electronic clinical data fit for use?**

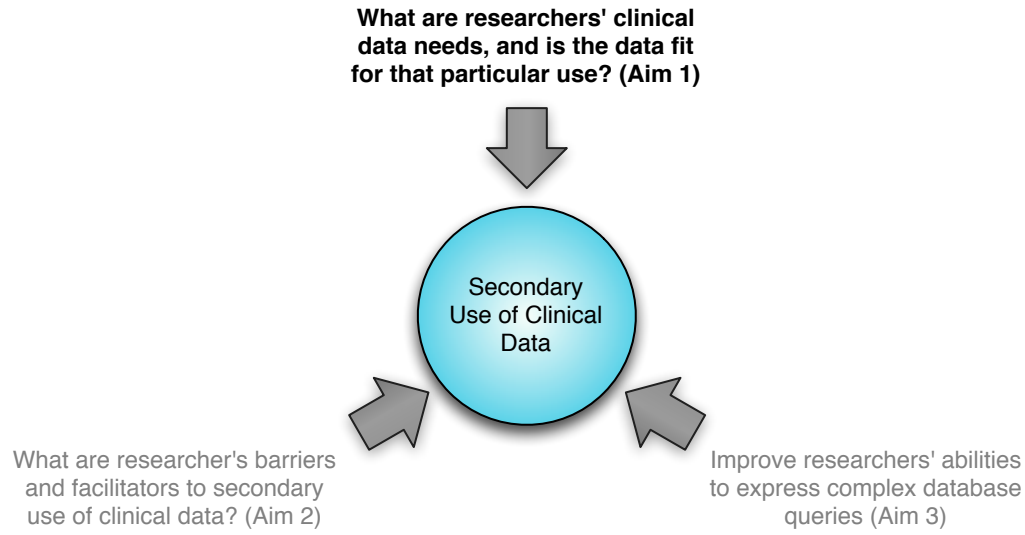


Figure 4.1: Aim 1 focused on **understanding researchers' clinical data needs and whether the data fit for that particular use**, in the context of this dissertation's overarching question: How can we improve researchers' abilities to use electronic clinical data for clinical and epidemiological research?

4.1.1 Assessing Data Quality: Fitness for Use

Clinical Data Repositories (CDR) are databases that aggregate disparate clinical data sources into a consolidated one. This source can be later queried for multiple types of secondary use. In the biomedical sciences CDRs are considered useful sources of data for researchers [105] and multiple organizations around the world have adopted them for research purposes [143, 145, 146, 147]. However, the usefulness of the data they contain greatly depends on it being accessible, extractable, and usable. As mentioned in chapter 1, researchers frequently encounter mismatches between what they need and what a CDR is able to deliver.

A researcher's need for clinical data is instantiated as a *clinical data request* [148]. A clinical data request embodies the set of clinical elements that a researcher wishes to extract from a clinical database. For example, the University of Minnesota Trans-

plant Information System offers researchers an on-line form to describe a clinical data request [149]. That form collects the requester's objectives, the human subjects' protection approval code, the type of organ involved (this is a transplant database), the list of study variables, the study population, the timeframe, and the data sources. Similarly, the i2b2 platform (Informatics for Integrating Biology and the Bedside), a CDR developed by Partners Healthcare lets users create a clinical data request based on Boolean combinations of clinical attributes such as demography, clinical conditions, procedures and laboratory results [142]. Evidently, data requests can be even more diverse than researchers. The multitude of health conditions, exposures or interventions being studied, along with the large variety of study designs available, means that the number of different clinical data requests that are possible to express are endless. This diverse set of needs is the first component required to assess clinical data's 'fitness for use'.

On top of having diverse needs, researchers are embedded in diverse organizations. These organizations might have different information technology infrastructures, different levels of EMR implementation, different degrees of information integration, as well as different levels of knowledge regarding the extraction of electronic clinical data for secondary use [150]. Thus, the second component of assessing whether electronic patient data is 'fit for use' is the ability of an organization's clinical data infrastructure to meet researchers' needs. In sum, assessing 'fitness for use' requires the consideration of both (1) users' needs and (2) the organization's resources such as data sources and infrastructure (see figure 4.2). Although not an absolute concept, when there is a large mismatch between these two, the data can be considered not fit for use.

The causes behind the mismatch between what researchers need and what they are able to extract from clinical databases—in other words, why the data is not 'fit for use'—have not yet been systematically explored. The vast majority of research

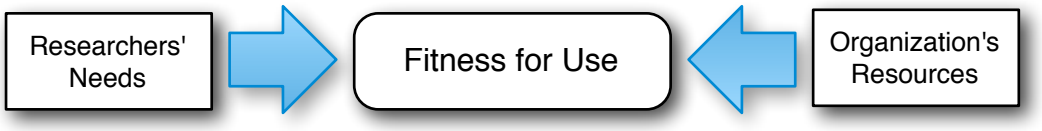


Figure 4.2: "Fitness for use" is a result of the combination of (1) researchers' needs and (2) the ability of the organization's data resources and infrastructure to meet those needs. Under this broader model of data quality, when researchers' needs are not being met because of problems with intrinsic data quality, inadequate formats, barriers to access, etc. the data is deemed to be not fit for use.

focuses on intrinsic data quality, this mostly means the accuracy of the data. In a study conducted by Stein et al. the authors investigated the concordance between textual and coded data inside a CDR [82]. Interestingly, the authors had to manually review all the text-fields inside the CDR, a very labor-intensive process, which is a potent signal that intrinsic data quality is not the only relevant component of 'fitness for use'. In this case, even when the data was concordant—an attribute of intrinsic data quality—the data was accessible only through a manual process, thus limiting its utility. On top of that, and as I described in chapter 3, researchers are a diverse set of users so the commonly used approach of analyzing data quality by lumping researchers into a single broad category of homogeneous users, without considering their diverse needs, is not adequate for the purpose of determining "fitness for use".

The goal of this dissertation aim is to systematically assess this dyad: researchers' clinical data needs along with the resources available to meet them. To this day there is no tool or instrument available to assess researchers' data needs and the complexity involved in meeting those data needs given the resources available. Although several tools to assess information systems' fitness for use exist [86], this broad view of data quality has been scarcely applied in health care and has not been applied in the

domain of secondary use of clinical data. This chapter describes the use of a Delphi process to develop such a tool. The resulting tool is then used to assess the 'fitness for use' of available clinical data sources, given a collection of unique and diverse individual clinical data requests. A flow diagram describing all stages of this study is shown in figure 4.3.

4.2 Methods

4.2.1 Delphi Method

The Delphi method has been defined as an "...iterative, multistage process designed to combine opinion into group consensus" [151]. It is a research method that seeks to collect consensus expert opinion about an issue where there is significant uncertainty or there is not enough information available. The RAND Corporation developed the Delphi Method in the 1950s for a study sponsored by the US Air Force [152]. The study involved obtaining expert consensus on military decision making issues but has been applied to a multitude of research domains.

To reach expert consensus the Delphi method uses a series of questionnaires along with the provision of controlled feedback to the study participants until consensus is reached. Although there are many ways of conducting a Delphi method, the basic structure is the following [153]:

Initial round: questions regarding the topic being discussed are presented to a panel of experts. This round usually includes open-ended questions that allow experts to elaborate on their opinions.

Data analysis: data provided during the initial round is analyzed—qualitatively and quantitatively—and is summarized so the results can be later presented back to the panel of experts. The results are anonymized and aggregated prior to

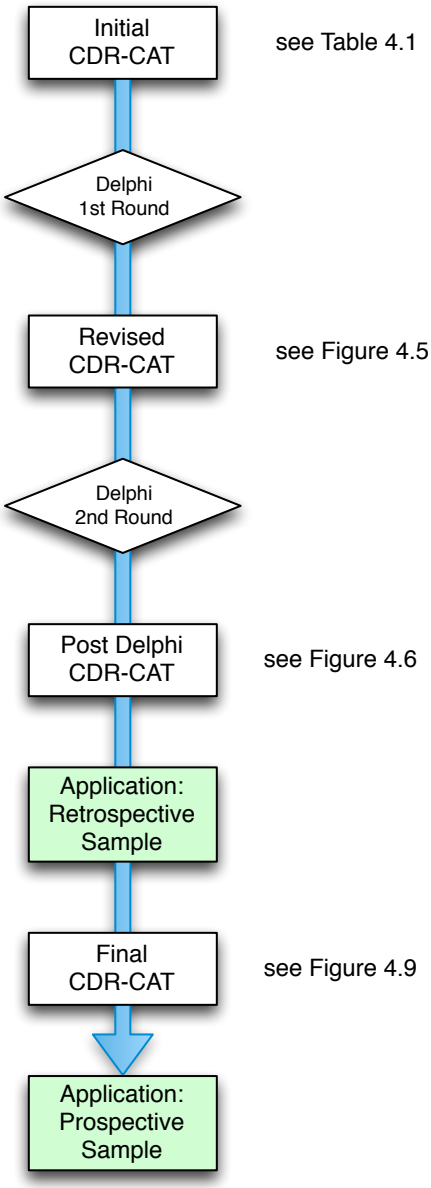


Figure 4.3: Flow diagram describing all stages of this study. CDR-CAT = Clinical Data Request Complexity Assessment Tool.

presentation in order to prevent the prioritization of one expert's opinion over others. In words of Linstone et al. [154]:

> *"...the tendency to judge those developments suggested by the most notable panelists [are] eliminated by virtue of anonymity".*

In addition to this, anonymity is considered a key component of a Delphi method since it *"allows the participants to freely express their opinions without undue social pressures to conform from others in the group. Decisions are evaluated on their merit, rather than who has proposed the idea"* [155].

Second round : the summarized results are presented back to the panel of experts along with a second set of questions. This time, questions are less open than during the first round.

Depending on the results obtained and whether consensus has been reached, a Delphi method can have more than two rounds. Since they are based on the results obtained during previous rounds, in each subsequent round the questions are more specific.

4.2.1.1 Sample Selection

A second characteristic is that the Delphi method does not select participants using random samples of individuals. It employs experts or specialists in the field based on their knowledge of the subject [156]. Some researchers have expressed concerns about these purposeful samples, however, by definition the goal of a Delphi process is to collect expert opinions and reach expert consensus on a specific topic.

4.2.1.2 Uses of the Delphi Method

Since its initial description, the Delphi method has gained popularity. In a 1994 report by Mackenna et al. the authors state that the method has been used in over 1000

publications, most of them in the area of social policy [156]. Its use has continued to explode since that report. For example, the Delphi method has been used as a forecast tool, as well as a tool to identify and prioritize relevant issues. Kendall et al. used the method to forecast the role of system analysts in the 21st century [157]. More recently, and in the domain of framework development, Soon et al. used the same method to develop and validate a farm food safety risk assessment tool [**?**].

4.2.1.3 Uses of Delphi Methods in Health Sciences Research and Health Informatics

Just as it has occurred in other disciplines, the use of the Delphi method has also increased in health sciences research, especially in public health policy and nursing. Normand et al. reported their use of the technique to identify a set of performance measures in cardiovascular medicine [158]. Steele et al. used a Delphi approach to establish research priorities in pediatric palliative care [**?**]. A search in the PubMed database of biomedical literature reveals more than 2,000 publications indexed with the structured search term "Delphi Technique"[Mesh].

The domain of health informatics has also used the method extensively. A search in the PubMed database retrieves more than 130 studies indexed using the structured terms ("Medical Informatics Applications"[Mesh]) AND "Delphi Technique"[Mesh]). Some examples include a study by Blumenthal et al. that used a Delphi method to reach expert consensus and establish a list of essential functionalities that define an electronic medical record [159]. Similarly, Riedman et al. conducted a study using a Delphi method to establish a set of attributes that would improve the delivery of medication alerts inside an electronic medical record [160].

4.2.1.4 Why a Delphi Method is Appropriate

As we can see from the examples provided above, the Delphi method can be effectively used to collect expert opinion and reach consensus in matters where there is

significant uncertainty. Consequently, given the lack of information available on the issue of assessing researchers' needs, this method is well suited for the purposes of (1) identifying the attributes that make a clinical data request complex to fulfill, and to (2) develop of a tool to systematically collect such attributes, while preserving the unique characteristics of each researcher's needs.

The benefit of applying the Delphi method to an area where significant uncertainty exists is the establishment of some ground rules that allow a more detailed exploration of the subject. For example, the report by Blumenthal et al. commissioned by the Robert Wood Johnson Foundation to establish list of EMR functionalities made possible a continuous and systematic assessment of the penetration of electronic medical records of different degrees of complexity, both in the US and elsewhere [161, 162, 150]. Similarly the development of a tool to assess researchers' needs and the complexity involved in meeting those needs should be a significant step towards improving the secondary use of clinical data.

4.2.2 Initial framework

The first step in this study was developing a preliminary list of issues or data attributes that can create difficulties when extracting electronic clinical data for research. This list was constructed with data collected during an initial literature review, the initial assessment of clinical data requests for research sent to the Institute of Translational Health Sciences in 2010 as described in chapter 1, and a set of interviews with domain experts from the University of Washington. This preliminary list can be found in table 4.1. This list constituted the first version of a Clinical Data Request Complexity Assessment Tool (CDR-CAT) (CDR-CAT) and would be the starting point for the Delphi process and subsequent stages, which will in turn refine and modify the tool into its final form.

Attribute	Domain	Description
1. Full database Schema Match (a structured field is available for the required data element)	1.1 Accuracy	The data frequently misrepresents the reality
	1.2 Consistency	The database contains conflicting versions for the value for one data element
	1.3 Objectivity	The data element requires subjective evaluation, as opposed to entering raw data.
	1.4 Timeliness	Element is not available in a timely fashion.
	1.5 Complex Query	Element can be extracted, but requires the use of a complex database query.
2. Not a database schema match (element is not contained in structured fields)	2.1 Non-computable	Element is present in a non- computable format.
	2.2 Not available	Element is not available in the database.
3. Post Processing		An initial database extract needs to be processed or validated through manual or statistical methods.
4. External or unavailable data for interpretation		There is a need to use external or unavailable data to interpret the requested data element.
5. Cannot be extracted		Element is not extractable through the query language in use or the query is too complex to be written using locally available knowledge.

Table 4.1: Initial framework: set of attributes used during the first round of the Delphi process.

This list was designed so it could be applied to each clinical element of an EMR data request. By clinical element I mean every clinically relevant component of a data request that could be individually assessed. For example, if a researcher requested an EMR data extract containing medical records belonging to adult patients with a diagnosis of type 2 Diabetes Mellitus, who had 1 or more clinic visits during the previous two years, without kidney failure and normocalcemia, the clinical elements of that request would be the following:

- Adult (>18 years old)
- Type 2 Diabetes Mellitus
- Normal kidney function
- Normal Calcium

The distinction between *clinical element* and *database element* is a key issue. Since this tool was designed with the concept of 'fitness for use' described in section 3.3.1.2 in mind and, thus, the tool was be constructed from the researchers' point of view as the unit of analysis. In this case, the clinical data request was the definition of a researcher's *use* and it was decomposed in its constitutional (clinical) parts. The data available inside the database, along with its attributes, may or may not be adequate to meet the researcher's needs and the complexity involved in doing so may vary significantly.

For this Delphi process, I invited experts to participate given their experience in developing clinical data warehouses and secondary uses of clinical data. The Delphi process was implemented using a web-based platform in which experts could anonymously express their judgment and opinions. It began with a general introduction to the issues around secondary use of clinical data and a description of what constituted a clinical data request.

4.2.3 Delphi: First Round

The first round consisted of a series of questions to assess each item on the preliminary Clinical Data Request Complexity Assessment Tool (CDR-CAT) (see table 4.1). Each question consisted of a formal definition of the item, and a request for each expert to:

1. Rate the relevancy of each item on the initial CDR-CEF using a 9-item Likert scale, with 1 meaning *'not relevant at all, should be removed from the assessment tool'* and 9 meaning *'absolutely relevant, it should be included in the assessment tool'.*

2. Provide comments on the item's label or its definition.

A sample question can be found in figure 4.4. The full questionnaire can be found in appendix A on page 219.

This first round also included at the end an open-ended question that allowed researchers to mention attributes they express any additional item they would consider adding to the assessment tool or any other suggestion or concerns they wished to express.

The results were anonymously collected using the on-line platform for later analysis. For each item on the assessment tool, the results of the question about its relevancy were analyzed in the following manner:

- If experts *homogeneously* though that the item was relevant (with a score >= 7), the item was kept in the assessment tool.

- If experts *homogeneously* thought that the item was not relevant (with a score <= 3), the item was removed from the assessment tool.

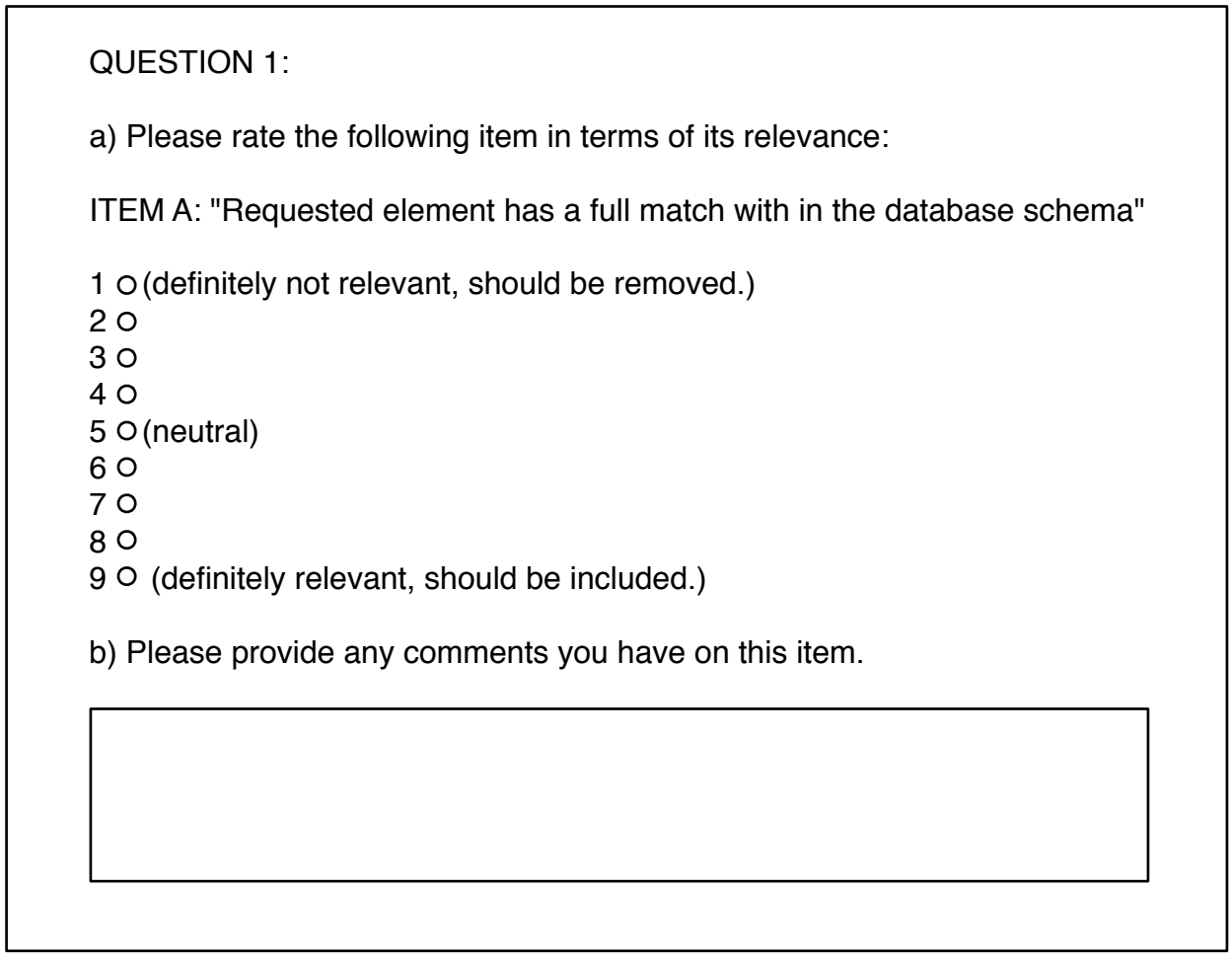


Figure 4.4: Sample question used in the first round of the Delphi process.

- If expert opinions showed *heterogeneity*, the item was reassessed in the second round of the Delphi process after incorporating the suggestions expressed by experts for that individual item.

Since the sample size did not allow for statistical testing of homogeneity/heterogeneity, responses were defined to be heterogeneous when 1/3 of the responses fell outside the 3-point range in which the responses' median fell. For example, if the median was 6, it fell in the middle interval (between 4 and 6) and responses would be considered heterogeneous if 3 or more responses were lower than 4 or higher than 6. This is consistent with what others have reported [153].

The responses to open ended questions were carefully reviewed and analyzed, using content analysis to identify themes or suggesting the need to modify each item's label, definition, or the tool itself. The output of the first round was a refined version of the CDR-CAT according to expert opinion and suggestions. Results of this round can be found on section 4.3.1.

4.2.4 Delphi: second round

The second round was implemented using the same on-line platform to ensure anonymity of the responses. In this round, experts were again provided with a general description of the problem, the definition of a clinical data request, and a revised version of the CDR-CAT.

Questions this time included all items that were selected for re-assessment during the first round, with modified labels and definitions according to the experts' suggestions. Again, experts were asked to respond regarding the relevance of each item individually. In addition, experts were asked to rank the top attributes that made extracting a clinical data element complex. Results were analyzed using the same

methods used in the previous round.

The output of the second round was a new version of the CDR-CAT, which we will denominate post-Delphi. This post-Delphi CDR-CAT was then used in the next stage of this study. Results of the second round can be found in section 4.3.3.

4.2.5 Application of the post-Delphi CDR-CAT

The objective of this stage was to test the applicability of the post-Delphi CDR-CAT to real-world clinical data requests sent to the local clinical data warehouse. This stage involved the collection of a retrospective sample of EMR data extraction requests sent by researchers to the University of Washington's Institute of Translational Health Sciences (ITHS). The ITHS is one of 60 Centers for Translational Science Awards funded by the US National Institutes of Health (NIH) [51]. These centers were created to provide *"infrastructure support to facilitate translational research, promoting training and career development for translational researchers, and developing innovative methods and technologies to strengthen translational research"* [163], which includes developing resources to support translational research. Clinical data repositories are one example of such resources. In the case of the ITHS, the center has numerous clinical data resources including an institution-wide clinical data warehouse and a structured method to collect and process requests for clinical data extractions made by local researchers. Clinical data requests received between July and December of 2011 were collected and retrospectively analyzed.

Each clinical data request was divided into it's constituting clinical elements as described in section 4.2.2. The post-Delphi CDR-CAT was then applied to each clinical element and the results were stored in a spreadsheet for later analysis. The application of the post-Delphi CDR-CAT had two objectives. The first one was to assess whether the tool was able to capture the database attributes that made extracting

clinical elements complex. For this first objective, the tool was applied by the clinical database expert that routinely fulfills researcher's requests at the University of Washington's ITHS. The second objective was to systematically characterize the set of clinical data requests.

During this stage, the tool was further modified according to its ability to adequately capture the complexity of a data request. The output of this stage was the final version of the CDR-CAT.

4.2.6 Prospective Sample

The final stage of this study consisted of the application of the final CDR-CAT to a prospective sample of clinical data requests received by the ITHS over a period of 6 months, from January to June of 2012.

4.3 Results

This study was conducted between January and June of 2011. A total of 10 experts were invited and 9 agreed to participate, a summary description of the participants can be found on table 4.2. Experts were involved in all rounds of the Delphi process.

4.3.1 Delphi: First Round

4.3.1.1 Qualitative Results

Experts' opinions and comments on the initial framework were analyzed and summarized into the following categories or themes:

Better description of the consultation process: Two experts asked to clarify whether the tool was assessing the data requests or the database queries.

	Location	Institution
Participant 1	USA	University of California - Davis
Participant 2	USA	Ohio State University
Participant 3	USA	University of Pennsylvania
Participant 4	USA	University of Washington
Participant 5	USA	University of Washington
Participant 6	USA	University of Washington
Participant 7	USA	Group Health Research Institute
Participant 8	USA	Columbia University
Participant 9	UK	University of Surrey

Table 4.2: Summary description of experts that participated in the Delphi process.

Clarify that the tool evaluates the fit between the data request and the database: Experts noted that, for some attributes, it was not clear whether they were evaluating the data contained in the database or one particular clinical element of the data request.

Provide examples: One expert requested for specific examples

Tool not sequential: Two experts noted that not every attribute should be applied to every element of a clinical data request. For example, if one clinical data element was not available in the database, then it is impossible to assess other data quality attributes for that same element. This resulted in transforming the tool from a table in which the attributes are assessed linearly (see table 4.1 in page 70) into a workflow with branching paths (see figure 4.5 in page 79).

Multiple sources: Two experts noted that the tool did not specify how to proceed when there was more than one source for a single clinical data element (i.e. more than one source from where to obtain kidney function).

Granularity of the categories under 'not available': One expert suggested additional sub-categories within the 'not available' category. For example, further specifying into 'free text' (which could be suitable for natural language processing), 'non-computable' (such as scanned notes and PDF documents), and simply 'not available'.

4.3.1.2 Relevance and Heterogeneity

Of the 12 attributes assessed, experts assigned scores $\geq$7 to 11 of them. The only one that was scored lower than 7 was the attribute 'cannot be extracted', however the low score was not homogeneous among participants, which determined that the attribute should be reassessed during the next round after incorporating experts' comments. Of the 11 attributes that were scored $\geq$7, for five of them the assessments were homogeneous and thus were included in the final tool. The remaining 6 attributes had heterogeneous assessments and thus were also reassessed in the second round. Results are summarized in table 4.3.

4.3.2 CDR-CAT: Revised Version

The results of the first round of the Delphi process were used to revise the CDR-CAT. The most important change was switching from a linear version into a branching version in order to reflect the fact that once a data element is not available in a structured format, it did not make sense to further assess the data quality attributes. In addition, the language was rephrased to reflect the experts' comments. The revised version is depicted in figure 4.5. It is worth noting that this revised version of the CDR-CAT has four domains: (1) *data not computable/not available*, (2) *data quality issues*, (3) *complexity of the query*, and (4) *post processing needed*.

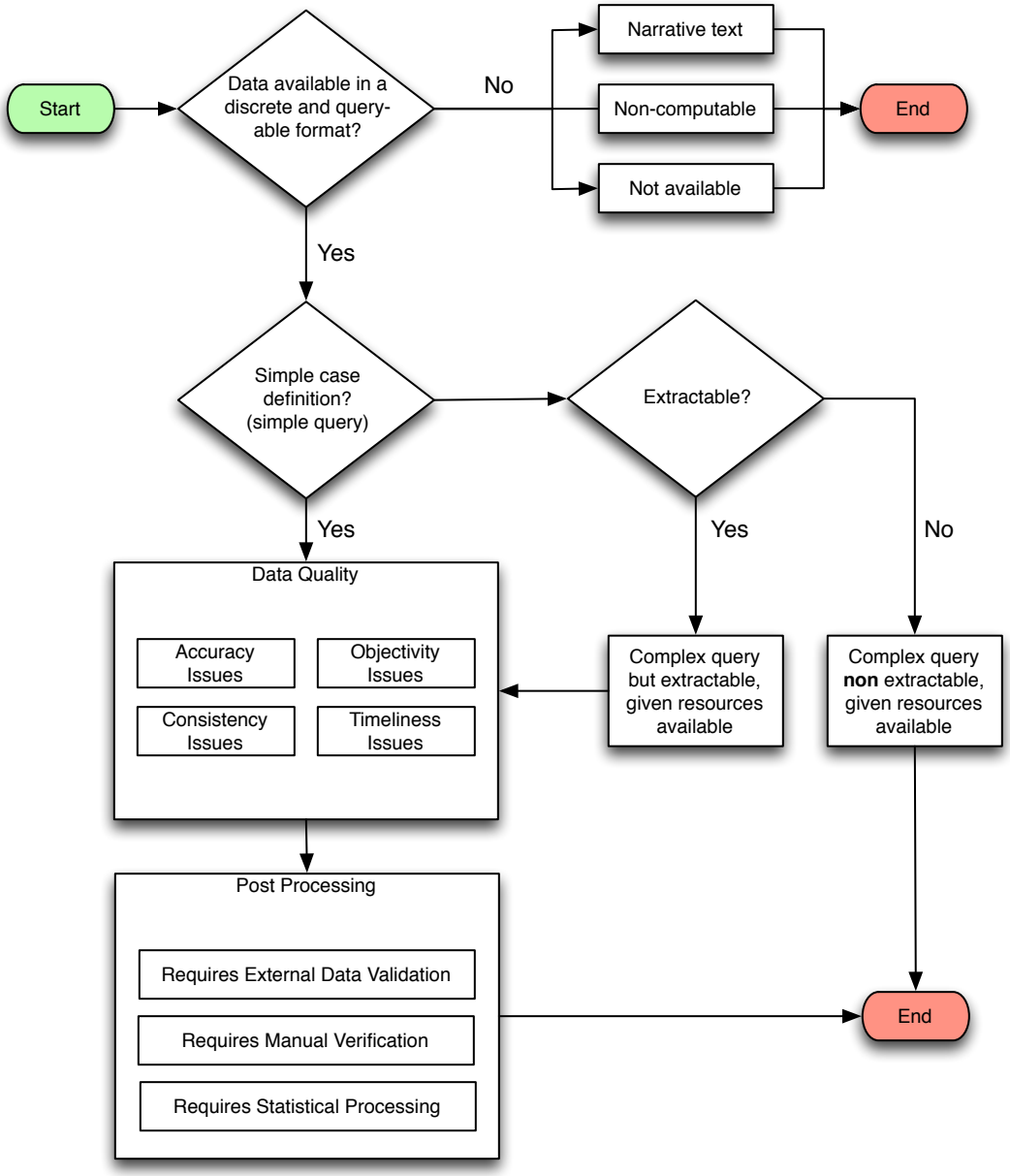


Figure 4.5: Clinical Data Request Complexity Assessment Tool (CDR-CAT): revised version after the first Delphi round. Note that the tool now has four domains: (1) data not computable/not available, (2) data quality issues, (3) complexity of the query, and (4) post processing needed

Attribute	Median Score (range)	Heterogeneity	Result
1. Full database schema match	7 (2-9)	Yes	Reassess
1.1 Accuracy	**9 (4-9)**	**No**	**Include**
1.2 Consistency	**9 (4-9)**	**No**	**Include**
1.3 Objectivity	7 (4-9)	Yes	Reassess
1.4 Timeliness	7 (5-9)	Yes	Reassess
1.5 Complex Query	8 (3-9)	Yes	Reassess
2. Not a database schema match	7 (1-9)	Yes	Reassess
2.1 Non-computable	**9 (2-9)**	**No**	**Include**
2.2 Not available	**9 (5-9)**	**No**	**Include**
3. Post Processing	**8 (5-9)**	**No**	**Include**
4. External or unavailable data for interpretation	7(4-9)	Yes	Reassess
5. Cannot be extracted	5(2-9)	Yes	Reassess

Table 4.3: Delphi process results: first round

4.3.3 Delphi: Second Round

For the second round, experts were presented with the revised version of the CDR-CAT, which included relabeling some attributes to clarify the concepts. They were then asked to rate in a 1-to-5 Likert scale the clarity of the new labels. Overwhelmingly, experts agreed that the new labels were a definite improvement. Next, and in the same way as in the first round, experts were asked to provide their assessment of the relevance of the re-labeled attributes. Finally, experts had to rank the

four domains of the tool in terms of their contribution to making a data extraction complex.

4.3.3.1 Qualitative Results

Since the second round consisted of fewer and more precise questions, open-ended questions were limited. However we obtained relevant information regarding the post-processing component of the assessment tool. In the words of one expert:

> "*I think post processing is too generic. Almost all queries will need post process[ing]*"

Additionally, most feedback on this item was related to the need for manual extraction or verification of clinical data from medical records when some request element was not accessible. Again, in the words of a participating expert:

> *"In the CABG example as given, it seems like the conclusion is that the query could not be done because of the non-computable nature of the radiocontrast exposure. However, in my view the query can be completed with the other two criteria and then create a cohort from which operative notes could be drawn to manually look for the radiocontrast dye. That is more cumbersome, but still may result in a successful study. "*

Given this feedback, the issue of manual verification/extraction was included in the tool.

4.3.3.2 Relevance and Heterogeneity

In this second round, 6 attributes were assessed. Of them, experts assigned scores of $\geq$7 to all of them. However, only '*Objectivity*', '*Timeliness*' and '*Non-extractable*'

received high scores homogeneously. Complex Query, External Validation and Statistical Processing received heterogeneous results and, since they had already been assessed in the same way on the previous round, they were excluded from the tool. Results are summarized in table 4.4.

Attribute	Median Score (range)	Heterogeneity	Result
Objectivity	**7 (6-9)**	**No**	**Include**
Timeliness	**8 (7-9)**	**No**	**Include**
Complex Query	7 (5-9)	Yes	Exclude
Non-extractable	**8 (6-9)**	**No**	**Include**
External validation	7 (4-9)	Yes	Exclude
Statistical processing	7 (1-9)	Yes	Exclude

Table 4.4: Delphi process results: second round

When asked to rank the four domains of the CDR-CAT in terms of their contribution to making a clinical data element complex to extract, experts expressed that the most relevant domains were, jointly, *data not computable* and *data quality*, followed by *complexity of the query* and, in last place, they ranked the *need for post-processing*.

4.3.4 CDR-CAT: post-Delphi version

After the second round, the CDR-CAT was further revised to reflect experts' opinions and consensus. Therefore, *external validation* and *statistical processing* were eliminated from the tool. Although there was no consensus on the *complex query* attribute, it was included in the tool, not as an attribute to assess but as a decision node in the algorithm. The post processing domain was ranked low in its contribution to a

complex data extract, nevertheless, the qualitative results pointed out that the specific issue of manual validation/extraction was very relevant, so it was included as the only component of the post-processing domain. The second version of the CDR-CAT is depicted in figure 4.6. Considering the results of this second round, a third round was deemed unnecessary. The post-Delphi CDR-CAT was used to characterize a set of clinical data requests.

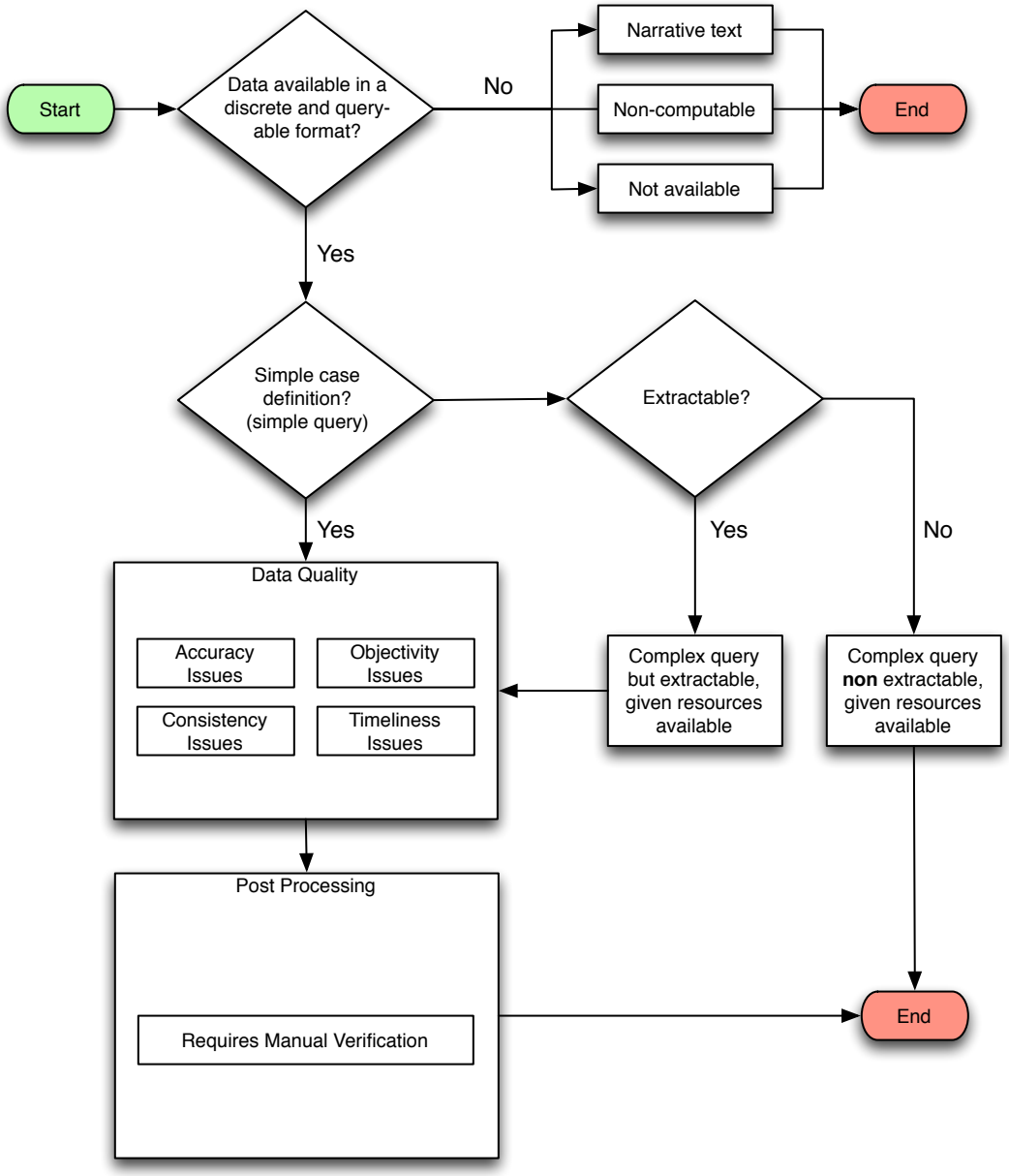


Figure 4.6: Clinical Data Request Complexity Assessment Tool (CDR-CAT): revised version after the second Delphi round.

4.3.5 Retrospective Analysis of Clinical Data Requests

All clinical data extraction requests received by the University of Washington's Institute of Translational Health Sciences (ITHS) from July to December 2011 were analyzed using the post-Delphi CDR-CAT.

During the six-month period the ITHS received 12 unique requests for clinical data for research, which included a total of 206 clinical elements. As expected, data requests included ranged from very simple to complex ones. The most simple data request consisted of only two clinical elements (obtaining counts of ICD-9 diagnostic codes associated with Diabetes type 1 and 2 over a predefined range of dates). The most complex data request sought to identify patients with HIV hospitalized in the intensive care unit and consisted of 59 distinct clinical elements. These clinical elements involved very simple database queries (i.e. CD4 count) or more complex ones (i.e. all relevant preexisting co-morbidities).

Of all the clinical elements requested, the largest proportion corresponded to laboratory results (n = 39, 19%), followed by diagnostic codes (n = 32, 16%), medications (n = 29, 14%) and a myriad of clinical elements contained in free-text clinical notes or reports (n = 29, 14%). A full description of the characteristics of all clinical data requests can be found in section 4.3.8.

As described in section 4.2.2, the complexity of extracting each clinical element was assessed using the post-Delphi CDR-CAT algorithm (Figure 4.6). To illustrate this, an example of the application of the CDR-CAT to three sample clinical data elements can be observed in figure 4.7. The full characterization of the complexity of extracting all clinical elements can be found in section 4.3.8.

			Data Not Computable/Available			Query Complexity		Data Quality Issues				Post-Processing
Clinical Element	Type	Discrete/ Queryable	Not Computable	Not Available	Free-text	Simple Case	Extrac-table	Accuracy	Consis-tency	Timeliness	Objectivity	Manual Verification
COPD	Dx	YES				YES	YES	UNK	UNK	NO		YES
FEV-1	Lab	NO		YES								YES
>48h in ICU	ADT	YES				NO	YES	UNK	UNK	NO		YES

Figure 4.7: Example of the application of the Clinical Data Request Complexity Assessment Tool (CDR-CAT) to a set of three clinical elements contained in a clinical data request. In the first row, we can see that a researcher requested the identification of patients with Chronic Obstructive Lung Disease (COPD) represented in the database as individuals with a specific set of ICD-9 diagnostic codes. In the second row, a researcher requested a specific spirometric result. Despite the forced expiratory volume in 1 second (FEV1) being a numeric result, they are not currently stored inside the clinical data warehouse and thus they are considered non available. The second row describes a researchers' request for patients that stayed more than 48 hours in the ICU. In this case, the clinical element is not represented as such in the database but as an element that can be calculated from the date the patient was admitted to and discharged from the ICU. These are discrete and queryable database fields. Since it is a relative date range, the database query required to calculate it was deemed complex to elaborate but extractable, given the resources available. COPD = chronic obstructive lung disease; FEV1 = forced expiratory volume in 1 second; ICU = intensive care unit; Dx = ICD-9 diagnostic code; Lab = laboratory result; ADT = admission, discharge, transfer; UNK = unknown.

4.3.5.1 Revision of the post-Delphi CDR-CAT

One issue that was evident during the application of the post-Delphi CDR-CAT to this set of clinical data requests was our inability to assess the attribute '*objectivity*'. This attribute assesses whether the clinical element being extracted was objectively or subjectively entered into the clinical database, as it is considered a threat to data quality when subjective criteria are utilized to enter data elements into a database. For example, a database field containing temperature readings (37.5 degrees Celsius) would be considered objective, as opposed to a different temperature assessment (warm) would be considered subjective. When presented with a clinical element to assess, database administrators had no way of judging the presence or absence of this attribute without considering the unique conditions in which the data element was entered into the database. For example, if a clinical data request was seeking for anemia and a diagnostic code for anemia was found, database administrators were not able to determine whether that code was entered by a medical coder as a consequence of a clinician mentioning paleness in a clinical note or whether it was derived from a formal hemoglobin measurement. As a consequence, it was excluded from the final version of the CDR-CAT.

4.3.6 CDR-CAT: Final Version

After retrospectively applying the post-Delphi CDR-CAT to a set of clinical data requests, the item "*objectivity*" was removed from the tool, giving way to the final version of the CDR-CAT. The final version was later applied to a prospectively collected set of clinical data requests and is shown in figure 4.8.

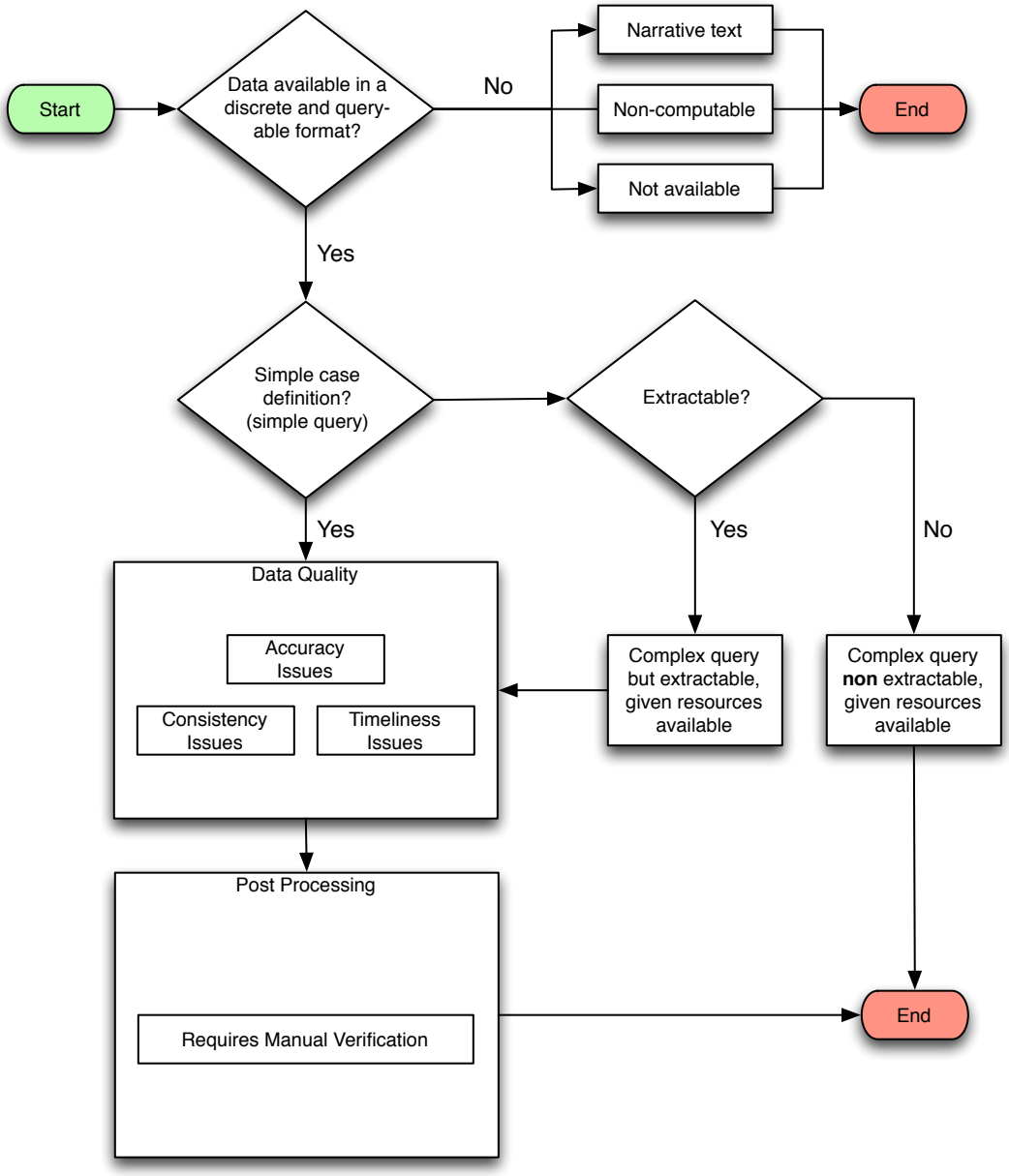


Figure 4.8: Clinical Data Request Complexity Assessment Tool (CDR-CAT): final version.

4.3.7 Analysis of a Prospective Sample of Data Requests

During the final six-month period the ITHS received 9 unique requests for clinical data for research, which included a total of 49 clinical elements. Like in the retrospective analysis, data requests included ranged from very simple to complex ones. The most simple data request consisted of only three clinical elements (obtaining counts pregnant women with positive AntiRh antibodies). The most complex data request sought to identify patients with malignant hyperthermia and consisted of 14 unique clinical elements. These clinical elements involved very simple database queries (i.e. patients receiving Dantrolene) or more complex ones (i.e. retrieve all patients with more than three instances of masseter spasm appearing in clinical notes).

Of all the clinical elements requested, the largest proportion corresponded to diagnostic codes (n = 11, 22%), followed by clinical notes (n = 10, 20%), laboratory (n = 8, 16%) and admission, discharges and transfer data (n = 5, 10%). A full description of the characteristics of all clinical data requests can be found in section 4.3.8.

4.3.8 Summary of the Characteristics of Clinical Data Requests

4.3.8.1 Characteristics of Clinical Data Requests

Clinical data requests varied broadly in their complexity. As mentioned in sections 4.3.5 and 4.3.7, they ranged from simple requests for counting the occurrences of certain diagnostic codes to very complex ones with more than 50 individual clinical elements of interest. The key characteristic of clinical elements that were complex to extract was that they consisted of what I will call **high-level abstractions**. By high-level abstractions I mean clinical elements that are not stored as such inside the database and need to be abstracted from raw data. One example of a high-level concept is a request to identify patients with post-surgical fever. In this case, the concept of post-surgical fever is not stored as such inside the database but there are

several elements that allow us to abstract the concept from raw data: (a surgical date) AND (a series of elevated body temperatures) AFTER (the surgical date). Evidently, extracting these higher-level concepts requires building more complex database queries. Overall, 83 of the 255 clinical elements identified in both the retrospective and prospective analyses (32.5%) were considered high-level abstractions. Some examples of these included: '*acute kidney injury*', '*never treated with antibiotics*', '*Glasgow Coma Scale at admission*', '*Acute Respiratory Distress Syndrome*'. This is one of the reasons why most researchers search for specific database elements and then abstract high-level abstractions through manual chart abstractions.

One characteristic of these searches for higher-level concepts is that they frequently included *relative temporal relations*. Querying for relative temporal relations involve searching for database elements that are temporally related but without a specific date range. For example, a relative temporal query would involve searching a database to find patients who presented fever *during* a course of treatment with the drug phenytoin. In this case, the temporal relation—*during*—is not an absolute date (or date range) but it is relative to the specific timeframe while the patient was taking the drug. In total, 66 of the 255 clinical elements (25.9%) contained some type of relative temporal relation. Examples of these included "Bronchoalveolar lavage within 48 hours of admission", "postoperative need for vasopressors" or "performance status at the moment of diagnosis".

4.3.8.2 Characteristics and Ease of Extraction of the Clinical Elements Requested

The clinical elements most frequently requested were laboratory results, diagnostic codes and medications administered/prescribed. A summary of the frequency can be found in table 4.5.

Type of Element Requested	6-month Retrospective Assessment	6-month Prospective Assessment	Total - N (%)
Laboratory result	39	8	47 (18.4)
Diagnostic code	32	11	43 (16.9)
Clinical Notes	29	10	39 (15.3)
Medications	29	1	30 (11.8)
Admission, discharge, transfers	22	5	27 (10.6)
Multiple	15	2	17 (6.7)
Procedures	8	7	15 (5.9)
Vital Signs	14	1	15 (5.9)
Respiratory therapy	5	1	6 (2.4)
Survival status	5	1	6 (2.4)
Clinical outcome	3	0	3 (1.2)
Image	3	0	3 (1.2)
Tubes	2	0	2 (0.8)
IV fluids	1	0	1 (0.4)
Oxygen administration	1	0	1 (0.4)
Total	206	49	255

Table 4.5: Frequency distribution of the types of clinical elements observed during both the retrospective and prospective assessment clinical data requests submitted to the University of Washington's ITHS. ITHS = Institute of Translational Health Sciences.

The analysis performed showed that 187 (73.3%) of all clinical elements were stored in discrete and queryable database fields. Of those, 149 (79.7%) were extractable using a simple query, reflecting a simple case definition, and 38 (20.3%) using a complex database query. An example of a simple case definition would be a patient's insurance company, in which the database query would be simple to construct. On the other hand, a complex case definition would be the identification of patients with Acute Respiratory Distress Syndrome (ARDS), for which the database might contain all data elements in a structured format but the database query to extract such patients would be rather complex. Of the 38 clinical elements that required complex database queries, only 4 were considered extractable—given the resources available for such purpose—and 28 were considered not extractable.

In spite of a majority of clinical elements available for extraction as discrete and queryable database field, the intrinsic data quality (accuracy and consistency) of 172 (91.98%) of them was unknown. Moreover, researchers still deemed necessary to conduct a manual chart abstraction to confirm the presence or absence of 148 of the 255 clinical elements studied (58%).

Sixty-eight of the 255 clinical elements (26.7%) were not stored in a discrete and queryable database field. Of those, 56 (82.4%) were stored in narrative text, 3 (4.4%) were stored in a non-computable format and 5 (7.8%) were not available in the database. Examples of clinical elements stored in free-text were ventricular function (ejection fraction, stored as free-text within a echocardiography report) and pathology reports. An example of a clinical element that was not available in the data warehouse was pulmonary function (spirometry reports are not included in the database). Finally, all non-computable elements consisted of imaging studies. A graphic description of the distribution of all clinical elements studied and the complexity to extract them can be seen in figure 4.9.

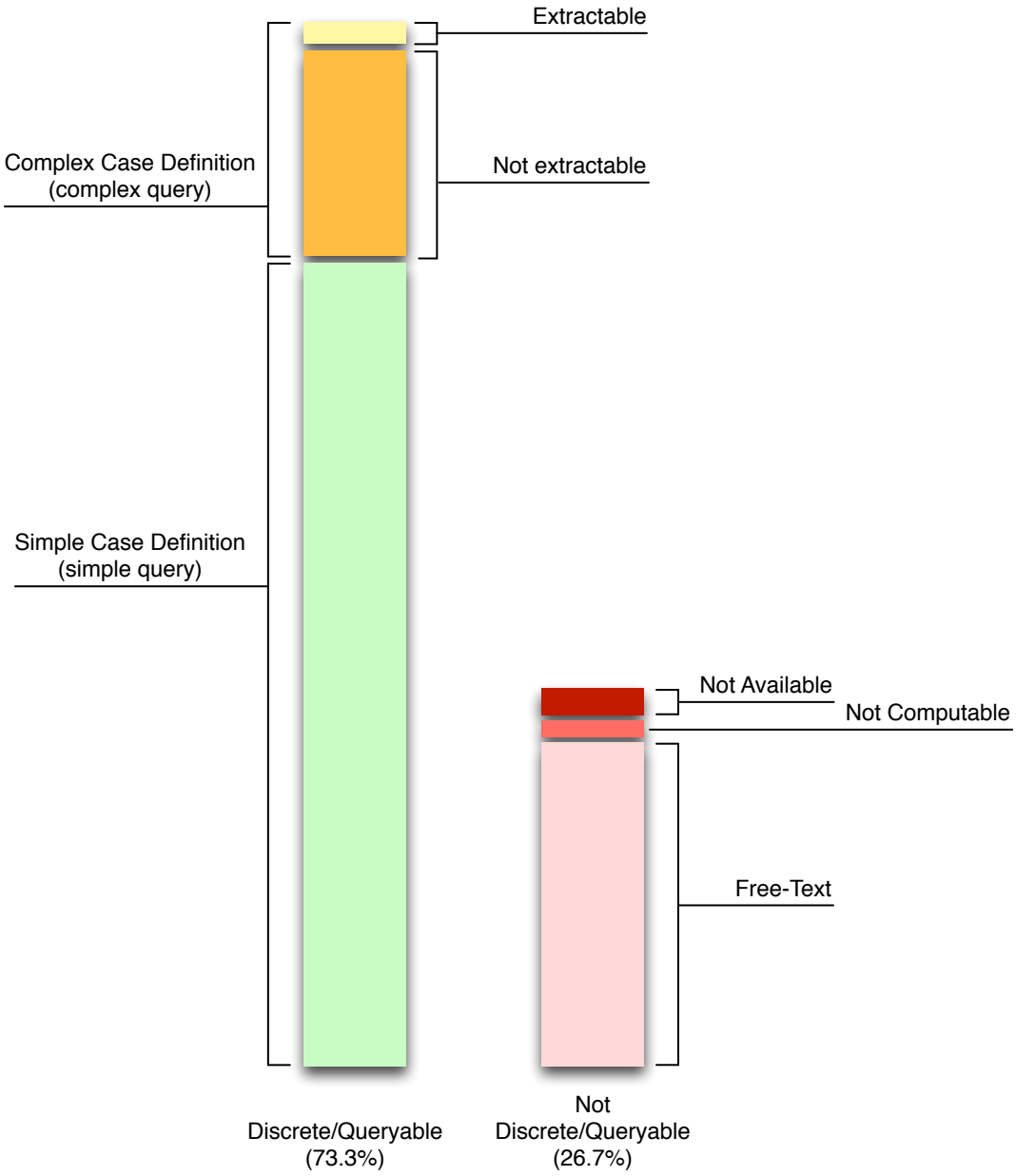


Figure 4.9: Graphical distribution of the 255 clinical elements contained in clinical data requests expressed by University of Washington researchers over a 6-month period and the complexity involved in extracting them from a clinical data warehouse. A vast majority of the clinical elements that are labeled as 'simple case definition' (green), had inadequate or unknown data quality and required verification against manual chart abstractions.

Overall, of the 255 clinical elements studied, only 99 (38.8%) were considered extractable (given the resources available) and researchers felt that they did not need to conduct additional manual verifications on the extracted dataset. Laboratory results, diagnostic codes and medications administered constituted almost 60% of the clinical elements that did not require manual verifications. This means that 61.2% of all clinical elements requested by researchers were either stored in non-queryable/non-computable formats (for example free-text or images), were too complex to extract, the data quality did not allow trusting the results so manual extractions were still needed, or were simply not available, thus limiting the usefulness of this particular CDR and highlighting the amount of work required to improve this situation.

4.4 Summary of Findings and Conclusions

This chapter presented the work done to characterize research clinical data requests and the complexity involved in fulfilling them. The chosen approach included the adoption of a broader understanding of data quality as not just the intrinsic data quality but as a broader construct that needs to consider the needs of the users and the resources available to meet those needs, from an information standpoint. This is what I have referred to as "fitness for use".

Since there was no tool available to systematically assess fitness for use, and this was a domain of significant uncertainty, I chose a Delphi process to elaborate such a tool. Experts in clinical data warehousing from different academic organizations in the USA and the United Kingdom provided their input and knowledge to create this tool. After two rounds, consensus was reached and the Clinical Data Request Complexity Assessment Tool (CDR-CAT) was established.

The most significant finding through the Delphi process was the identification of the idea of a clinical data element being *'extractable, given the resources available'*.

This idea helps identify the—very common—situation in which the data might be available inside a clinical data repository but the resources (time, knowledge, etc.) required to extract it might be insufficient, rendering the information inaccessible.

After the Delphi process, the CDR-CAT applied to a previously collected set of clinical data requests to assess whether it really captured the complexities involved in fulfilling a clinical data request. After this initial application, small modifications were made. The final version of the tool was then applied to a prospectively collected set of clinical data requests.

The results of applying the CDR-CAT to a set of clinical data requests can be divided into two components: (1) characteristics of the clinical data requests, and (2) characteristics and ease of extraction of clinical elements.

Clinical data requests were very heterogeneous in their complexity. Some clinical data requests only included simple extraction of diagnostic codes or procedures. Others included a large number of clinical elements, each of them with various degrees of difficulty to extract. A common characteristic of these complex clinical data requests was that they included what I have called *high-level abstractions*.

These high-level abstractions are clinical elements that are not represented as such inside the database but need to abstracted from a combination of database fields and their relations. For example, a high-level concept requested by a researcher was the concept of acute respiratory distress syndrome (ARDS), which is not stored as such (a database field for ARDS: yes or no), but could be abstracted by a combination of database fields that will classify the patient as having ARDS or not (acute onset, bilateral pulmonary infiltrates on a chest x-ray, no evidence of elevated left atrial pressure and PaO2/FiO2 $\leq$ 200 mmHg [164]). Ten out of the 21 clinical data requests

included at least one clinical element that was considered a high-level abstraction.

One frequent characteristic of these high-level abstractions was the presence of relative temporal relations. Relative temporal relations are defined by a temporal relation that is not absolute relative to time–such as an exact date range—but is relative to the occurrence of another event. For example a request for "all patients with chest pain admitted in 2011" would be an absolute temporal relation between patients with chest pain and time. In contrast, a request for "all patients with chest pain admitted after 48 hours of being discharged" is a relative temporal relation between patients with chest pain and the time when they were previously discharged. These relative temporal relations are complex to identify inside clinical databases and require complex database queries. Overall, 1/4 of all clinical elements requested contained relative temporal relations.

This high frequency of high-level abstractions within clinical data requests is an area that needs to be further explored. The ability to easily identify those high-level abstractions should significantly improve researchers' ability to use clinical databases for research or other secondary purposes.

The analysis of the clinical elements requested provided additional insights into researchers' needs and the ability of a given CDR to meet them. Altogether, almost 40% of all the clinical elements studied were accessible through a database query and did not require manual chart abstraction to verify the results. This number was higher than what we had previously anticipated given our past experience with secondary use of clinical data. In a previous study [Black et al. not published] this proportion was 14%. In a similar analysis conducted on a database being considered for surgical quality improvement, 17% of all clinical elements were accessible through a database query. These differences can be a signal that researchers are self-filtering

before requesting a clinical data extraction and the ones we received were the ones that researchers thought they might have a chance of obtaining.

Whether it is closer to 40% as reported here or 20%, as previously found, this means that a great majority of the requested clinical elements were not accessible (stored in non-queryable/non-computable formats, require complex queries and not enough resources are available to build them), the data quality was unknown or inadequate, or were simply not available. This highlights that a majority of researchers' needs are not being met, meaning that an important proportion of electronic clinical data is not fit for use. This key finding helps answer the question addressed in this chapter: What kind of clinical data do researchers need and is electronic clinical data is fit for that particular use? Furthermore, through these findings allow the identification of areas where progress can be made to improve researchers' ability to use electronic clinical data for secondary purposes, the overarching theme of this dissertation.

Some limitations of this study need to be acknowledged. In first place, the number of experts recruited for the Delphi process was small, which could have limited the diversity of opinions. However, recruiting experts from different settings could partially compensate for this. Second, the analysis of clinical data requests was done in a single academic medical center. This might limit the transferability of the results, but the University of Washington is one of the few academic medical centers that is actively and systematically involved in comparative effectiveness research in the USA so, in case of significant bias, the direction should suggest that in other less experienced sites the situation could be worse. The logical next step is to disseminate this tool and analyze a broader set of organizations involved in secondary use and compare the results. This will allow us to reach a better understanding of the fitness for use of electronic clinical data.

The reasons behind why, with considerable frequency, the available resources do not allow the extraction of some clinical data elements were not explored in this study. In particular, the barriers that lie at the organizational level can play a significant role. These barriers—and facilitators—will be explored in Chapter 5 (Aim 2). In addition to this, Chapter 6 (Aim 3) will describe the development of a temporal data abstraction and query system that may allow overcoming some of the difficulties identified in this study.

Chapter 5

AIM 2: UNDERSTANDING RESEARCHERS' BARRIERS AND FACILITATORS TO SECONDARY USE OF CLINICAL DATA

5.1 Introduction

In chapter 3 I described the known barriers to secondary use of clinical data. However, the published literature does not account for some barriers identified during the preliminary inquiry leading to this dissertation presented in chapter 1. Among them I is the complicated process of working with database programmers to design adequate data extraction processes. As a consequence, to better address this dissertation's overarching question—how can we improve researchers' abilities to use electronic clinical data for clinical and epidemiological research?—this chapter will formally assess the question: **What are the barriers and facilitators faced by researchers when using clinical data for secondary purposes?**

In the previous chapter, I presented a systematic characterization of clinical data requests submitted by University of Washington researchers to a clinical data warehouse. This analysis highlighted that researchers are heterogeneous users and face multiple data and database-related barriers when extracting clinical data for research. Here, I present the methods and results of a qualitative study conducted to identify the barriers and facilitators faced by researchers when using clinical data for secondary purposes and the implications of the findings from an organizational perspective (see figure 5).

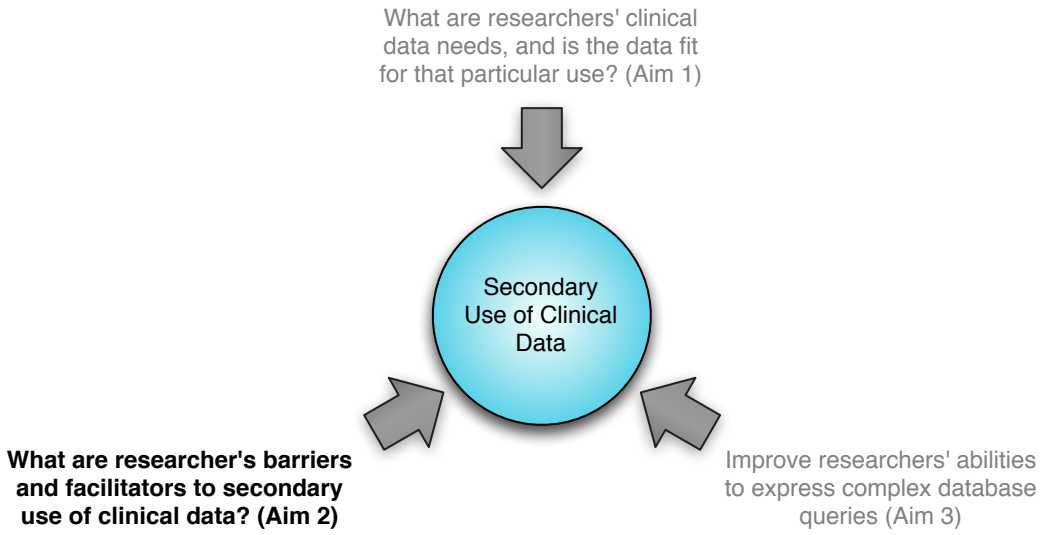


Figure 5.1: Aim 2: Understanding researchers' barriers and facilitators for secondary use of clinical data

5.2 Researchers as a Diverse Set of Users

As described in chapter 2, researchers using clinical data seek to answer a multitude of questions such as [41]:

- How to screen for diseases
- How to prevent, treat, ameliorate, or rehabilitate health problems
- How to predict the course of disease
- How to measure the burden of illness, quality of life and the effects of health services innovations
- How to increase the quality of health care and improve its outcomes
- How to systematically summarize evidence from research

This list illustrates a wide diversity of research questions, spanning multiple dimensions and scales. From studying diseases to studying health care systems—from studying individuals to studying populations. As a consequence of these multiple

dimensions and scales, the characteristics of data requests will be as varied as the research questions they try to answer. This is also true for other disciplines that study the delivery of health care, such as Health Services Research [165].

This diversity is further highlighted by the types of research that can benefit from the secondary use of clinical data. Examples of these kinds of research are (for a more detailed description see chapter 2):

- Clinical Epidemiology - Clinical Trials: this type of research studies can benefit from an expedited patient recruitment process [43] and outcomes assessment [44] using electronic patient data.

- Comparative Effectiveness Research: one possible form of comparative effectiveness research is to utilize large patient databases to identify cohorts and study the relative effectiveness of competing therapeutic or diagnostic strategies [48].

- Translational Research: a cornerstone of translational health is the integration of biological information with clinical and population information. Electronic patient data is a key component of such integration [50].

Adding to what was known in terms of the diversity of users and uses for clinical data in research endeavors, results presented in chapter 4 further advance our understanding of such diversity. Data requests submitted by University of Washington researchers ranged from obtaining simple counts of patients with Type 2 Diabetes for an annual report on the number of visits, to complex data requests seeking identify intensive care unit patients with a diagnosis of ventilator associated pneumonia and the results of invasive procedures performed on them. The knowledge, resources, and institutional support to conduct such data extractions can differ widely as well.

5.3 Diverse Organizations Engaged in Secondary Use of Clinical Data

In addition to the diversity of research questions and the corresponding clinical data requests, the organizations that host researchers are also very diverse. The US health care system, as well as health care systems in many other countries, is a mixture of public and private, academic and non-academic, for profit and non-profit organizations [166]. This means that—from the researcher's point of view—the missions and objectives and the institutional organization to achieve them can also vary broadly.

Most U.S. health care delivery organizations receive fee-for-service reimbursement based on the volume and complexity of services provided or procedures performed. Conversely, health maintenance organizations (HMOs) and other managed care organizations receive pre-payment for providing needed services to a defined population of patients (aka, enrollees, members) for a defined period of time. Payments are typically structured on a per member per month (PMPM) basis. Profitability thus depends on reducing the number of services provided through, at least ideally, keeping the covered population healthy.

Combinations of these two also exist. The reasons to use clinical data for secondary purposes may differ dramatically in these two types of organizations. Fee-for-service organizations might want to identify patients that are due a prostate cancer screening in order to increase the number of prostate cancer-related services provided; managed care organizations may want to identify excessive use of antibiotics for acute otitis in children in order to reduce unnecessary treatments [167]. Furthermore, both kinds of organizations may be interested in studying the quality of care related to surgical interventions [40].

Another layer of variability that resides at this level is the amount of '*organiza-*

tional learning' that takes place within a specific organization. From the organizational theory's perspective, a learning organization is one that is [168]:

> "*. . . skilled at creating, acquiring, and transferring knowledge, and at modifying its behavior to reflect new knowledge and insights.*"

Organizations that are able to learn are in a better position to adopt an innovation. In a systematic review by Greenhalgh et al. [169] the authors found strong and direct evidence suggesting that organizations that are able to learn and absorb new knowledge are more likely to '*assimilate new innovations*'. In this case, secondary use of clinical data is learning process in itself, as described in chapter 2, and is requires the adoption of an innovation, as described in chapter 3. As a consequence, a direct look at the degree of learning is key to understand the degree of engagement on secondary uses of clinical data.

5.4 Barriers Faced by Researchers

This variety of research questions and diversity of organizational settings, suggest that researchers might face multiple and assorted barriers to using EMR data for research. Knowledge about such barriers and facilitators could inform the design of processes and systems to facilitate the use of clinical data for clinical epidemiology and health services research.

As I described in chapter 3, there are multiple known barriers to secondary use of clinical data that researchers face. Among the most studied ones are:

Inadequate data quality: issues such as data inaccuracy, inconsistencies and incomplete data plague current clinical databases. These can have a significant impact on researchers' ability to use the data for secondary purposes [82, 83].

Lack of data standardization and integration: the fragmentation of the health care delivery system generates fragmentation of health care information. Without adequate data integration—which requires adequate data standards—the analysis of electronic clinical data can be severely biased [101].

Patient privacy issues: researchers need to take additional precautions to ensure that patient privacy is protected during secondary use. Some of those protections are perceived as barriers to researchers' effective use clinical data for research [129].

Most of these barriers affect any kind of secondary user, but some of them—especially patient privacy—can particularly affect researchers since the required privacy safeguards are different from other kinds of secondary use [127]. Still, except for patient privacy, data quality, and data fragmentation, these barriers have not been systematically studied from a researchers' perspective. Moreover, reports have addressed this issue without taking into consideration the previously described heterogeneity of uses and organizational circumstances in which researchers are embedded when obtaining and using EMR data for clinical research [137].

In a study conducted by Safran et al. [170] the authors convened a panel of 36 experts to discuss the challenges associated with secondary use of clinical data. This panel did not explicitly include secondary users of clinical data; most of the participants came from the clinical and translational informatics domain. Similarly, Embi et al. conducted a group interview and a follow-up survey with key stakeholders to understand the existing challenges and opportunities in the field of clinical research informatics [171]. Since clinical research informatics includes the secondary use of clinical data, Embi included this item in the topics being investigated. However, the conclusions regarding secondary use are remarkably vague, with the following paragraph presenting the single finding related to it:

> *"Study participants felt that a multitude of organizational, policy-based, and practical factors collectively made the ability to access comprehensive and/or integrative data sets throughout the clinical Research Spectrum, difficult to achieve. This issue was particularly evident in discussions surrounding the secondary use of clinical data in support of research activities."*

Given the paucity of information regarding barriers and facilitators to the secondary use of clinical data from the researchers' perspective, additional research is required. Since further investigation of this issue requires an exploration using a broad perspective, in particular to identify barriers and facilitators at the organizational level, a qualitative approach seems the most appropriate way to advance [172]. In this chapter I will present a qualitative inquiry across three institutions to understand the barriers and facilitators faced by researchers when using clinical data for research.

5.5 Qualitative Approach

Qualitative research methods—also known as naturalistic—developed primarily in anthropology, sociology and other social science disciplines are appropriate for studies that aim to identify and describe complex phenomena. Qualitative methods, unlike quantitative methods, do not seek to answer questions about how much, how many, or in what proportion; they seek to answer questions about what is the nature of this, why does this happen, or what does this phenomenon mean. As described by Pope [173]:

> *"...they do not primarily seek to provide quantified answers to research questions ... The goal of qualitative research is the development of concepts which help us to understand social phenomena in natural (rather than*

> *experimental) settings, giving due emphasis to the meanings, experiences, and views of all the participants".*

Qualitative research includes combinations of tools and strategies to systematically collect, organize, and interpret textual material obtained through observation, interviews and the collection of artifacts [174]. Given the complexity that emerges from the entities that participate in health care systems and their interaction, qualitative research methods are well-suited to study, describe and understand such complexities. Rich descriptions of complex phenomena can be achieved through a qualitative approach.

Frequently, qualitative methods are seen as the opposite of quantitative methods [173]. This results from the idea that the frameworks from which they draw their points of view, constructivism—the idea that there are multiple constructed realities—and positivism—the idea that there is only a single tangible reality– are irreconcilable. However, current views of these methods also pose that qualitative and quantitative methods are different—and potentially complementary—instruments to understand the world. As Paley and Lilford argue [175]:

> *"Qualitative and quantitative methods are only alternative tools, used for different tasks in research (as saws and screwdrivers are alternative tools used for different tasks in carpentry). They have no philosophical implications. They are fit for purpose in a variety of situations".*

That view is the one that underlies this study.

5.5.0.3 Qualitative Research in Health Care

Since the 1990's, qualitative methods have been progressively incorporated into health care research, especially health services research. Although not immediately embraced

[176], qualitative research has gained a significant place in health care research and the number of publications using such methods has been steadily increasing [175], especially to uncover underlying realities and meaning for complex phenomena. For example, in a study conducted by Bloor et al., the authors sought to uncover the reasons that explained geographical variations in adenoidectomy and tonsillectomy in Scotland, which could not explained epidemiologically. The authors were able to identify explaining characteristics such as the amount of relative weight that general practitioners gave to the patient's history versus the findings on the physical examination [177]. Similarly, Mort et al. described the psychosocial impact of the 2001 foot and mouth disease epidemic in the UK and the implications for future disaster management [178]. These examples illustrate the variety of uses for qualitative research methods in health care.

5.5.0.4 Qualitative Research on Health Information Systems

In a similar fashion, qualitative research methods have been progressively incorporated into the set of tools used to understand the context around information systems in health care. Diane Forsythe and Bruce Buchanan argued, back in 1991, that "*For the evaluation of non-technical aspects of system functionality and acceptability, the methods of qualitative social science are more suitable (than quantitative and experimental approaches). Such unobtrusive methods as participant observation and interviewing can provide systematic data on patterns of thought and behavior in natural workplace settings*". This is especially relevant when we consider that a significant proportion of information system implementations fail, not due to technical factors, but due to human or organizational ones [179].

Following that premise, a significant amount of work has been conducted to better understand how people interact with information systems in the context of health care and the consequences of such interactions. Ash et al. carried out a study to

understand errors related to patient care information systems, concluding that these systems might be causing some of the same errors they were designed to prevent, such as errors in information entering and retrieval and errors related to communication and coordination [180]. Weir et al. studied issues around clinical electronic documentation and were able to describe a significant lack of trust in the system, which could undoubtedly have an influence over the effects of implementing such a system [181].

In the specific domain of the interaction of biomedical researchers and information systems, there has been significant work conducted using qualitative methods that have provided deep insights how these entities interact. Weng et al. used ethnographic methods to understand the issues around designing collaborative information systems for clinical researchers [182]. Similarly, Anderson et. al studied the effects on workflow, collaboration and information management of implementing a bioinformatics analysis system [183]. In another study by Anderson et al. the authors explored the needs and barriers faced by researchers when dealing with data management issues [184].

Concordant with the assessments presented here, and given the complexity of organizations involved both in clinical care and research, I propose a qualitative approach to understand and characterize the barriers and facilitators faced by researchers when using routinely collected clinical data for research.

5.6 Methods

Many different qualitative research methods are available to study a problem like the one I propose, however, some characteristics are common to all methods. These include a data collection stage—which occurs in the participants' natural setting—, iteratively entwined with the data analysis stage, during which the researcher organizes and makes sense of the collected data. During this stage, a researcher might decide

to expand the sample of subjects—purposefully rather than randomly—to cover interesting aspects that can emerge during data analysis [185], as well as contract the sample when saturation is reached. This is also the case of the study I am presenting, but for the purpose of clarity, I will present data collection and data analysis sections as separate ones.

5.6.1 Semi structured interviews

Semi-structured interviews were used as the data collection instrument. These interviews consist of open-ended questions to guide the interviewee into the areas that require exploration while, at the same time, giving him or her the freedom to "diverge to pursue an idea or response in more detail" [185].

In this study, I interviewed 12 researchers with experience in using electronic clinical data for research from three different research institutions. Interviews were conducted privately, at the interviewees' location of choice, and were recorded for posterior transcription.

A baseline conceptual framework developed through a literature review and informal preparatory interviews guided the interview questionnaire. For a description of the conceptual framework, see section 5.6.2. Example questions asked included:

- What kind of research do you conduct?
- Please describe the last time you tried to obtain clinical data from an electronic medical record database for your research.
- How did you ensure the protection of patient privacy?
- Please describe the procedures/resources you used to extract the data from the database
- What kind of external (to your lab) support did you receive, in order to extract

the data from the database.

Interviews were transcribed verbatim and were later individually checked for accuracy against the original recordings. Transcriptions were loaded into the qualitative data analysis software Atlas.ti version 6.2 [186].

5.6.2 Preliminary Conceptual Framework

To guide the qualitative inquiry I first developed a conceptual framework based on the literature review presented in chapters 2 and 3. This literature review was conducted through a broad search of the biomedical literature for using relevant search terms ('secondary use of clinical data' 'secondary use* of clinical data' 'comparative effectiveness research' 'translational research' 'personalized medicine') as well as reviewing key publications in the domain and their bibliographies. In addition to the literature review, informal interviews of domain experts were conducted to refine preliminary versions of this conceptual framework. The overview of the framework is depicted in figure 5.2.

This baseline conceptual framework assumes that there are three factors that must be present in order to use EMR data for research purposes:

1. Ability to access electronic medical record data
2. Ability to query a clinical data repository
3. Ability to interpret and analyze the extracted data

In turn, the ability to access clinical data is influenced by the availability of EMR data as well as clinical data from other sources, privacy issues, data ownership and data use agreement issues. The ability to query a clinical database is influenced by the availability of adequate tools, skills, financial resources and a supporting infrastructure. Finally, the ability to interpret the data is influenced by the type of healthcare

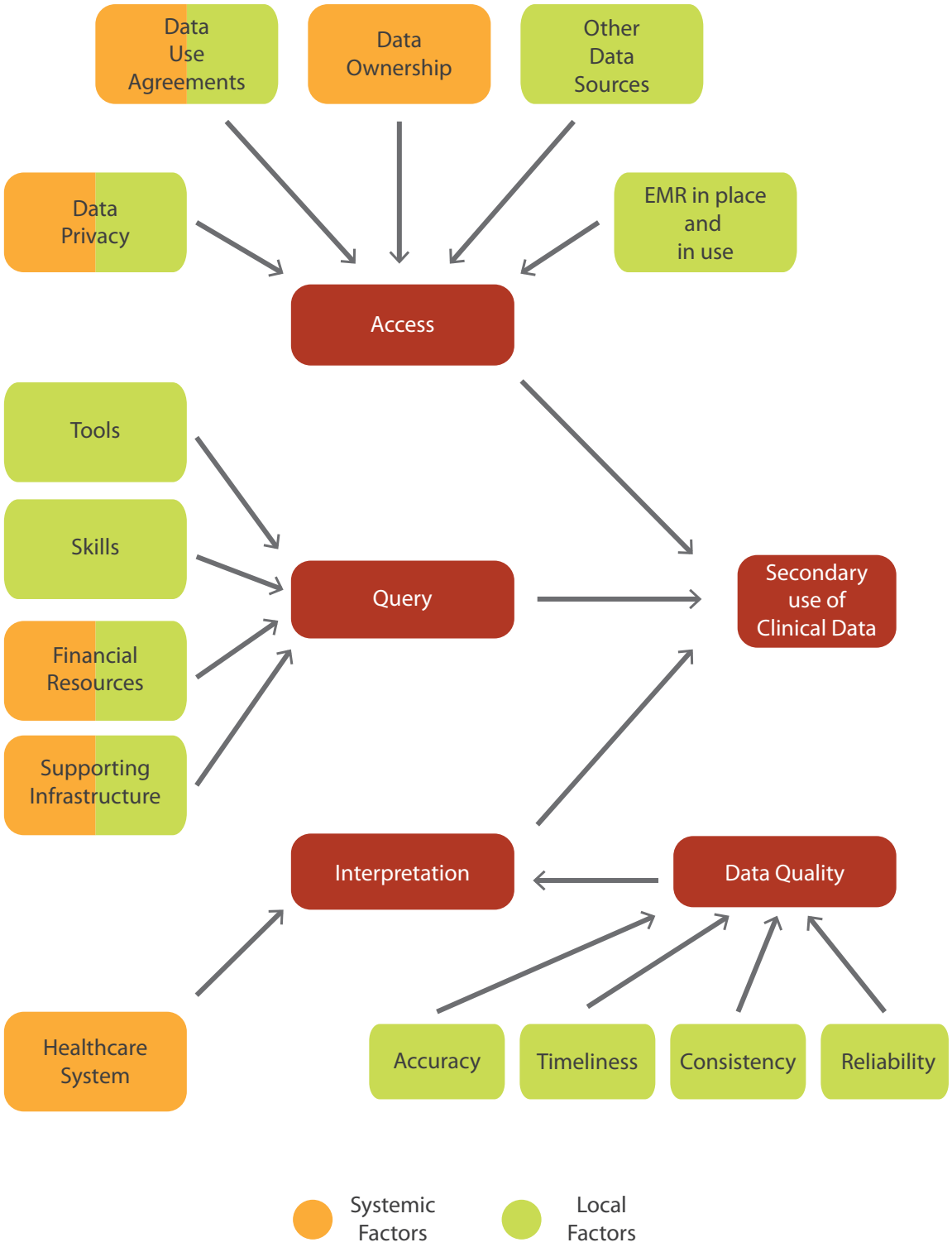


Figure 5.2: Baseline Conceptual Framework

system and the underlying data quality. Furthermore, the influencing factors can be classified into local factors—within the researcher's organization—and systemic factors—outside the researcher's organization—, and factors that are both local and systemic.

5.6.3 Study Sites

Participants were selected from different organizations engaged in the secondary use of clinical data: the University of Washington, Group Health Research Institute and the Veterans Affairs' Northwest Center for Outcomes Research in Older Adults.

The University of Washington (UW) is a public research university with a long history of high-impact biomedical and clinical research. Its health care facilities, grouped under UW Medicine, include the university-owned University of Washington Medical Center and three affiliate hospitals, Harborview Medical Center, Northwest Hospital and Medical Center, and Valley Medical Center. UW Medicine is a combination of public and private not-for-profit institutions that receive reimbursement primarily under fee-for-service model. In addition to these inpatient facilities, the University of Washington's health care network includes several ambulatory clinics. In terms of the clinical information systems, the UW has also a long history of developing clinical data warehouses available for research [187] and has recently embarked in the implementation of an institution-wide clinical data warehouse to support clinical, research and operational information needs. As a research institution, the conduction of basic biomedical and clinical research is a core component of its mission and there are hundreds of researchers and clinicians conduct research at a given time.

The Group Health Research Institute (GHRI) is a research institution affiliated to the Group Health Cooperative, a private, member-owned, not-for-profit health maintenance organization (HMO). This institute conducts clinical and health ser-

vices research to fulfill the mission of its parent organization, the improvement of its beneficiaries' health. Like the UW, the GHRI has a long history of successfully using clinical data for research purposes. It is also part of an HMO research consortium called the HMO Research Network that conducts clinical research using routinely collected clinical data from 19 HMOs across the USA and Israel [188]. The fact that Group Health Cooperative is not a research institution has two immediate implications. First, the institute does not conduct basic biomedical research, and, second, research is conducted exclusively inside GHRI, thus limiting the number of active researchers and having most clinicians not involved in research activities.

The Veterans Affairs' Northwest Center for Outcomes Research in Older Adults is a national Center of Excellence for Health Services Research and Development, a component of the Veterans Affairs Office of Research and Development. Located in Seattle, this center conducts research using clinical data generated across the US during routine care that occurs inside Veterans Affairs health care facilities administered by the Veterans Affairs Healthcare Administration. The Veterans Affairs Healthcare Administration is a public, tax-payer-funded organization that provides health care for US veterans. In this way, it is similar to Group Health in that research is not at the core of its mission but is mean to improve care of its affiliates. Likewise, research is fundamentally performed within the Office of Research and Development and most clinicians are not involved in research.

The three sites were selected since they offered an insight into local, regional and national organizations, with capitated and fee-for-service payment systems, and with different degrees of commitment to research.

5.6.4 Study Participants

To be included, participants had to meet the following inclusion criteria:

- 3 or more years of a successful experience conducting clinical or health services research. Success was defined as having one or more publications in the domain.
- History of use or attempted use of routinely collected clinical data for clinical research.

I conducted 12 semi-structured interviews with researchers in the three institutions included between May 2011 and March 2012. Researchers included spanned different levels of experience with EMR database extracts and seniority inside the organization. Descriptions of the interviewees can be seen on table 5.1. Interviews lasted between 20 and 45 minutes and were digitally recorded for verbatim transcription. Interviews were anonymized for transcription by an external transcription service. Interviews and transcriptions were stored on an encrypted drive using 128-bit AES encryption throughout the duration of the study [189].

The interviews produced a corpus of 158 pages that were analyzed using Atlas.ti. Analysis proceeded as described in the previous section.

Initial access to researchers was gained through existing relations with the Department of Biomedical Informatics and Health Education at the University of Washington. Additional researchers were identified through a snowball sampling approach [190] and contacted through email. Recruitment was continued until saturation was reached.

The University of Washington's Institutional Review Board approved this study.

5.6.5 Data Extraction Consult Process

Although the process of obtaining electronic patient data collected during routine clinical data had some variations within the three institutions studied, a set of shared stages can be described.

	Gender	Institution [a]	Experience (years)
Participant 1	Male	UW	10
Participant 2	Male	UW	4
Participant 3	Female	UW	3
Participant 4	Male	UW	6
Participant 5	Female	GHRI	5
Participant 6	Male	GHRI	>20
Participant 7	Male	GHRI	10
Participant 8	Female	GHRI	7
Participant 9	Male	VA	>20
Participant 10	Male	VA	15
Participant 11	Male	VA	6
Participant 12	Male	VA	6

[a]UW = University of Washington, GHRI = Group Health Research Institute, VA = Veterans Affairs' Northwest Center for Outcomes Research in Older Adults

Table 5.1: Summary of participants interviewed for the study.

- Case definition: during this stage, the researcher defines the type of patients that need to be identified from the clinical database. This includes defining a list of clinical inclusion and exclusion criteria. This set of inclusion and exclusion criteria constitutes a *clinical data request*.

- Definition of corresponding data elements: during this stage the researcher, with or without help of a database programmer, defines the database elements that correspond to the clinical data request established in the previous stage.

- IRB approval: once the data elements are established, the researcher elaborates

an IRB application, which includes the data elements requested and data protection measures, to seek approval to examine identifiable patient information.

- Build database query: in this stage the researcher, usually with the help of a database programmer, identify the database tables that contain the elements established in the previous stage and elaborate a database query that will extract the requested information.

- Test/run database query: once the database query is defined, it is run against the patient database.

- Assess data extract: once the data is extracted, the researcher and its team assess whether the data meets the requirements. If not, the previous stages can be repeated as necessary to obtain the desired dataset.

- Post processing: once the data is available and meets the researcher's requirements, it can be post processed to meet additional needs such as integration with other datasets, specific statistical analysis software definitions, et..

- Data analysis: once data post processing is finalized, the researcher conducts data analysis as planned.

It is important to note that there can be significant overlap between the stages as well as revisions and modifications of previously completed stages to ensure the complete process is adequate. A summary of the process can be seen in figure 5.3.

5.6.6 Data Analysis and Interpretation

Qualitative content analysis in general involves the iterative analysis of textual documents, media or artifacts. This process leads to progressive data reduction and 'distillation' that ultimately leads to the identification of themes and, sometimes,

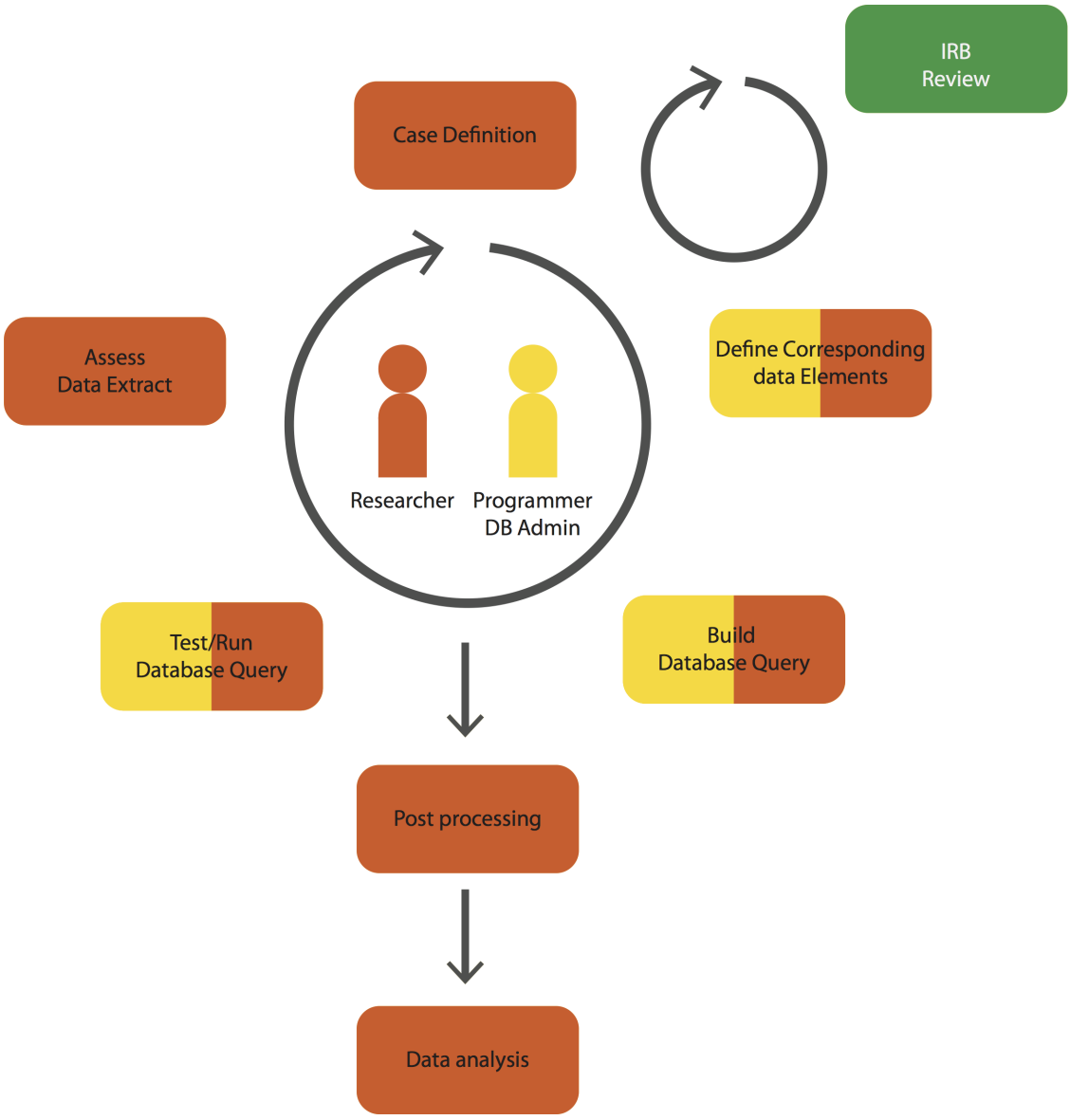


Figure 5.3: Summary diagram describing the stages involved in the extraction of electronic patient data for research in three institutions

theories that describe the phenomenon being studied [citation from Miles and Huberman]. This general process can be conducted in several ways. As Pope describes, there are three broad approaches frequently used in health related qualitative research [185]:

Thematic Analysis is the most frequently encountered type of qualitative data analysis in healthcare. In this case, researchers analyze data by grouping into relevant themes, iteratively reviewing documents to ensure that each theme is thoroughly described. In the words of Given et al. [191]:

> "*Thematic Analysis is a data reduction and analysis strategy by which qualitative data are segmented, categorized, summarized, and reconstructed in a way that captures the important concepts within the data set. Thematic analysis is primarily a descriptive strategy that facilitates the search for patterns of experience within a qualitative data set; the product of a thematic analysis is a description of those patterns and the over-arching design that unites them*"

In addition, the analysis can also include describing how the themes are connected to each other. Thematic analysis usually begins with a set of anticipated themes, obtained through a literature review. However, the researcher must be open to identify unanticipated themes that might emerge from the data.

Glaser and Strauss [192] defined *Grounded Theory* as an "*inductive process of coding incidents in the data and identifying analytical categories as they 'emerge from' the data (developing hypotheses from the 'ground' or research field upwards rather than defining them in advance)*"[185]. One of its main characteristics is that it is cyclical; data analysis informs further data collection. This concept is known as theoretical sampling and consists of the purposeful selection of participants or settings to test emergent concepts or categories. Data is analyzed until saturation is reached using

different coding approaches such as open, axial and selective coding. Through this iterative process, the researcher progressively builds explanations for the observed phenomena, tests the findings and elaborates explanatory theories.

The National Centre for Social Research in the United Kingdom developed the *Framework* approach and is a more deductive approach since it starts with a set of objectives defined in advance, usually in the context of a funding agency requesting a qualitative assessment of a specific issue, frequently related to policy. Data collection and analysis is similar to thematic analysis but it "*tends to be more explicit and more strongly informed by a priori reasoning*" [185].

Since this is an initial exploration of the barriers and facilitators faced by researchers when using clinical data for research purposes, I opted for thematic analysis as the data analysis approach.

After verbatim transcription and quality control, the resulting text documents were analyzed. Initially, the materials were repeatedly reviewed and initial themes and possible categories were captured as memos. A baseline codebook was developed using information from the literature review and the baseline conceptual framework described above. The codebook can be found in Appendix B on page 228. This codebook was used to code all transcriptions for segmentation and easy retrieval in later stages. To increase the trustworthiness of the analysis, two researchers iteratively conducted this initial stage. One researcher has formal training in qualitative research methods (DC) and the other researcher has formal training and significant experience in the conduction of qualitative studies in the domain of health informatics (JT) ; results were continuously compared until sufficient confidence that all significant codes and their manifestations were identified in the dataset.

After this initial stage, all text segments were iteratively and openly coded, following a thematic analysis approach, to reflect the emergent themes, as well as the similarities and contrasts with the baseline conceptual framework. Again, a second researcher reviewed the identified themes to ensure the correspondence with the original interview transcripts. Coding differences were discussed until consensus was reached and open coding was conducted until saturation was reached.

In order to further organize emergent themes, a thematic network analysis approach was undertaken. As described by Attride-Stirling, "Thematic analyses seek to unearth the themes salient in a text at different levels, and thematic networks aim to facilitate the structuring and depiction of these themes". Thematic network analysis seeks to organize emergent themes into web-like maps that include three levels of themes, *basic*, *organizing* and *global* themes. Quoting Attride-Stirling [193]:

> **Basic Theme**: This is the most basic or lowest-order theme that is derived from the textual data. It is like a backing in that it is a statement of belief anchored around a central notion (the warrant) and contributes toward the signification of a super-ordinate theme. Basic Themes are simple premises characteristic of the data, and on their own they say very little about the text or group of texts as a whole. In order for a Basic Theme to make sense beyond its immediate meaning it needs to be read within the context of other Basic Themes. Together, they represent an Organizing Theme.
>
> **Organizing Theme**: This is a middle-order theme that organizes the Basic Themes into clusters of similar issues. They are clusters of signification that summarize the principal assumptions of a group of Basic Themes, so they are more abstract and more revealing of what is going on in the texts. However, their role is also to enhance the meaning and significance

of a broader theme that unites several Organizing Themes. Like Toulmins warrants, they are the principles on which a super-ordinate claim is based. Thus, Organizing Themes simultaneously group the main ideas proposed by several Basic Themes, and dissect the main assumptions underlying a broader theme that is especially significant in the texts as a whole. In this way, a group of Organizing Themes constitutes a Global Theme.

Global Theme: Global Themes are super-ordinate themes that encompass the principal metaphors in the data as a whole. A Global Theme is like a claim in that it is a concluding or final tenet. As such, Global Themes group sets of Organizing Themes that together present an argument, or a position or an assertion about a given issue or reality. They are macro themes that summarize and make sense of clusters of lower-order themes abstracted from and supported by the data. Thus Global Themes tell us what the texts as a whole are about within the context of a given analysis. They are both a summary of the main themes and a revealing interpretation of the texts. Importantly, a set of texts may well yield more than one Global Theme, depending on the complexity of the data and the analytic aims; however, these will be much fewer in number than the Organizing and Basic Themes. Each Global Theme is the core of a thematic network; therefore, an analysis may result in more than one thematic network.

To further support the trustworthiness of findings, they were shared with a two interviewees of different organizations, for verification and member checking, after saturation was achieved.

5.7 Findings

Overall, emergent themes describing barriers and facilitators were grouped into three global themes: **Re-usable knowledge**, **Organizational Structure**, and **Organizational Support and Resources**.

5.7.1 Global Theme: Re-usable Knowledge

The concept of re-usable knowledge was the global theme that emerged with greatest strength during the whole study. Recognizing that secondary use of clinical data requires a large degree of understanding of the data available and the process that needs to be followed to access that data seemed critical for every researcher interviewed. This fundamental role of the knowledge needed to access, extract and use clinical data for research was supported by two organizing themes: process-related knowledge and data-related knowledge. See figure 5.4.

5.7.1.1 Organizing Theme: Process-related Knowledge

Process-related knowledge describes the effect that possessing knowledge on how to access and extract clinical data exerts on the final ability to use clinical data for research. This process-related knowledge was supported by two basic themes, knowledge related to obtaining IRB approval and knowledge about the available resources and data extraction processes.

- Basic Theme: Knowledge related to IRB processes and requirements for obtaining approval for the secondary use of clinical data

During interviews, several researchers described how knowing how to seek IRB approval for reusing clinical data for research, as opposed to other kinds of human subjects research, facilitated the process of accessing data. This knowledge was fre-

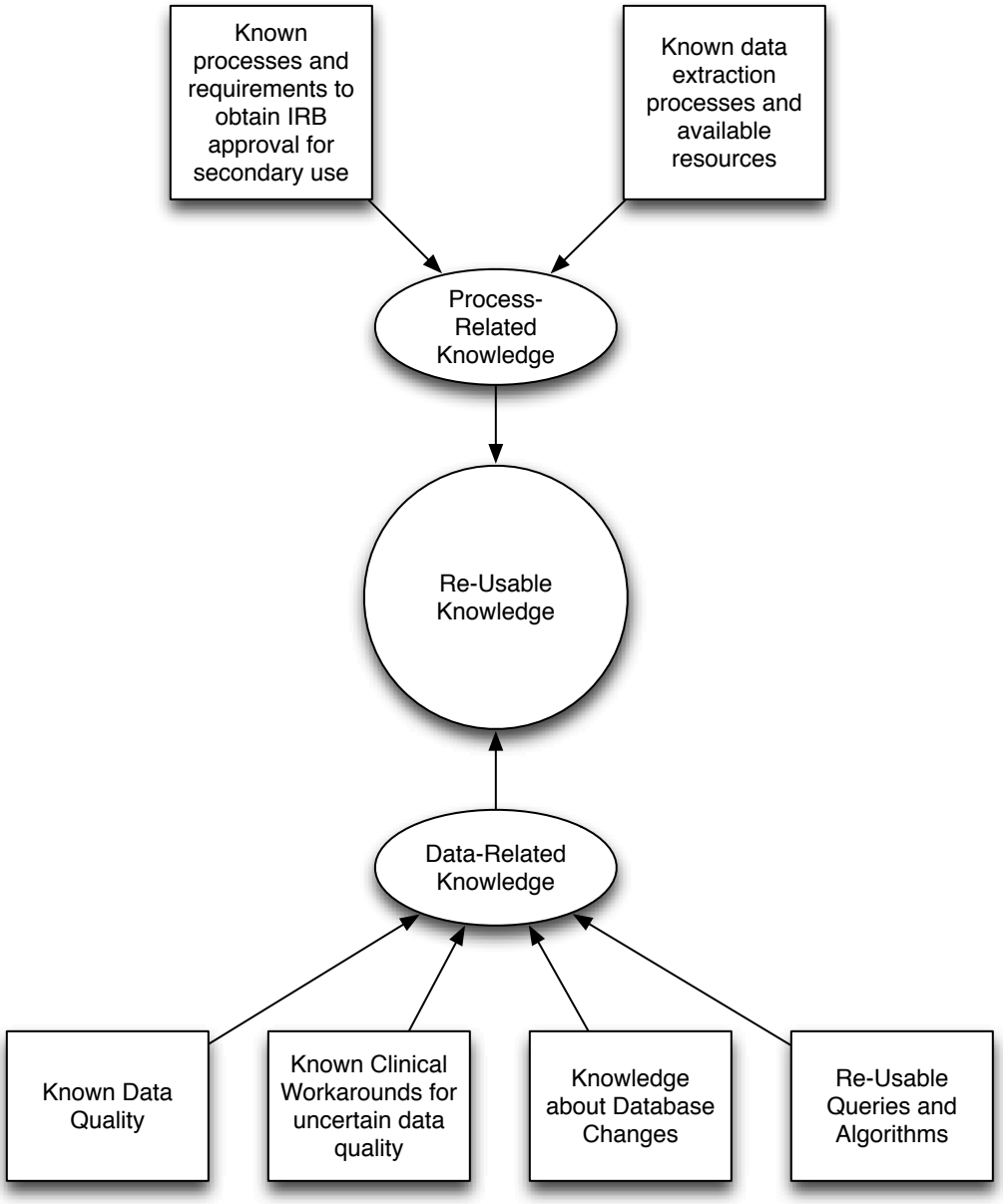


Figure 5.4: Thematic Network: Re-usable knowledge

quently implicit and gained through years of interaction with the local IRBs. For example, in one participant's words:

> *"...part of it is that they [IRB personnel] are, they are in-house and they're just very responsive and the woman who runs it is very savvy and then also...this is not new for us, we have a long history of doing this kind of work [IRB application], and we understand how to do this kind of work."*

Additionally, specific knowledge about the specific IRB data protection requirements and local policies for working with previously collected clinical data was viewed as a significant asset, which could become a barrier when absent:

> *"Like people who work with me, and health services research group at the [organization], has, you know, some boilerplate information to help you. They have experts who can help you put the [IRB] proposals together. I'm loosely affiliated with them. I worked with them a lot in the past and, you know, so they have data security policies that you can refer to but most people, you know who are not affiliated with them probably would be at a significant disadvantage."*

One researcher even described the need for specific 'buzzwords' in the IRB application that could help getting it approved:

> *"I know specific individuals at the [local] IRB. But there's, you know, certain buzzwords that they're looking for in terms of data security, you know, how you're gonna, you know, what kind of data elements you are looking for, what data elements will raise red flags, like HIV status and things like that. And then, locked file cabinets, locked rooms, policies for data security and I think a new investigator would have a hard time with."*

It is also worth noting that researchers did not perceive that requiring IRB approval or data privacy regulations were a barrier by themselves. This is contrary to what was described in chapter 3, where I describe the extensive literature addressing patient data privacy protections as a significant barrier for secondary use. In this case, researchers felt that this was part of the process, that it was a valuable step to ensure patient privacy.

> *"They're very, the IRB administrators and people are very approachable and the want to solve problems. They don't want to put up barriers."*

In addition to knowledge pertaining the IRB application process, researchers described the process of accessing the data itself as a potential barrier.

- Basic Theme: Knowledge related to available resources and data extraction processes.

Although a general process of accessing and extracting clinical data can be perceived—like the one depicted in figure 5.3—, each institution had its own set-up and procedure—either implicit or explicit—that needed to be followed to extract clinical data. Researchers felt that the knowledge describing that process was not always readily available and, when absent or inaccessible, it could become a significant barrier. This rather young investigator describes his struggle with finding the adequate resources.

> *"I think a lot of people still don't know exactly what [local resource] is and what it has to offer. So the first hurdle was that"*

And he continues

> *"We just didn't know how to extract! My programmer was, still is not sophisticated. My team, it was a team created because they're good with people not so much good with programming. And you know... we couldn't do it!"*

5.7.1.2 Organizing Theme: Data-related Knowledge

Even more central than process-related knowledge were data-related knowledge issues. Knowledge about the characteristics of the data, such as how it is collected, its quality, and, in particular, specifics algorithms to access clinical data were seen as factors that influenced their ability to extract data and are the basic themes supporting this organizing theme.

- Basic Theme: Knowledge related to data quality

Starting their research projects without knowing the quality of the data they were dealing with and having to figure it out each time was described as a significant barrier.

> *"We get bad forms. We get from upstairs and they're incomplete where they have more events in them than we are interested in. That becomes quite inefficient."*

Also, researchers seem to spend a significant amount of time understanding the quality of the data they are dealing with:

> *"I mean you really go through with every variable and sort of think about, you know, when we look at frequencies. And we do really detailed, sort of, quality control on the variables and if they don't look like they do well we don't use them."*

> *"We found that using a [Dementia] ICD-9 code mentioned at least, I think, 4 or 5 times and/or the prescription of a dementia drug like Aricept gives us about 80% sensitivity at the best and that's in exchange for about 75% specificity so few want an electronic extraction of those cases this is what you're going to get."*

"So we pull, we look at some positive and some negatives, so people who have liver cancer by ICD-9 codes and people that do have cancer by ICD-9 codes, we then reviewed the charts lined it to that status and try to see if there's any evidence of liver cancer, and you say yes or no, and then you compare the 2 in a 2 x 2 table, essentially and see how accurate the ICD-9 codes are, and we found that, you know, a few people with cancer don't have ICD-9 codes, and a few people with ICD-9 codes don't have cancer. So it works both ways, where they have a different kind of cancer not a primary liver cancer, metastatic cancer and they picked the wrong code."

Having a method for continuously monitoring data quality seemed to help during the grant-writing phase:

"Well, I think, you know, what we do is we try to have in our research grant writing preparation, we tried to take some of the critical variables and drill down on them so we can say, you know, we know that 98% of all prescriptions filled, drugs that people take, are filled at [institution], you know. And so somebody goes out and actually validate that by, in fact we do that about every 2 to 3 years."

Also, knowing the way certain data elements are collected from the beginning helped understand the quality of the data:

"The [institution] is not a fee-for-service system, so everything is not going to be coded like it does in [other institution] where procedures are being billed for. And everything has got to be billed. It doesn't work that way in the VA, so you're always a little concerned about how complete the coding is. Not just for diagnosis but for procedures as well"

Some of that knowledge can be very unique to an individual institution:

You really have to understand quirks of your own system. [Hospital] had all these homegrown codes for labeling different chemo agents and somebody had to sit down and map them to the standard codes so [researcher] has a lot of information about that."

"Well, you know, first of all you have to look at the data sources to see if they meet your purpose and so, and in the [institution] people will know how good it is and what's in it."

One of the resources that researchers used the most to understand the quality of the data they were working with was clinical knowledge. A clinician-researcher describes:

"And is also just as a physician who sees patients there are some conditions where I just know that this is one where doctors have a hard time deciding if is it present or not... and that the coding is wrong because I know how bad the records are that I see when I take care of patients."

Other clinicians reinforced this concept:

"...unless you have sort of the clinical exposure to know how that information got into the record and who was writing it down and in what way were they asking those questions... We know that from earlier experience, when I was an intern, I know how interns wrote notes and I know how I got that information. So that insight helps me understand the validity of the data."

"If you don't have clinical knowledge then, you know, you're dealing with just ICD-9 codes, safety codes, drug names, and you don't, you can't put it together necessarily. Right? There are errors in all of these databases and so a clinician can do a little bit of reality testing."

- Basic Theme: Knowledge related to changes in the databases

In addition to having to deal with uncertain or unknown clinical data, researchers usually had to deal with changes in the databases that they were not aware of.

> *"One challenge is, you know, health care. They have programmed... you know, the EMRs in our data systems are forever changing, right? And it's hard to keep up with it so something that you think worked before, might not work now. Right? And, you know, you don't know because no one has told you."*

Frequently, those changes are only apparent after the fact and researchers need to adjust their queries:

> *"When, you know, you go along and you see for example that 4, 5, 6, 7 months there's a certain number of myocardial infarctions and all of a sudden there's a 50% drop and you say no... that doesn't happen. And so we will drill down and try to figure out what's the difference between May and June or April through May and June, and often times we will find there has been a change in some billing or other kinds of coding factors."*

> *"Sadly, a generic property of EMRs to date is that they're always changing, for everything, from the way that data is collected, you know. One day it's a free text entry, the next day it's a drop-down menu, one day, there is a range limitation, the next day there isn't, so there are some kind of constant issues that you can document and look out for, and then there are other listed issues that can arise so you just have to be on the lookout for it."*

Intrinsic data attributes posed a significant barrier to secondary use, and the knowledge to overcome such barriers appears to have a significant impact in a researcher's ability to use clinical data for secondary purposes. On top of this, knowledge regarding the elaboration of database queries and algorithms appeared to have an ever greater one.

- Basic Theme: Re-usable queries and algorithms

As we described in figure 5.3, there is significant effort involved in building the right database query that is able to pull out the data that researchers are looking for. This frequently entails identifying the list of data elements and relations between those elements that meet the researcher's needs, routinely arriving at a complex and convoluted database query.

> *"Right. So we use an algorithm. We had–and I am not as familiar with the cardiotoxicity part of that study–but we basically had 5 different branches in different combinations of where ICD-9 codes for heart failure and cardiomyopathy could appear in our data. So, it could be, you know, one or... I think if they had one code as an inpatient diagnosis and that was heart failure, or if they had like 2 or more codes for an outpatient diagnoses or something like that, you know, and emergency diagnosis code, and emergency visits plus a combination of codes at another visit..."*

> *"No, it was [an algorithm] specific to the study. We have some general algorithms for things like the [study specific score]. But something like these case-control studies, which were really hard to do with administrative data. We had to write our own routines. And those got complicated, especially in terms of selecting the controls. That really gets mind-boggling actually."*

And even though researchers recognize the fact that some data requests are unique, they feel that there is knowledge that could be re-used by other researchers, which will streamline the process of data extraction.

> *"Gosh. You know, I think there's still too much starting from scratch. So I worked on a number of breast cancer studies in [institution] and every time we start a new study, it's always - you always are redefining your population of breast cancer cases and I wish there was—we started to do this but—I wish there's more development of macros that a programmer can take and knows if that macro works and, you know, maybe you can change some of the parameters like the years of diagnosis or your age or stage of diagnosis."*

The same investigator suggests that storing the knowledge involved in building a successful database query might facilitate them in the future:

> *"But I think for a future study, it's making sure that those [queries] are recorded and documented somewhere, as kind of a 'lessons learned' so that somebody else doesn't have to start from scratch again and figure out all the same limitations that we already have noted."*

> *"So, I would like to think that the programmers ultimately end up with that knowledge and they maintain a data wiki that anybody can access and they try to update it regularly. So that has information about, you know, when certain datasets start, what the limitations might be, what kind of information you can actually get out of them. I think that most programmers would probably say that the data wiki is very complete and up-to-date. I'm not sure that most investigators would agree with that statement. So I think we feel that there are always improvements that can*

be made, that there are always instances where someone has knowledge in their head that's not written down anywhere. And if that person gets hit by a bus, where does it go?"

The metaphor of 'getting hit by a bus', although commonplace, highlights the fragility that some researchers perceived in the way knowledge was managed, in the sense that it resided within individuals rather than within the organization.

5.7.2 Global Theme: Organizational Structure

The second global theme that emerged with strength during interviews was related to how the research organization is structured, how the research teams are conformed, and how the team members—broadly defined—interact. This global theme is supported by the much more specific organizing team labeled stable interacting teams. Furthermore, this organizing theme was strongly supported by themes associated with the manner in which researchers interact with programmers and the local IRB. See figure 5.5

5.7.2.1 Organizing Theme: Stable Interacting Teams

This organizing theme describes a common issue raised by most researchers interviewed. They described that knowing and interacting with the research team, broadly defined, which included database programmers and IRB personnel reviewing their applications was a great facilitator. This organizing theme is supported by: face-to-face communication, long-term relationships with this broader research team, and fluent communications with IRB personnel.

- Basic Theme: Face-to-face iterative communication

The role of having the opportunity to have a face-to-face interaction with database programmers was deeply valued by researchers. They mentioned that this was key to building successful database queries that included all their requirements.

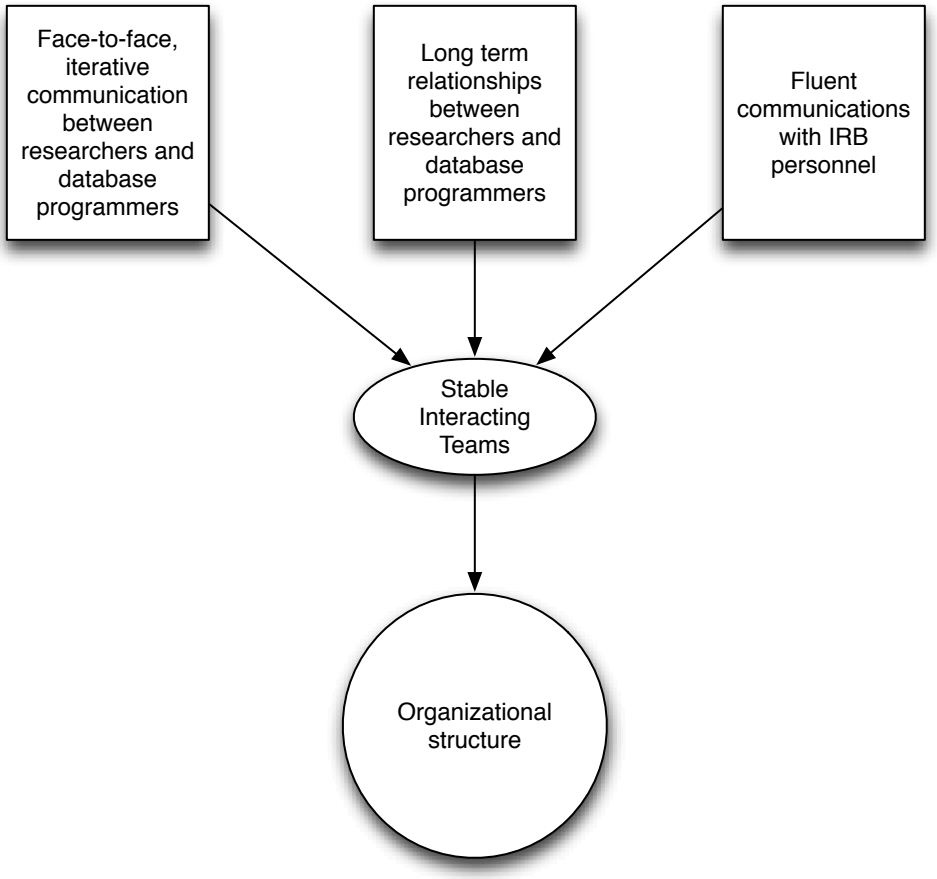


Figure 5.5: Thematic Network: Organizational Structure

> *"The first time was with [database programmer]. So we talked with [database programmer]. We met two or three times in person with [database programmer] and [informatics leadership]. It was at one meeting and those, one of the gentlemen who were there at one meeting as well and we looked over the eligibility criteria. We work together to identify how those were coded within the EMR. It was a very important process; and like I said, I essentially had no idea how things work behind the front end of the computer. And [database programmer] is wonderful. I don't think she knows too much about diabetic kidney disease. So we would have completely different skill sets but just sitting, it really took sitting down in the same room and trying to explain, I'd explain what I was trying to do. [Database programmer] had to explain to me how the system works and what can or can't be done to the system."*

Another researcher also stresses the importance of face-to-face meetings:

> *"You know, I think that having like face-to-face communication and a contact person you can talk with and go to make things a lot easier."*

Some researchers considered database programmers as an integral part of the research team, although they administratively depend from other units. Again, active participation on the team meetings was considered valuable:

> *"So they are a fundamental part of our team, so they come to all the meetings and this is you know, a part in the process and it goes on simultaneously with many other parts of the process so they generally attend the meetings they get a sense of what we're wanting."*

- Basic Theme: Long-term relationships between researchers and database programmers

In addition to face-to-face interactions, study participants described the importance of having long term relationships with database programmers. Without such relationships, researchers felt that it was complicated to achieve meaningful results:

> *"Yeah, I mean, I can't imagine how people get anything done if they don't have a close relationship with this skilled programmer. I know there are people at [other organization] who, I guess, have very strong programming skills themselves and do some extraction. But to me it's more efficient to have a programmer who really knows the [local database] very well and who works with me everyday and knows how I'd like to do things."*

This long term relationship was seen as an opportunity to bi-directionally transfer clinical and database knowledge that will in time, improve their ability to extract clinical data for research:

> *"So the programmer needs to have some sense of what data is available and needs to educate the clinician. And the clinician needs to educate the programmer."*

There were several instances in which the organizations in which researchers worked did not provide the organizational structure to foster these long-term relationships. In these cases, researchers employed a series of strategies to force such structures. One of those strategies was to establish 'connections' with database programmers:

> *"That is essential!! Right, so without sort of a champion for your projects, within informatics, sort of that database world, I don't know that you can get stuff done. But someone like that makes it possible.*
>
> *I know there's other folk's similar situations without that sort of connection and that is felt like there are some barriers to finding the right person to answer these questions."*

Another researcher, describing the situation in which a programmer was co-located with his research team but not administratively part of it, further illustrates this phenomenon:

> *"I've used it [organization's programming resources] once to get some information and worked just fine, mainly because one of the programmers, one of the main guys, is actually in our office. He's a central office employee but he's actually working out here in Seattle. So he helps us out a lot with that."*

A researcher describes hiring a database programmer as part of his team to maintain this long-term relationship despite the fact that the programmer was now providing the same services as a consultant within the organization.

> *"[Database programmer] worked for me first, actually. I take some credit for where [database programmer] is now because [database programmer's] first exposure to biomedical informatics was working for me. His initial exposure was working with the [local] database... So he worked for me for about 3 years, 4 years, and then, took the job with [institution's IT services]... So we managed to retain some of his talent."*

He further stresses the importance of having the database programmer inside his research team:

> *"Well, without him [database programmer], we would get nothing. I think its the short answer right now, I think, you know, part of the shift to [new data warehouse] has created more bottleneck to data collection for research, and that bottleneck is both an issue with access and an issue of understanding I think [database programmer] has both, and so, you know, I think hens been able to facilitate it that way."*

- Basic Theme: Fluent communications with IRB personnel

Given the significance of patient privacy regulations in this domain, researchers also highlighted the concept that having a fluent interaction with the local IRBs was a significant facilitator for accessing clinical data for research.

> *"It depends on you and the project, but a lot of times what I will do is write something up and, if I have any questions, I can e-mail them ahead of time and, a lot of times I might send it to them ahead of time and say: can you look at this application and see if there's any problem? Can you see what might come up? And so, then, they will pre-screen it and tell me, you know, what they think needs more detail, and what might raise any flags, and then, by the time I put it in, it's pretty much ready to review and, you know, it's it usually won't face any hurdles."*

> *"I think probably because we were close with them and are just down the hall, so I can, we can easily talk with them. They're very... the IRB administrators are people... are very approachable and the want to solve problems. They don't want to put up barriers."*

5.7.3 Global Theme: Organizational Support and Resources

Complementing the organizational structure around research teams engaged in secondary use of clinical data is the global theme labeled organizational support and resources. This theme encompasses the resources that the institution has and provides researchers to be able to access and use electronic clinical data for research. This global theme is supported by two organizing themes: data resources and tools. See figure 5.6

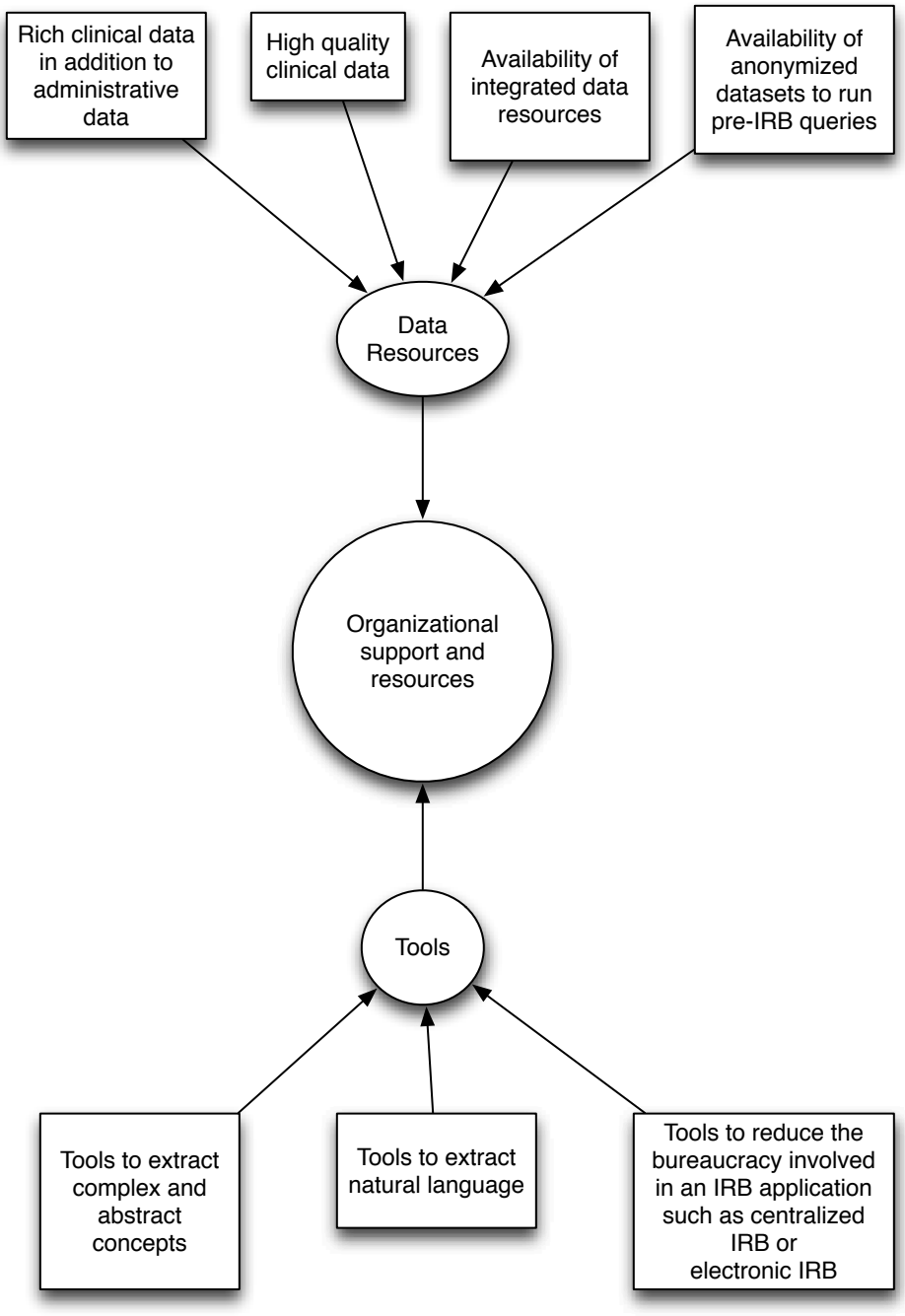


Figure 5.6: Thematic Network: Organizational Support

5.7.3.1 Organizing Theme: Data Resources

The amount of data generated during routine clinical care, and the different formats it is generated in, set up a complex scenario in which researchers need access to a broad variety of data resources in order to successfully use electronic clinical data for research. The basic themes supporting this organizing theme are: availability of rich clinical data sets, high-quality clinical data, integrated data resources and anonymized data sets.

- Basic Theme: Availability of rich clinical data sets

Clinical care can generate data that can be classified into two broad categories: administrative data and clinical data. The first type encompasses data-related to visit dates, payments, reimbursements, prescription fillings, and so on, and it is mainly used for managing the health care delivery organizations like any other enterprise. On the other hand, clinical data encompasses data that describes the clinical conditions suffered by patients, diagnoses, treatments and procedures, clinical assessments, etc.

Although there can be significant overlap—visits are frequently associated with diagnostic codes that describe the reasons for the visit—administrative data tends have a limited ability to provide a rich description of clinical conditions or physiological states, when compared to clinical data. However, a critical problem with clinical data is that a significant portion of the information about a patient is contained in the form of free text, which is not as accessible, from a computation point of view, as structured data. This contrast between administrative data and clinical data should help us understand the barriers and facilitators associated with this theme.

As one researcher explains it, having access to richer clinical data allowed him to find more complex patterns in the data:

"One thing about the SQL database, actually called [database name], they're actually importing vitals, other measures that weren't previously available in the [previous database]. So, things like measure of vitals, like blood pressure and heart rate. Well, maybe down the road, they may have better ways of measuring patient comorbidities. Things that might become available that are not available right now are like audits for determining alcohol consumption. Possible measures of depression, accepted valid measures of these kinds of behaviors that would be directly input into the database rather than us having to search all over the place for them. Or even compute them..."

Another researcher mentions the same issue when describing the reason to tap into physician notes:

"So mainly physician notes and we're trying to—this project—we're trying to look at issues around diabetes and the thing that we're trying to understand are more characteristics of the diabetes that isn't necessarily coded very well in the records like type of diabetes, like the time they've been on insulin, all of those types of things. So that's the project that we're working on now, that the structured is data is not rich enough."

The problem arises when researcher realize that most of the relevant clinical data is stored in the form of free text and need to resort to labor intensive methods of data extraction, such as manual chart abstractions. This arises as a significant barrier for secondary use:

"Administrative data is one thing. Trying to get deep, rich, clinical data is a whole different complex task. So that's why he was doing his manual review. And then the DEXA results—some lab results you could try to

get programming the computer to do it. But I envision that some DEXA results are in a tabular form somewhere, and text form, and if you're trying to... then there is many different results. So there's that. Result at the femur, the results of the spine, and so one would require... it would take a lot of work to get a computer to try to tease that out and it's just easier to do it by brute force."

The amount of work involved in manual abstraction can be quite large. The same researcher describes:

*"**Interviewee**: So it's a data warehouse that includes more clinical data than just administrative data. It has labs, x-ray results, things like that. And so we use [the local database] data and identified predictors of anyone with DEXA scan. Now, DEXA results were not in [the local database] but we know if the DEXA was done, then we had to go to the chart to look at the actual DEXA results in the radiology files.*

***Interviewer**: And when you wanted to confirm whether the patient had IBD, was that a manual process? like someone reading that chart?*

***Interviewee**: That was a manual process.*

***Interviewer**: Okay. How many charts?*

***Interviewee**: I think he looked at about 2,000 charts."*

- Basic Theme: High-quality clinical data

Complementing the availability of rich clinical datasets is the need for high-quality clinical data. Inadequate data quality was frequently mentioned as a barrier for secondary use. Furthermore, and using a broader definition of data quality such as the ability to provide what the user needs, researchers frequently find that the data is

inadequate to identify complex or abstract concepts. In this quotation the researcher describes the influence of data quality in extracting a high level concept such as health failure:

> *"...that sets one level of quality...it is just missing? Or is it obviously garbage. Of the 2nd level of quality is can the [local database] actually measure what you think? So you want to measure congestive heart failure, a notorious example where we know that the administrative codes for it that doctors assign during visits are bad. They have, you know, like maybe 50% of positive protective value or less, and that's just because doctors or coders don't do a very good job of writing down the code for a disease that it is a little hard to come up with a very clear diagnostic criteria for anyway."*

'Readiness to change' as an example of another high level concept that is not adequately represented in the form of structured clinical data:

> *"Another one is, you know, patient's readiness to change. You know, those types of things that are in the text field often but aren't in the structured data."*

And when the data is not useful to extract high-level concepts, researchers usually rely on manual chart abstraction:

> *"I mean everybody does it differently. But I think it grows out of the question that you're asking. So one choice is to ask yourself: is this, you know, is this a question that can either be answered well with our [local database's], administrative data? and some questions can...or is this the question where we need to do chart review?"*

Some research groups have recognized that they cannot get away from doing manual abstraction and have set up systems to efficiently do so by training chart abstractors. However, the task is not trivial:

> *"It takes time. Our chart abstractors are well trained. We don't go through a lot of them so they stick with us for a long time and we've estimated it takes a good 6 months to a year to get somebody trained up on doing these... They basically have to memorize it for the most part."*

- Basic Theme: Integrated data sources

Another organizational resource that was consistently described as a facilitator for secondary use of clinical data was the availability of integrated data sources. One researcher clearly states:

> *"So at [institution], you know, I don't know how long we've had it, we've seen a couple of [clinical data warehouses], which is really a bunch of really wonderful electronic data repositories. One thing that has all the billing codes, so it has to ICD-9 codes assigned by doctors or the billers for the visits, and then it also has the laboratory results, and it has prescriptions that were filled, and has demographics and all that stuff."*

Conversely, when not available, integrating data sources could become a significant hurdle:

> *"There is an inpatient database, outpatient database, pharmacy database, there's a fee database for when the [institution] contracts outside the [institution], you know. And so, if you're trying to find everybody who had a colonoscopy, for instance, if you just look in the outpatient files you miss all the inpatients. Do the outpatient and inpatient, you miss the people who the [institution] paid for private practitioners who did the colonoscopy."*

- Basic Theme: Anonymized Data Sets

The fourth theme describing data-related organizational resources required by researchers were anonymized data sets. Despite the fact that researchers did not consider privacy protections as significant barriers for secondary use, having access to anonymized data sets was considered a valuable asset, especially when preparing grant proposals.

One researcher describes her experience using an anonymized database to obtain patient counts:

> *"I can just open up the data counter on a desktop. I don't need a IRB approval I can tell it the ICD-9 codes am interested in, I can tell it some drugs I'm interested in, and it will count for me how many prescriptions were filled, and inpatients and outpatients, and men and women, and different age groups in different years for the drug, how often were that those diagnosis codes assigned."*

And she further highlights the idea:

> *"I can just cross tab a woman of a certain age, to a certain year, who has these diagnoses, what drugs did they, you know, did they use this drug? So, I mean, that's fabulous and it takes—I haven't used it that much—I think it takes about 5 or 10 min., it's very easy!"*

Another researcher who did not have access to anonymized data sets highlights the potential use for such a resource:

> *"I think those things will be helpful for, sort of, the hypothesis generating sort of issues. So, for example, when we look at clinical data through [the*

local database], we could not be doing it without the permission and with the right human subjects approval. But often you don't know whether, you know, if it's even worth while to do the study unless you can look at the sample size or, you know, that there were 200 patients that that had CHF that actually came in last year and so that there is enough people here to even study."

5.7.3.2 Organizing Theme: Tools

A second set of basic themes related to resources available within an organization are the ones supporting the organizing theme labeled tools. Since some of these tools could be either available locally, inside the researcher's team, or for the entire organization, we considered the concept of the organization broadly, including both. However, as we will discuss, some tools necessarily need to be provided as an organization-wide resource.

- Basic Theme: Tools to extract complex and abstract clinical concepts

As previously identified (see section 4.3.8 on page 89), researchers are frequently interested in extracting abstract concepts that sometimes are not represented as such inside a clinical database; or at least not accurately enough. The complexity of doing so and the unavailability of tools to do so is a recurring theme in this study.

One researcher describes his unsuccessful experience trying to identify patients with Dementia in one data set:

"We found that using a [Dementia] ICD-9 code mentioned at least, I think, 4 or 5 times and/or the prescription of a dementia drug like Aricept gives us about 80% sensitivity at the best and that's in exchange for about 75% specificity. So if you want an electronic extraction of those cases, this is what you're going to get."

Another researcher describes the task of finding temporal relations in a dataset. In this case, he and his research group ended up doing a manual chart abstraction to find the temporal relations:

> *"**Interviewee**: You could definitely go in and identify the cancer. Is it esophageal cancer? You can identify your cases and my guess is they would have probably more complete information about procedures in the history, and maybe when they were originally diagnosed and that information. Whereas the ICD-9 code, you have to go back to the farthest you can go back, and say, here's the first example of esophageal cancer, the first incidence, so maybe we will call that the start date.*
>
> ***Interviewer**: Do you do that manually? Go back and look at the 1st incidence of ICD-9 code?*
>
> ***Interviewee**: Yes, you essentially go back and you have to set up say this study is in 2000, I'll go back 15 years to check the diagnosis, so we'll go back to 1985. We just go through all the records and you just search for that code for esophageal cancer."*

As this researcher did—similarly to what happens with low data quality—it is frequent that they end up resorting to time consuming manual data extractions to deal with complex case definitions that are not accurately represented as such in the database.

- Basic Theme: Tools to extract concepts embedded in free-text

Researchers readily recognize that large proportion of the clinical information describing a patient lies inside free text and is not easily accessible through database queries. As described in the previous section, most use manual chart abstractions to deal with free text. However, some are beginning to tap into this source of clinical data using natural language processing (NLP) tools.

"A lot of people are getting involved in... I guess I would call it natural language processing. So we have a project that we've been working on for a year. We are trying to teach the computer to make sense of radiology reports from chest x-rays. And so, we use a system called [proper name], we fit in the text of their radiology reports and that classifies all the words and it extracts concepts and then we roll that out to make variable definitions."

Another researcher recalls using NLP tools but also recognizes its limitations when the text they are dealing with is less structured:

"We tried some natural language processing. The problem with pathology reports is that the language that's used isn't necessarily standard. Some of the pieces are, so... some of the results like, you know, estrogen receptor positive, or negative, the can be standardized.

Where we have been more successful with it is actually with the radiology reports. Would did a really small pilot for just getting some data for a grant where we wanted to get findings from radiology reports that were different... It turns out that text is pretty standard. I mean, if you search on calcifications, you will find calcifications, and you will not not find calcifications."

- Basic Theme: Bureaucracy Related to IRB applications

Although researchers overwhelmingly recognize IRB and HIPAA regulations as something necessary, they did highlight the perception that some procedures and formalities imposed locally, within the organization, exceeded legal requirements and constituted an additional burden when using clinical data for research. This was especially relevant when interviewing researchers working at the Veterans' Administration.

One VA researcher describes:

"Yeah. You would search the Medicare data files to see if they're good. If... you know, the biggest problems for the last 3 or 4 years are bureaucratic ones, were we haven't been able to get the Medicare data... but it has to do a lot... 5 or 6 years ago, with the data... potential data loss, and the VA became like, I think, you know, tremendously preoccupied with issues of privacy and security."

He continues:

"And so, one of the things they did was block off essentially the system for 3 or 4 years. Kind of, sort of, as they got on top of security. I hope they now start to ease a bit. One of the big losses was Medicare med profile, decision file... and so CMS and the VA couldn't use those data any longer, so it was blocked."

A second VA researcher describes the situation as regulatory entities not understanding the work they were doing:

"Even issues about accessing data and privacy rules. They think of everything in terms of kind of clinical work and when fact we're doing research work. Maybe it's just our skewed perspective that were actually different from clinical work but I think research is different from clinical work."

A second barrier described by many researchers was the need to deal with multiple IRBs when using data from multiple organizations.

"We went out and got birth certificate data from Washington State to link to the group health data and it's been a mess because we had to get approval from the Washington State Department of Health IRB and we also have to get approval from the Group Health IRB, and we had to tell each of them

what part they owned and could be responsible for, so we didn't want the Department of Health telling us what we could do with her birth certificate data Sweet had to separate IRB approvals for that and that's pretty, it's annoying."

On the other hand, researchers were able to describe the tools that help overcome such barriers. One of them was the availability of electronic IRB applications:

"Oh I think we are much faster than anywhere else that's kind of, see, part of it is that we are using electronic IRB now."

Or to have agreements with organizations to mutually recognize IRB applications:

"And in fact what we've with done with the [research network] is developed a process for data-only studies, where if it's just pulling data out of electronic files an you're going to 5 different members of the[research network], one side will see the review to the other four."

5.8 Summary of Findings

This study was able to bring up several issues that can facilitate or impede researchers' ability to use clinical data for their investigations. Three global themes, namely *re-usable knowledge*, *organizational structure* and *organizational support* emerged as themes during the interviews and the subsequent data analysis. Furthermore, these global themes are supported by organizing and basic themes. A table summarizing the basic, organizing and global themes can be seen in table 5.2.

Global Theme	Organizing Theme	Basic Theme
Re-usable knowledge	**Process-related knowledge**	•Knowledge related to IRB processes and requirements for obtaining approval for the secondary use of clinical data •Knowledge related to available resources and data extraction processes
	Data-related knowledge	•Knowledge related to data quality •Knowledge related to changes in the databases •Re-usable queries and algorithms
Organizational structure	**Stable and interactive teams**	•Face-to-face iterative communication •Long-term relationships between researchers and database programmers •Fluent communications with IRB personnel
Organizational support	**Data resources**	•Availability of rich clinical data sets •High-quality clinical data •Integrated data sources •Anonymized Data Sets
	Tools	•Tools to extract complex and abstract clinical concepts •Tools to extract concepts embedded in free-text •Bureaucracy Related to IRB applications

continued on next page

continued from previous page

Global Theme	Organizing Theme	Basic Theme

Table 5.2: Summary of identified global themes, organizing themes and basic themes

5.9 Discussion

To use electronic clinical data for research, researchers need to interact with various entities and people within and outside their organizations, and need to apply a diverse set of skills and tools. Most of the barriers and facilitators identified were related to those interactions and tools, as well as the knowledge required to navigate through the complex network that emerges from them.

Of the three global themes discussed, *re-usable knowledge* was the theme that emerged with the greatest strength. Researchers acknowledge that there is a significant amount of study-specific knowledge that goes into successfully extracting electronic clinical data for research. However, they also recognize that there is significant generalizable knowledge accumulated, within their organizations, on how to perform that task. This knowledge spanned the whole process of extracting clinical data, including knowledge to streamline the IRB application process, re-use previously successful database queries and knowledge about data quality.

The idea of re-usable knowledge is a core component of the knowledge management literature, that states that an organization's knowledge and its ability to store and maintain such knowledge is a key element for organizational success [194]. Research does not seem to be an exception and, as described in section 5.3, it is likely that research organizations with higher levels of organizational learning might be in advantage.

This need for re-usable knowledge is further supported by previous research that showed similar results in a different setting. In a study by Myneni and Patel, the researchers also encountered several knowledge management issues, for example they found that knowledge frequently resided within an individual and not within the organization, thus limiting knowledge transferability [195]. Further support for this finding comes from the activities currently happening within the eMERGE network, in which they are developing a knowledge base of shareable, effective, database queries or algorithms to extract patients with certain clinical conditions [196]. There has been limited research around the issue of knowledge management within the research enterprise, and—to our knowledge—this is the first report on knowledge management barriers as a limiting factor for researchers' secondary use of clinical data, and should be added to the set of barriers presented in chapter 3.

The second global theme was the organizational structure to support secondary use of clinical data. It is important to note that this theme was almost exclusively focused on the composition and functioning of the team of people involved in secondary use. This broad definition of the team includes not only the researchers themselves but also the database programmers and the IRB personnel. Researchers deeply valued the possibility of establishing long term-relations with them, and the possibility of interacting personally and fluently with them. This organizational structure is well aligned with the re-usable knowledge theme. It is well known from other industries, including health care, that successful teams are the ones able to communicate well, share knowledge, and learn from each other [197].

The third global theme was the organizational support. Two aspects of this were perceived. The first one was related to data resources. Researchers valued having access not only to administrative data generated during routine clinical care but also to high-quality rich clinical data that would better describe a patient's condition.

In addition to this, being able to work with databases that integrated internal and external data sources were regarded highly, as well as the possibility of—at least during early stages of research projects—accessing anonymized datasets to gain a preliminary idea of potential sample sizes. These findings are concordant with data-related barriers presented in section 3.3.1 of chapter 3.

These data resources, however, cannot be adequately exploited without adequate tools. The most remarkable issue related to tools was researchers' almost universal need to extract high-level concepts from clinical databases. These high-level concepts frequently result from an abstraction of many elementary and interrelated data elements contained in a database, as well as embedded in poorly structured free text. As a consequence, and not always explicitly, researchers frequently described the need for tools to create these complex queries more easily. On top of that, they constantly mentioned the need for tools to extract data from the free text inside medical records. When these tools were not available, as was often the case, researchers resorted to manual data extraction, a costly and time-consuming process. These findings are consistent with what has been reported in the literature about the need for natural language processing tools and manual data extractions. They also fully support the results presented in chapter 4 when I discussed the complexity of extracting clinical data. Similarly, in a study conducted by Anderson et al., a majority of the researchers interviewed expressed that the *"limited availability of institutionally provided expertise and systems"* constituted a significant barrier for data management tasks [184].

Besides these, tools to streamline the IRB process were deemed as necessary. It is important to call the attention to the fact that researchers did not mention IRB or HIPAA regulations as barriers for secondary use. This finding contrasts the ideas frequently discussed in the literature and presented in section 3.4.2. On the contrary, most researchers described these regulations as necessary and valuable. What they

did raise as a barrier was the fact that some organizations had regulations that went beyond IRB or HIPAA, or they implemented burdensome bureaucracy that delayed the process of obtaining approval.

5.9.1 Implications

The findings presented here have numerous implications for organizations currently engaged or planning to engage in re-using clinical data for research. Using a commonly employed framework, for clarity these can be organized into four aspects: Process, People, Data and Tools. These four aspects need to be present in order to adequately foster secondary use of clinical knowledge. Here I present a list of the most significant implications that emerge from these findings. See figure 5.7.

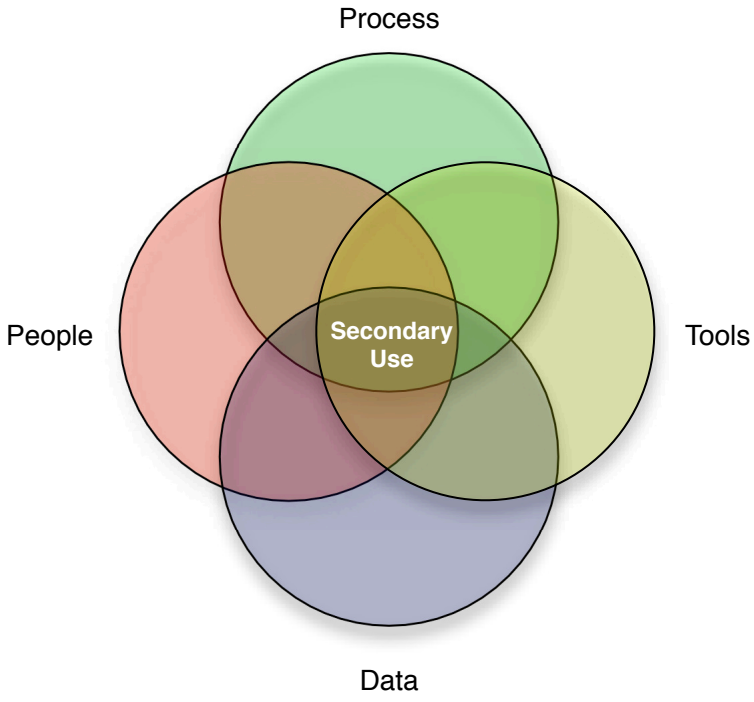


Figure 5.7: Adequate interaction between processes, people, data and tools enable secondary use of clinical data

5.9.1.1 People

The global theme **re-usable knowledge** is the one that has the greatest impact on recommendations involving human resources that we have labeled '*people*' implications. During this investigation it was easily apparent that knowledge currently resides within individual people, and organizations need to engage in efforts aiming at institutionalizing and disseminating such knowledge.

- Establish inter/transdisciplinary research teams able to bridge research, IT, and regulatory domains.
- Formalize and disseminate knowledge about previously successful database queries
- Formalize and disseminate knowledge about the quality of clinical data and accrued knowledge on how to overcome data quality issues
- Formalize and disseminate knowledge regarding IRB requirements and standard data security measures for secondary use of clinical data

5.9.1.2 Process

Closely related to the '*people*' recommendations are recommendations regarding '*processes*'. In spite of its heavy use of information systems and data sources, the processes involved in extracting clinical data for research are very people-intensive. As a consequence, the '*process*' implications arise from two global themes identified: **organizational structure**, which mostly referred to the way teams were organized within organizations, and *re-usable knowledge*, which, as I just described, refers to the way knowledge is managed by people and teams of people within an organization.

- Establish formal IT consultation processes for researchers, with well-defined procedures and expectations when accessing clinical data for research.
- Foster long-term relationships between programmers and researchers, with frequent opportunities for interaction, to enable mutual education on the complexities of extracting clinical data for research.

- Establish processes to ensure fluent interactions with Institutional Review Boards and regulatory stakeholders.
- When possible, establish centralized IRB and data usage agreements for projects using multi-institutional datasets.

5.9.1.3 Data

The global theme labeled **organizational support** has deep *data* and *tools* implications. In terms of the *data*, they refer to improving the access to high-quality clinical data and not just administrative data. This idea of high quality is not trivial since it implies a continuous monitoring of users' needs and intrinsic data quality.

- Provide access to clinically rich, structured and integrated clinical data
- Continuous assessment and report of data quality for secondary use
- Provide access to de-identified datasets

5.9.1.4 Tools

Finally, the *tools* implications refer to the set of tools to allow researchers to access, query and analyze clinical datasets in a meaningful way. These resources are needed by a broad range of researchers and should be available at an institutional level, as opposed to provided by individual research teams. This was also clearly identified through the **organizational support** global theme.

- Integrated data warehouses
- Intuitive tools to elaborate queries able to extract complex/abstract concepts inside clinical databases
- Natural Language Processing tools to enable the extraction of information from unstructured clinical text
- Implement tools to store and reuse previously successful queries or algorithms
- Electronic IRB applications

5.9.2 Limitations

The main limitations of this study derives from the setting in which it was conducted. Although an effort was made to include a diversity of organizations, there are several traits that were not represented in this set of organizations and might influence its generalizability.

First, this study only included not-for-profit organizations. As discussed in the Methods section of this chapter, the motivations and resources available for secondary use of clinical data offered by for-profit organizations might be different from the ones presented here. However, since the focus of this study was on the researchers and not on any kind of secondary use, this limitation is attenuated.

Second, all three organizations had long histories of using clinical data for research. This may limit the generalization of the findings to other institutions that are new to this domain. Nevertheless, the sample included researchers with a broad range of experiences, spanning from novice to seasoned researchers, in order to account for this limitation. In addition, for the most part researchers felt that the barriers and facilitators discussed here had influenced their ability to use clinical data for research since the beginning, which could signal that the findings might be applicable to institutions just starting this process.

5.10 Summary

This chapter presents a formal assessment that allows us to answer the question: **What are the barriers and facilitators faced by researchers when using clinical data for secondary purposes?** This, in turn, improves our overall understanding of the processes involved in secondary use of clinical data and advances the knowledge required to improve researchers' abilities to use electronic clinical data for clinical and epidemiological research, the main focus of this dissertation.

Several factors, most of them at the organizational level, were identified as barriers and facilitators for the secondary use of clinical data for research. Some of these factors were already known from the literature review presented in chapter 3. Furthermore, this chapter fully supports the findings presented in chapter 4, the idea that researchers frequently need to extract abstract concepts from clinical databases and, when the adequate tools are not available, they resort to manual chart abstractions. On the other hand, several others had not been adequately described yet, such as the need for knowledge management related to the extraction of clinical data, and the need to organize research teams in a manner that fosters that knowledge management.

Finally, on top of establishing methods to manage knowledge and organize teams, researchers are in need of tools that enable them to extract abstract concepts from clinical databases. Chapter 6 will further explore the idea of tools to easily extract higher-level clinical abstractions.

Chapter 6

AIM 3: QUERYING TEMPORAL PATTERNS IN CLINICAL DATABASES

6.1 Introduction

In chapter 4, I presented a systematic description of the kind of electronic patient data that researchers need and the complexities involved in obtaining such data. Chapter 5 provided a qualitative description the main barriers and facilitators that researchers face when using clinical data or research purposes. These two chapters presented findings that help us answer some aspects of this dissertation's overarching question: **How can we improve researchers' abilities to use electronic clinical data for clinical and epidemiological research?** The following findings are relevant to this chapter:

- Researchers are usually interested in high-level clinical concepts that are not necessarily represented as such inside clinical databases (see section 5.9)
- When faced with the need to identify cohorts of patients based on those high level concepts, researchers need to engage in a lengthy process that might eventually lead to the elaboration of a complex database query that is able to represent those high levels concepts. This frequently fails and researchers resort to time-consuming manual chart abstractions (see section 5.7.3.2).
- Queries that include temporal relations are complex to elaborate (see section 4.3.8 on page 89).
- Researchers need easy to use tools that will help them elaborate complex database queries to identify patients based on high-level clinical concepts 5.7.3.2).

Those findings suggest that a tool that allows researchers to query a clinical database using temporal relations might improve their ability to extract electronic patient data for research purposes. In this chapter I will describe—after a discussion of the context and the work done by other in this domain—the development and initial testing of such a tool (see figure 6.1).

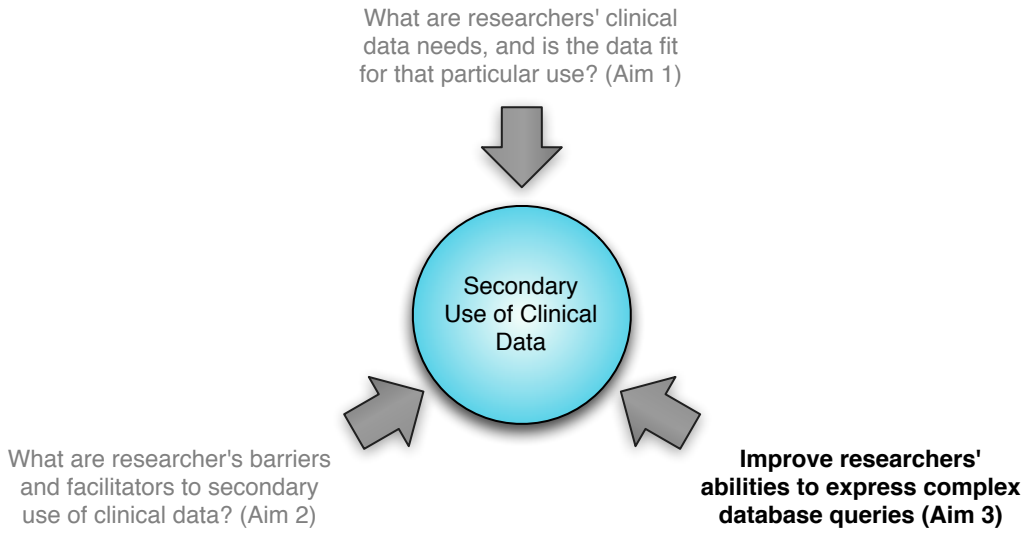


Figure 6.1: Aim 3: Improve researchers' abilities to express complex database queries: querying temporal patterns in clinical data

6.2 Time in Medicine and the Problem of Temporal Queries

Time is a key attribute of clinical data [198]. The act of taking a patient's history in order to make a clinical diagnosis involves describing a series of events—signs and symptoms—in a particular temporal order. The pattern of events over time is then matched against known patterns and their corresponding causal conditions. The well-known concept of prognosis also consists of the expected course of a clinical condition over time. In this age of widespread adoption of electronic medical records, time gains additional importance. Clinical data is being stored in medical record databases using

time stamps to characterize the time an event occurred or was recorded (Abdominal CT scan performed on 11/30/2011). Querying clinical data based on its temporal attributes is one of the key aspects that can enable the reutilization of data routinely collected during patient care.

Temporal attributes can be categorized into absolute or relative temporal ones [199, 200]. *Absolute temporal attributes* involve querying a clinical database within an absolute timeframe, for example extracting patients with a diagnosis of heart failure between 1/1/2000 and 12/31/2010. These queries are not complex to write in standard database query languages and pose no additional complexity when the data is time stamped. Almost all clinical data requests described in chapter 4 contained absolute temporal attributes.

On the other hand, *relative temporal relations*—also present in a significant number of clinical data requests—are significantly more complex to write. Relative temporal relations involve searching for elements that are temporally related but without a specific date range. An example of a relative temporal relation would be searching a database to find patients that presented a positive blood culture while having a central venous catheter (CVC) in place. In this case, the temporal relation—*while*—is not an absolute date or date range but relative to the timeframe when the patient had a CVC in place.

Querying a clinical database using temporal relations is a core component of secondary use. In the domain of quality improvement, for example, when estimating the rate of adequate prophylactic antibiotic administration in elective surgical patients—one element of the Surgical Care Improvement Project (SCIP) Core Measures [201]—the extraction must not only include patients undergoing a surgery and having received a dose of antibiotics, but the time of antibiotics administration must be temporally related with the surgical incision time.

In the area of research there is also an abundance of examples. For instance, to study the incidence of drug-induced Steven-Johnson's Syndrome, researchers must make sure that the symptoms started after the first drug exposure [202]. In a study to characterize patients with radiocontrast induced kidney failure will need to identify patients that received a dose of radiocontrast before presenting a deterioration of kidney failure within a specified period of time [203].

This pervasiveness and relevance of temporal attributes when querying databases for secondary use of clinical data has led to multiple efforts to model, reason with and query on temporal attributes of clinical data. This dissertation aim seeks to develop a system to allow researchers to express complex temporal queries and run them against a clinical relational database.

6.3 Previous Work

6.3.1 Modeling Temporal Data

Modeling temporal data refers to "*... the definition or the adoption of a set of basic concepts that enable a description of a time-oriented clinical world in a sound and unambiguous way*" [204]. Here I paraphrase the same review by Combi, who provides a nice summary of the elements required to expressively model time:

Time Domain This is the collection of elements or entities that allow the representation of time.

Intervals and Instants They are the basic entities used to represent time. Instants represent events that happen instantaneously, such as the time of a blood pressure measurement. Intervals represent events that happen over a period of time, such as a cardiopulmonary by-pass, which can occur over several hours.

Linear, Branching and Circular Times Overwhelmingly time is represented lin-

early, with events occurring in an ordered fashion. When prediction is needed, it may be necessary to use branching time to represent the different possible outcomes in time. In addition, circular time allows representing events that happen periodically.

Absolute and Relative Times As Combi describes, "the position on the time axis of an interval or of an instant can be given as an absolute position, such as the calendric time, when mapped to the time axis used", for example on September, 21 2007. Relative time refers to time mapped in relation to a different time, for example within 30 days of an abdominal surgery. As mentioned in section 6.2, database queries including relative times are more complex to express.

Temporal Relationships The most widely used framework to model temporal relationships is Allen's interval algebra [205]. Allen describes thirteen different temporal relationships between two intervals. Examples of these are during, before, after, and so on. This will be further discussed in section 6.5.2.

Time Granularities According to Combi, This refers to the "level of abstraction at which the [temporal] information is expressed". Granularity can be very precise [2009-06-28 10:00:00] or very coarse [in 2004].

Indeterminacy This refers to temporal uncertainties that are present when there is not enough knowledge available. This is extremely common in medical information. For example, when taking a patient's history, a clinician might record, using free text, a statement describing that the patient started with night sweats between four and six weeks ago. In this case the uncertainty is implicit, since the clinician does not necessarily mean exactly four or exactly six weeks ago.

6.3.2 Representing Temporal Data

In terms of adequately representing temporal data inside clinical databases, efforts have focused on enriching the temporal description of clinical events, adding several temporal dimensions to it. Clinical relational databases usually describe the temporal attributes of data using time stamps, system generated instant definitions of when a new data element was stored in the database. I will refer to these as *time stamped clinical databases*. Temporal clinical databases augment this by adding other temporal attributes. Snodgrass defined three kinds of time that need to be represented in a database to enable rich temporal definitions: transaction time (TT), valid time (VT) and user defined time (UDT) [206]. Transaction time is an attribute of a relation stored in the database, like the time stamp just mentioned. When the database is updated, a new relation—with a new time stamp—is stored and all past values are kept in the database. When using valid times, databases do not store a sequence of states for a given relation. They store a single relation but add a the valid time, the time when *"the stored information models reality"* as Snodgrass says. This approach does not keep track of past states.

These two types of times, transaction and valid time, can be combined into what has been called temporal databases, which can store historical states and the time periods in which those states were valid. In addition to these types of time, Snodgrass adds the idea of user-defined times, which can add richer temporal descriptions to data not captured by TT and VT. In the clinical domain, the concept of user-defined times has been used to add other temporal attributes to clinical data, such as the event time (ET) [207]. An event time is the time describing when an event happened in reality, as opposed to when it was entered or validated inside the database. An example of a data element annotated with multiple temporal attributes would be the result of a blood culture, with a time stamp for when the order was placed, when the blood was drawn, when the culture was obtained positive and when the result was finally

reported. I will refer to these clinical databases that use these multi-dimensional descriptions of time as *temporal clinical databases*.

Despite the utility of this richer way to describe temporal attributes of clinical data, they demand additional effort at data entry time and considerable maintenance of valid times. Furthermore, except for some instances of laboratory results, currently clinical databases are not modeled in such way, thus, making difficult the exploitation of these richer temporal attributes [208].

6.3.3 Querying Temporal Data

Most efforts leading to the ability to query databases belong to two categories: (1) querying temporal clinical databases and (2) querying time stamped clinical databases. The first category has been explored to a greater extent.

TQuel and TSQL2 were developed by a group of researchers lead again by Snodgrass [209, 210]. Both languages query on temporal databases. These initiatives have also included efforts to include such temporal query languages into internationally recognized standards. TSQL2 was initially incorporated into the SQL standard in 1995 but was finally removed in 2001[211]. Other examples of temporal database query languages include S-WATCH-QL [212], GCH-OSQL [213] and T4SQL [214]. These query languages have not been broadly adopted since the underlying data models require rich temporal descriptions, which are not the current standard of clinical databases.

The second category, querying time stamped clinical databases, has been explored in a less systematic way. Generic database languages such as SQL are able to express complex temporal queries, but the complexities and intricacies of such queries limits their adoption by non-experts such as clinical researchers [215]. Most of the work trying to facilitate temporal queries for non-experts has been developed as per-system solutions and not as generalizable solutions. Examples of these systems are DXtrac-

tor, developed at Boston Children's Hospital [208], and ArchiMed developed at the Universty of Vienna [200].

DXtractor—as described by Nigrin—is a *"clinician oriented data retrieval and mining tool"* which includes functions to query time stamped clinical databases. Its main advantage is that it allows building temporal queries using a fairly intuitive natural-language-like query language. The main disadvantages of this system are that it does not allow the explicit definition of time intervals, and it does not allow expressing nested queries. Nested queries are such that the result of one query is the input for the next query. Interval-based querying and nested queries facilitate the expression of more complex temporal relations. ArchiMed, which includes the temporal querying language AMAS, was also developed on top of a time stamped clinical database. Similar to the previous case, this system enables temporal queries using a reasonably simple interface. The main limitation of this system is the fact that it uses a specific temporal data model, in which each document must be part of a case, which are entities created to address a patient's unique medical problem. This system-specific data model—with a *problem*, *case*, and *document* hierarchy—limits its generalizability to other clinical databases. For both these systems, as with most temporal querying systems, the specifications provided by the authors are so insufficient that their generalizability is greatly limited.

6.3.4 Visualizing Temporal Clinical Data

In addition to storing and querying temporal clinical data, additional work has been conducted in the area of human computer interaction and temporal queries. This area has been developed given the complexity involved in building temporal queries for non-experts [200]. The first aspect of this domain has been developing meaningful user interfaces to visualize temporal clinical data stored in a relational database[216, 217, 218]. In addition, Combi[219]has recently described and evaluated

a user interface framework to visually create clinical queries based on time intervals and their relations.

6.3.5 Temporal Abstraction

A central element involved in temporal data modeling, querying and visualization is the concept of *temporal abstraction*. As illustrated in chapter 4, researchers need to identify patients based on high-level concepts such as "all patients with COPD". However, clinical databases usually do not store patient data in such way, but store very detailed data on an individual patient such as a collections of symptoms, clinical findings, and laboratory and imaging results. The process of organizing raw, detailed, patient data into higher-level concepts such as a diagnosis of COPD is what is called data abstraction [220]. In the case of temporal clinical data, this process of organizing temporal data into higher temporal concepts is what is called temporal data abstraction [221]. Temporal abstraction has been consistently included as a fundamental step in the abstraction of clinical data, either manually or automatically.

6.3.6 Knowledge Based Temporal Abstraction

To this date, the most comprehensive theory for temporal modeling and reasoning is the Knowledge Based Temporal Abstraction (KBTA) framework proposed by Shahar et al. in 1997 [222]. This theory describes the temporal abstraction task, five domain-independent sub-tasks required for such temporal abstraction, and four domain-specific knowledge types. This theory allows the modeling, querying and reasoning with temporal data and has been applied to several medical domains [223, 224]. However, this theory relies on significant domain-specific knowledge that needs to be acquired from domain experts and represented in order to be applied to each abstraction task. Although, once represented, the knowledge may be reused, Shahar describes the process of acquiring domain knowledge as a task that would require several hours. Moreover, previously acquired knowledge must be maintained to assure its currency.

In addition, domain-specific knowledge required to perform the five temporal abstractions might vary enormously depending on the use case, even within the same clinical domain. For example, the time distance between two intervals with the same attributes (i.e. an interval representing a normal and stable creatinine) that allows a safe interpolation to build a larger interval can vary whether we are working with ambulatory, hospitalized or critical patients. These differences can definitely be captured in Shahars interpretation context, however, one can easily envision that the number of interpretation contexts is immense, even within a specific clinical domain.

KBTA concepts, all or parts of it, have been implemented in clinical contexts, within a few clinical domains[225, 226]. One such example is the implementation of the Process-Oriented Temporal Analysis (PROTEMPA) system [227]. Its advantage is that this implementation does not use every component of the KBTA theory, thus reducing the amount of domain-specific knowledge acquisition and representation. In addition, it can be coupled to a standard relational clinical database, which vastly increases its generalizability. However, according to the published literature on this system, it relies mainly on two elements: (1) a context, defined as the presence or absence of a certain condition, such as a diagnosis, and (2) the abstraction of intervals constituted by identical instants. For example in the use-case described by the authors, intervals are abstracted from a list of platelet counts. This does not allow the abstraction of intervals based on non-identical instants. An example of the latter would be an interval describing a hospitalization, which would start with an admission and end with a discharge. Since its initial communication, this system has been used to transform clinical data into study-specific representations for clinical research [228] but has not been further used for cohort discovery or outcome assessment, critical components of secondary use.

6.4 Objectives

Given the current state of temporal abstraction and query systems, the goal of this aim was to develop a temporal abstraction and query system based on the strengths of most the KBTA principles with a reduced knowledge engineering and representation effort. It was developed according to the following principles:

- Designed to query database results for multiple patients at a time
- Domain-specific knowledge and context definitions are only required at query time
- Able to abstract intervals constructed from identical and non-identical instants
- Implemented on top of a time stamped clinical relational database

In the following sections I present the work conducted to achieve these goals.

6.5 Methods

6.5.1 Temporal Abstraction Task Selection

The first step was the selection of components of the KBTA framework necessary to query a patient relational database. The KBTA considers a temporal abstraction ontology, which describes the full set of entities required by the framework (such as time stamps, intervals, events, interpretation contexts), temporal abstraction tasks, and the temporal abstraction mechanisms, which describe the mechanisms that solve each temporal abstraction task. According to our requirements, and since this system was designed to identify cohorts of patients, the context is only defined at query time by means of other intervals or by a baseline SQL query. A consequence of this is the elimination of the temporal abstraction tasks that require a pre-defined temporal context. Given these constraints, the following temporal abstraction tasks were finally selected:

Contemporaneous Abstraction involves a simple abstraction from a single contemporaneous element. For instance, a single laboratory result (Hemoglobin; 15g/dL; 2011-10-2) can be abstracted into a contemporaneous instant (Hemoglobin; 15g/dL; 2011-10-2; NORMAL).

Temporal Interpolation involves the creation of intervals by joining adjacent intervals, considering the biological plausibility of that interpolation. For example, in an intensive care context, two identical blood pressure readings separated by a one-minute distance, can be interpolated into a single blood pressure interval ranging for a minute. However, the two identical blood pressure separated by a 48 hour period cannot be interpolated in a similar manner. The user defines the knowledge required for such interpolation at query time.

Temporal Pattern Matching involves the definition of a temporal pattern made of a set of intervals and their relations, with which the query system will find the cohort of interest.

In addition, since the clinical knowledge required for the temporal abstraction tasks is only used at query time, we removed the temporal abstraction knowledge acquisition tool from this implementation.

6.5.2 Temporal Abstraction Element Representation

Once we defined the temporal abstractions that needed to be implemented, we defined the adequate representation of the required elements needed.

Instant: the first level of abstraction, reached through the contemporaneous abstraction temporal task is an instant. Instants are characterized as elements with a single time stamp and a collection of attributes. In the case of a single laboratory result, the attributes are type, result, evaluation (LOW, NORMAL, ELEVATED) and the time stamp.

```
public class Instant {

    public int Id;
    public String Name;
    public Evaluation Eval;
    public LinkedList<IntervalResult> Results;
    public Date Date;

    ...}
```

Figure 6.2: Instant class definition

Interval: after the Instant, the next level of abstraction is the Interval. As mentioned in section 2.2, Intervals are constructed through the temporal interpolation task using instants. Intervals are characterized by a start and end dates, a type, and a series of attributes depending on the type. In the case of intervals constituted by identical elements (such as a series of plasma sodium values over time), the attributes are the Interval's evaluation (LOW, NORMAL or ELEVATED) and a list of all results that constitute the interval. In the case of intervals constituted by non-identical elements (such as a tracheal intubation interval, which starts with an intubation and ends with an extubation) it only has the basic attributes type and start and end dates. The implementation of the Interval class can be seen in Figure 6.3. To enable the third temporal abstraction task, we defined the representation of a *Relation* between two intervals. The implementation of a *Relation* will also be discussed in a few moments.

IntervalList: the fundamental data structure of the system is a list of Intervals, what we will refer to as the IntervalList. Through a standard SQL query to the database, the system retrieves all the necessary time stamped elements to build

```
public class Interval {

    public int Id;
    public String Name;
    public Evaluation Eval;
    public LinkedList<IntervalResult> Results;
    public Date Start;
    public Date End;

    ...}
```

Figure 6.3: Interval class definition

a specific type of interval. Depending on the type of intervals, the system then abstracts instants—which are also stored in a list—and then finally abstracts the intervals—following the parameters defined at query time—storing them in the IntervalList.

IntervalTree and Node: the best-known data structure to store intervals is the interval B-tree [229]. Interval B-trees are best designed to optimize the insertion and deletion of intervals and to perform what has been denominated stabbing queries. Stabbing queries are queries to retrieve all intervals that intersect an instant or an interval. In our case, the main task involves searching for intervals that are before/after a specific instant, hence stabbing queries are not ideal. Likewise, the intervals are only inserted once in the IntervalTree and no deletions are anticipated. As a consequence, we decided to implement a different version of interval trees.

The basic element of the IntervalTree is a Node. The node has a slot for one

interval and three slots for additional nodes denominated BeforeNode, OverlapNode and AfterNode. The Node class implementation can be seen in Figure 6.4. When the IntervalTree is populated, the first interval of an IntervalList is stored in the first node, which becomes the root of the tree. Since the root is already occupied by an interval, the second interval is stored in the corresponding node depending on its relation to the interval stored in the root node. This process is repeated for each interval in an IntervalList until there are no more intervals left. See Figure 6.5. To populate the tree we implemented the PopulateTree method.

```
public class Node {

    public Interval CurrentInterval;
    public Node Before;
    public Node Overlaps;
    public Node After;

    ...}
```

Figure 6.4: Node class definition

The **PopulateTree** method receives an IntervalList and returns an IntervalTree. This method first creates an instance of a Node. To populate the IntervalTree, it recursively checks whether a node is already occupied by an interval. If it is not, then it stores the current interval inside that node. If it is occupied, it tests the temporal relation between the already stored interval and the to-be-stored interval and checks the BeforeNode, OverlapNode or AfterNode concordant to the result of the test. One advantage of this data structure and the populateTree methods is that chaining overlaps cannot occur. This means that the Overlaps relation is not transitive, meaning that if interval A overlaps

interval B, and interval B overlaps interval C, interval C will not be in the Overlap branch of interval A unless it really overlaps interval A.

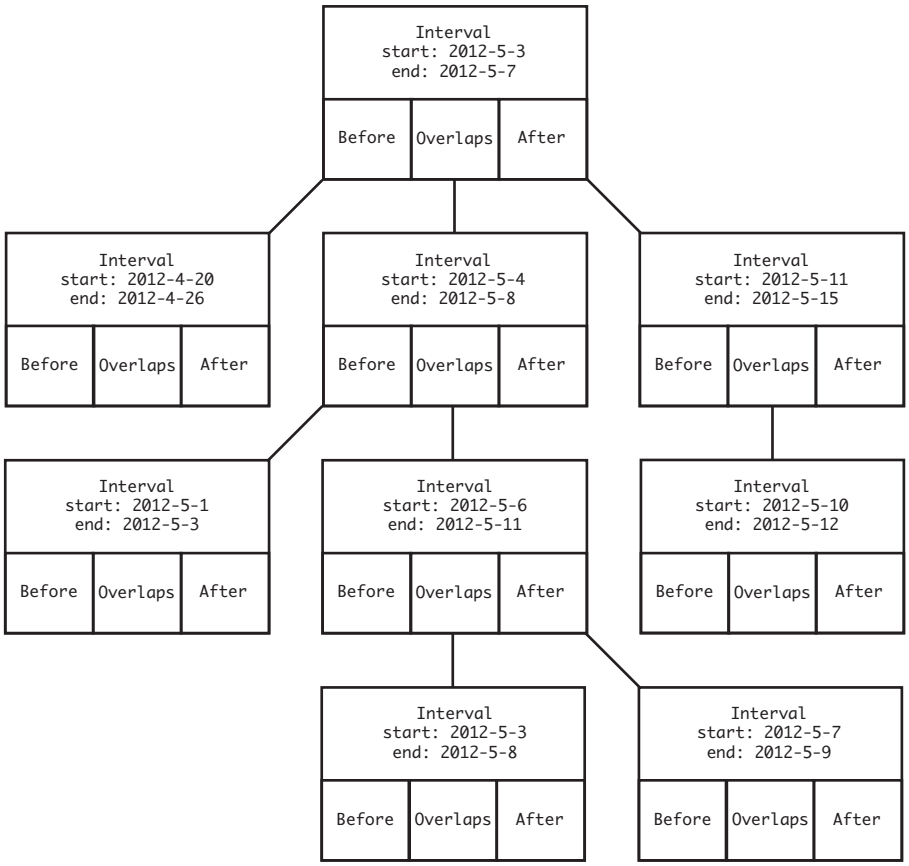


Figure 6.5: Structure of an IntervalTree

Relation: a Relation is a data structure used to store the information needed to test the temporal relationship between two intervals. It has a type (BEFORE, AFTER, OVERLAPS, DURING), in the case of testing two numerical intervals of the same type—such as two intervals of blood pressure readings—it has a direction (INCREASE, DECREASE), attributes that describe the magnitude of the direction (ABSOLUTE-INCREASE, PERCENT-INCREASE), and for all types of intervals, attributes that describe the temporal distance between the two intervals. See Figure 6.6.

```
public class Relation {

    public RelationType TemporalRelation;
    public long Min;
    public long Max;
    public long MaxStart;
    public long MaxEnd;
    public DifferenceDirection Direction;
    public int Percentage;
    public BigDecimal Absolute;

    ...}
```

Figure 6.6: Relation class definition

6.5.3 Query Language

After defining the elements to represent and store temporal clinical data, we defined the query language structure. This query language was implemented in a way that allowed build queries based on a pattern of intervals and their relationships.

The basic query element is a QueryRelation, constituted by two intervals and a relation between them (first interval, relation, second interval). This query returns a set of second interval that fulfill the specified relation with first interval. The second query element is the NextRelation, constituted by a relation and another second interval. In this way, a series of NextRelation can be concatenated to a QueryRelation. The result of a QueryRelation becomes the first interval that is concatenated with the following NextRelation and can be processed in the subsequently in the same fashion. See figures 6.8 and 6.7.

```
BASIC DATA STRUCTURES
-------------------------
<lab_result> ::= <name> <result> <date> <upperlimit> <lowerlimit>
<name> ::= <string>
<result> ::= <float>
<date> ::= <int>
<upperlimit> ::= <float>
<upperlimit> ::= <float>

<instant> ::= <id> <name> <result> <evaluation> <date>
<evaluation> ::= low | normal | elevated
<id> ::= <int>

<intervalResult> ::= <result> <date>

<interval> ::= <id> <name> <evaluation> <results> <startdate> <enddate>
<results> ::= {<intervalResult>}*
<startdate> ::= <date>
<enddate> ::= <date>

<node> ::= <interval> <node> <node> <node>

<relation> ::= <relationtype> <min> <max> <direction> [<startmax> <endmax>]  [<percent> <absolute>]
<min>::= <int>
<max>::= <int>
<startmax> ::= <int>
<endmax> ::= <int>
<direction> ::= <increase> | <decrease>
<percent> ::= <int>
<absolute> ::= <float>

<relationtype> ::= before | overlaps | after | during
<searchinterval> ::= <name> <evaluation>
<patient>::= <id> {<lab_result>}* {<instants>}* {<intervals>}*



THE QUERY
-------------------------
<query> ::= <QueryRelation> {<NextRelation>}*
<QueryRelation> ::= <searchinterval> <relation> <searchinterval>
<NextRelation> ::= <relation> <searchinteval>
```

Figure 6.7: Extended BNF Grammar of the query language

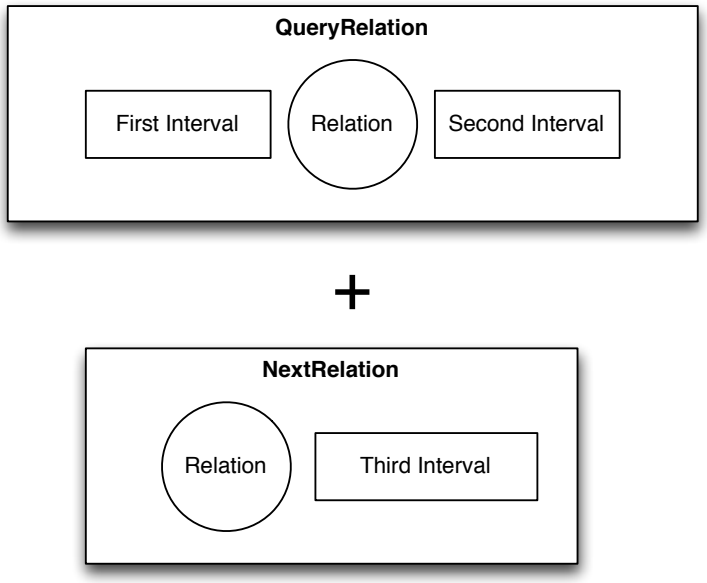


Figure 6.8: The QueryRelation and NextRelation structures that enable the concatenation of successive intervals and relations

Search Methods: the three search methods implemented are: FindAfter, Find Overlaps, and FindDuring. The three methods receive an IntervalList, an Interval-Tree, and a Relation; they return an IntervalList. The search methods go through the IntervalList and for each interval in the list, searches the interval tree for those intervals for which the relationship is true—including the temporal distances and changes in magnitude defined at query time—and stores them in a new IntervalList. See Figure 6.9.

The fact that the input and the output of these search methods are IntervalLists allows us to construct nested queries, in which the output of one query is the input for the next query, just like it occurs when using standard database query languages such as SQL [230].

```
public LinkedList<interval> FindAfter
(Node node, Interval interval, Relation rel, LinkedList<interval> list) {

if (node.CurrentInterval != null)
   {
      if (node.CurrentInterval.Start.getTime() - interval.End.getTime() >= 0)
         {
            double distance = node.CurrentInterval.Start.getTime() -
                                        interval.End.getTime();
            if ((relationSatisfied)
               {
                  list.add(node.CurrentInterval);
               }
               FindAfter(node.LeftNode, interval, rel, list);
               FindAfter(node.CenterNode, interval, rel, list);
               FindAfter(node.RightNode, interval, rel, list);
               return list;
          }
       else
          {
             if (node.RightNode != null)
                {
                   FindAfter(node.RightNode, interval, rel, list);
                }
          }
   }
         return list;
}
```

Figure 6.9: The FindAfter method

6.5.4 Implementation

The implementation of the temporal abstraction and database query system involved the definition of the relevant data structures and methods and the initial data extraction from the relational database. We chose to implement the system using Java 1.6 in order to make it easy for other researchers to adopt. Furthermore, the proposed implementation does not utilize libraries or features not present in a standard Java distribution.

The data required for the initial abstraction of instants and intervals is fairly simple and should be available in any relational clinical database. In the case of intervals constituted by identical intervals (such as numeric laboratory results), the data should be inserted in a database table containing tuples in the following structure:

```
(id, name, date, result, lower-level, upper-level)
```

For example:

```
(12345, 2011-5-13, Creatinine, 1.8, 0.6, 1.1)
```

In the case of intervals constituted by non-identical intervals (such as a hospitalization) should be inserted in a database table containing tuples in the following structure:

```
(id, name, date)
```

For example:

```
(12345, admit, 2011-8-1)

(12345, discharge, 2011-8-12)
```

6.5.5 Testing

To test our implementation we sought to identify patients that had presented an acute kidney failure during a hospitalization in the medical intensive care unit in our hospital, using a previously de-identified dataset. According to the definition proposed by the Acute Kidney Injury Network (AKIN)[15] an acute kidney failure can be defined, in terms of creatinine values, as an absolute increase of serum creatinine $\geq$ 0.3mg/dL or a relative increase of serum creatinine of $\geq 50\%$, both from the baseline. In this case, we can define this with following temporal patterns:

```
(Creatinine LOW)
(BEFORE, INCREASE, 0.3, 50%, 0, 2)
(Creatinine LOW)
          or
(Creatinine LOW)
(BEFORE, INCREASE, 0.3, 50%, 0, 2)
(Creatinine NORMAL)
          or
(Creatinine LOW)
(BEFORE, INCREASE, 0.3, 50%, 0, 2)
(Creatinine ELEVATED)
          or
(Creatinine NORMAL)
(BEFORE, INCREASE, 0.3, 50%, 0, 2)
(Creatinine NORMAL)
```

```
          or
(Creatinine NORMAL)
(BEFORE, INCREASE, 0.3, 50%, 0, 2)
(Creatinine ELEVATED)
or
(Creatinine ELEVATED)
(BEFORE, INCREASE, 0.3, 50%, 0, 2)
(Creatinine ELEVATED)
```

In this case, we created a table containing all creatinine results for patients hospitalized in the medical intensive care unit and stored them in a table according to the previously described specifications and ran the temporal pattern queries. To estimate the sensitivity and specificity of our system, we compared the results against a gold standard, which consisted of a manually annotated sample of 100 randomly selected medical records. Manual annotation was performed by a medical doctor, utilizing the AKIN criteria, and unaware of the results of the temporal abstraction and query. In addition, we estimated the sensitivity and specificity of the annotation conducted by the medical billers in our institution using the discharge ICD-9 codes assigned to each patient, again compared to the mentioned gold standard.

6.6 Results

Overall, the system classified 87% of patients correctly, compared to a 76% of correct classifications by medical billers. The misclassification was equally distributed among false positives (7%) and false negatives (6%). Detailed results can be observed in Table 6.1. Furthermore, the query performed better with more complex combinations of intervals. For example, to identify patients with a normal creatinine before a transient elevation of creatinine, which then returned to normal, according to the following pattern:

```
(Creatinine NORMAL)
(BEFORE, INCREASE, 0.3, 50%, 0, 2)
(Creatinine ELEVATED)
(BEFORE, DECREASE, 0.3, 50\%, 0, 2)
(CREATININE NORMAL)
```

Overall, the system classified 94% of patients correctly, with the misclassification depending entirely on false positives. Detailed results can also be observed in Table 6.1.

	Billers ICD-9 codes vs. Manual Annotation (Acute Kidney Failure)	Temporal Abstraction and Query vs. Manual Annotation (Acute Kidney Failure)	Temporal Abstraction and Query vs. Manual Annotation (Transient Elevation)
Sensitivity	0.76	0.89	1.00
Specificity	0.76	0.83	0.93
LR (+)	3.13	5.48	14.16
LR (-)	0.31	0.13	0.0

Table 6.1: Performance of the temporal abstraction and query system.

6.7 Summary and Conclusions

In this chapter I have presented the arguments supporting the idea that the ability to elaborate temporal queries against time stamped clinical databases is a key component of secondary uses of clinical data. This idea gains additional relevance given the findings presented in chapters 4 and 5 in which researchers expressed their need to extract high-level concepts—in which relative temporal relationships are a fundamental aspect—, and without the necessary tools they end up resorting to time-consuming manual chart abstractions. As a consequence, there is the need to provide tools that are flexible enough to allow the expression of complex temporal patterns while, at the same time, being simple to understand by non-experts in temporal databases is desirable.

There has been a significant amount of research conducted in the area of temporal representation and reasoning with clinical data, however, most efforts include adding additional temporal dimensions to clinical data. The fact that most clinical databases consist of time stamped clinical databases and not temporal clinical databases limits the generalizability of most of the presented systems. Temporal abstraction of higher-level concepts from time stamped clinical data is the most suitable approach, given the limitations of current clinical databases. A few systems able to query time stamped clinical databases using temporal abstraction concepts have been developed, but their use is still limited since they are mostly proprietary.

The Knowledge Based Temporal Abstraction framework is a key contribution to the domain of temporal abstraction and reasoning since it allows the definition and implementation of the necessary entities and tasks to meaningfully query clinical databases. However, its full implementation requires significant efforts to instantiate the contextual knowledge required for its adequate performance. Although generalizable, this characteristic poses a significant barrier to extending it to new domains.

The objective of this aim was to develop a system according to these requirements:

- Designed to query database results for multiple patients at a time
- Domain-specific knowledge and context definitions are only required at query time
- Able to abstract intervals constructed from identical and non-identical instants
- Implemented on top of a time stamped relational database

I proposed the selection of a subset of the KBTA tasks to allow researchers to use the knowledge they already posses to specify the required context at query time. This led to the selection of the Contemporaneous Abstraction task, the Temporal Interpolation task and the Temporal Pattern Matching task. In terms of the data structures, the system was developed to support instants, intervals, and relations.

Testing the system demonstrated its ability to execute a clinically relevant query using a simple query language. For the simple creatinine pattern, the system was able to correctly classify 87% of patients, performing better than that standard manual coding of clinical records using ICD-9 codes. For the more complex pattern, the system classified correctly 94% of all patients. The misclassification was mainly due to false-positives, with no false-negatives observed. A detailed review of the false-positives revealed that most of them were probably due to measurement errors. For example, a patient with consistently normal daily creatinines that shows an acute elevation to 3.0mg/dL in the morning, and an hour later shows a normal value again is probably a measurement error, but the system would consider that a positive result for a transient elevation of the patient's creatinine.

These results show that it is possible to develop systems that can identify high-level concepts from clinical databases—such as an acute kidney failure—without having to rely on diagnostic or procedure codes that are frequently inexact and are usually not

created for this purpose. Given the barrier that the inability to query for high level concept poses on researchers—as described in chapter 3—and the frequency of data requests that include relative temporal attributes—as described in chapter 4—this system could be a significant addition to the set of tools available for secondary use.

In the next chapter I will summarize and synthesize the findings across all three aims, in the context of this dissertation's overarching question: **How can we improve researchers' abilities to use electronic clinical data for clinical and epidemiological research?** and in relation to what was previously known and was discussed in chapters 2 and 3.

Chapter 7
CONCLUSIONS AND FUTURE DIRECTIONS

The use of routinely collected clinical data for purposes other than direct patient care has numerous potential benefits for health care systems. The potential benefits of secondary use of clinical data were discussed in chapter 2. Improvements in the effectiveness and quality of the delivery of health care, the possibility of conducting research in 'real world' settings, and fostering new discoveries through the integration of clinical and genomic data are examples of these potential benefits. Our ability to access and analyze electronic patient data meaningfully serves as the foundation of the learning healthcare system. However, there are significant barriers to secondary use of clinical data.

Previously known barriers to secondary data use can be classified as data-related barriers and societal and organizational barriers. These barriers were discussed in chapter 3. Briefly, data-related barriers include inadequate data quality, fragmented data, and data captured in non computable formats. These data-related barriers limit researchers' ability to use the data they can access. On the other hand, societal and organizational barriers, such as a health care system that has incentives that are not aligned with some of the objectives of secondary use, difficulties in the adoption of tools that enable secondary use and patient privacy concerns, affect researchers' ability to access clinical data. Despite the numerous articles describing barriers to secondary use, and the key role that research plays in the learning health care system, researchers' perspective on the barriers to secondary use have been insufficiently studied.

Given these known barriers, the overarching question for this dissertation was: **How can we improve researchers' abilities to use electronic clinical data for clinical and epidemiological research?** In turn, the sub-questions addressed in each aim were:

- What are researchers' clinical data needs and is the data fit for this particular use?
- What are the barriers and facilitators faced by researchers when using clinical data for secondary purposes?
- Is it possible to build a clinical database query system that has the potential to overcome some of the barriers frequently faced by researchers?

To address these questions, I approached the problem from the three different aspects depicted in figure 7.1.

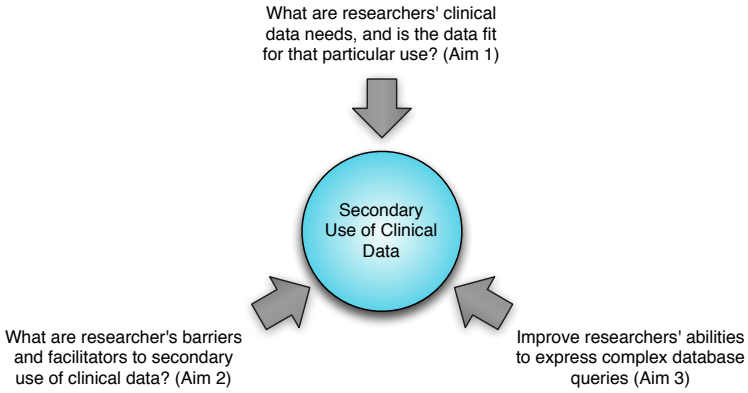


Figure 7.1: Three aspects of secondary use of clinical data addressed in this dissertation.

The next sections contain a summary of the key findings for each aim.

7.1 Aim 1: A Systematic Characterization of Researchers' Clinical Data Requests and of Electronic Clincal Data's Fitness for Use

Aim 1 presented the need to assess not only the intrinsic quality of clinical data, but to expand the view of data quality to include the concept of 'fitness for use'. Although the concept of 'fitness for use' has been applied to other domains, it has not been applied to secondary use of clinical data in the context of research, as expressed in chapter 3. A literature review revealed the lack of an instrument to assess 'fitness for use' in this context. As a consequence, the first objective of this dissertation aim was to develop such a tool.

Through a Delphi process, involving experts from the USA and the UK, concepts contained in previously described frameworks to assess data quality were assessed for pertinence to secondary use. After two rounds, a near final version of this tool—denominated Clinical Data Request Complexity Assessment Tool (CDR-CAT)—was produced. This tool was then systematically applied to a set of clinical data requests to test its ability to capture the attributes of a clinical data request that make it complex to extract under the constraints of the available information systems. After further revising the tool, a final version was produced and it was applied to a prospective sample of consecutive clinical data requests at the University of Washington. The product of applying this tool was a systematic assessment of the complexities involved in fulfilling clinical data requests.

Overall, 25% of all the clinical elements that compose a clinical data request did not have data available in a queryable or computable format. Examples of these were clinical elements that were stored as free-text inside clinical notes or scanned images. On top of that, for most of the clinical elements that did find a structured and queryable database field the intrinsic quality of that data was not known. This combination of not available data with unknown data quality might explain

that most researchers still resorted to manual chart extractions—a very resource-consuming task—despite having all the information stored inside a database.

One common characteristic of clinical elements that were complex to extract was that they frequently involved identifying high-level concepts that are not stored as such inside the database. An example of this would be a request to identify patients with a Glasgow Coma Scale (GCS) of lower than 12 for patients admitted through the emergency room. In this case, despite it being a numerical score, the GCS was not stored as such but had to be abstracted from a complex combination of neurological symptoms. In this sample, approximately 33% of all clinical elements matched this description of high-level abstractions. Interestingly, one of the most common attributes of high-level clinical elements was the presence of relative temporal relations. This supports the preliminary findings described in chapter 1 that were encountered during the early developmental stages of this dissertation. This particular barrier was further confirmed in Aim 2 and was explicitly addressed in Aim 3.

Aim 1 provided key insights into the first research question: *What are researchers' clinical data needs and is the data fit for this particular use?* Specifically, this work revealed that most electronic clinical data is not fit for researchers' intended uses and that is a reason that might explain their extensive use of manual chart abstractions despite having most data in electronic form. Moreover, researchers' frequent need to abstract high-level concepts from raw data provides grounds to the work developed in aim 3.

7.2 Aim 2: Understanding Researchers' Barriers and Facilitators for Secondary Use of Clinical Data

Aim 2 presented a qualitative study of the barriers and facilitators to secondary use of clinical data experienced by researchers at three different research institutions. The three institutions were selected because the offered insights into organizations

with very diverse research settings. These elements of diversity included different reimbursement methods and different degrees of commitment to research. The data was analyzed using a thematic network approach. The analysis identified three global themes and their corresponding organizing and basic themes.

The most relevant theme that emerged during the analysis was the concept of *re-usable knowledge*. Re-usable knowledge refers to the compendium of knowledge required to access and extract electronic patient data for research. This knowledge was further categorized into process-related knowledge—required to navigate the intricacies of the data extraction process—and data-related knowledge—required to understand data quality, to informally assess data quality when formal assessments are not available, and the ability to store and re-use previously successful database queries. This finding is concordant with organizational theories that state that organizations that are able to learn and manage knowledge are more likely to succeed and innovate.

The second theme was the concept of *organizational structure*. This theme refers to the way research teams are organized and how they collaborate. The concept of research teams used was not restricted to researchers and their supporting staff, but extended to programmers accessing databases and IRB personnel and other data-access regulators. Supporting this theme was the organizing theme of *stable and interacting teams*. Research teams that were able to foster face-to-face, long-term interactions among their members—researchers, clinicians, programmers, IRB staff—appeared to have better chances at succeeding in extracting electronic clinical data.

The third identified theme was the concept of *organizational support*. This theme refers to the set of data resources and tools available for researchers. Among the relevant data resources that were considered relevant were the availability of high-quality clinical data (as opposed to just administrative data), the availability of integrated

data resources such as data warehouses and the ability to access anonymized datasets to run pre-IRB queries in preparation for research. The tools that were considered relevant were those that enabled researchers to extract high-level clinical concepts, to extract information from free-text and to reduce the bureaucracy involved in obtaining IRB approval. It is worth calling the attention to the fact that two of the tree types of tools that researchers thought were important for secondary use address issues identified in Aim 1, namely the abundance of information contained in free-text and the frequency with which they need to extract high-level clinical concepts from clinical databases, which is formally addressed in Aim 3.

It is important to note that patient privacy issues were not usually considered barriers *per-se* as the literature frequently reports and was discussed in chapter 3. Researchers felt that patient privacy protections were necessary, and only perceived them as a barrier when local implementations of the Common Rule and HIPPA regulations were unnecessarily bureaucratic or imposed limitations beyond what those mandated by these regulations.

The implications of these findings are numerous, and can be organized using a conceptual framework that describes systems as dynamic entities comprising people, processes, data, and tools components. People implications involve the organization of human resources in the form of interdisciplinary teams in which team members are able to interact and learn from each other, and managing knowledge that is accumulated during the process of extracting clinical data so it can be easily shared and re-used by other teams within the organization. Process implications involve the formalization of the interaction between researchers and database administrators as well as fostering long-term collaborations. Data implications include an active commitment to provide access to high-quality clinical data and anonymized datasets that will accelerate the pre-IRB phases of research. Finally, tools implications included

the provision of integrated data sources, tools to query databases for high-level clinical concepts and tools to support the aforementioned data extraction and knowledge sharing processes.

Aim 2 formally addressed this dissertation's second research question: *What are the barriers and facilitators faced by researchers when using clinical data for secondary purposes?*. The findings confirm the results obtained in aim 1 and also support the need for specific tools for secondary use of clinical data. Aim 3 describes the development and testing of such a tool.

7.3 Aim 3: Querying Temporal Patterns in Clinical Databases

Since this dissertation's preliminary inquiry it was apparent that researchers frequently encountered the need to extract high-level concepts from clinical databases. This idea was further confirmed during Aims 1 and 2. Aim 1 also supported the idea that a significant contributor to these high-level concepts was the presence of relative temporal relations between clinical data elements. Aim 3 presented the work conducted to implement a simplified version of a knowledge-based temporal abstraction system that allows researchers to elaborate complex queries and search clinical databases using high-level concepts containing temporal components.

The first part of this aim was a review of the published literature around representing, modeling and querying temporal clinical data. The output was the identification of the Knowledge Based Temporal Abstraction (KBTA) framework as the most comprehensive tool to model and query temporal clinical data. The downside of this comprehensiveness is that it requires a fair amount of knowledge representation before actually querying a database. The second part of this aim was to identify the components of this framework—the temporal abstraction tasks—that would allow for representing clinical knowledge required for the temporal abstraction at query time.

The next part was the elaboration of data structures, search methods and a query language to express temporal abstraction tasks and temporal queries.

The resulting system was able to execute temporal queries based on combinations of clinical intervals. Testing the system showed that, for the selected measure, it outperformed manual annotation of medical records routinely done by medical billers. The system performed even better when searching for more complicated patterns of time intervals. These results are encouraging since they suggest that it could be used to overcome some of the barriers identified in aims 1 and 2.

Aim 3 addressed the dissertation's third research question: *Is it possible to build a clinical database query system that has the potential to overcome some of the barriers frequently faced by researchers?*. Although the system has not been tested with real users yet, it brings a powerful tool to researchers' arsenal with the potential to facilitate the expression of complex database queries—like the ones identified in aim 1—and reduce the need for manual chart abstractions.

7.4 Contributions

The major contributions of this dissertation are the following:

- Development of a tool to assess the 'fitness for use' of clinical data given the individual researchers' needs and the information resources available (Aim 1).
- Provision of an initial estimation of the overall 'fitness for use' of clinical data at a single research institution using a set of diverse clinical data requests (Aim 1).
- Provision of an initial estimation of the frequency with which researchers request high-level clinical concepts (Aim 1).
- Identification of barriers and facilitators to secondary use of clinical data faced by researchers (Aim 2).

- Identification of recommendations for organizations seeking to improve their use of electronic patient data for research (Aim 2).
- A temporal abstraction and temporal query language to allow researchers to elaborate database queries to identify high-level concepts inside clinical databases (Aim 3).

7.5 Limitations and future work

As any research project, there are some limitations in the applicability of some of these findings that are a consequence of the research methods selected for this dissertation.

Aim 1 was conducted using clinical data requests and the information infrastructure of a single, high-intensity research organization. This may limit the generalizability of results to other research or non-research organizations. However, the Delphi process included experts from multiple organizations from within and outside the USA, which may help attenuate this limitation. The logical future step for this aim is to disseminate the CDR-CAT, to apply it in different settings and compare the results obtained.

Transferability of qualitative results can always be a limitation. Multiple variables may positively or negatively influence the transferability of results, including the researcher being the research tool, and the settings of choice and aim 2 is no exception. Nevertheless, during the course of this study, several measures were taken to ensure rigor and increase the potential for transferability of findings. Inclusion of a diverse set of researchers from a purposely diverse set of organizations, independent coding by two researchers, and engaging participants in member checking (face validity verification) of findings suggest that the results are robust. However, the most significant sign that these results may be transferable to other settings is that they were consistent with previously elaborated organizational theories and previous results of studies investigating the interactions between researchers and information systems. Possible

next steps may include replicating this study in different settings and re-designing organizational structures and processes inside a research organization to improve the secondary use of electronic patient data.

Aim 3 was also conducted using data from a single organization. However, the simplicity of the data structures required to feed the system should make it transferable to any organization with a clinical relational database. One component that was not fully studied during this aim was the idea of heterogeneous intervals. Heterogeneous intervals are those that are composed of non-identical instants (for example the interval HOSPITALIZATION begins with an ADMISSION and ends with a DISCHARGE) as opposed to identical instants (an interval of NORMAL BLOOD PRESSURE is constituted exclusively by BOOD PRESSURE instants). Given the heterogeneity of information standards used by health care organizations it is easy to visualize that there will be multiple ways in which an admission or a discharge can be represented inside a database. This limiting the applicability of this system to other organizations and again transfers the burden to researchers to figure out the local coding standards, which is undesirable. To address this issue, the logical next step would be the development of a *library of heterogeneous intervals* based on the types of intervals that researchers frequently need. Examples of heterogeneous intervals are tracheal intubation (INTUBATION - EXTUBATION), surgical procedure (INCISION - CLOSURE), mechanical ventilation (CONNECTION - DISCONNECTION), and so on. Such a library could improve the applicability of this system to other settings. Additional future work might include the development of a user interface to eliminate the need to write a query language or an evaluation of the usability of the current query language or the eventual user interface.

7.6 Final Remarks

The availability of electronic medical records around the world has exploded during the last few years and, given the current set of incentives to adopt them, they will continue to increase. However, it is remarkable to realize that when the time to use the collected data to improve the health care system and advance scientific knowledge—hallmarks of the learning health care system—arrives, data is frequently not accessible, not reliable or non-existent. This is even more significant when we consider the accumulating evidence that suggests that health information systems, by themselves, will not improve health care [231]. This needs to draw our attention a few steps back and start thinking, not only about meaningful use [37], but about how do we meaningfully implement health information systems, with the complete set of users of such systems in mind—including researchers.

BIBLIOGRAPHY

[1] Institute of Medicine. The Learning Healthcare System. Olsen L, Aisner D, McGinnis M, editors. Workshop summary. Washington D.C.: The National Academy Press; 2007.

[2] Capurro D, Kalet I. Representation of a Simple and Unambiguous Clinical Guideline in a Computer Interpretable Language. In: 2009 AMIA Annual Symposium. San Francisco, CA; 2009. .

[3] Calonge N, Petitti D, DeWitt T, Dietrich A, Gregory K, Harris R, et al. Screening for colorectal cancer: US Preventive Services Task Force recommendation statement. Annals of Internal Medicine. 2008 Nov;149(9):627–37.

[4] Bungard TJ, Ghali WA, Teo KK, McAlister FA, Tsuyuki RT. Why do patients with atrial fibrillation not receive warfarin? Archives of Internal Medicine. 2000;160(1):41–46.

[5] Go A, Hylek E. Anticoagulation therapy for stroke prevention in atrial fibrillation. how well do randomized trials translate into clinical practice? Journal of the American Medical Association. 2004;290(26):2685–2692.

[6] Hart RG, Pearce LA, Aguilar MI. Meta-analysis: antithrombotic therapy to prevent stroke in patients who have nonvalvular atrial fibrillation. Annals of Internal Medicine. 2007;146(12):857–867.

[7] Hu J, Wang Q, Pashos C. Utilization and Outcomes of Minimally Invasive Radical Prostatectomy. Journal of Clinical Oncology. 2008;26(14):2278–2284.

[8] Hu JC, Gu X, Lipsitz SR, Barry MJ, D'Amico AV, Weinberg AC, et al. Comparative effectiveness of minimally invasive vs open radical prostatectomy. Journal of the American Medical Association. 2009 Oct;302(14):1557–1564.

[9] The Dartmouth Atlas of Health Care: Understanding of the Efficiency and Effectiveness of the Healthcare System; [cited July 23, 2012]. Available from: `http://www.dartmouthatlas.org/`.

[10] World Health Organization. Fact File: 10 facts on patient safety; [cited July 23, 2012]. Available from: http://www.who.int/features/factfiles/patient_safety/patient_safety_facts/en/index.html.

[11] Institute of Medicine. To Err is Human: Building a Safer Health System. Kohn L, Corrigan J, Donaldson M, editors. Committee on Quality of Health Care in America. Washington D.C.: National Academy Press; 2000.

[12] Davis P, Lay-Yee R, Briant R, Schug S, Scott A, Johson S, et al. Adverse events in New Zealand public hospitals: principal findings from a national survey. HIC 2003 RACGP12CC [combined conference]: Proceedings. 2003;p. 522.

[13] Baker GR, Norton PG, Flintoft V, Blais R, Brown A, Cox J, et al. The Canadian Adverse Events Study: the incidence of adverse events among hospital patients in Canada. Canadian Medical Association Journal. 2004;170(11):1678.

[14] Percent Annual Increase in National Health Expenditures (NHE) per Capita vs. Increase in Consumer Price Index (CPI), 1980-2009; 2011 [cited July 23, 2012]. Available from: http://facts.kff.org/chart.aspx?ch=212.

[15] Organization for Economic Cooperation and Development. Value for Money in Health Spending. OECD Health Policy Series; 2010.

[16] World Health Organization. Building blocks for action innovative care for chronic conditions: Global Report 2002. Noncommunicable Diseases and Mental Health. Geneva, Switzerland: World Health Organization; 2002.

[17] Reinhardt UE. Does The Aging Of The Population Really Drive The Demand For Health Care? Health Affairs. 2003 Nov;22(6):27–39.

[18] Rettig RA. Medical innovation duels cost containment. Health Affairs. 1994;13(3):7–27.

[19] Sackett DL, Rosenberg W, Gray J, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. British Medical Journal. 1996;312(7023):71.

[20] Eisenstein EL, Lemons PW, Tardiff BE, Schulman KA, Jolly MK, Califf RM. Reducing the costs of phase III cardiovascular clinical trials. American Heart Journal. 2005;149(3):482–488.

[21] Rothwell PM. External validity of randomised controlled trials: "To whom do the results of this trial apply?". The Lancet. 2005 Jan;365(9453):82–93.

[22] Ioannidis J, Lau J. Completeness of safety reporting in randomized trials. Journal of the American Medical Association. 2001;285(4):437.

[23] Singh JA, Sperling J, Buchbinder R, McMaken K. Surgery for shoulder osteoarthritis. Cochrane Database of Systematic Reviews. 2010;(10):CD008089.

[24] Villar H, Saconato H, Valente O, Atallah A. Thyroid hormone replacement for subclinical hypothyroidism. Cochrane Database Systematic Reviews. 2007;3:CD003419.

[25] MacPherson H. Pragmatic clinical trials. Complementary Therapies in Medicine. 2004 Jun;12(2-3):136–140.

[26] Tunis SR, Stryer DB, Clancy CM. Practical clinical trials. Journal of the American Medical Association. 2003;290(12):1624.

[27] The ALLHAT Officers and Coordinators for the ALLHAT Collaborative Research Group. Major cardiovascular events in hypertensive patients randomized to doxazosin vs chlorthalidone: the antihypertensive and lipid-lowering treatment to prevent heart attack trial (ALLHAT). Journal of the American Medical Association. 2000 Apr;283(15):1967–1975.

[28] Center for Reviews and Dissemination, University of York. Systematic Reviews: CRD's gudance for undertaking reviews in health care. York, UK: CRD; 2009.

[29] Higgins JPT, Green S, The Cochrane Collaboration. Cochrane Handbook for Systematic Reviews of Interventions. The Cochrane Collaboration; 2011.

[30] Laupacis A, Wells G, Richardson S, Tugwell P. Users' Guide to the Medical Literature: V. How to Use and Article about Prognosis. Journal of the American Medical Association. 1994 Jul;272(3):234–237.

[31] Aspinall MG, Hamermesh RG. Realizing the promise of personalized medicine. Harvard Business Review. 2007;85(10):108.

[32] British Columbia Medical Association. Warfarin Therapy Management; 2010 [cited July 23, 2012]. Available from: `http://www.bcguidelines.ca/pdf/warfarin_management.pdf`.

[33] Epstein RS, Moyer TP, Aubert RE, O Kane DJ, Xia F, Verbrugge RR, et al. Warfarin genotyping reduces hospitalization rates: results from the MM-WES (Medco-Mayo Warfarin Effectiveness Study). Journal of the American College of Cardiology. 2010;55(25):2804–2812.

[34] Rosell R, Moran T, Queralt C, Porta R, Cardenal F, Camps C, et al. Screening for epidermal growth factor receptor mutations in lung cancer. New England Journal of Medicine. 2009;361(10):958–967.

[35] Glaser J, Henley DE, Downing G, Brinner KM. Advancing personalized health care through health information technology: an update from the American Health Information Community's Personalized Health Care Workgroup. Journal of the American Medical Informatics Association. 2008;15(4):391–396.

[36] Department of and Human Services. Medicare and Medicaid Programs; Electronic Health Record Incentive Program; Final Rule. 2010 Jul;p. 1–276.

[37] Blumenthal D, Tavenner M. The" Meaningful Use" Regulation for Electronic Health Records. New England Journal of Medicine. 2010 Aug;363(6):501–504.

[38] Graham DJ, Campen D, Hui R, Spence M, Cheetham C, Levy G, et al. Risk of acute myocardial infarction and sudden cardiac death in patients treated with cyclo-oxygenase 2 selective and non-selective non-steroidal anti-inflammatory drugs: nested case-control study. The Lancet. 2005;365(9458):475–481.

[39] Bresalier RS, Sandler RS, Quan H, Bolognese JA, Oxenius B, Horgan K, et al. Cardiovascular events associated with rofecoxib in a colorectal adenoma chemoprevention trial. New England Journal of Medicine. 2005;352(11):1092–1102.

[40] Kwon S, Thompson R, Florence M, Maier R, McIntyre L, Rogers T, et al. β-blocker Continuation After Noncardiac Surgery: A Report From the Surgical Care and Outcomes Assessment Program. Archives of Surgery. 2012 Jan;157(5):467–473.

[41] Haynes RB, Sackett DL, Guyatt GH, Tugwell P. Clinical Epidemiology. 3rd ed. How to do Clinical Practice Research. Lippincott Williams & Wilkins; 2005.

[42] Sinackevich N, Tassignon JP. Speeding the critical path. Applied Clinical Trials. 2004;13(1):42–48.

[43] Li L, Chase H, Patel C, Friedman C, Weng C. Comparing ICD9-encoded diagnoses and NLP-processed discharge summaries for clinical trials pre-screening: a case study. AMIA Annual Symposium Proceedings. 2008;p. 404–408.

[44] Asche C, McAdam-Marx C, Shane-McWhorter L, Sheng X, Plauschinat C. Association between oral antidiabetic use, adverse events and outcomes in patients with type 2 diabetes. Diabetes Obesity and Metabolism. 2008 Aug;10(8):638–645.

[45] Morgan SG, McMahon M, Mitton C, Roughead E, Kirk R, Kanavos P, et al. Centralized Drug Review Processes In Australia, Canada, New Zealand, And The United Kingdom. Health Aff (Millwood). 2006 Mar;25(2):337–347.

[46] Wilensky GR. Developing a center for comparative effectiveness information. Health Aff (Millwood). 2006 Oct;25(6):w572–85.

[47] American Recovery and Reinvestment Act of 2009. Public Law 111-5. 111th United States Congress;.

[48] Institute of Medicine US Committee on Comparative Effective Research Prioritization. Initial national priorities for comparative effectiveness research. Washington D.C.: National Academy Press; 2009.

[49] Woolf SH. The Meaning of Translational Research and Why It Matters. Journal of the American Medical Association. 2008 Jan;299(2):211–213.

[50] Chen R, Sigdel T, Li L, Kambham N, Dudley J, Hsieh S, et al. Differentially Expressed RNA from Public Microarray Data Identifies Serum Protein Biomarkers for Cross-Organ Transplant Rejection and Other Conditions. PLoS Computational Biology. 2010 Sep;6(9):e1000940.

[51] NCATS: National Center for Advancing Translational Sciences. CTSA Funded Institutions; [cited July 23, 2012]. Available from: `http://www.ncats.nih.gov/research/cts/ctsa/about/institutions/institutions.html`.

[52] Pain E. European Programs Offer Translational Training; 2007 [cited July 23, 2012]. Available from: `http://sciencecareers.sciencemag.org/career_development/previous_issues/articles/2007_08_17/caredit_a0700119`.

[53] RFA-TR-12-006: Institutional Clinical and Translational Science Award (U54); 2012 [cited July 23, 2012]. Available from: `http://grants.nih.gov/grants/guide/rfa-files/RFA-TR-12-006.html`.

[54] Sung NS, Crowley WF, Genel M, Salber P, Sandy L, Sherwood LM, et al. Central challenges facing the national clinical research enterprise. Journal of the American Medical Association. 2003 Mar;289(10):1278–1287.

[55] Dougherty D, Conway PH. The "3T's" Road Map to Transform US Health Care: The "How" of High-Quality Care. Journal of the American Medical Association. 2008 May;299(19):2319–2321.

[56] Lee TH. Eulogy for a quality measure. New England Journal of Medicine. 2007 Sep;357(12):1175–1177.

[57] Institute of Medicine Committee on Quality of Health Care in America. Crossing the Quality Chasm: A New Health System for the 21st Century. Washington D.C.: National Academy Press; 2001.

[58] Pyzdek T, Keller P. Quality Engineering Handbook. vol. 60. Tucson, Arizona: CRC Press; 2003.

[59] Wiseman B, Kaprielian V. What is Quality Improvement?; [cited July 23, 2012]. Available from: http://patientsafetyed.duhs.duke.edu/module_a/module_overview.html.

[60] Weed J. Factory Efficiency Comes to the Hospital. The New York Times. 2010 Jul;.

[61] The Virginia Mason Production System. Contact: News and Information for the Clinician Community from Virginia Mason Medical Center. 2008;2(1).

[62] Medicare Hospital Compare Quality of Care; [cited July 23, 2012]. Available from: http://hospitalcompare.hhs.gov/.

[63] Quality Initiatives - General Information. Centers for Medicare and Medicaid; [cited July 23, 2012]. Available from: http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityInitiativesGenInfo/index.html.

[64] The Joint Commission. Specifications Manual for National Hospital Inpatient Quality Measures; 2012 [cited July 23, 2012]. Available from: http://www.jointcommission.org/specifications_manual_for_national_hospital_inpatient_quality_measures.aspx.

[65] Asch SM, Sloss EM, Hogan C, Brook RH, Kravitz RL. Measuring underuse of necessary care among elderly Medicare beneficiaries using inpatient and outpatient claims. Journal of the American Medical Association. 2000 Nov;284(18):2325–2333.

[66] McGlynn EA. Six challenges in measuring the quality of health care. Health Affairs. 1997;16(3):7–21.

[67] Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. Journal of Clinical Epidemiology. 2005;58:323–337.

[68] Winslow CEA. The untilled fields of public health. Science. 1920 Jan;51(1306):23.

[69] American Public Health Association. 10 Essential Public Health Services; [cited July 23, 2012]. Available from: http://www.apha.org/programs/standards/performancestandardsprogram/resexxentialservices.htm.

[70] National Center for Health Statistics Center for Disiase Control and Prevention. National Health and Nutrition Examination Survey; [cited July 23, 2012]. Available from: http://www.cdc.gov/nchs/nhanes.htm.

[71] The Information Centre for Health and Social Care National Health Services. Health Survey for England; [cited July 23, 2012]. Available from: http://www.ic.nhs.uk/statistics-and-data-collections/health-and-lifestyles-related-surveys/health-survey-for-england.

[72] Pontificia Universidad Católica de Chile and Universidad Alberto Hurtado. Encuesta Nacional de Salud 2009-2010. Ministerio de Salud, Gobierno de Chile; 2011 [cited July 23, 2012]. Available from: http://www.minsal.gob.cl/portal/url/item/bcb03d7bc28b64dfe040010165012d23.pdf.

[73] Notifiable conditions and the health care provider. Washington State Department of Health; [cited July 23, 2012]. Available from: http://www.doh.wa.gov/Portals/1/Documents/5100/210-001-Poster-HCP.pdf.

[74] Hepatitis A, acute. Notification Form. Washington State Department of Health; [cited July 23, 2012]. Available from: http://www.doh.wa.gov/Portals/1/Documents/5100/210-030-ReportForm-HepA.pdf.

[75] Brabazon ED, O'Farrell A, Murray CA, Carton MW, Finnegan P. Underreporting of notifiable infectious disease hospitalizations in a health board region in Ireland: room for improvement? Epidemiology and Infection. 2008 Feb;136(2):241–247.

[76] Influenza-like Illness Surveillance Program (ILINet). New York State Department of Health; [cited July 23, 2012]. Available from: http://www.health.ny.gov/diseases/communicable/influenza/surveillance/ilinet_program/.

[77] Schirmer P, Lucero C, Oda G, Lopez J, Holodniy M. Effective Detection of the 2009 H1N1 Influenza Pandemic in U.S. Veterans Affairs Medical Centers Using a National Electronic Biosurveillance System. PLoS ONE. 2010 Mar;5(3):e9533.

[78] Health Information Technology: Standards, Implementation Specifications, and Certification Criteria for Electronic Health Record Technology, 2014 Edition; Revisions to the Permanent Certification Program for Health Information Technology; Proposed Rule; [cited July 23, 2012]. Available from: http://www.gpo.gov/fdsys/pkg/FR-2012-03-07/html/2012-4430.htm.

[79] Scannapieco M, Catarci T. Data quality under a computer science perspective. Archivi & Computer. 2002;2:1–15.

[80] Orr K. Data Quality and Systems Theory. Communications of the ACM. 1998 Feb;41(2):66–71.

[81] Pipino L, Lee Y, Wang R. Data quality assessment. Communications of the ACM. 2002;45(4):211–218.

[82] Stein HD, Nadkarni P, Erdos J, Miller PL. Exploring the Degree of Concordance of Coded and Textual Data in Answering Clinical Queries from a Clinical Data Repository. Journal of the American Medical Informatics Association. 2000 Jan;7(1):42–54.

[83] Weaver F, Hatzakis M, Evans C, Smith B, Lavela S, Wallace C, et al. A comparison of multiple data sources to identify vaccinations for veterans with spinal cord injuries and disorders. Journal of the American Medical Informatics Association. 2004 Sep;11(5):377–379.

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

[85] Tayi GK, Ballou DP. Examining data quality. Communications of the ACM. 1998;41(2):54–57.

[86] Batini C, Cappiello C, Francalanci C, Maurino A. Methodologies for data quality assessment and improvement. ACM Computing Surveys (CSUR). 2009;41(3):16.

[87] Tate AR, Williams T, Puri S, Beloff N, Van Staa T. Developing Quality Scores for Electronic Health Records for Clinical Research: A Study using the General Practice Research Database. In: First International Workshop on Managing Interoperability and compleXity in Health Systems MIX-HS'11. Glasgow, UK: ACM Press; 2011. p. 35.

[88] Siegler EL. The evolving medical record. Annals of Internal Medicine. 2010 Nov;153(10):671–677.

[89] McCahill LE, Single RM, Aiello Bowles EJ, Feigelson HS, James TA, Barney T, et al. Variability in Reexcision Following Breast Conservation Surgery. Journal of the American Medical Association. 2012 Jan;307(5):467–475.

[90] Wang Z, Shah AD, Tate AR, Denaxas S, Shawe-Taylor J, Hemingway H. Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning. PLoS ONE. 2012;7(1):e30412.

[91] Hutchins J. The first public demonstration of machine translation: the Georgetown-IBM system, 7th January 1954. AMTA conference. 2004;.

[92] Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. Journal of the American Medical Informatics Association. 2011 Aug;18(5):544–551.

[93] Harkema H, Chapman WW, Saul M, Dellon ES, Schoen RE, Mehrotra A. Developing a natural language processing application for measuring the quality of colonoscopy procedures. Journal of the American Medical Informatics Association. 2011;18(Suppl 1):i150–i156.

[94] Hripcsak G, Austin JHM, Alderson PO, Friedman C. Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports. Radiology. 2002 Jul;224(1):157–163.

[95] Embi PJ, Yackel TR, Logan JR, Bowen JL, Cooney TG, Gorman PN. Impacts of computerized physician documentation in a teaching hospital: perceptions of faculty and resident physicians. Journal of the American Medical Informatics Association. 2004 Jun;11(4):300–309.

[96] Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. Journal of the American Medical Informatics Association. 2011 Mar;18(2):181–186.

[97] Glikrich R, Dreyer N, editors. Registries for Evaluating Patient Outcomes: A User's Guide . 2nd ed. Rockville, MD: Agency for Healthcare Research and Quality; 2010.

[98] About NACCHO. National Association of County and City Health Officials; [cited July 23, 2012]. Available from: http://www.naccho.org/about/.

[99] Bourgeois FC, Olson KL, Mandl KD. Patients treated at multiple acute health care facilities: quantifying information fragmentation. Archives of Internal Medicine. 2010 Dec;170(22):1989–1995.

[100] Cwinn MA, Forster AJ, Cwinn AA, Hebert G, Calder L, Stiell IG. Prevalence of information gaps for seniors transferred from nursing homes to the emergency department. Canadian Journal of Emergency Medicine. 2009 Sep;11(5):462–471.

[101] Wei WQ, Leibson CL, Ransom JE, Kho AN, Caraballo PJ, Chai HS, et al. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. Journal of the American Medical Informatics Association. 2012 Feb;19(2):219–224.

[102] Murphy SN, Lowe H, Bernstam EV, Belazzi R, Ohno-Machado L, Tarczy-Hornoch P. ACMI Bridge Day Panel: Integrating Genomic and Clinical Data in Electronic Health Records and Biomedical Repositories: Challenges, Solutions and Opportunities. In: Cimino JJ, editor. 2011 Summit on Clinical Research Informatics. March 7-9. San Francisco, CA; 2011. .

[103] Vest JR, Gamm LD. Health information exchange: persistent challenges and new strategies. Journal of the American Medical Informatics. 2010;17(3):288–294.

[104] HMO Research Network; [cited July 23, 2012]. Available from: http://www.hmoresearchnetwork.org/.

[105] Louie B, Mork P, Martin-Sanchez F, Halevy A, Tarczy-Hornoch P. Data integration and genomic medicine. Journal of Biomedical Informatics. 2007 Feb;40(1):5–16.

[106] HMO Research Network: Collaboration Toolkit; [cited July 23, 2012]. Available from: http://www.hmoresearchnetwork.org/resources/collab_toolkit.htm.

[107] Anderson N, Abend A, Mandel A, Geraghty E, Gabriel D, Wynden R, et al. Implementation of a deidentified federated data network for population-based cohort discovery. Journal of the American Medical Informatics Association. 2012;19:e60–e67.

[108] A New Generation of American Innovation. The White House; 2004 [cited July 23, 2012]. Available from: http://georgewbush-whitehouse.archives.gov/infocus/technology/economic_policy200404/chap3.html.

[109] Committee ICSSC. IEEE standard computer dictionary: a compilation of IEEE standard computer glossaries. ANSI / IEEE Std. Institute of Electrical and Electronics Engineers; 1990.

[110] Grossmann C, Powers B, McGinnis J. Digital Infrastructure for the Learning Health System: The Foundation for Continuous Improvement in Health and Health Care. Workshop Series Summary. Washington D.C.: National Academy Press; 2011.

[111] Standards and Interoperability Framework. The Office of the National Coordinator for Health Information Technology; 2011 [cited July 23, 2012]. Available from: http://healthit.hhs.gov/portal/server.pt/gateway/PTARGS_0_12811_953732_0_0_18/Standards_and_Interoperability_Framework_Data%20Sheet.pdf.

[112] Garde S, Knaup P, Hovenga E, Herd S. Towards Semantic Interoperability for Electronic Health Records. Methods of information in Medicine. 2007;46(3):332–343.

[113] Yu AC. Methods in biomedical ontology. Journal of Biomedical Informatics. 2006 Jun;39(3):252–266.

[114] Smith B, Ceusters W. HL7 RIM: an incoherent standard. Studies in Health Technologies and Informatics. 2006;124:133–138.

[115] Hasan R, Sion R, Winslett M. Preventing history forgery with secure provenance. ACM Transactions on Storage (TOS). 2009;5(4):12.

[116] Clark C, Taylor R, Shore A. The difference in blood pressure readings between arms and survival: primary care cohort study. British Medical Journal. 2012 Mar;344.

[117] Casson PR, Krawetz SA, Diamond MP, Zhang H, Legro RS, Schlaff WD, et al. Proactively establishing a biologic specimens repository for large clinical trials: an idea whose time has come. Systems Biology in Reproductive Medicine. 2011;57(5):217–221.

[118] Simeon-Dubach D, Perren A. Better provenance for biobank samples. Nature. 2011 Jul;475(7357):454–455.

[119] Dinov I, Lozev K, Petrosyan P, Liu Z, Eggert P, Pierce J, et al. Neuroimaging study designs, computational analyses and data provenance using the LONI pipeline. PLoS ONE. 2010;5(9).

[120] HIT Standards Committee, Office of the National Coordinator for Health Information Technology(ONC). HIT Standards Committee Final Transcript 6/22/2011; 2011 [cited July 23, 2012]. Available from: http://healthit.hhs.gov/portal/server.pt/gateway/PTARGS_0_10741_955142_0_0_18/2011-06-22_standards_thttp://healthit.hhs.gov/portal/server.pt/gateway/PTARGS_0_10741_955142_0_0_18/2011-06-22_standards_transcript_final.pdf.

[121] Allison V, Shefali M. Metadata and Meaningful Use. Journal of AHIMA. 2012 Feb;83(2):32–38.

[122] Hsaio WC. Abnormal economics in the health sector. Health Policy. 1995 Apr;32(1-3):125–139.

[123] Williams C, Mostashari F, Mertz K, Hogin E, Atwal P. From The Office Of The National Coordinator: The Strategy For Advancing The Exchange Of Health Information. Health Affairs (Millwood). 2012 Mar;31(3):527–536.

[124] The Good Stewardship Working Group, Aguilar I, Berger ZD, Casher D, Choi RY, Green JB, et al. The "Top 5" Lists in Primary Care: Meeting the Responsibility of Professionalism. Archives of Internal Medicine. 2011 Aug;171(15):1385–1390.

[125] Miller RH, Miller BS. The Santa Barbara County Care Data Exchange: What Happened? Health Affairs. 2007;26(5):w568–w580.

[126] Vest JR. More than just a question of technology: Factors related to hospitalsâ^{TM} adoption and implementation of health information exchange. International Journal of Medical Informatics. 2010 Dec;79(12):797–806.

[127] Health Insurance Portability and Accountability Act of 1996. Public Law 104-191. 104th United States Congress;.

[128] Summary of the HIPAA Privacy Rule. Department of Health and Human Services. hhsgov. 2003 May [cited July 23, 2012];Available from: http://www.hhs.gov/ocr/privacy/hipaa/understanding/summary/privacysummary.pdf.

[129] Nass S, LA L, Gostin L, editors. Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health through Research. Institute of Medicine. 2006. National Academies Press: Washington DC; 2009.

[130] Ness RB, Joint Policy Committee, Societies of Epidemiology. Influence of the HIPAA Privacy Rule on health research. Journal of the American Medical Association. 2007 Nov;298(18):2164–2170.

[131] Department of Health and Human Services. Title 45. Code of Federal Regulations. Protection of Human Subjects. 2009;.

[132] Secondary Use of Personal Information in Health Research: Case Studies. Ottawa: Canadian Institutes of Health Research; 2002.

[133] Health Information and Quality Authority. International Review of Secondary Use of Personal Health Information; 2012 [cited July 23, 2012]. Available from: http://www.hiqa.ie/system/files/Review-Secondary-Use-Health-Info.pdf.

[134] Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH. Automatic deidentification of textual documents in the electronic health record: a review of recent research. BMC Medical Research Methodology. 2010;10(1):70.

[135] Ross SE, Schilling LM, Fernald DH, Davidson AJ, West DR. Health information exchange in small-to-medium sized family medicine practices: Motivators, barriers, and potential facilitators of adoption. International Journal of Medical Informatics. 2010 Feb;79(2):123–129.

[136] The DGI Data Governance Framework; [cited July 23, 2012]. Available from: http://datagovernance.com/dgi_framework.pdf.

[137] MacKenzie S, Wyatt M, Schuff R. Practices and perspectives on building integrated data repositories: results from a 2010 CTSA survey. Journal of the American Medical Informatics Association. 2012;19(e1):e119–e124.

[138] Ramamurthy KR, Sen A, Sinha AP. An empirical investigation of the key determinants of data warehouse adoption. Decision Support Systems. 2008 Mar;44(4):817–841.

[139] Rogers E. Diffusion of Innovations. 4th ed. New York: The Free Press; 1995.

[140] Klein K, Conn A, Sorra J. Implementing Computerized Technology: An Organizational Analysis. Journal of Applied Psychology. 2001 Jan;86(5):811–824.

[141] Schubart J, Einbinder J. Evaluation of a Data Warehouse: Understanding Users' Needs. Proceedings of the AMIA Symposium. 2000;p. 1131.

[142] i2b2: Informatics for Integrating Biology & the Bedside. A National Center for Biomedical Computing; [cited July 23, 2012]. Available from: `https://www.i2b2.org/`.

[143] Horvath MM, Winfield S, Evans S, Slopek S, Shang H, Ferranti J. The DEDUCE Guided Query tool: Providing simplified access to clinical data for research and quality improvement. Journal of Biomedical Informatics. 2011 Apr;44(2):266–276.

[144] Kamal J, Liu J, Ostrander M, Santangelo J, Dyta R, Rogers P, et al. Information Warehouse–A Comprehensive Informatics Platform for Business, Clinical, and Research Applications. AMIA Annual Symposium Proceedings. 2010;2010:452.

[145] Clinical Practice Research Datalink; [cited July 23, 2012]. Available from: `http://www.cprd.com/intro.asp`.

[146] Weber GM, Murphy SN, McMurry AJ, Macfadden D, Nigrin DJ, Churchill S, et al. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. Journal of the American Medical Informatics Association. 2009 Aug;16(5):624–630.

[147] Health Search. Instituto di Ricerca della Società Italiana di Medicina Generale; [cited July 23, 2012]. Available from: `http://www.healthsearch.it/`.

[148] CRDW Data Request Workflows; [cited July 23, 2012]. Available from: `http://cri.uchicago.edu/?page_id=954`.

[149] Transplant Information System Clinical Data Request. University of Minnesota [cited July 23, 2012];Available from: `http://tis.ahc.umn.edu`.

[150] Jha A, DesRoches C, Campbell E, Donelan K, Rao S, Ferris T, et al. Use of electronic health records in US hospitals. New England Journal of Medicine. 2009;360(16):1628.

[151] Hasson F, Keeney S, McKenna H. Research guidelines for the Delphi survey technique. Journal of Advanced Nursing. 2000;32(4):1008–1015.

[152] Dalkey N, Helmer O. An experimental application of the Delphi method to the use of experts. Management Science. 1963;9(3):458–467.

[153] Keeney S, Hasson F, McKenna H. The Delphi Technique in Nursing and Health Research. Wiley-Blackwell; 2011.

[154] Linstone HA, Turoff M. Delphi Method: Techniques and Applications. Boston, MA: Addison Wesley Educational Publishers Inc.; 1975.

[155] Skulmoski G, Hartman F, Krahn J. The Delphi method for graduate research. Journal of Information Technology Education. 2007;6(1).

[156] McKenna HP. The Delphi technique: a worthwhile research approach for nursing? Journal of Advanced Nursing. 1994 Jun;19(6):1221–1225.

[157] Smithson S, Angell I, Kendall J, Kendall K. A divergent methodology applied to forecasting the future roles of the systems analyst. Human Systems Management. 1992;11(3):123–135.

[158] Normand SLT, McNeil BJ, Peterson LE, Palmer RH. Eliciting expert opinion using the Delphi technique: identifying performance indicators for cardiovascular disease. International Journal for Quality in Health Care. 1998;10(3):247–260.

[159] Blumenthal D, DesRoches C, Donelan K, Ferris T, Jha A, Kaushal R, et al. Health information technology in the United States: the information base for progress. Robert Wood Foundation. 2006;.

[160] Riedmann D, Jung M, Hackl WO, Ammenwerth E. How to improve the delivery of medication alerts within computerized physician order entry systems: an international Delphi study. Journal of the American Medical Informatics Association. 2011 Nov;18(6):760–766.

[161] Jha A, Ferris T, Donelan K, DesRoches C. How Common Are Electronic Health Records In The United States? A Summary Of The Evidence. Health Affairs. 2006 Jan;.

[162] Jha AK, Doolan D, Grandt D, Scott T, BATES DW. The use of health information technology in seven nations. International Journal of Medical Informatics. 2008 Dec;77(12):848–854.

[163] About the CTSA Program. National Institutes of Health; [cited July 23, 2012]. Available from: http://www.ncats.nih.gov/research/cts/ctsa/about/about.html.

[164] Bernard GR, Artigas A, Brigham KL, Carlet J, Falke K, Hudson L, et al. The American-European Consensus Conference on ARDS. Definitions, mechanisms, relevant outcomes, and clinical trial coordination. American Journal of Respiratory and Critical Care Medicine. 1994;149:818–824.

[165] Lohr KN, Steinwachs DM. Health services research: an evolving definition of the field. Health Services Research. 2002 Feb;37(1):7–9.

[166] Gaynor M, Haas-Wilson D. Change, consolidation, and competition in health care markets. Journal of Economic Perspectives. 1999;13(1):141–164.

[167] Sanders S, Glasziou PP, Del Mar CB, Rovers MM. Antibiotics for acute otitis media in children (Review). Cochrane Database of Systematic Reviews. 2004;(1).

[168] Garvin D. Building a Learning Organization. Harvard Business Review. 1993;71(4):78–91.

[169] Greenhalgh T, Robert G, Macfarlane F, Bate P, Kyriakidou O. Diffusion of innovations in service organizations: systematic review and recommendations. The Milbank Quarterly. 2004;82(4):581–629.

[170] Safran C, Bloomrosen M, Hammond W, Labkoff S, Markel-Fox S, Tang P, et al. 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. 2007;14(1):1.

[171] Embi PJ, Payne PR. Clinical Research Informatics: Challenges, Opportunities and Definition for an Emerging Domain. Journal of the American Medical Informatics Association. 2009 May;16(3):316–327.

[172] Bradley EH, Curry LA, Devers KJ. Qualitative data analysis for health services research: developing taxonomy, themes, and theory. Health Services Research. 2007 Aug;42(4):1758–1772.

[173] Pope C, Mays N. Qualitative research: reaching the parts other methods cannot reach: an introduction to qualitative methods in health and health services research. British Medical Journal. 1995;311(6996):42–45.

[174] Malterud K. The art and science of clinical knowledge: evidence beyond measures and numbers. The Lancet. 2001;358(9279):397–400.

[175] Paley J, Lilford R. Qualitative methods: an alternative view. British Medical Journal. 2011;342.

[176] Pope C, Mays N. Opening the black box: an encounter in the corridors of health services research. British Medical Journal. 1993;306(6873):315.

[177] Bloor MJ, Venters GA, Samphier ML. Geographical variation in the incidence of operations on the tonsils and adenoids. The Journal of Laryngology & Otology. 1978;92(10):883–895.

[178] Mort M, Convery I, Baxter J. Psychosocial effects of the 2001 UK foot and mouth disease epidemic in a rural population: qualitative diary based study. British Medical Journal. 2005;331(7527):1234–1239.

[179] Lyytinen K, Robey D. Learning failure in information systems development. Information Systems Journal. 1999;9(2):85–101.

[180] Ash JS. Some Unintended Consequences of Information Technology in Health Care: The Nature of Patient Care Information System-related Errors. Journal of the American Medical Informatics Association. 2003 Nov;11(2):104–112.

[181] Weir C, Nebeker J. Critical Issues in an Electronic Documentation System. AMIA Annual Symposium Proceedings. 2007;p. 786.

[182] Weng C, McDonald DW, Sparks D, McCoy J, Gennari JH. Participatory design of a collaborative clinical trial protocol writing system. International Journal of Medical Informatics. 2007 Jun;76 Suppl 1:S245–51.

[183] Anderson NR, Ash JS, Tarczy-Hornoch P. A qualitative study of the implementation of a bioinformatics tool in a biological research laboratory. International Journal of Medical Informatics. 2007 Nov;76(11-12):821–828.

[184] Anderson NR, Lee ES, Brockenbrough JS, Minie ME, Fuller S, Brinkley J, et al. Issues in biomedical research data management and analysis: needs and barriers. Journal of the American Medical Informatics Association. 2007 Jul;14(4):478–488.

[185] Pope C, Mays N, editors. Qualitative Research in Health Care. 3rd ed. Oxford: Blackwell Publishing - British Medical Journal Books; 2006.

[186] ATLAS.ti v6.2. atlasticom [cited July 23, 2012];Available from: http://www.atlasti.com/index.html.

[187] Tarczy-Hornoch P, Kwan-Gett T, Fouche L, Hoath J, Fuller S, Ibrahim K, et al. Meeting clinician information needs by integrating access to the medical record and knowledge resources via the Web. Proceedings of the AMIA Annual Fall Symposium. 1997;p. 809.

[188] Platt R, Davis R, Finkelstein J, Go AS, Gurwitz JH, Roblin D, et al. Multicenter epidemiologic and health services research on therapeutics in the HMO Research Network Center for Education and Research on Therapeutics. Pharmacoepidemiology and Drug Safety. 2001;10:373–377.

[189] US National Institute for Standards and Technology Federal Information Processing Standards Publication 197. Advanced Encryption Standard (AES). 2001 Nov;.

[190] Bernard HR, Ryan GW. Sampling. In: Analyzing qualitative data. Thousand Oaks, CA: Sage Publications, Inc; 2009. .

[191] Thematic Coding and Analysis. In: The Sage Encyclopedia of Qualitative Research Methods. Thousand Oaks, CA: Sage Publications; 2008. .

[192] Glaser BG, Strauss AL. The Discovery of Grounded Theory. Strategies for Qualitative Research. New Brunswick, USA: Aldine Transactions; 1967.

[193] Attride-Stirling J. Thematic networks: an analytic tool for qualitative research. Qualitative Research. 2001;1(3):385–405.

[194] Alavi M, Leidner DE. Review: Knowledge management and knowledge management systems: Conceptual foundations and research issues. MIS quarterly. 2001;p. 107–136.

[195] Myneni S, Patel VL. Organization of biomedical data for collaborative scientific research: A research information management system. International Journal of Information Management. 2010 Jun;30(3):256–264.

[196] Library of Phenotype Algorithms; [cited July 23, 2012]. Available from: https://www.mc.vanderbilt.edu/victr/dcc/projects/acc/index.php/Library_of_Phenotype_Algorithms.

[197] Carroll J, Edmondson A. Leading organisational learning in health care. Quality and Safety in Health Care. 2002 Jan;11(1):51–56.

[198] Shahar Y. Dimension of time in illness: an objective view. Annals of Internal Medicine. 2000;132(45):45–53.

[199] Augusto J. Temporal reasoning for decision support in medicine. Artificial Intelligence In Medicine. 2005;33(1):1–24.

[200] Dorda W, Gall W, Duftschmid G. Clinical Data Retrieval: 25 Years of Temporal Query Management of the University of Vienna Medical School. Methods of information in medicine. 2002;41(2):89–97.

[201] The Joint Commission. Surgical Care Improvement Project Core Measure Set; [cited July 23, 2012]. Available from: `http://www.jointcommission.org/assets/1/6/Surgical%20Care%20Improvement%20Project.pdf`.

[202] Strom BL, Carson JL, Halpern AC, Schinnar R, Snyder ES, Shaw M, et al. A Population-Based Study of Stevens-Johnson Syndrome. Archives of Dermatology. 1991;127(6):831–838.

[203] McCullough PA, Wolyn R, Rocher LL, Levin RN, O'Neill WW. Acute renal failure after coronary intervention: incidence, risk factors, and relationship to mortality. American Journal of Medicine. 1997 Nov;103(5):368–375.

[204] Combi C, Keravnou-Papailiou E, Shahar Y. Temporal Modeling and Temporal Reasoning. In: Temporal Information Systems in Medicine. New York: Springer; 2010. p. 1–36.

[205] Allen J. An interval-based representation of temporal knowledge. Proc 7th International Joint Conference on Artificial Intelligence, Vancouver, Canada. 1981;p. 221–226.

[206] Snodgrass R, Ahn I. A taxonomy of time databases. ACM Sigmod Record. 1985;14(4):236–246.

[207] Combi C, Montanari A. Lecture Notes in Computer Science. vol. 2068 of Lecture Notes in Computer Science. Dittrich KR, Geppert A, Norrie MC, editors. Berlin, Heidelberg: Springer Berlin Heidelberg; 2001.

[208] Nigrin D, Kohane I. Temporal expressiveness in querying a time-stamp-based clinical database. Journal of the American Medical Informatics Association. 2000 Jan;7(2):152–163.

[209] Snodgrass R. The temporal query language TQuel. ACM Transactions on Database Systems (TODS). 1987;12(2):247–298.

[210] Snodgrass R. The TSQL2 Temporal Query Language. London; 1995. Springer-Verlag.

[211] Combi C, Pozzi G. Querying Temporal Clinical Databases on Granular Trends. Journal of Biomedical Informatics. 2012;45(2):273–291.

[212] Combi C, Missora L, Pinciroli F. Supporting temporal queries on clinical relational databases: the S-WATCH-QL language. In: Proceedings of the AMIA Annual Fall Symposium; 1996. .

[213] Combi C, Pinciroli F, Cavallaro M, Cucchi G. Querying temporal clinical databases with different time granularities: the GCH-OSQL language. Proceedings of the Annual Symposium on Computer Applications in Medical Care. 1995;p. 326–330.

[214] Combi C, Montanari A, Pozzi G. The t4sql temporal query language. Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. 2007;p. 193–202.

[215] Das AK, Musen MA. A comparison of the temporal expressiveness of three database query methods. Proceedings of the Annual Symposium on Computer Application in Medical Care. 1995;p. 331.

[216] Plaisant C, Lam S, Shneiderman B, Smith M, Roseman D, Marchand G, et al. Searching Electronic Health Records for Temporal Patterns in Patient Histories: A Case Study with Microsoft Amalga. AMIA Annual Symposium Proceedings. 2008;2008:601.

[217] Wang T, Plaisant C, Quinn A, Stanchak R, Murphy S, Shneiderman B. Aligning temporal data by sentinel events: discovering patterns in electronic health records. Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems. 2008;p. 457–466.

[218] Tao C, Wongsuphasawat K, Clark K, Plaisant C, Shneiderman B, Chute CG. Towards event sequence representation, reasoning and visualization for EHR data. Proceedings of the 2nd ACM SIGHIT symposium on International health informatics. 2012;p. 801–806.

[219] Combi C, Oliboni B. Visually defining and querying consistent multi-granular clinical temporal abstractions. Artificial Intelligence In Medicine. 2011;54:75–101.

[220] Clancey WJ. Heuristic classification. Artificial intelligence. 1985 Dec;27(3):289–350.

[221] Combi C, Keravnou-Papailiou E, Shahar Y. Abstraction of Time-Oriented Clinical Data. In: Combi C, Keravnou-Papailiou E, Shahar Y, editors. Temporal Information Systems in Medicine. New York: Springer; 2010. p. 1–46.

[222] Shahar Y. A framework for knowledge-based temporal abstraction. Artificial Intelligence. 1997;90(1-2):79–133.

[223] Shahar Y, Miksch S, Johnson P. The Asgaard project: a task-specific framework for the application and critiquing of time-oriented clinical guidelines. Artificial Intelligence In Medicine. 1998;14(1):29–51.

[224] Klimov D, Shahar Y, Taieb-Maimon M. Intelligent visualization and exploration of time-oriented data of multiple patients. Artificial Intelligence in Medicine. 2010;49(1):11–31.

[225] Larizza C, Bellazzi R, Riva A. Temporal abstractions for diabetic patients management. Artificial Intelligence in Medicine. 1997;p. 319–330.

[226] Shahar Y, Goren-Bar D, Boaz D, Tahan G. Distributed, intelligent, interactive visualization and exploration of time-oriented clinical data and their abstractions. Artificial Intelligence in Medicine. 2006;38(2):115–135.

[227] Post A, Harrison Jr J. Protempa: A method for specifying and identifying temporal sequences in retrospective data for patient selection. Journal of the American Medical Informatics Association. 2007;14(5):674–683.

[228] Post A, Kure T, Overcash M, Cantrell D, Eckerson K, Tsui C, et al. A Temporal Abstraction-based Extract, Transform and Load Process for Creating Registry Databases for Research. AMIA Summits on Translational Science Proceedings. 2011 Mar;2011:46–50.

[229] Ang CH, Tan KP. The interval B-tree. Information Processing Letters. 1995;53(2):85–89.

[230] Subquery Syntax. MySQL. Oracle Corporation; [cited July 23, 2012]. Available from: `http://dev.mysql.com/doc/refman/5.0/en/subqueries.html`.

[231] Black AD, Car J, Pagliari C, Anandan C, Cresswell K, Bokun T, et al. The Impact of eHealth on the Quality and Safety of Health Care: A Systematic Overview. PLoS Med. 2011 Jan;8(1):e1000387.

Appendix A

DELPHI QUESTIONNAIRE

INTRODUCTION

Researchers compose clinical data requests to extract data from a clinical data warehouse in order to use the data for non-clinical purposes. Each request needs to be translated into a structured query, which is then executed against the data warehouse and returns a dataset. This data is then sent to the requester. Often, a researcher receives data that is different from what he or she was expecting. While many things can explain this gap, this area has not yet been systematically studied.

The process in which you are participating seeks to modify and comment on a tool that will allow us to evaluate clinical data requests in terms of the properties that make them complex to fulfill, which might help explain the aforementioned gap. Our current framework is based on preliminary work done at University of Washington (unpublished data). It will be the basis for the first round of this Delphi process.

First, you will be led through an application of the tool to a sample clinical data request. You will then be asked evaluate each of the tool's items, in terms of its relevance to clinical data requests in general, comment on each item, propose additional items to be included in the final framework, suggest modifications or any other ideas you would like to discuss.

Again, thank you for participating in this process. We greatly value the time you are devoting to this project!

The following table contains a preliminary version of the Clinical Data Request Complexity Evaluation Tool, with the items and a description of them.

Item	Description
A. Requested data element has a full match to the database schema	There is a structured database field for the required data element
A.1 Accuracy	One or more database fields frequently misrepresents the reality
A.2 Consistency	The database contains different and conflicting versions of the value of one variable
A.3 Objectivity	The data element requires subjective evaluation before entering the data, as opposed to entering raw data.
A.4 Timeliness	The data element is not available in a timely fashion.
A.5 Complex query	The data element can be extracted, but requires the use of a complex database query.
B. Request does not fully match the database schema	The data element is not contained in structured fields
B.1 Non-Computable	The data element is present in a non-computable format, including Natural Language
B.2 Not available	The data element is not available in the database.
C. The data element needed requires post processing	In order to extract the requested data element the database extract needs to be extracted through manual or statistical methods, or manually validated.
D. The data element needed requires external or unavailable data to be adequately interpreted	In order to interpret the data element correctly, the researcher needs to use external data such as medication orders from a different provider.
E. The data element cannot be extracted using the query language in use	The data element is not extractable through a SQL query or the currently used query language; the SQL query is too complex to be built with local knowledge.

In the following sections you will be asked to evaluate each item in terms of its relevancy, I will also ask you to comment on each one of them if you believe it is necessary. Finally, at the end of the survey, you will be able suggest additional attributes and descriptions to be included in the tool.

This preliminary version of the tool was designed to be used on an individual clinical data request. To do this, a clinical data request is divided in parts, and then each item of the tool is applied to every part individually.

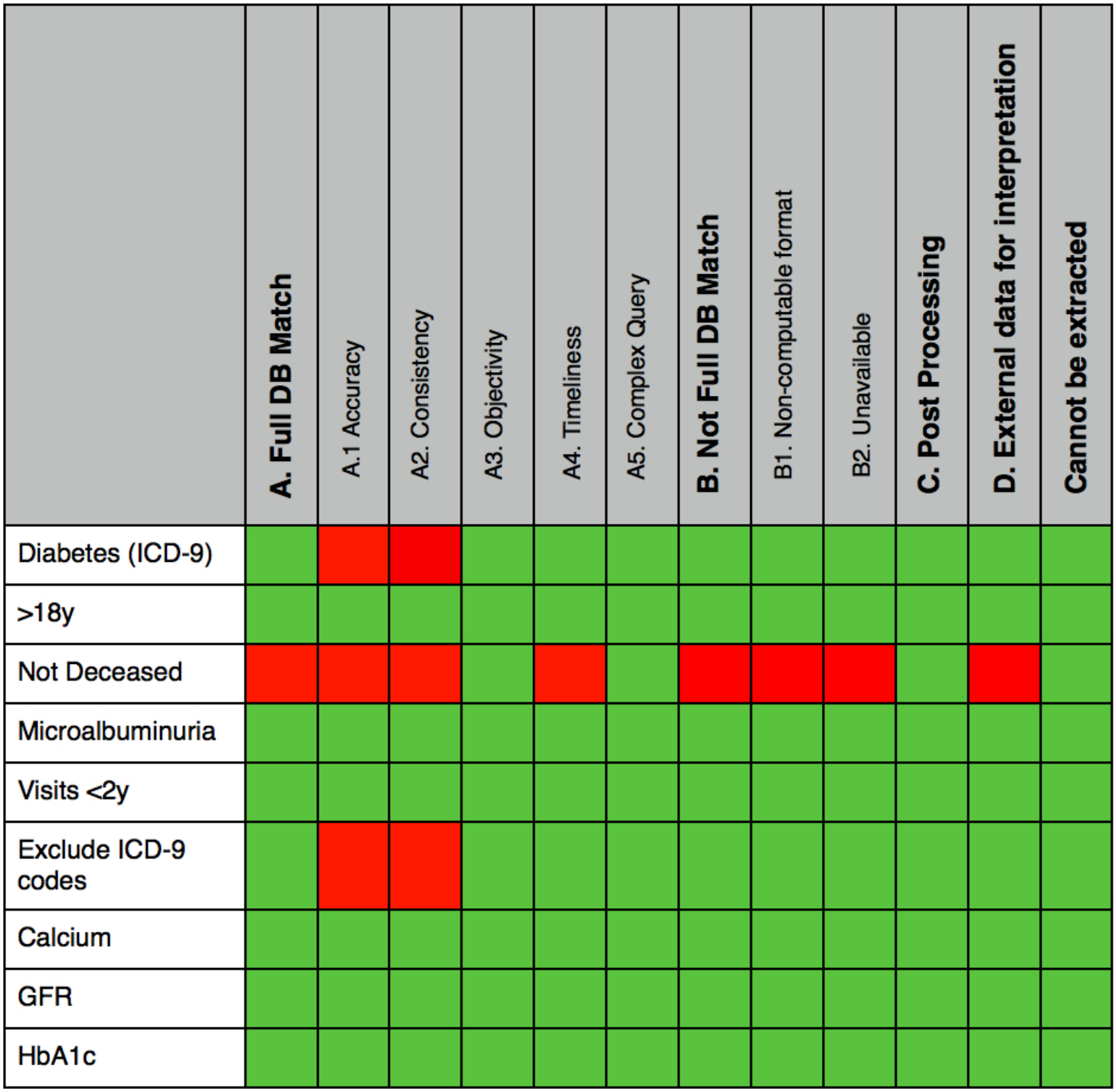
Let's look at an example using the following request:

Researcher's request: retrieve all adult diabetic patients, with microalbuminuria or Albumin/Creatinine index >30, that have visited our center in the past two years. Exclude patients with Calcium >10.5, Glomerular filtration rate <55 and HbA1c >11. Exclude patients with the following ICD9 codes (xxxxxx).

Data elements:
Diabetes
>18 years old
Not deceased
Microalbuminuria >30 or Albumin/Creatinine index >30

Visits within past 2 years
Exclude ICD9 codes (xxxxxxxx)
Exclude: Calcium >10.5 OR GFR <55 OR HbA1c >11

As you can see, the data request is divided by data elements, each elements of the framework should be applied to each data elements, in this case, green means not complex and red means complex. The following table depicts the results of applying the tool to this example clinical data request:



	A. Full DB Match	A.1 Accuracy	A2. Consistency	A3. Objectivity	A4. Timeliness	A5. Complex Query	B. Not Full DB Match	B1. Non-computable format	B2. Unavailable	C. Post Processing	D. External data for interpretation	Cannot be extracted
Diabetes (ICD-9)												
>18y												
Not Deceased												
Microalbuminuria												
Visits <2y												
Exclude ICD-9 codes												
Calcium												
GFR												
HbA1c												

Question 1.
Please rate the following item in terms of its relevancy:

ITEM A: "Requested data element has a full match the database schema".

DESCRIPTION: There is a structured database field for the required data element.
Despite full match, there can still be other issues such as inadequate data quality, which

will be considered in the next sections.

- 1 = Definitely not relevant, should be removed from this Framework
- 2
- 3
- 4
- 5 = Neutral
- 6
- 7
- 8
- 9 = Definitely relevant, should be included in this framework

Question 2.
Please add any comments you have on this item.

Question 3.
Even if there is a full match between the Clinical Data Request and all fields that can answer the query are present in the database schema, there can be additional issues that make extracting the relevant data difficult.

Please rate the following item in terms of its relevancy:

ITEM A.1: Data Accuracy

DESCRIPTION: The database field that contains this data element frequently misrepresents the reality. An example of this could be an ICD9 diagnostic code that was entered by billing coders, which may be inaccurate.

- 1 = Definitely not relevant, should be removed from this Framework
- 2
- 3
- 4
- 5 = Neutral
- 6
- 7
- 8
- 9 = Definitely relevant, should be included in this framework

Question 4.
Please add any comments you have on this item.

Question 5.

Please rate the following item in terms of its relevancy:

ITEM A.2: Consistency

DESCRIPTION: The database contains different and conflicting versions of the value of one variable. For example, there could be a Creatinine Clearance measured by two different methods with conflicting results.

- 1 = Definitely not relevant, should be removed from this Framework
- 2
- 3
- 4
- 5 = Neutral
- 6

- 7
- 8
- 9 = Definitely relevant, should be included in this framework

Question 6.
Please add any comments you have on this item.

Question 7.
Please rate the following item in terms of its relevancy:

ITEM A.3: Objectivity

DESCRIPTION: The data element requires subjective evaluation before entering the data, as opposed to entering raw data. For example, to enter a diagnosis of dementia a clinician might need to consider multiple subjective variables, thus introducing subjectivity in the entered data element.

- 1 = Definitely not relevant, should be removed from this Framework
- 2
- 3
- 4
- 5 = Neutral
- 6
- 7
- 8
- 9 = Definitely relevant, should be included in this framework

Question 8.
Please add any comments you have on this item.

Question 9.
Please rate the following item in term of its relevancy:

ITEM A.4: Timeliness

DESCRIPTION: The requested data element is not available in a timely fashion.

- 1 = Definitely not relevant, should be removed from this Framework
- 2
- 3
- 4
- 5 = Neutral
- 6
- 7
- 8
- 9 = Definitely relevant, should be included in this framework

Question 10.
Please add any comments you have on this item.

Question 11.
Do you believe we should provide a more specific definition of *timely*? If so, what do you think the definition should include?

Question 12.
Please rate the following item in term of its relevancy:

ITEM A.5: Complex Query

DESCRIPTION: The requested data element can be extracted, but requires the use of a complex database query

- 1 = Definitely not relevant, should be removed from this Framework
- 2
- 3
- 4
- 5 = Neutral
- 6
- 7
- 8
- 9 = Definitely relevant, should be included in this framework

Question 13.
Please add any comments you have on this item.

Question 14.
Do you believe we should provide a more specific definition of *Complex Query*? If so, what do you think the definition should include?

Question 15.
Please rate the following item in term of its relevancy:

ITEM B: Requested element does not match the database schema

DESCRIPTION: The requested data element is not available in the database as a structured field.

- 1 = Definitely not relevant, should be removed from this Framework
- 2
- 3
- 4
- 5 = Neutral
- 6
- 7
- 8
- 9 = Definitely relevant, should be included in this framework

Question 16.
Please add any comments you have on this item

Question 17.
Please rate the following item in term of its relevancy:

ITEM B.1: Non-computable format

DESCRIPTION: The requested data element is present in the database but in a non-computable format such as free text (natural language) or scanned notes.

- 1 = Definitely not relevant, should be removed from this Framework
- 2
- 3

- 4
- 5 = Neutral
- 6
- 7
- 8
- 9 = Definitely relevant, should be included in this framework

Question 18.
Please add any comments you have on this item

Question 19.
Please rate the following item in term of its relevancy:

ITEM B.2: Not available

DESCRIPTION: The requested data element is not available at all in the database.

- 1 = Definitely not relevant, should be removed from this Framework
- 2
- 3
- 4
- 5 = Neutral
- 6
- 7
- 8
- 9 = Definitely relevant, should be included in this framework

Question 20.
Please add any comments you have on this element

Question 21.
Please rate the following item in term of its relevancy:

ITEM C: Post processing

DESCRIPTION: In order to extract the requested data element, the initial database extract needs to be processed through manual or statistical methods or needs to be manually validated.

- 1 = Definitely not relevant, should be removed from this Framework
- 2
- 3
- 4
- 5 = Neutral
- 6
- 7
- 8
- 9 = Definitely relevant, should be included in this framework

Question 22.
Please add any comments you have on this item.

Question 23.
Please rate the following item in term of its relevancy:

ITEM D: The data element needed requires external or unavailable data to be adequately interpreted

DESCRIPTION: In order to interpret the data element correctly, the researcher needs to use external data such as medication orders from a different provider. For example, to obtain a list of current medications a patient is taking, the researcher might need to obtain a list of filled prescriptions and its duration from the state's drug stores.

- 1 = Definitely not relevant, should be removed from this Framework
- 2
- 3
- 4
- 5 = Neutral
- 6
- 7
- 8
- 9 = Definitely relevant, should be included in this framework

Question 24.
Please add any comments you have on this item

Question 25.
Please rate the following item in term of its relevancy:

ITEM E: The data element cannot be extracted using the query language in use

DESCRIPTION: The data element is not extractable through a SQL query or the currently used query language; the SQL query is too complex to be built with the locally available knowledge.

- 1 = Definitely not relevant, should be removed from this Framework
- 2
- 3
- 4
- 5 = Neutral
- 6
- 7
- 8
- 9 = Definitely relevant, should be included in this framework

Question 26.
Please add any comments you have on this item

Question 27.
If appropriate, please add other comments you would like to make about this tool. For example, please let me know if you believe we should add additional items to this tool.

Appendix B

CODEBOOK

Code 1.1	
Brief Definition	Accuracy
Full Definition	Data contained in the database does not accurately reflect reality
Use	Use in sentences or paragraphs describing situations in which data elements of the database do not reflect the actual reality, such as medication orders instead of real medication lists or diagnosis entries instead of a list of current and active diagnoses.
Do not Use	For situations in which there is a missing data element that might be present in a different database. See HEALTHSYS under interpretation for this code.
Example:	*"The accuracy for the vital signs was 85%. The accuracy corresponding to an exact match of the risk class was 80%, and the accuracy for obtaining a match differing by at most one category was 100%".*

Code 1.2	
Brief Definition	Consistency
Full Definition	Database contains inconsistent information
Use	Use for sentences or paragraphs where the researcher describes the complexity of using a database given data inconsistencies. For example a problem list might define a patient as diabetic but the ICD-9 discharge diagnosis list does not have the same information.
Do not Use	For missing data
Example:	

Code 1.3	
Brief Definition	Timeliness of Access
Full Definition	Data is not available in a timely fashion

Use	Use for sentences or paragraphs where the researcher describes that data is missing as a consequence of delayed entry or delayed transfer from another database.
Do not Use	For data that is not in the database because it is not part o the database schema. See DATAAVAIL.
Example:	*"Fifth, the administrative data set used in this study is created monthly, not daily. As a result, "real time" CQI projects are not feasible".*

Code 1.4	
Brief Definition	Ease of Access
Full Definition	Narrative that describes situations in which it is easier to access the EMR
Use	Paragraphs or sentences describing situations when there is an advantage in having an EMR because it allows ubiquitous access
Do not Use	
Example:	*"The other thing that is great is that, you know, in theory a chart reviewer could do this remotely so, in theory a chart reviewer could be at her house doing this, she could be in her office doing this, I mean that is really great"*

Code 1.5	
Brief Definition	Availability in DB
Full Definition	Data is not available because it is not part of the database or is present but in a non-computable format
Use	Use for sentences or paragraphs where the researcher describes a data element that is not available in the database or that is available in a different database that is not integrated. Also for data that might be in non-computable formats such as scanned notes or a PDF.
Do not Use	For data that is inconsistently entered into the database. See DATAINC For data that is in the database as free text, see FREETXT
Example:	*"Because the electronic clinical data do not include vital signs, we could not use re- corded temperature on admission as a variable with which to identify target encounters".*

Code 1.6	
Brief Definition	Free Text
Full Definition	Data is not available because it is not part of the database or is present but in a non-computable format
Use	Use for sentences or paragraphs where the researcher describes a data element that is only available as free text.
Do not Use	For data that is not part of the database of is in non-computable formats such as scanned notes or PDF, see DQAVAIL
Example:	*"Whether these textual notes are ultimately fully parseable by natural language processing or can be successfully entered using structured data entry remains to be seen"* *"Because the electronic clinical data do not include vital signs, we could not use recorded temperature on admission as a variable with which to identify target encounters".*

Code 1.7	
Brief Definition	Manual Verification
Full Definition	Although the database can be queried for a data element, the researcher still needs to conduct manual verification of the medical records to make sure the data is accurate.
Use	Use for sentences or paragraphs where the researcher describes a manual chart review to validate a database query or to extract/abstract data contained in the medical record as free text. Can overlap with FREETXT when the manual verification is conducted on free text or DQAVAIL when the manual verification is conducted on non-computable data such as scanned notes, PDF or paper charts
Do not Use	For other kinds of post-processing
Example:	*"If you want to go back in time beyond around 2005, we still have to pull paper charts; so even the last year, so even the last month we have obtained paper charts to supplement of electronic medical record data."*

Code 1.8	
Brief Definition	Data Fragmentation

Full Definition	Narrative that describes data that is located in different, non-integrated locations.
Use	Paragraphs or sentences describing the work involved in pulling data from different non-integrated sources
Do not Use	
Example:	*"We did a project where we wanted to study birth outcomes in babies and we had to go up to… we had to drive to different locations where the inpatient charts were kept"*

Code 2: Privacy

Code 2.1	
Brief Definition	Institutional Access Controls
Full Definition	The data might be used with restrictions because the same institution has placed limitations to its use, because of the existence of data use agreements or state/federal regulations (HIPAA, IRB).
Use	Use in sentences or paragraphs that describe situations in which the researcher's home organization has set up limitations on the use of certain elements of EMR databases. Some examples of this might be HIV status, mental disorder diagnosis.
Do not Use	Do not use to describe limitations because a different institution has placed limitations to its use. See DATAAGREE for this purpose.
Example:	*"Yeah. I don't remember it being that big a deal, believe it or not. So we had to state explicitly what we were doing. I'm sure the IRB asked a question or two about it. This was a HIPPA waiver, of course. But I think that this particular protocol wasn't very problematic from IRBs standpoint."*

Code 2.2	
Brief Definition	Organizational Privacy/Strategic Sensitivities
Full Definition	Description of limitations of data usage because the results of the research might cause effects that collide with institutional interests, especially at the business level.
Use	Use in sentences or paragraphs where the researcher describes situations in which he or she has had to limit his use of data because the results of a particular study may affect the researcher's home institution interests. Examples of this could be reporting a higher than average mortality for a particular condition.

Do not Use	For limitations of data use that are explained by other reasons. See DATAOWN, DATAAGREE.
Example:	*"This misalignment of incentives represents perhaps the single most important barrier to moving ahead and is especially problematic in the outpatient sector".*

Code 2.3	
Brief Definition	Patient Privacy
Full Definition	Narrative describing the complexities of dealing with patient privacy, such as HIPAA regulations and IRB approvals.
Use	For sentences or paragraphs that describe the difficulties that a researcher might have with gaining access to data as a consequence of patient privacy concerns or HIPAA regulations. Also for sentences and paragraphs describing difficulties that are consequence of the need to receive explicit consent by patients to share their data.
Do not Use	For privacy issues not involving patients.
Example:	*"To produce de-identified data from the* [institution] *node, we remove all forbidden HIPAA defined fields, translate dates into integer years for patients below 90 and into decades for patients older than 90, and replace zip codes with the three most significant digits or less in order to comply with the "minimum of 20,000 inhabitants" from the HIPAA requirement. Finally, we scrub all of the narrative text to eliminate identifying numbers, addresses, and proper names and so on".*

Code 2.4	
Brief Definition	Physician Privacy
Full Definition	Narrative describing inability to conduct/report desired research because of physician privacy concerns.
Use	For sentences or paragraphs that describe the difficulties or inability to access data as a consequence of concerns of Physician privacy. Examples of this could be the inability to study physician performance comparisons.
Do not Use	For privacy issues not involving physicians.
Example:	No examples were found in the dataset

Code 3: Query Capabilities

Code 3.1	
Brief Definition	Query Language
Full Definition	Narrative describing the complexities of using a database query language
Use	For sentences or paragraphs describing the complexities for non-expert researchers to use structured database query languages such as SQL, MQL, TSQL.
Do not Use	To describe complexities of using query software. See QUERYTOOLS
Example:	*"This limitation makes it difficult to map a clinician's clinically relevant queries into the query language of the DBMS. As a result, the clinician often must provide data processing above the level of the database. Relational DBMS users, for instance, might need to decompose their time-oriented queries manually into simpler, a temporal Structured Query Language (SQL) statements, and then to manipulate the results to select data that satisfy the desired temporal constraints".*

Code 3.2	
Brief Definition	Query Tools
Full Definition	Narrative describing the complexities of using specific query softwares such as business intelligence applications or NLP engines.
Use	Paragraphs or sentences describing the complexities/ease for non-expert researchers to use business intelligence software, data mining tools or NLP engines.
Do not Use	To describe complexities of using database query languages. See QUERYLANG.
Example:	*"...so we use a system called* [proprietary name]*, we fit in the text of the radiology reports and that classifies all the words, and it extracts concepts and then we roll that out to make the variable definitions"* *"I think it [NLP] is good enough now that I would, you know, the*

	next time that I get a grant funded I will use it as a supplement. I don't think [NLP] is ready to replace manual review"

Code 3.3	
Brief Definition	Infrastructure/Resourcing
Full Definition	Narrative describing the lack of financial resources to perform data extractions.
Use	Paragraphs or sentences describing the financial constraints perceived by researchers to hire or contract the adequate personnel or to buy the adequate software to perform data extraction from clinical databases.
Do not Use	To describe complexities of using query languages or tools.
Example:	*"Interviewer: And… you didn't have the technology because it was not available in your team or was it not available at the* [Institution]*?* *Interviewee: Well, it was -- Yes. So it was not available in my team. It was not available in my division. I don't believe it was available in my department…"*

Code 3.4	
Brief Definition	Insider Knowledge
Full Definition	Narrative describing the need of someone with "insider knowledge" on how the database works and its quirks and twists.
Use	Paragraphs or sentences describing the usefulness of working with someone that has deep knowledge on how the database works
Do not Use	To describe an individual researcher's lack of financial resources
Example:	*"…because getting started, there are a lot of hurdles including just knowing what's feasible and what's not feasible and having someone accessible to bounce that off and particularly someone you know and can trust is useful."*

Code 3.5	
Brief Definition	Clinical Knowledge

Full Definition	Narrative describing the need of someone with "clinical knowledge" to understand the quality of the data.
Use	Paragraphs or sentences describing the need to rely on clinicians that understand how the data is produced as a proxy of data quality
Do not Use	
Example:	*"The other side of things as being a clinician helps you design research because, you know the clinical questions that are important. So most of the -- most people who were doing this research, you know, have some affiliation with their clinician one way or another in this* [type of researche] *research who are clinicians like myself."*

Code 3.6	
Brief Definition	Institutional Knowledge
Full Definition	Narrative describing the presence or absence of institutional knowledge (such as log books, wikis, etc.) to store past experiences querying the database
Use	Paragraphs or sentences describing how every query is like "reinventing the wheel" because previous query experiences are not being stored. In contrast, it can be used for narratives describing the re-utilization of previous queries or knowledge about previous queries to facilitate the process of finding patients.
Do not Use	
Example:	*"So, I would like to think that the programmers ultimately end up with that knowledge and they maintain a data wiki that anybody can access and they try to update it regularly. So that has information about, you know, when certain datasets start, what the limitations might be, what kind of information you can actually get out of them. I think that most programmers would probably say that the data wiki is very complete and up-to-date."*

Code 4: Interpretation

Code 4.1	
Brief Definition	Healthcare System
Full Definition	Narrative describing the complexities imposed by a non-integrated healthcare system in the interpretation of data.

Use	Paragraphs or sentences describing the difficulties analyzing data contained in clinical databases that are a consequence of a fragmented healthcare system or a healthcare system where organizations do not share data. In those situations, if an outcome is not registered in the database (for example a myocardial infarction) it might be difficult to interpret is absence since it could be because it happened elsewhere or that id didn't happen at all.
Do not Use	To describe difficulties analyzing data or to describe missing data.
Example:	"Each data form also asked for a yes/no response from the patient to the question ''*After you are discharged from this visit, if you get worse, will you return to Carolinas Medical Center for a recheck*? [to only include patients that will be taken care inside this system]".

Code 4.2	
Brief Definition	Analysis Complexities
Full Definition	Narrative that describes the complexities of analyzing EMR data after the data has been extracted
Use	Paragraphs or sentences describing the data post processing procedures the researchers have to undertake in order to make the data interpretable. Examples of data post processing include forcing one row per patient in the data structure, performing complex calculations, manual verification of cases.
Do not Use	For descriptions of difficult data visualizations.
Example:	*"The typical data mining process involves transferring data originally collected in production systems into a data warehouse, cleaning or scrubbing the data to remove errors and check for consistency of formats, and then searching the data using statistical queries, neural networks, or other machine learning methods"*

Code 4.3	
Brief Definition	Visualization Complexities
Full Definition	Narrative that describes the complexities of visualizing EMR data after the data has been extracted
Use	Paragraphs or sentences describing the complexities of creating visual displays of the data such as heat maps, dendrograms.

Do not Use	For description of complexities regarding data analysis that are not visual.
Example:	No examples were found in the dataset

Code 4.4	
Brief Definition	Information overload
Full Definition	Narrative that describes situations in which making sense of the data is complex because of information overload.
Use	Paragraphs or sentences describing situations when there is too much data or much more that when paper records were used and it is much harder to analyze/interpret the data
Do not Use	
Example:	"*The downside is that it just seems somehow like there is just a lot more being recorded than was ever before, and I don't know if that would have happened, maybe, if we would have stayed with paper charts*"

Code 5: Other

Code 5.1	
Brief Definition	Awareness
Full Definition	Narrative that describes situations in which a researcher's awareness of what the EMR contains influences his/her ability to query data inside the EMR
Use	Paragraphs or sentences describing situations in which a researcher's knowledge (implicit or explicit) has effects (positively or negatively) the way they use the electronic medical record.
Do not Use	
Example:	*"... the* [institutional database], *which is really a bunch of wonderful electronic data repository; one thing that it has, it has all the billing codes, so it has the ICD-0 codes assigned by doctors or the biller for the visits and then it also has the laboratory results and it has prescriptions that were filled and has demographics and all that stuff. So, I do that all the time… I don't think of that as an electronic medical record because when… I guess you could… when I think of an electronic medical record I think of a version of a patient's chart with free text an all*"