Approaches and Strategy for Cancer Research and Surveillance Data: Integration, Information Pipeline, Data Models, and Informatics Opportunities

Paul A. Fearn

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Peter Tarczy-Hornoch, Chair

Sean D Mooney

Meliha Yetisgen

Program Authorized to Offer Degree:

Biomedical and Health Informatics

University of Washington

Abstract

Approaches and Strategy for Cancer Research and Surveillance Data: Integration, Information Pipeline, Data Models, and Informatics Opportunities

Paul A. Fearn

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

The advancement of cancer research, patient care and public health currently rely on acquisition of data from a variety of sources, information-processing activities, and timely access to data that is of acceptable quality for investigators, clinicians and health officials. With cancer patients living longer and undergoing multiple rounds of treatment, as well as the rise of molecular data that characterize individual patient tumors, there are challenges across all aspect of cancer data collection, integration and delivery. Although there have been advances in deployment of electronic medical records (EMRs) and use of data from EMRs and related systems to support cancer research and patient care, most data needs are still met through costly project specific manual abstraction and project specific databases.

This dissertation builds on my previous work on the Caisis cancer research database at Memorial Sloan-Kettering Cancer Center, and my assessment of trends in information technology (IT) and informatics through site visits and interviews at 60 cancer centers. My hypothesis for this dissertation was that new tools and methods from biomedical informatics could improve the availability of data for cancer research if they were applied thoughtfully and strategically. Within the context of experimenting with the application of selected informatics tools and methods in a cancer center, my overarching research question was: how can we improve access to clinical and related data about cancer patients for research?

The first aim of this dissertation was to develop and assess a modern integrated data platform to support a wide variety of cancer research, which explored the following questions: What are the challenges and opportunities for informatics at Fred Hutch? Do these challenges and opportunities align with my previous work and lessons learned? Are there tools and methods that others or I have used before that could be successfully applied at Fred Hutch? How can we enable new types of research, faster results, and better quality of research data at Fred Hutch? How can we improve access to data with tools and methods that are scalable, extensible to new projects, sustainable and portable to other centers? What is the impact of the data platform developed at Fred Hutch?

The second aim was to develop and characterize a clinical data processing pipeline that is scalable to the cancer center enterprise level, is well-supported and sustainable, and can complement or streamline existing manual data abstraction and information-processing activities. It explored the following questions: How can we improve the quality, speed, and economics of the acquisition, processing, and delivery of clinical data to support cancer researchers? How can we make clinical data processing a core competency of Fred Hutch at an enterprise scale?

The fourth aim was to develop, model, and assess database frameworks for cancer, which explored the following question: How can we best characterize and compare database models and big-data technologies to inform a sensible strategy for cancer research database design and technology?

The fifth aim was to characterize the big-data needs and informatics opportunities for cancer surveillance at the national level, which explored the following questions: What informatics tools and methods have already been applied successfully in the cancer surveillance domain? What are the current and coming opportunities for applying or developing new tools and methods for advancing biomedical informatics in this domain?

The research includes four related papers. The first paper describes the design, development and results of the Hutch Integrated Data Repository and Archive (HIDRA), a modern data platform that provides data feeds, a high security operational and hosting environment, and a self-service data access tool for exploring clinical data as well as associated biospecimen, study and molecular or other assay data. An additional chapter following the first paper describes the impact of HIDRA. The second paper describes the development, implementation and usage of an enterprise pipeline to facilitate the transition from manual data collection and information processing to broad use of clinical data processing and machine-learning methods. The fourth paper characterizes cancer database approaches and big-data technologies to support current and next-generation cancer research. The third paper reviews and summarizes the need for informatics, clinical data processing and machine learning to advance cancer registries at the hospital, state and national level, and reviews current informatics research related to cancer surveillance. The results and contribution of this dissertation are new examples of approaches to data platforms, data models and a pipeline for clinical data processing for

cancer research and cancer surveillance, as well as explanations of underlying motivations, concepts, and tradeoffs for these informatics tools and methods. The contribution of the fourth paper is a potential roadmap for application of informatics, clinical data processing and machine-learning tools and methods to cancer registries and national cancer surveillance.

The generalizable contributions of Chapter 2 are a working a comprehensive database model and associated web-based tool for data abstraction that is temporally organized and has the ability to stack into an analytic structure for predictive modeling. This system is freely distributed under an open-source license, meets common requirements for IT security, extensibility and supportability, and it has already been adopted and extended by numerous other cancer centers in the United States and internationally.

The generalizable contributions of Chapter 3 are the following. First, the volume and variety of data elements that can practically be collected through clinical templates is limited. Second, given the importance of research and collaboration networks, cancer centers should adopt or at least be interoperable with common platforms like REDCap, i2b2, OpenSpecimen and OnCore so that we can wrestle with common issues as a community. Third, due to limited and variable funding for research, solutions need to scale down to affordable levels for individual researchers and labs. Fourth, site visits and active cross-pollination of tools and methods across center must extend deeper into all levels of IT and informatics staff rather than just connecting senior IT leaders and informatics researchers. Finally, centers should spend time and effort resolving social and organizational barriers to progress in informatics and IT.

The generalizable contributions of Chapters 4–5 are the following: First, the legal, IRB, and security framework for HIDRA is relevant to other centers and has been applied to at least two similar efforts. Second, HIDRA provides an example of leveraging a clinical data repository

at a broader academic medical center to support a cancer-specific data repository. Third, HIDRA provides an example of adopting and extending the IT and informatics work of other groups to solve local issues economically. Fourth, HIDRA provides an example of an overall strategy for clinical data acquisition, processing, storage and self service data access. Fifth, HIDRA identified the need for a realistic and de-identified testing dataset to facilitate software development and system implementation. Sixth, the HIDRA work found that lack of federated security for a consortium or matrix cancer center is a critical barrier to progress on an integrated data repository. Finally, the HIDRA project found that the Agile approach to software engineering and system implementation was critical for project momentum and success.

The generalizable contribution of Chapter 6 is a tool for shifting the work of manual data abstraction so that it generates training and validation data suitable for development of automated clinical data processing algorithms (e.g., statistical algorithms, NLP, machine learning). This tool was built with a technology partner, LabKey Software, so that it can be portable to other clients, scalable, and extensible to different clinical data sources and databases. Other groups are already adopting this tool.

The generalizable contributions of Chapter 8 are the following. First, it provides a description of cancer registries and cancer surveillance from an informatics perspective, including the case for automation. Second, it contributes a review of informatics tools and methods applied to cancer registries that indicates potential for automation of clinical data processing. Third, it identifies cancer registries and cancer surveillance as an area for funding and advancing biomedical informatics research.

Table of Contents

Chapter 1 Executive Summary 1

1.1 Overview 1

1.2 Motivation for this Dissertation 2

1.3 Research Questions 4

1.4 Outline of this Dissertation 6

1.4.1 Chapter 2. Preliminary Work: Caisis 6

1.4.2 Chapter 3. Preliminary Work: Recurring Themes in Informatics and IT from 60 Cancer Centers 7

1.4.3 Chapter 4. Hutch Integrated Data Repository and Archive (HIDRA): Data Platform and Clinical Research Informatics Strategy for a Consortium Cancer Center 8

1.4.4 Chapter 5. Evaluation of HIDRA 8

1.4.5 Chapter 6. Scalable Clinical Data Pipeline for a Cancer Center 9

1.4.6 Chapter 7. Comparison of Database Models for Cancer Research 9

1.4.7 Chapter 8. Informatics and Cancer Surveillance: Literature Review and Vision 10

1.4.8 Chapter 9. Conclusions 10

1.5 Contributions 10

Chapter 2 Preliminary Work: Caisis 16

2.1 Context 16

2.2 Overview of Caisis Work 16

2.3 Issues with Caisis 18

2.4 Conclusion 20

2.5 Synthesis 20

Chapter 3 Preliminary Work: Recurring Themes in Informatics and IT from 60 Cancer Centers 23

3.1 Context 23

3.2 Abstract 23

3.3 Introduction 24

3.4 Objective 25

3.5 Methods 25

3.6 Results 27

3.6.1 Clinical Systems Issues 27

3.6.2 Research Systems Issues 30

3.6.3 Technology Trends, Platforms and Tools 36

3.6.4 Social and Organizational Issues 39

3.7 Discussion 43

3.7.1 Limitations of This Research 43

3.7.2 Clinical Systems Implications 44

3.7.3 Research Systems Implications 45

3.7.4 Technology Trends, Platforms and Tools Implications 46

3.7.5 Social and Organizational Implications 48

3.7.6 Rise of Design and Visualization 49

3.8 Conclusion 50

3.9 Acknowledgements 51

3.10 Synthesis 51

Chapter 4 Hutch Integrated Data Repository and Archive (HIDRA): Data Platform and Clinical Research Informatics Strategy for a Consortium Cancer Center 55
4.1 Context 55
4.2 Abstract 57
4.3 Background and Rationale 57
4.4 Methods 61
4.5 Results 69
4.6 Discussion 74
4.7 Conclusion 78
4.8 Acknowledgements 78
4.9 Synthesis 78
Chapter 5 Evaluation of HIDRA 82
5.1 Context 82
5.2 Background and Rationale 82
5.3 Methods 83
5.3.1 Evaluation of Features 83
5.3.2 Evaluation of Performance 84
5.3.3 Evaluation of Usability 86
5.3.4 Evaluation of Outcomes and Impact 87
5.4 Results/Findings 87
5.4.1 Results of Requirements and Feature Evaluation 87
5.4.2 Results of Performance Evaluation 87
5.4.3 Results of Usability Evaluation 90

5.4.4 Results of Outcomes and Impact Evaluation 90
5.5 Discussion 91
5.6 Acknowledgements 92
5.7 Synthesis 92
Chapter 6 Scalable Clinical Data Pipeline for a Cancer Center 94
6.1 Context 94
6.2 Abstract 96
6.3 Background and Rationale 97
6.4 Objective 103
6.5 Approach 104
6.6 Results 111
6.7 Discussion 117
6.8 Conclusion 122
6.9 Acknowledgements 123
6.10 Synthesis 123
Chapter 7 Comparison of Database Models for Cancer Research 126
7.1 Context 126
7.2 Abstract 127
7.3 Background and Rationale 127
7.4 Objective 131
7.5 Methods 131
7.6 Results/Findings 135
7.6.1 Common Issues with Database Models in Cancer Research 135

7.6.2 Underlying Assumptions and Concepts 144
7.6.3 Describe Tradeoffs between Different Modeling Approaches 149
7.7 Discussion 153
7.7.1 Lessons Learned about Data Modeling Approaches 153
7.7.2 Future Research/Next Steps 163
7.8 Conclusion 164
7.9 Acknowledgements 167
7.10 Synthesis 167
Chapter 8 Informatics and Cancer Surveillance: Literature Review and Vision 170
8.1 Context 170
8.2 Abstract 173
8.3 Background and Rationale 174
8.4 Objective 178
8.5 Methods 178
8.6 Results 183
8.6.1 False Positives and Exclusions 183
8.6.2 Identifying Reportable Cancer Cases 185
8.6.3 Natural Language Processing 187
8.6.4 Data Linkages 188
8.6.5 Integrated Platforms 189
8.6.6 Automation 190
8.6.7 Race and Ethnicity Algorithms 190
8.6.8 Security 191

8.6.9 Software 192
8.6.10 Other Findings 192
8.7 Discussion 193
8.7.1 Implications of Findings 193
8.7.2 Next Steps and Opportunities 194
8.8 Conclusion 195
8.9 Acknowledgements 196
8.10 Synthesis 196
Chapter 9 Conclusions 200
9.1 Conclusions from Chapter 2. Preliminary Work: Caisis 200
9.2 Conclusions from Chapter 3. Preliminary Work: Recurring Themes in Informatics and IT from 60 Cancer Centers 203
9.3 Conclusions from Chapter 4. Hutch Integrated Data Repository and Archive (HIDRA): Data Platform and Clinical Research Informatics Strategy for a Consortium Cancer Center 205
9.4 Conclusions from Chapter 5. Evaluation of HIDRA 211
9.5 Conclusions from Chapter 6. Scalable Clinical Data Pipeline for a Cancer Center 211
9.6 Conclusions from Chapter 7. Comparison of Database Models for Cancer Research .. 215
9.7 Conclusions from Chapter 8. Informatics and Cancer Surveillance: Literature Review and Vision 218
9.8 Limitations 222
9.9 Overall Conclusions and Avenues for Future Research 223

List of Figures

Figure 3.1. Routes of driving (30,000 miles) and flying (40 flights)..... 26

Figure 4.1. Conceptual diagram of the legal and IRB framework for HIDRA.. 66

Figure 4.2. Ability for user to search by clinical parameters in Argos self-service application... 72

Figure 5.1. Ranked user requirements and features for the Argos self-service data access tool .. 84

Figure 6.1. Conceptual diagram of a clinical data processing pipeline as deployed at Fred Hutch cancer center 112

Figure 6.2. General architecture of the clinical data processing engine 114

Figure 6.3. Detailed workflow diagram of the clinical data processing pipeline mediated through LabKey Server 115

Figure 6.4. Data abstraction, annotation, and review interface 116

Figure 7.1. Venn diagram of relevant/documented and retrieved/abstracted information 134

Figure 7.2. Measures of accuracy and completeness..... 135

Figure 7.3. Some database models are equivalent for user interface, data storage and reporting; others compartmentalize and translate between these functions. 136

Figure 7.4. Clinical data in relation to underlying model of cancer as a chronic, progressive disease. 145

Figure 7.5. Natural evolution of database technology, adapted from Kalakota (119)..... 154

Figure 7.6. Normalized relational database model may be difficult to query and require multiple joins to integrate data across subtables. 155

Figure 7.7. Data complexity in ontology or terminology and start schema (e.g., i2b2). 156

Figure 7.8. Data complexity in the field names and relationships between tables (e.g., OMOP). 157

Figure 7.9. Data complexity in structured documents (e.g., MongoDB or JSON features in relational databases). 158

Figure 8.1. Flow of cancer case reporting, from hospitals, labs, and other healthcare facilities, to central cancer registries and to aggregated databases 175

List of Tables

Table 3.1. Semistructured interview agenda 27

Table 4.1. As of December 2015, count of patients for each cancer for which they had related events in the medical record (in bold on diagonal), and counts of patients associated with at least two cancer types. 70

Table 5.1. Sample of performance criteria matrix for the Argos self-service data access tool (times in seconds) 85

Table 5.2. Example of results from performance evaluation (times in seconds) 89

Table 8.1. Cancer registry search terms explored in PubMed. From Feb 7, 2016. 180

Table 8.2. Summary of informatics articles retrieved by individual terms and combined. 181

Acknowledgements

I would first like to thank all the members of my dissertation committee, especially Peter Tarczy-Hornoch, who has given me unwavering support and outstanding guidance as my primary advisor and chair of my dissertation committee. I would also like to acknowledge all of my talented colleagues from Fred Hutch cancer center, University of Washington, Seattle Cancer Care Alliance, Seattle Children's Hospital, and LabKey Software that have provided me with the opportunity to envision and shepherd the development of the Hutch Integrated Data Repository and Archive (HIDRA), the Argos self-service tool for data access and exploration and the Fred Hutch/LabKey Server clinical data processing pipeline. I would especially like to thank my Fred Hutch colleagues Mary Gardner, Karen Hansen, James Riddle, Gerianne Sands and Sarah Ramsay for their foundational work on the security, IRB, legal and operations underpinnings of these efforts, David Sharp for outstanding planning and program management, and Emily Silgard for being an awesome collaborator in the development of the clinical data processing pipeline as well as advancing NLP at both the Hutch and the National Cancer Institute (NCI). I would like to thank my mother, Mimi Fearn, and Kathleen Shannon-Dorcy for their encouragement to finish this dissertation. Thanks to Jordan, Kaitlyn and the friendly staff of Woods Coffee who provided me with caffeine and high-speed WiFi access. Last but not least, I would like to thank Lynne Penberthy and the team at the NCI Surveillance Research Program for the tremendous opportunities to apply everything I have learned about biomedical and health informatics from UW and Fred Hutch to the big-data challenges in cancer registries and cancer surveillance.

This manuscript was prepared in the cloud using Google Docs and Paperpile, with finishing touches in Microsoft Word for Mac 2011 and figures generated in Microsoft PowerPoint.

Dedication

To Peter Scardino and Mike Kattan, who provided so many opportunities for me to learn and a foundation for my career in cancer research and biomedical informatics.

List of Acronyms and Abbreviations

ACS	American College of Surgeons
ADaM	Analysis Data Model
AI	Artificial Intelligence
AJCC	American Joint Committee on Cancer
ALK	Anaplastic Lymphoma Kinase
AMIA	American Medical Informatics Association
AML	Acute Myeloid Leukemia
API	Application Programming Interface
AWS	Amazon Web Services
ASCO	American Society of Clinical Oncology
BAA	Business Associate Agreement
Brat	Brat Rapid Annotation Tool
BRIDG	Biomedical Research Integrated Domain Group
caBIG	Cancer Biomedical Informatics Grid
CancerLinQ	Cancer Learning Intelligence Network for Quality
CDASH	Clinical Data Acquisition Standards Harmonization
CDC	Centers for Disease Control and Prevention
CDISC	Clinical Data Interchange Standards Consortium
CDR	Clinical Data Repository
CFR	Code of Federal Regulations
CI4CC	Cancer Informatics for Cancer Centers

CINA	Cancer in North America
CNS	Central Nervous System
CoC	Commission on Cancer
CODI	Consortium Oncology Data Integration
CPT	Current Procedural Terminology
CRN	Cancer Research Network
CSS	Cancer Surveillance System
CSV	Comma Separated Values
CT	Computed Tomography
cTAKES	Clinical Text Analysis and Knowledge Extraction System
CTMS	Clinical Trial Management System
CTSA	Clinical and Translational Science Award
DBMS	Data Base Management System
DMS	Data Management System
DTUA	Data Transfer and Use Agreement
EAV	Entity-Attribute-Value
EDC	Electronic Data Capture
EDW	Enterprise Data Warehouse
EGFR	Epidermal Growth Factor Receptor
EHR	Electronic Health Record
EMR	Electronic Medical Record
eMerge	Electronic Medical Records and Genomics
EOD	Extent of Disease

ETL	Extract-Transform-Load
FDA	Food and Drug Administration
FISMA	Federal Information Security Management Act
FN	False Negative
FP	False Positive
FTE	Full-Time Equivalent
GATE	General Architecture for Text Engineering
GB	Gigabyte
GE	General Electric
Gyn	Gynecology
HDFS	Hadoop Distributed File System
HICOR	Hutchinson Institute for Cancer Outcomes Research
HIDRA	Hutch Integrated Data Repository and Archive
HIPAA	Health Insurance Portability and Accountability Act
HIV	Human Immunodeficiency Virus
i2b2	Informatics for Integrating Biology and the Bedside
ICD	International Classification of Diseases
ICD-O-3	International Classification of Diseases for Oncology, 3rd Edition
IRB	Institutional Review Board
IT	Information Technology
J2EE	Java 2 Platform Enterprise Edition
JSON	JavaScript Object Notation
LIS	Laboratory Information System

LOINC	Logical Observation Identifiers Names and Codes
LS-DAM	Life Sciences Domain Analysis Model
MDS	Myelodysplastic Syndrome
MeSH	Medical Subject Headings
MIT	Massachusetts Institute of Technology
MOU	Memorandum of Understanding
MRI	Magnetic Resonance Imaging
MSKCC	Memorial Sloan-Kettering Cancer Center
NAACCR	North American Association of Central Cancer Registries
NCDB	National Cancer Data Base
NCI	National Cancer Institute
NCIt	National Cancer Institute Thesaurus
NDI	National Death Index
NIH	National Institutes of Health
NHS	National Health Service
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
NPCR	National Program of Cancer Registries
NPV	Negative Predictive Value
NoSQL	Not Only SQL
NPCR	National Program of Cancer Registries
ODM	Operational Data Model
OMOP	Observational Medical Outcomes Partnership

OpenDMAP	Open Source Direct Memory Access Parser
ORIEN	Oncology Research Information Exchange Network
OWL	Web Ontology Language
PHI	Protected Health Information
PPV	Positive Predictive Value
PRO	Patient Reported Outcome
QI/QA	Quality Improvement/Quality Assurance
REDCap	Research Electronic Data Capture
RFI	Request for Information
RLIMS	Research Laboratory Information System
S3	Simple Storage Service
SAS	Statistical Analysis Software
SCCA	Seattle Cancer Care Alliance
SCH	Seattle Children's Hospital
SEER	Surveillance, Epidemiology and End Results
SLA	Service Level Agreement
SNOMED CT	Systematized Nomenclature of Medicine--Clinical Terms
SPIN	Shared Pathology Informatics Network
SPORE	Specialized Program of Research Excellence
SQL	Structured Query Language
SDTM	Study Data Tabulation Model
STRIDE	Stanford Translational Research Integrated Database Environment
SVM	Support Vector Machine

TAB	Tabular
TN	True Negative
TP	True Positive
TSV	Tab Separated Values
UICC	Union for International Cancer Control
UIMA	Unstructured Information Management Architecture
UML	Unified Modeling Language
UW	University of Washington
US	United States
WA	Washington
XML	Extensible Markup Language

Chapter 1 Executive Summary

1.1 Overview

Cancer research is fueled by data. The advancement of cancer prevention, diagnostics and treatments depend on clinical research and quality improvement efforts, as well as translational and broader epidemiologic research. These research and quality improvement efforts, in turn, depend on the timely acquisition and processing of increasingly detailed data about patients. In the ongoing cycle of advancing patient care and research, the volume, variety and detail of data that characterizes patients with cancer and their diseases is growing in multiple dimensions as each medical specialty (e.g., radiology, pathology, radiation oncology, surgery, laboratory medicine) expands its repertoire of tests and procedures. In addition, we now have access to reams of public and commercially available demographic, environmental and consumer data that can potentially be used to characterize factors outside of the healthcare delivery system that may be associated with cancer development, progression and treatments outcomes.(1)

Although this growth in biomedical and healthcare data about patients with cancer has been emerging for decades, and numerous advances in biomedical informatics have been published, the actual level of technology used for the acquisition, processing and management of data in cancer centers has not kept pace, at least not at an enterprise scale. The National Cancer Institute's cancer Biomedical Informatics Grids (caBIG) initiative from 2004 through 2010 attempted to spur the development of scalable and interoperable informatics infrastructure for all cancer centers, however this program eventually failed to improve informatics infrastructure for most centers.(2) In the post-caBIG years, the big-data challenges at cancer centers have continued to grow; however, the value of big cancer informatics initiatives is often met with

skepticism. How can we as informatics researchers and professionals help to translate useful tools and methods from our field into practice at individual centers? How can we facilitate advancements in cancer informatics that are scalable and portable across centers without creating another caBIG?

1.2 Motivation for this Dissertation

The motivation for this dissertation came from my previous work in developing and extending data management systems for cancer research at Baylor College of Medicine and Memorial Sloan-Kettering Cancer Center (MSKCC), described in Chapter 2, and from my visits to other cancer centers and discussions with clinicians, researchers, and informatics colleagues, described in Chapter 3. After rewriting and supporting numerous databases that seemed to overlap and wrangling data from those disparate systems into formats suitable for predictive modeling, I prototyped and shepherded the development of a more reusable solution to cancer data management at MSKCC, the Caisis system.(3) That work naturally led into research on data integration, terminologies, user interface design, clinical workflow integration, extensibility and other common informatics topics.

My hypothesis for this dissertation was that new tools and methods from biomedical informatics could improve the availability of data for cancer research if they were applied thoughtfully and strategically. Specifically, I thought the focus areas to extend my previous work should include the following:

- Clinical natural language processing (NLP), because this technology has been maturing rapidly, and there seemed to be practical limits to clinical data acquisition and processing through templated documentation, data feeds and manual abstraction efforts.

- High security, because multiple grants, contracts and data transfer and use agreements that I had seen were trending toward requiring HIPAA business associate agreements, FISMA/NIST 800-53 level security controls and thorough documentation of security policies, standards and procedures. Moreover, my previous work at MSKCC and with other centers had proven to me the value of proactively meeting and exceeding anticipated security requirements.
- A well-designed database model and strategic approach to database technology, because as databases scale in multiple dimensions, the limits and tradeoffs of their data models tend to rank among the top frustrations of users.
- Self-service and intuitive access to aggregated data through an elegantly designed user interface, because cancer researchers want better access to browse, search and visualize data for their projects. Also, in previous work at MSKCC, I had seen the value of working with professional software developers and user interface artists.

Over the past few years, I have had the opportunity to apply everything I learned from previous work, from discussions with colleagues from other cancer centers, and from coursework and training in biomedical informatics, computational biology and computational linguistics at the University of Washington (UW) toward the advancement of databases and systems to support cancer research. The primary motivation for this work has been to learn how to successfully apply these tools, methods and lessons to the ever-increasing big-data problems at individual cancer centers and across multiple cancer centers.

1.3 Research Questions

Within the context of experimenting with the application of selected informatics tools and methods in a cancer center, my overarching research question has been "How can we improve access to clinical and related data about cancer patients for research?"

Each of the following chapters addresses subquestions related to the overall research question:

- Chapter 2 briefly summarizes my previous work on the Caisis system and aims to answer these two questions: What aspects of Caisis are applicable today to current and future informatics platforms for cancer research? What are the limits of Caisis that would need to be addressed with new tools and methods?
- Chapter 3 summarizes my previous work in visiting more than 60 cancer centers to explore information technology (IT) and informatics trends. In this work, I sought to determine the portability of lessons learned from Caisis and at MSKCC to other centers, and answer the following question: What are the current opportunities for the strategic application of biomedical informatics tools and methods in cancer centers?
- Chapter 4 describes the application of previous lessons learned from both Caisis and my cancer center visits, and the implementation of a new strategy for an integrated data repository at a consortium cancer center. The aim of this chapter was to develop and assess a modern integrated data platform to support a wide variety of cancer research, which explores the following questions: What are the challenges and opportunities for informatics at Fred Hutch? Do these challenges and opportunities align with my previous work and lessons learned? Are there tools and methods that others or I have used before that could be successfully applied at Fred Hutch? How can we enable new types of

research, faster results, and better quality of research data at Fred Hutch? How can we improve access to data with tools and methods that are scalable, extensible to new projects, sustainable and portable to other centers?

- Chapter 5 is an extension of Chapter 4 that answers the following question: What is the impact of the data platform developed at Fred Hutch? It completes the assessment aspect of Aim 1 (to develop and assess a modern integrated data platform to support a wide variety of cancer research) that was started in Chapter 4.
- Chapter 6 is also an extension of Chapter 4 that describes a critical component of the HIDRA informatics platform and strategy, a scalable pipeline for clinical data processing. The aim of this chapter was to develop and characterize a clinical data processing pipeline that is scalable to the cancer center enterprise level, is well-supported and sustainable, and can complement or streamline existing manual data abstraction and information-processing activities. It explores the following questions: How can we improve the quality, speed, and economics of the acquisition, processing, and delivery of clinical data to support cancer researchers? How can we make clinical data processing a core competency of Fred Hutch at an enterprise scale?
- Chapter 7 returns to a more fundamental exploration of database models. It draws from my previous work designing the Caisis database model, anticipated growth of molecular data, and the challenges in the previous three chapters of implementing different database models and technologies to support data acquisition, information processing, and self-service data exploration. Given the rise of marketing and discussion of big-data tools in cancer centers and their emerging big-data struggles, the aim of this chapter was to develop, model, and assess database frameworks for cancer. It explores the following

question: How can we best characterize and compare database models and big-data technologies to inform a sensible strategy for cancer research database design and technology?

- Chapter 8 explores the challenges in cancer surveillance, particularly the scalability to much greater volume and variety of data than any single cancer center would experience. The emerging challenges of big data across cancer registries, national cancer surveillance efforts, and networks of cancer centers are orders of magnitude greater than those faced by any one center. The aim of this chapter was to characterize the big-data needs and informatics opportunities for cancer surveillance at the national level, which explores the following questions: What informatics tools and methods have already been applied successfully in the cancer surveillance domain? What are the current and coming opportunities for applying or developing new tools and methods for advancing biomedical informatics in this domain?
- Chapter 9 summarizes the conclusions from previous chapters and returns to the overall research question: How can we improve access to clinical and related data about cancer patients for research?

1.4 Outline of this Dissertation

This section briefly describes the contents of each chapter.

1.4.1 Chapter 2. Preliminary Work: Caisis

This chapter describes key points from my work on the design and implementation of the Caisis database at MSKCC and my assistance with implementing Caisis at other cancer centers. The issues arising in this work are relevant to the strategy and development of HIDRA (Chapters 4–5,

focused on the aim to develop and assess a modern integrated data platform to support a wide variety of cancer research), the clinical data processing pipeline (Chapter 6, focused on the aim to develop and characterize a clinical data processing pipeline that is scalable to the cancer center enterprise level, is well-supported and sustainable, and can complement or streamline existing manual data abstraction and information-processing activities), and the discussion of database models (Chapter 7, focused on the aim to develop, model, and assess database frameworks for cancer).

1.4.2 Chapter 3. Preliminary Work: Recurring Themes in Informatics and IT from 60 Cancer Centers

Building on the work at a single freestanding cancer center summarized in the prior section 1.4.1, this chapter describes findings from my 2008–2009 visits to 60 cancer centers around the United States and meetings with 394 people at these centers to explore issues and trends in informatics and IT. The information gathered from these discussions was organized using a qualitative analysis tool, and a number of distinct patterns became evident. This preliminary work includes findings regarding EMRs and clinical data repository implementations, the selection and implementation of clinical research systems, struggles with biospecimen information management, the curation and technical support of research databases, trends in database and web development platforms, and social/organizational issues. The findings of this preliminary work informed the strategy and implementation of HIDRA (Chapters 4–5, focused on the aim to develop and assess a modern integrated data platform to support a wide variety of cancer research), and the pipeline for clinical data processing (Chapter 6, focused on the aim to develop and characterize a clinical data processing pipeline that is scalable to the cancer center enterprise

level, is well-supported and sustainable, and can complement or streamline existing manual data abstraction and information-processing activities).

1.4.3 Chapter 4. Hutch Integrated Data Repository and Archive (HIDRA): Data Platform and Clinical Research Informatics Strategy for a Consortium Cancer Center

Building on the lessons learned from a single freestanding cancer center and the Caisis system summarized in section 1.4.1 and across multiple cancer centers as summarized in section 1.4.2, this chapter describes HIDRA, an initiative to develop an integrated and modern data platform to support current and next-generation clinical and translational research for a consortium cancer center. The aim of this chapter was to develop and assess a modern integrated data platform to support a wide variety of cancer research. This chapter describes the vision and strategic goals for HIDRA, the management approach to planning and implementing the system, several key decisions such as developing a legal and IRB framework, a high security environment, an enterprise clinical data processing pipeline, and lessons learned. HIDRA currently contains information on more than 300,000 patients, a self-service application for data exploration, a data request service for access to data through analysts, and a scalable, extensible platform that is portable to other centers. Because this chapter has been submitted as a conference paper for the 2017 AMIA Annual Meeting, the co-authors are included after the title.

1.4.4 Chapter 5. Evaluation of HIDRA

This chapter extends the assessment for Chapter 4. It evaluates the outcomes and impact of HIDRA at Fred Hutch and in the broader cancer research community, as well as the measures of progress such as performance and completion of required features.

1.4.5 Chapter 6. Scalable Clinical Data Pipeline for a Cancer Center

Building on the HIDRA platform work summarized in the prior section (1.4.3), this chapter describes the strategy and implementation of a scalable pipeline for processing clinical documents into desired discrete data elements through both clinical data processing and manual data abstraction. This pipeline is a core component of data acquisition and processing for the HIDRA platform described in Chapter 4. The aim of this chapter was to develop and characterize a clinical data processing pipeline that is scalable to the cancer center enterprise level, is well-supported and sustainable, and can complement or streamline existing manual data abstraction and information-processing activities. This enterprise clinical data processing pipeline will serve to expedite access to usable data that are needed to advance cancer research and improve healthcare operations. By reducing redundancy in abstraction and information-processing efforts, tracking and reducing variation or bias of the interpretation of data points, and making patient history data as up-to-date and complete as possible, this work aims to ease the burden of manual abstraction and improve the timeliness, quality, and availability of clinical data. This chapter was written to submit for publication, and the co-authors are included after the title.

1.4.6 Chapter 7. Comparison of Database Models for Cancer Research

Building on the prior work summarized in sections 1.4.1 and 1.4.2, this chapter describes database models selected or designed to support cancer research and the challenges in their implementation, usability, and sustainability. With a steady increase in the volume and variety of incoming data for cancer research, and the explosion of big-data tools and available database platforms, many centers struggle to discuss and understand the issues and tradeoffs between different approaches or technologies and to develop attainable, sustainable strategies for data management to support research. The aim of this chapter was to develop, model, and assess

database frameworks for cancer. This chapter describes database and data quality concepts, common issues with a variety of database models in cancer research, and lessons learned about the tradeoffs of different modeling and technology approaches. It recommends a general approach to thinking about and evaluating database model and technology options.

1.4.7 Chapter 8. Informatics and Cancer Surveillance: Literature Review and Vision

Building on all of the prior summarized sections, this chapter describes the cancer surveillance data flows, current data collection and information-processing activities, and big-data challenges. The aim of this chapter was to characterize the big-data needs and informatics opportunities for cancer surveillance at the national level. It is primarily a literature review of informatics tools and methods that have been applied to cancer registries, and the identification of opportunities for further research or the application of adjacent research to this domain. This review is a natural extension of my previous work at Fred Hutch on HIDRA and on the pipeline for clinical data processing, as well as the database models work, however it takes these from the individual center level back out to a national scale across multiple centers.

1.4.8 Chapter 9. Conclusions

This chapter summarizes the findings from each of the previous chapters, extrapolates to conclusions that span multiple chapters and describes interesting areas for further research related to or derived from this work.

1.5 Contributions

This dissertation makes a number of contributions to biomedical informatics and cancer research:

- Chapter 2 (Caisis) describes by preliminary work on the Caisis database. The generalizable contributions of Chapter 2 are a working a comprehensive database model and associated web-based tool for data abstraction that is temporally organized and has the ability to stack into an analytic structure for predictive modeling. This system is freely distributed under an open-source license, meets common requirements for IT security, extensibility and supportability, and it has already been adopted and extended by numerous other cancer centers in the United States and internationally.
- Chapter 3 describes recurring themes in informatics and IT from 60 cancer centers. The generalizable contributions of Chapter 3 are the following: First, the volume and variety of data elements that can practically be collected through clinical templates is limited. Second, given the importance of research and collaboration networks, cancer centers should adopt or at least be interoperable with common platforms like REDCap, i2b2, OpenSpecimen and OnCore so that we can wrestle with common issues as a community. Third, due to limited and variable funding for research, solutions need to scale down to affordable levels for individual researchers and labs. Fourth, site visits and active cross-pollination of tools and methods across center must extend deeper into all levels of IT and informatics staff rather than just connecting senior IT leaders and informatics researchers. Finally, centers should spend time and effort resolving social and organizational barriers to progress in informatics and IT.
- Chapter 4 (HIDRA) delivers a cancer data integration platform that is modern, scalable, extensible, portable to other centers, and sustainable through a technology partner, LabKey Software. The lessons learned around the HIDRA legal and IRB framework have been applied to develop IRB files for the Hutchinson Institute for Cancer Outcomes

Research (HICOR) regional data platform and the SCH Research Informatics Platform. The self-service data exploration tool developed for HIDRA, called Argos, is being adopted by at least two other large initiatives. The aim of this chapter was to develop and assess a modern integrated data platform to support a wide variety of cancer research, which explores the following questions: What are the challenges and opportunities for informatics at Fred Hutch? Do these challenges and opportunities align with my previous work and lessons learned? Are there tools and methods that others or I have used before that could be successfully applied at Fred Hutch? How can we enable new types of research, faster results, and better quality of research data at Fred Hutch? How can we improve access to data with tools and methods that are scalable, extensible to new projects, sustainable and portable to other centers? The generalizable contributions of Chapter 4 are the following: First, the legal, IRB, and security framework for HIDRA is relevant to other centers and has been applied to at least two similar efforts. Second, HIDRA provides an example of leveraging a clinical data repository at a broader academic medical center to support a cancer-specific data repository. Third, HIDRA provides an example of adopting and extending the IT and informatics work of other groups to solve local issues economically. Fourth, HIDRA provides an example of an overall strategy for clinical data acquisition, processing, storage and self service data access. Fifth, HIDRA identified the need for a realistic and de-identified testing dataset to facilitate software development and system implementation. Sixth, the HIDRA work found that lack of federated security for a consortium or matrix cancer center is a critical barrier to progress on an integrated data repository. Finally, the HIDRA project found

that the Agile approach to software engineering and system implementation was critical for project momentum and success.

- Chapter 5 (assessment of the impact of HIDRA) provides an evaluation of the impact of the work done on the HIDRA platform at Fred Hutch, described in Chapter 4. It answers the following question: What is the impact of the data platform developed at Fred Hutch?
- Chapter 6 (clinical data processing pipeline) delivers a unique pipeline for both clinical data processing that can be integrated with any database platform. It is extensible, portable to other centers, scalable, and sustainable through partnership with LabKey Software. This pipeline technology is anticipated to be adopted by at least two other groups and gradually replace existing electronic data capture (EDC), data review, and document annotation tools at Fred Hutch once it is completed and deployed into production in summer 2016. The aim of this chapter was to develop and characterize a clinical data processing pipeline that is scalable to the cancer center enterprise level, is well-supported and sustainable, and can complement or streamline existing manual data abstraction and information-processing activities. It explores the following questions: How can we improve the quality, speed, and economics of the acquisition, processing, and delivery of clinical data to support cancer researchers? How can we make clinical data processing a core competency of Fred Hutch at an enterprise scale? The generalizable contribution of Chapter 6 is a tool for shifting the work of manual data abstraction so that it generates training and validation data suitable for development of automated clinical data processing algorithms (e.g., statistical algorithms, NLP, machine learning). This tool was built with a technology partner, LabKey Software, so that it can

be portable to other clients, scalable, and extensible to different clinical data sources and databases. This tool is already being adopted by other groups.

- Chapter 7 (database models) delivers a comparison of commonly discussed database models and big-data tools for the cancer research domain. It serves as a framework to inform informatics and IT professionals, guide the discussion about these technologies, and recommend pilots or overall strategy for big-data platforms at a cancer center. The aim of this chapter was to develop, model, and assess database frameworks for cancer, which explores the following question: How can we best characterize and compare database models and big-data technologies to inform a sensible strategy for cancer research database design and technology?
- Chapter 8 (informatics for cancer surveillance) delivers a description of cancer registry operations, a review of informatics that have been applied to automate and facilitate cancer registration, and a discussion of the current opportunities for informatics research and application in cancer surveillance. The aim of this chapter was to characterize the big-data needs and informatics opportunities for cancer surveillance at the national level, which explores the following questions: What informatics tools and methods have already been applied successfully in the cancer surveillance domain? What are the current and coming opportunities for applying or developing new tools and methods for advancing biomedical informatics in this domain? The generalizable contributions of Chapter 8 are the following. First, it provides a description of cancer registries and cancer surveillance from an informatics perspective, including the case for automation. Second, it contributes a review of informatics tools and methods applied to cancer registries that indicates potential for automation of clinical data processing. Third, it identifies cancer

registries and cancer surveillance as an area for funding and advancing biomedical informatics research.

Overall, these contributions significantly extend my previous work on data platforms for cancer research and create opportunities to accelerate the application of informatics, clinical data processing and machine learning in cancer centers and cancer surveillance communities.

Chapter 2 Preliminary Work: Caisis

2.1 Context

Chapter 2 and the work I describe here serve as background and preliminary to the primary work of my dissertation in Chapters 4 to 8. The work described in this chapter was a first attempt to begin to explore my hypothesis and to answer the overarching question: how can we improve access to clinical and related data about cancer patients for research? It was the genesis of my dissertation hypothesis and research question. The questions I sought to answer with the work in this chapter were the following: What aspects of Caisis are applicable today to current and future informatics platforms for cancer research? What are the limits of Caisis that would need to be addressed with new tools and methods?

2.2 Overview of Caisis Work

Caisis is a web-based application and relational database built to manage information about patients with cancer.(4) Its original intent was to facilitate predictive modeling in prostate cancer using clinical data organized temporally and with consistent coding of variables. The details of the system and lessons learned from its implementation have been presented and published in conference proceedings and a book chapter.(3,5,6)

The original concept and prototypes of Caisis that I developed in the late 1990s at the Memorial Sloan-Kettering Cancer Center (MSKCC) Department of Urology resembled some recently developed analytic database models (e.g., OMOP, i2b2).(7,8) The Caisis database was organized temporally for scalable and consistently coded patient histories that could be easily fed into algorithms that would process patient histories for a variety of predictive modeling projects.

At the time it was developed, populating the Caisis database for research required manual data abstraction and coding of information from medical records. The original Caisis database prototypes were conceived and designed for computer information processing and analysis, and were not intuitive for data entry staff. Several iterations of the database model and user interface were prototyped in Microsoft Access, and over time the more analytic database model and data entry application were refactored to align with the way medical records data were organized (e.g., grouped into pathology, procedures, medical therapy, radiation therapy, lab tests, diagnostic imaging). The adapted database model and user interface were tested and refined by importing and integrating datasets from all cancer treatment modalities (e.g., surgery, radiation, chemotherapy, active surveillance), and over time the database model and user interface were extended for all cancer types.

Importantly, although the core tables in Caisis were aligned with clinical data systems, the database model retained the ability to stack easily into a temporally organized data model for computer algorithms and analytics. This stacking ability was enabled through the consistent definition of fields, the application of naming conventions, and the use of data types. The stackable fields were defined as required fields and were coded consistently across the Caisis database model. To allow extensibility and customization to manage data for new diseases, diagnostics, and treatments, the core database model was decorated with disease- and treatment-specific subtables and forms. A central vocabulary management system was built into Caisis, and the application became increasingly driven by metadata to keep pace with the needs for configuration and customization.

Over time the Caisis database was scaled up to handle a higher volume of cases, to characterize a greater variety of cancer types, and to keep up with the pace of data requests from

investigators. Data feeds for discrete demographic, appointment, and laboratory information from the MSKCC enterprise data warehouse were developed to alleviate data entry. Also, the application was integrated with the clinical practice workflow within the Department of Urology, allowing physicians and clinic staff to collaboratively enter data into web-based forms that would both generate clinical notes and populate the database, thereby reducing the work and time required for manual abstraction. The system was built with robust security controls and audit logs to exceed the requirements for implementation in a HIPAA-regulated environment.(9)

My experience in conceiving, prototyping and leading the development of Caisis at MSKCC from 1998 to 2008 was the foundation for much of the work in this dissertation. Fred Hutch cancer center was an early adopter of the Caisis system. For the Hutch Integrated Data Repository and Archive (HIDRA) project (described in Chapter 4), Caisis was adopted as a stepping stone for a comprehensive cancer database model with a fully functional and extensible web-based application for data entry and management. This decision to adopt and adapt Caisis allowed the HIDRA project to initially bypass the need to design and develop its own database model and data entry application while working instead to implement data feeds, data access tools, and a clinical data processing pipeline (described in Chapter 6).

2.3 Issues with Caisis

Over time, several weaknesses in the Caisis database model and application became evident. As disease- and treatment-specific subtables and fields were added to accommodate different disease- and treatment-specific data points, the model became more difficult for new users to understand intuitively, and many of the new fields were sparsely populated as research projects and priorities shifted. The ability to add “virtual fields” to any table by configuring metadata was added to the database and web application, similar to an Entity-Attribute-Value (EAV) database

design.(10) Unfortunately, querying data from virtual fields has not been intuitive for Caisis or other EAV databases and may not perform well at scale.(11)

There were practical limits to the integration of Caisis with clinical workflows. Although it was possible to capture standardized templated data for both research and clinical practice using Caisis, this feature required re-engineering clinical workflows and was difficult to sustain because it potentially competes with hospital-wide clinical system advances. Moreover, there appeared to be practical limits to the use of templated notes in increasingly busy clinics.

In addition to the limits to clinical integration and templating, the Caisis database model remained difficult to query. As a highly normalized, temporally organized relational database model with an expanding number of subtables and fields, common complaints were that queries returned multiple rows (as a consequence of having normalized data), and many queries required procedures to iterate over rows sorted by time rather than return simple set-based queries.

As described above, the Caisis database model was designed to be stacked into a simple temporally organized analysis model suitable for algorithmic processing and analysis. Originally, a set of algorithms was distributed with Caisis and used to generate datasets for analysis. These algorithms were similar to the Jigsaw tool for the OMOP common data model.(12) However, the domain knowledge and technical skill required to implement these algorithms were too complex for most users or developers. Over time, users stopped using algorithms and slipped back to developing simple SQL queries to fulfill data requests from Caisis. Without simple tools to query the Caisis model and handle the complexities of flattening temporal and normalized data into a denormalized view, many potential users opted for simpler solutions to data management.

In addition to issues with the querying the Caisis database model, most users were more familiar and comfortable with entering data into flat database models like Excel, surveys for

epidemiology research, or case report forms for clinical trials. Although dedicated data entry staff were often able to adapt well to and prefer the Caisis web application, new users and many researchers still preferred flatter, customized data entry forms. To fulfill the need for flatter and more streamlined data entry for both clinical workflows and research staff, Caisis "EForms" were developed. These web-based forms were highly customized, requiring configuration and sometimes additional programming. The forms stored data temporarily in an XML format. When an EForm was finalized and submitted, Caisis would parse the XML and place each data point in the database model. It was an alternative interface, but not an alternative database model.

2.4 Conclusion

As a stable and extensible system, Caisis continues to be used, adopted, and extended by multiple centers. It has proven useful for data collection and viewing or working with data about one patient at a time. However, without sufficient tools to facilitate queries and self-service access to data from a rather complex relational database model, it is limited. As the volume and variety of available clinical data increases, the Caisis database model is proving a bit inflexible, requiring database model changes to accommodate new data elements per disease or organization. Also, Caisis was not designed to handle high-dimensional molecular data and queries of that data. Staff domain knowledge and technical skill are also limiting factors in its adoption and optimal use.

2.5 Synthesis

This chapter (sections 2.2 to 2.4) provided an overview of my previous work in designing and implementing the Caisis database for managing clinical and research data about patients with

cancer, and identified some unresolved issues. This work began to address my overall research question: how can we improve access to clinical and related data about cancer patients for research? The two questions related to this chapter were [1] What aspects of Caisis are applicable today to current and future informatics platforms for cancer research? and [2] What are the limits of Caisis that would need to be addressed with new tools and methods? The answer to these questions is that the underlying Caisis relational database model and web-based interface are applicable to current and future informatics platforms for cancer research in the short term. The platform and self-service data access limits of Caisis are addressed in Chapter 4 (HIDRA). The limits to the web-based interface for data abstraction are addressed at least partly by the pipeline for clinical data processing described in Chapter 6. The limits to the Caisis database model and opportunities to improve it are explored further in Chapter 7.

My experience with dozens of other cancer centers that evaluated or adopted Caisis inspired the preliminary work in Chapter 3. Cancer research and informatics appeared to have evolved over the decade that Caisis was designed and developed at MSKCC (1998–2008), and I wanted to look broadly at IT and informatics trends and lessons learned across multiple centers before rethinking the strategy for another single organization.

Findings from this preliminary work on Caisis are relevant to other cancer centers addressing the need to acquire, process and store data about patients with cancer.

First, there working is a comprehensive and working database model. Its temporal structure and ability to stack into an analytic structure for use in predictive modeling is as relevant today as it was when the Caisis database model was developed. Moreover, the meaningful clinical data elements in the Caisis database model (roughly 2000 unique data elements) that were developed from years of bottom-up data import and integration efforts at

MSKCC map pretty closely to the clinical data elements developed through a top-down comprehensive analysis of cancer data elements conducted at Fred Hutch.(13) Details of the alignment of the Caisis database model and the clinical data element analysis at Fred Hutch are outside the scope of this dissertation, but could perhaps inform or be included in future work.

Second, the Caisis system has a working, extensible web-based interface for data entry and editing. No matter which database model is designed or adopted by a cancer center, the ability for clinicians, researchers and staff to add, edit and view clinical data will be needed. My previous work on the Caisis system provides an interim solution (or even a long-term solution) to data acquisition, processing and storage using data entry staff and traditional ETL data feeds.

Third, the Caisis system is freely available under an open-source license. My experience has been that funding for research systems, especially for departments and smaller cancer centers, is limited and variable. Most researchers cannot afford an expensive, recurring software license and support team. A system like Caisis that is freely available, open-source licensed, and can be ported to other centers with little or no technical support fits the needs of many centers. Caisis has already had broad impact in the United States and international cancer research communities, and it is used by dozens of institutions and groups. Many of these users are listed on the Caisis web site(14) and even recently I have spoken with new users who are currently adopting or planning to adopt this system. Many adopters of Caisis value the existing system because the data structure allows them to easily collaborate with other Caisis users, and because most local adaptations of the system require minimal configuration and programming. Finally, in Caisis user conferences and in-person discussions (from work in Chapter 3), Caisis adopters have told me that its robust security features enable them to pass requirements from their institutional IT security offices that may prevent use of other similar systems.

Chapter 3 Preliminary Work: Recurring Themes in Informatics and IT from 60 Cancer Centers

3.1 Context

This chapter is my first attempt to answer the question "How can we improve access to clinical and related data about cancer patients for research?" and to test my overall hypothesis that new tools and methods from biomedical informatics could improve the availability of data for cancer research if they were applied thoughtfully and strategically.

The question I sought to answer in this chapter was "What are the current opportunities for strategic application of biomedical informatics tools and methods in cancer centers?" The work that I describe in this chapter was background and preliminary to the work of my dissertation (covered in Chapters 4 to 7). The net result of my visits to other centers and analysis of my notes from discussions was the realization that I had preliminary data to answer my question and validate my hypothesis, but that it would take a dissertation to fully explore the hypothesis and question. This chapter was written as a standalone paper, so it has a separate acknowledgements section. However, it is not currently intended for publication outside of this dissertation.

3.2 Abstract

The National Institutes of Health (NIH) is investing considerable funding in National Cancer Institute (NCI)-designated cancer centers and Clinical Translational Science Awards (CTSAs). Biomedical informatics and IT efforts are often expensive and significantly impact organizations, yet leaders and practitioners often know little about the context, efforts and innovations from

other centers. From 2008 to 2009, I visited 60 cancer centers around the United States and met with 394 people to explore issues and trends in informatics and IT.

The information gathered from these meetings was organized using a qualitative analysis tool, and a number of distinct patterns became evident. This preliminary work included findings regarding EMR and clinical data repository implementations, the selection and implementation of clinical research systems, struggles with biospecimen information management, curation and technical support of research databases, the persistence of Access and Excel databases, the rise of caTissue (now OpenSpecimen), Research Electronic Data Capture (REDCap), and Informatics for Integrating Biology and the Bedside (i2b2), trends in database and web development platforms, and social/organizational issues.

3.3 Introduction

The NIH and cancer centers have invested considerable funding into developing informatics and IT infrastructure to support patient care and research. The NCI caBIG program invested at least $350 million in informatics ($20 million annually for the pilot phase between 2004 and 2006, $41.7 million in 2007, $45.8 million in 2008, $43.1 million in 2009, and $100 million for 2010).(2) The CTSA program also actively funds informatics efforts.(15) More than 75% of the current CTSA recipients are also affiliated with NCI funded cancer centers, and almost half of NCI-designated cancer centers have CTSA awards.(15,16) By 2009, $92 million of NIH stimulus funding went to support informatics projects, and many of the funded centers were implementing clinical systems per stimulus-funding incentives.(17)

Although the caBIG, CTSA, and other funded networks fostered greater collaboration and interaction between informatics leaders and practitioners over the past few years, there is

still much room for improvement at cancer centers. Explorations of the trends in biomedical informatics have mostly been anecdotal or conducted by surveys, literature reviews, collaborative authorship, or limited to interviews with local or professional networks.(18,19)

From summer 2008 to summer 2009, I set out to visit every NCI-designated cancer center in the United States. There were several motivations behind and goals for this exploration. This 10-month exploration was primarily conducted on my own time and at my own personal expense. It does not represent the views of MSKCC, Fred Hutch, UW or any other particular group or individual. The draft of this paper has been distributed to all of the people and sites I visited for their corrections, suggestions, and permissions where appropriate.

3.4 Objective

The objective of this work was to form a foundation for my own strategic thinking about the application of biomedical informatics in cancer centers. I wanted to take a systematic look across the United States, to find out what colleagues in cancer informatics, IT, healthcare and biomedical research were thinking, doing, and learning. What were the top issues, struggles, and trends in IT and informatics efforts among cancer centers in the United States? Also, I wanted to gather and synthesize details and perspectives to evaluate crosscutting programs such as caBIG, CTSA grants and the rise of tools like caTissue (now OpenSpecimen), i2b2 and REDCap.(8,20,21)

3.5 Methods

Between August 27, 2008, and June 29, 2009, I met with more than 394 people at 60 centers around the United States (see Appendix A: Sites Visited). A total of 49 of these sites were NCI-

designated cancer centers, including 32 of the 40 comprehensive cancer centers. I also visited colleagues at the NCI, and had phone conversations with informatics colleagues from 2 additional centers, Indiana and Yale. I was unable to visit 16 of the 65 NCI-designated cancer centers, including two in the southeast (Emory and South Carolina) that were added in 2009 after I had driven through that part of the country. Figure 3.1 shows the approximate routes of more than 30,000 miles of driving and 40 flights.

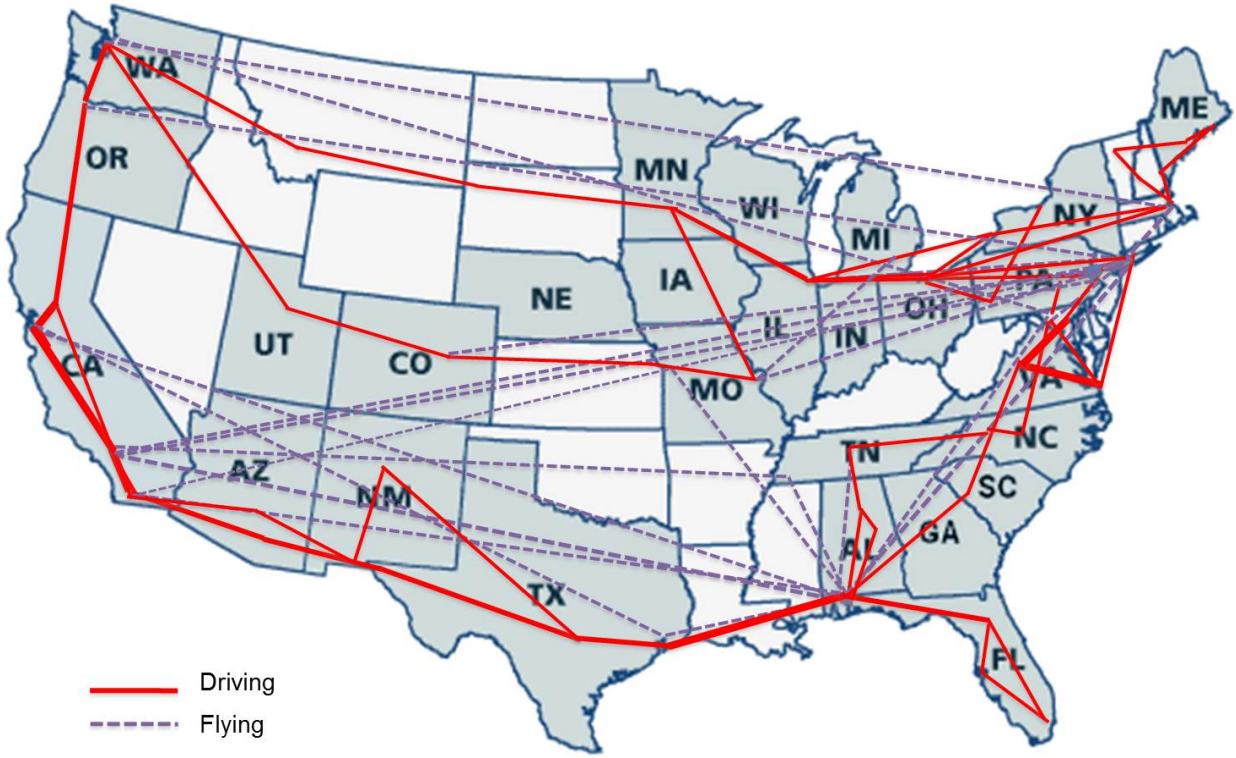

Figure 3.1. Routes of driving (30,000 miles) and flying (40 flights)

For the interviews, I used a semistructured agenda (Table 3.1). I did not record any interviews, but took semistructured notes in notebooks. Because this was an exploration rather than an investigation of a focused research question, I did not direct or confine the conversations to the agenda unless it was helpful to stimulate discussion.

Table 3.1. Semistructured interview agenda

1. IT/clinical systems	**3. Innovation**
1.1. Data center issues	3.1. Telemedicine
1.2. EMR, LIS, operational systems	3.2. Virtual worlds
1.3. Data warehouse/clinical data repository/de-identified repository	3.3. Web 2.0
2. Research informatics	**4. DBMS and web development platforms**
2.1. CTMS systems	**5. Organization/team structure**
2.2. Biorepository systems	**6. Center vision/goals**
2.3. CTSA/genomics data	**7. Local struggles/issues**
2.4. caBIG efforts	**8. Collaborations/networks**
2.5. SPOREs or large program grants	8.1. Data sharing
2.6. Bioinformatics tools, platforms, efforts, and strategy	8.2. Project tracking/productivity

The raw data from four notebooks, my Outlook calendar and email were coded into PersonalBrain 5.5.2 mind-mapping software to facilitate information retrieval and analysis.(22)

3.6 Results

The following findings are roughly organized around the semistructured agenda in Table 3.1:

3.6.1 Clinical Systems Issues

At the time of these site visits, most hospitals and cancer centers were still early in the process implementing electronic medical records (EMR) systems, with Epic(23) being the most common (20 centers). Most sites were also wrestling with the overall complexity of integrating "best of breed" clinical systems modules from a variety of leading vendors: Epic, Cerner, GE Centricity, Eclipsys (now Allscripts) and McKesson (see Appendix A: Sites Visited).(23–27) One

informatics leader stated that “clinical systems are hard to evaluate until they are in-house,” and it is well known that what works well in one organization may not translate well to other environments.(28) Another described their medical center as a “vendor graveyard.” Between EMR implementation teams at different sites, staff expertise and awareness of potential pitfalls varied greatly, and interaction or collaboration rarely occurred across sites to share successful practices or plan for interinstitutional or widespread interoperability.

For the most part, each site was in the process of implementing custom, templated clinical notes. The majority of EMR planning and implementation focused on technical system integration details and the capture of patient care data for billing and documentation, with little consideration of downstream secondary use of clinical data for research or EMR integration with research systems. Most sites did not want to “mess up clinical systems with research data.” However, a few sites were backtracking to attempt to integrate research data collection into clinical system templates or reengineering entire processes.(29) Although many sites were wrestling with similar issues, most teams did not seem to apply lessons from the clinical informatics literature on system implementations to assess or manage factors that are likely to determine success or failure of clinical systems.(30) There was a surprising amount of relearning going on rather than leveraging the research, experience and expertise from other similar groups.

Most sites were planning or implementing enterprise data warehouses (EDWs) or clinical data repositories (CDRs) for clinical systems, which would enable administrative and quality assessment reporting as well as provide access points for research systems to retrieve relevant clinical data from a single source.(31) While most sites were rolling out leading commercial clinical systems, MD Anderson, Dana Farber, and Vanderbilt stood out as significant exceptions, choosing to build their own homegrown systems to integrate research data capture with clinical

practice.(32) Although a couple sites, notably MD Anderson, Northwestern, and the University of Chicago, were striving to implement a service-oriented architecture, for the most part conventional data warehousing practices and extract-transform-load (ETL) data feeds were the dominant approach to EDWs and CDRs. ETL data integration efforts predominantly focused on the highly standardized and/or high-volume data sources such as demographics, laboratory results, scheduling, billing diagnoses, and procedure codes.

The collection of detailed patient medical history data through EMRs was much more problematic, and no emerging trends or innovative solutions were evident in practice. There was an even split between the advocates of structured data capture through standard templates versus primarily narrative data capture with post hoc data abstraction or clinical data processing employed to extract clinical data elements for reporting and research databases. Although most informatics and IT leaders seemed to prefer implementing structured templates in clinical systems, many found this approach untenable in their centers and thought that NLP was the only practical solution to gleaning structured data elements from EMRs.

In general, teams were focused more on EMR system selection and timely implementation than on longer-term research, evaluation, and engineering of clinical processes.(33) More than half of the centers I visited were in early stages (first couple years) of EMR and CDR implementation or reimplementation. In almost all cases, teams were pushing hard to get their systems implemented as scheduled, and considered clinical systems to be on separate tracks from research systems. Most clinical systems groups expected that researchers would be able to access and retrieve all of the clinical data they would need from an EDW or CDR, however a few strategies were different. From the Vanderbilt experience, Stead emphasized a focus on “iterative process of care improvement” over systems

implementation.(29) The Partners and MD Anderson teams stood out in their efforts toward integrating clinical informatics with research dataset production, using NLP and structured data collection, respectively. The different perspectives and focuses of IT and informatics within cancer centers was evident, with the former focused on system selection, implementation, and support and the latter leaning toward system evaluation, design, and development.(18,34,35)

3.6.2 Research Systems Issues

3.6.2.1 Serial Dating with Clinical Research Systems

Most sites were wrestling with and formalizing research systems, particularly for clinical trials and research data management. Implementing a clinical research data management, or clinical trials management system (CTMS), took much greater time, effort, and resources than anticipated, and many sites seemed to be "serial daters", moving toward new, promising solutions after frustrations with current systems. There were no clear winners between commercial, open-source, and homegrown systems. For every solution, there were champions and supporters, as well as critics and disgruntled users. The one research system that generally earned favorable reviews and appeared to gaining ground around the country was REDCap.(21) This tool, developed by Paul Harris and his team at Vanderbilt, looked like a simple disruptive innovation that was already overtaking more complex and cumbersome commercial CTMS software and custom CTMS applications.(36) The only frequent issues mentioned about REDCap were the nonstandard license and its lack of some high-end CTMS features. However, the rapid and widespread adoption to more than one-third of the sites visited may indicate that the license and simplicity of REDCap did not significantly detract from its appeal. Among commercial clinical research systems, more than 15 of the sites visited had experience with Velos(37) or were in the process of implementing it at an enterprise level. Several people from

10 different sites gave rave reviews of Forte OnCore, and mdlogix appeared to be gaining ground at 5 sites.(38,39) Several groups gave favorable reviews of Medidata and PhaseForward (now owned by Oracle) systems, but these solutions were generally considered too expensive for most cancer centers.(40,41) Eight sites had homegrown CTMSs, and a couple sites were using OpenClinica.(42) Many sites were working toward their second or third attempts at implementing a commercial enterprise CTMS, and no clear winners or leaders emerged except for REDCap.

Tracking of patients for clinical trial eligibility and recruitment were common issues mentioned by staff at 10 sites. Tracking of clinical research billing was another frequently mentioned problem for IT and informatics, as well as clinical research document management, IRB systems, adverse event reporting, and system compliance with policies and regulations. Clinical research appeared to be a huge problem domain that most centers wrestle with, as well as an opportunity for informatics research and solutions. The OCHRe effort led by Ida Sim at UCSF, the Prostate Cancer Clinical Trials Consortium, and the caBIG CTMS workspace were noteworthy efforts to develop shared informatics resources for clinical trials.(43–46)

3.6.2.2 Struggles with Biospecimen Information Management Systems

Most centers were struggling with a strategy and systems to support biorepositories and biospecimen tracking. At least 15 centers had primarily homegrown specimen management systems, predominantly in Microsoft Access. There were notable exceptions. Mayo Clinic was implementing the LabVantage research laboratory information system (RLIMS) in a broad and rigorous effort across their sites in Minnesota, Arizona, and Florida.(47) Their third attempt to implement enterprise-level biospecimen information management was well-funded, well-staffed, and included 6 months of interviewing and requirements analysis, a staged rollout, and formal

project management. The University of New Mexico and City of Hope had successfully implemented TissueMetrix and LabWare, respectively.(48) MD Anderson, MSKCC, and Stanford had notable homegrown systems. The Stanford STRIDE system was integrated with the clinical data repository, to facilitate sample discovery.(49) There were ambivalent views on caTissue (now OpenSpecimen) system at most sites. Nevertheless, caTissue seemed to be reaching an adoption tipping point and was considered a standard among cancer centers for biospecimen information management.(20) Twenty-four sites articulated a caTissue strategy, either working to implement it as a production system or as an external interface for data sharing.

Many sites had implemented or participated in specimen locators and specimen discovery tools, however most biospecimen networks had only a small number of participants.(50–55) Like EMR and CTMS implementations, the implementations of biospecimen information management systems were complex and required more resources and planning than most centers estimated. Although good commercial laboratory and biospecimen information management systems were available, these were generally financially out of reach or unsustainable for most cancer centers biorepositories. The open-source caTissue Suite looked likely to become the most widely adopted system among cancer centers.

3.6.2.3 Curation of and Technical Support for Research Databases

Many people talked about the resources and processes needed to support research databases, particularly those sites considering the implementation of new systems such as Caisis. Almost all of the 25 groups I talked with regarding Caisis plus several other sites were interested in discussing how many and what kinds of people to hire, and what skills and training would be needed to implement and support a research database. The data from clinical systems and clinical data repositories were deemed insufficient for research and were often organized and formatted

in ways that were difficult for the uninitiated to understand and query.(56) Tumor registry databases were mentioned several times as a data source for research, with widely disparate views on data quality and suitability of registry data for research. Clinicians generally perceived the clinical data in tumor registries to be inadequate for current clinical research. Pathology and laboratory-based researchers generally perceived tumor registry data to be a good source of clinical data.

Overall, the employment of research nurses, data managers, and data entry staff to abstract information from medical records and curate disease- or study-specific databases was common. Some investigators maintained their own databases, but this responsibility frequently fell to trainees, staff, and students, with relatively high turnover. Technical and content knowledge about databases was often embedded in the original developers. Many groups had hired an Access database developer or web programmer and then became dependent on that individual for upgrades and queries. Lack of use of (or lack of alignment with) existing standards, reinvention of terminologies, variety of design patterns, and divergent system implementations made it difficult to sustain and improve systems that would enable collaborative science. The development and curation of department-, disease- or study-focused systems to support particular research efforts was a need that did not seem to be met by either EDW or CDR efforts alone.(56)

3.6.2.4 Access to and Retrieval of Research Data

Access to and retrieval of data were chronic and vexing problems within most centers. Data repositories were not always optimally structured for research. Although many groups had developed or were developing clinical and research data repositories, the processes, methods, and tools they used for searching, browsing, and retrieving information typically required

knowledge and expertise beyond the level of most investigators or users of data. In some groups, statisticians performed the role of stewards of research datasets and preferred to maintain databases in spreadsheets or SAS files.(57) In other groups, departments or individual investigators supported their own separate data collection and information retrieval efforts.

Access to data was typically limited to technical staff that supported databases, either because of policies, lack of trust, or technical barriers to use. Several of the groups mentioned a desire for self-service tools to perform quick queries on databases, as well as for running simple statistics, survival curves, frequency tables, and graphs that required minimal technical skills.

3.6.2.5 Access and Excel Research Databases Will Not Go Away, but Perhaps There Are Substitutes

Although many informatics and IT experts expressed frustration with the widespread use of Excel spreadsheets and Access databases, they have continued to proliferate. Many investigators and staff stated that it was easier to work with spreadsheet interfaces for small projects, and to maintain some local control and malleability by maintaining data in Access.

> "There is a threshold in terms of easy to use."
>
> "People hate structured data entry; it's too slow."
>
> "Users are addicted to the quick fix with Excel."
>
> "The Excel user interface is attractive, at first..."
>
> "People get a little money and hire someone to build an Access database."
>
> "They want to manage it themselves."

While high-end and high-throughput departments and laboratories may have funding and may be able to reap the benefits of implementing sophisticated systems, there appeared to be a "long tail" of smaller operations and individuals who performed research well with low-cost, simple,

and malleable systems.(58) The smaller databases and spreadsheets approach to data management seemed to work fine for startup operations (e.g., small trials, experiments, and databases). However, when the volume and scope of data and operations increased, the processes and systems did not scale well, so the simpler solutions were untenable for growth. Unfortunately, few evaluations have been done to understand the role and performance of Access and Excel in the biomedical informatics and IT ecosystem. Excel and Access solutions for biospecimen management, research billing, retrospective registries of disease, and clinical trials management were quite common across cancer centers.

The prevalence of Excel and Access in research operations as opposed to clinical areas may be related to the relative size and budget of hospital IT and clinical systems support versus IT and informatics for laboratories and clinical research support. Larger research units were more likely to invest in and implement commercial or custom enterprise systems.

The biomedical and biological investigation community had recently recognized the inevitability of spreadsheets in practice and have published TAB-based formats for reporting experimental results.(59,60) The rapid and widespread adoption of REDCap seemed to be indicative of a need for simpler systems. At the time of the site visits, a few groups were experimenting with Ruby on Rails as a modern version of relatively easy and rapid system development (where Access has dominated in the past). For smaller cancer centers and laboratories, there still seemed to be a need for solutions that were locally developed, controlled, or both. Cross-platform, web-based, and open-source solutions that were cheap or free to labs and individual investigators were preferred.

3.6.3 Technology Trends, Platforms and Tools

3.6.3.1 Database and Web Development Platforms

When people were asked about IT and software development tools and platforms, every major platform had its proponents and detractors, but there were some trends. On the database side, the sites with Oracle site licenses preferred that platform because of its powerful tools and little additional costs for new applications. However, most other sites seemed to be shifting toward using Microsoft SQL Server due to lower licensing costs and acceptable performance. The next most common database platform mentioned was MySQL, and a few groups preferred Postgres. Most groups were attempting to move away from Access and Excel where possible, but there were still many workhorse Access applications in use, and Excel seemed to be around for the long haul, even winning over XML as a data exchange format.

In terms of web development, the Java/J2EE/Tomcat/Apache platform was strongly represented at 14 sites, and there was continued pressure to use that technology stack because of the caBIG and i2b2 tools built on it. The next most popular platform (at 7 sites) was Microsoft ASP.NET and C#. Python with Zope or Django frameworks followed in prevalence, mentioned by 5 sites. Ruby on Rails was mentioned by a couple sites for rapid prototyping of smaller applications, but most larger scale projects were aligning with either Java/Tomcat or ASP.NET/C# platforms. The use of PHP appeared to be dropping, with notable exceptions such as REDCap, and new development using Perl seemed to be falling relative to Python. For bioinformatics programming, Python, R, and BioConductor were the most well-regarded and frequently mentioned platforms.(61) Although there were many existing Perl and PHP applications and libraries, Python seemed to be replacing Perl for bioinformatics data wrangling,

and Ruby on Rails seemed to be slowly replacing PHP for rapid application prototyping and development.

3.6.3.2 Open-Source Systems to Promote Collaboration

The most prevalent open-source research systems were Research Electronic Data Capture (REDCap), the Cancer Biomedical Informatics Grid (caBIG), Informatics for Integrating Biology and the Bedside (i2b2), and Caisis.(36) Discussions about some of these platforms raised strong and mixed responses, both positive and negative. Some people predicted collisions ahead for caBIG and i2b2. There were some strong fans of caBIG tools and methods among cancer center informatics teams, and although many users and developers expressed disillusionment with caBIG, most would continue to participate as long as funding incentives for development and adoption were available. Two caBIG applications were well regarded enough to consider adoption regardless of financial incentives, caTissue (at 16 sites) and caArray (at 7 sites). Although many of the software tools developed through the caBIG program were needed by the cancer research community, most people found them incomplete and difficult to integrate with other systems. Also, most caBIG adopters were piloting these systems rather than "putting them on a critical path." Many people expressed the view that as long as funding and incentives for adoption and improvement of these tools continues, they would keep piloting and helping to improve the technology. Overall, the most valued contribution of the caBIG program was the establishment of a strong social and professional network among biomedical informatics and IT teams across cancer centers.

There was considerable overlap between cancer center informatics and CTSA informatics teams, and the collaborative network established through caBIG may have paved the way for subsequent success in CTSA informatics efforts. Although there were a handful of mixed views

of i2b2, overall the confidence regarding this toolset was very positive, and adoption by 15 CTSA sites was remarkable, especially without financial incentives for implementation. That said, REDCap, described above, was a clear market winner among CTSA sites.(36)

The Caisis system was widely implemented in cancer centers (16 sites), however it was often considered difficult to locally extend and difficult for smaller groups to support. The level of technical skill required to create custom EDC forms in Caisis was a major issue, as well as the lack of tools for browsing and querying data from within the web application. Overall, Caisis has a strong presence among cancer centers, but lack of interoperability with caBIG or i2b2 appeared to be stalling its adoption on the high end, and the simpler REDCap system appeared to be a substitute for smaller studies and groups.

Software licensing cost and local extensibility was a major issue for biomedical software applications regardless of technology platform or tool. Most people I spoke with strongly favored lower-cost and open-source technology. The most widely adopted applications, including caBIG, i2b2, Caisis, and BioConductor were all freely distributed under open-source licenses.

3.6.3.3 Data Storage

Several sites mentioned data storage issues. Lynn Vogel noted that data storage requirements for MD Anderson cancer center were growing at 30-40% per year.(32) There was only one surprising finding in terms of data storage, Isilon (Seattle, WA). Isilon was a high-performance data storage company with a strong presence in the life sciences and bioinformatics, and the company was mentioned repeatedly when discussions turned to data storage technology. Cloud computing for data storage did not appear common yet, probably because its price was still perceived as high relative to the cost of in-house data storage and computing.

3.6.4 Social and Organizational Issues

A significant part of many conversations pertained to social and organizational issues. caBIG and CTSA informatics efforts seemed to be catalysts for discussion and integration efforts across hospitals, universities, and cancer centers. Although variation among CTSA sites was considerable, most sites were exploring new matrix organizational structures and processes, and sometimes crossing organizational barriers that had not been crossed before.

Although many internal integration and coordination efforts were ongoing at CTSA sites, fewer efforts were made to coordinate with other cancer centers. The planning and attention given by individual sites to facilitate participation in research networks and extramural collaborations seemed to correlate with the ease that I experienced in contacting and scheduling site visits. In terms of logistics, it was sometimes difficult - but not impossible - to find the right contacts at each site, but most people contacted were responsive and welcoming. Some sites had web sites that were more transparent and organized with descriptions and contact information for IT and informatics staff. For these centers, it was much easier to find and connect with the appropriate people to interview. Also, internal coordination and communication varied between sites. Some highly organized visits were more the result of central and highly organized individual coordinators than of overall organizational integration, however several leading cancer centers were well organized for interaction with external visitors (e.g., Vanderbilt, Harvard, Mayo Clinic)

3.6.4.1 Software Development Patterns: Build, Adopt or Buy

It is common in informatics circles to proclaim that we are against "reinventing the wheel", but then we proceed to either do just that or to argue that “standards are essential”, as long as they are locally developed standards. Reinvention was very common for simple web-based surveys

and EDC systems. Most groups seemed to prefer to build their own solutions first, adopt open-source software second, and only buy commercial software for enterprise clinical systems. Larger collaborative and open-source software projects like caBIG, i2b2, Caisis, and REDCap had attracted attention and adopters, but still a lot of local software development went on with little awareness of, few designs around, or lack of plans to integrate with systems or standards developed commercially or at other centers. There was generally reluctance to learn foreign technology or applications. Fortunately, many informatics solutions in biomedical research and healthcare seemed to converge on common patterns because the people wrestling with different technology platforms faced the same issues and came up with similar solutions. Unfortunately, most of these locally developed solutions were still incompatible with solutions developed at other sites. Although informatics researchers and leaders may be aware of efforts at other centers, local development teams were generally not.

3.6.4.2 Lack of Widespread Cross-Pollination

One of the key factors that contributes to innovation is the cross-pollination of ideas. Although many researchers reported their ideas at meetings and in papers, practices looked very different in person and in context within organizations. In general, informatics and IT professionals were still locally focused and unaware of comparable efforts and people at other centers. After visiting more than 60 centers, I could not see that any one center leads in all areas of informatics and IT. Each group had strengths and weaknesses, and many projects seemed more impressive and innovative in presentations and papers than they did in person. As in *The Wizard of Oz*, there was always something going on behind the curtain.(62)

The caBIG effort stimulated professional networking and cross-site collaborations, and CTSA informatics leaders were catalyzing greater intersite awareness, exchange of tools and

ideas, and regional collaborations. However, many leaders and staff at all levels would gain valuable perspective from site visits and discussions with their counterparts. I have a much greater understanding and respect for the community as a whole and the work being performed at other centers than when I started these visits. Having this perspective earlier would have changed my strategy for Caisis development at MSKCC toward greater adoption of other ideas and tools from other centers.

3.6.4.3 Learning and Training Issues

One issue that appeared to work against cross-site collaboration and adoption was an initial (almost allergic) reaction by people to foreign systems. I visited a few sites several times, and spent some time training and working with data managers who were using or evaluating Caisis. I noticed that the first couple times they tried to use the system, they were overwhelmed by complexity and immediately saw flaws and gaps or wanted numerous changes and customizations to mimic their current interfaces, artifacts and processes. However, after entering about 5 to 10 cases together, the unfamiliarity and complexity of Caisis were no longer so overwhelming. After entering 50 to 100 cases, users became accustomed to and even attached to the system. This initial reaction and learning curve seemed to apply for clinical systems as well. Many physicians and staff stated that after 1 year of using a new EMR, they were just as efficient as they were before the system was implemented. Developers and project managers may be overreacting to initial complaints and frustrations with new clinical and research software, and it may be better to perform a little "informatics therapy" rather than immediately rejecting or re-engineering a solution.

However, there was still considerable divergence of views, lack of shared understanding in communications, frustration, and separation between IT, informatics, physicians, researchers,

and staff. Several groups were constrained by high expectations and ambitions, unrealistically short project timelines, and inadequate resources, rather than having realistic expectations and a strategy of starting small and compounding success over time.

IT and informatics skills and expertise varied widely among different groups, as did the measurement or enforcement of technical standards for staff. Although clinical research staff and data managers were frequently trained and credentialed according to professional groups such as Society of Clinical Research Associates (SoCRA),(63) fewer informatics and IT staff had comparable training and skill certifications. How can we hire the right people for complex technical projects and keep their IT and informatics knowledge and skills up-to-date? There appeared to be a need and role for broader and ongoing informatics training and certification (e.g., AMIA 10x10 program) as well as IT training and certification.(64)

3.6.4.4 Rise of Networks and Collaborations

Overall, there was a rise in active social and professional networking, the development of collaborations, and collaborative development. The caBIG and CTSA efforts were catalyzing a surge in informatics networks and collaborations within and across sites, and many people were connecting to their colleagues using social and professional networking web sites. However, this behavior seemed to occur mostly among researchers and leaders who participated in external meetings or extramural projects. Within organizations, local teams still demonstrated provincialism and a lack of active networking. Sometimes when I met with groups during a site visit, it was the first time that local IT and informatics staff had met each other. Perhaps site visits and active networking could be catalysts for innovation, adoption of foreign tools and ideas, and extramural collaborations.

The people that collaborated extensively noted that "collaboration takes energy." In the Caisis team at MSKCC, approximately 20% of the group's time and effort was dedicated to supporting the broader community and collaborations. Also, communication and cognitive load are associated with collaboration. Interinstitutional and highly collaborative teams and projects used tools to facilitate management and communication, especially web-based conference systems like GotoMeeting, WebEx or Adobe CONNECT, and information/document management systems such as Microsoft SharePoint or Google Groups. Unfortunately, only a handful of centers had enterprise-wide deployment of web conference systems. Individuals without enterprise options seemed to be acquiring their own laboratory or department accounts for web-based collaboration tools. A few people were using Skype for communication with other sites, but in some places, firewall issues were keeping this tool from becoming widespread.

3.7 Discussion

3.7.1 Limitations of This Research

The greatest weakness of this exploration was inadequate sampling. Because these visits were largely self-funded, and there were significant challenges in scheduling and planning both meetings and a travel itinerary in a pattern that would give me a chance to finish before July 1, 2009, I had to shorten or forego a few visits along the way. I only managed to visit 49 of the 65 NCI-designated cancer centers that I had intended to visit, and at many centers I did not meet a broad or representative sample of leaders and staff (see Appendix A: Sites Visited).

At many sites, I initially made contact through my own professional networks, so there was some bias and variation in the types of people encountered at each site. I did not reach all of

the people with comparable roles at each site due to organizational differences and lack of time. Instead, I talked with as many convenient and amenable contacts as possible.

With a few exceptions (i.e., City of Hope, Johns Hopkins) the visits were short, with little time for snowball sampling or multiple meetings. Also, due to travel and meeting constraints, I did not have sufficient time and energy to codify and process all of my notes after each visit as is customary for more rigorous qualitative research in biomedical informatics. Therefore, the observations herein may be somewhat broad rather than deep. However, the goal of this work was to explore and assess broad trends from a broad sample rather than to pursue a particular research question deeply. Informatics and IT at cancer centers were evolving rapidly and these results may already be somewhat dated, though hopefully still relevant.

3.7.2 Clinical Systems Implications

Although most sites were rapidly implementing EMRs, relatively little integrated planning for downstream research use of clinical data or interoperability with interinstitutional research networks existed. This did not bode well for the broad development of research data networks and the possibility of comparative effectiveness research across sites. Though it may be easier to align data structure and format across institutions that are using the same EMR vendor, most teams were only implementing locally customized templates. Given that path, we may still be looking at a massive data mapping effort to enable meaningful exchange and comparison of patient care and research data.

All of the centers I visited had a strong research mission, and there seemed to be a huge opportunity for cancer centers to partner with each other and for their information system vendors to develop common approaches and templates to facilitate data exchange, interinstitutional research collaboration and healthcare quality reporting. Several individuals

expressed interest in these types of efforts, along with frustration with the current narrow (nonresearch) focus of clinical systems implementations. There were opportunities for hospitals with similar clinical and research priorities and systems to actively network and cross-pollinate successful practices, ideas, innovations, evaluations, and lessons learned. Most centers were early in the EMR implementation process compared to the EMR trailblazers like Vanderbilt, Kaiser-Permanente, and Partners, and the majority of these technology adopters appeared to be relearning painful lessons. Reaching out to, visiting, and partnering with colleagues at other centers who are further ahead in aspects of their IT and informatics efforts could enhance planning and reduce the risks of current IT and informatics efforts.

3.7.3 Research Systems Implications

Overall, clinical research systems were somewhat of a mess. Although there were some good commercial research systems available that were suitable for industrial customers, most were financially out of reach of academic medical centers and individual departments, laboratories, and investigators.

Many investigators continued to gather data locally within laboratories and departments because data supply chains that crossed organizational divisions were neither efficient nor reliable. Many researchers could not trust other groups to respond quickly and provide data needed for their projects to meet tight deadlines and shifting priorities in a dynamic research environment.

In terms of systems and people, we may need to harmonize and formalize policies, standards, and procedures (e.g., consent documents, biorepository standards, data sharing policies, access procedures). Although much has been published, presented, and discussed around standards, data integration, and the semantic and syntactic alignment of systems, there

has been relatively little acknowledgement of (or research into) issues of integration and interoperability of people. As with clinical systems, research knowledge, processes, and workflows seemed to be largely and tacitly embedded in people rather than in information systems. People integrate research systems, but systems could also integrate people. Research workflow management and related methods and tools that enhance the integration of investigators and diverse research teams could be a fruitful investment at cancer centers.

Perspectives on large research information system projects like Caisis and caBIG seemed to disagree. From central vantage points, these projects looked much more useful and popular because most communications to their core teams come from people who are using or want to use the system. Most presentations and papers were biased toward views of promoters, developers, and adopters. For a more balanced perspective, developers and managers of larger projects like Caisis and caBIG should look externally and systematically for minimally biased feedback and input.

3.7.4 Technology Trends, Platforms and Tools Implications

Information systems are social entities in situ. A beautiful system designed and built in isolation may, like its author, work well in isolation and narrow scope, but such a system may not play well with other systems and be a social failure. Social systems (e.g., informatics to support interinstitutional collaboration) are inherently messy. Boundaries are drawn idiosyncratically around system modules in a similar manner to the way boundaries emerge around individuals and groups within organizations. There is no perfect technology or superior system, and the grass always looks greener on the other side - or when someone is presenting their work in public.

Informatics leaders and staff at all levels should invest time and effort in understanding and leveraging technology from informatics and IT colleagues at other centers, even if candidate

foreign solutions do not initially appear to perfectly meet our requirements. It is nearly impossible to make great achievements in informatics without standing on the shoulders of pioneers. New and foreign platforms and tools can be catalysts for change, even a change of requirements. Employing IT and informatics as a catalyst for change can be uncomfortable at first and may subject users and developers to learning curves that may take months to traverse.

In terms of tools, it looked like Java and C# based platforms were still the leading platforms for enterprise systems development, and they did not tend to coexist well in development teams. There was definitely a role in informatics for rapid prototyping using a good scripting language and framework like Ruby on Rails or Python on Django.

Most academic groups still could not afford professional software development. There was a strong preference for open-source and free tools among academic centers and informatics teams, though many would pay much more to develop in-house technology or adopt open-source software rather than purchase proprietary software or hire professional developers. Some grant or contract-funded development efforts at cancer centers outsourced or partnered with software development companies, and several of these companies have risen to prominence among cancer centers (e.g., ConvergeHealth,(65) 5AM Solutions,(50) and LabKey(66)). Many of their products are open-source. What are the respective roles of informatics and software engineering businesses versus academia in the current environment? How can we optimize this relationship to balance innovation with scalability and sustainability of systems? Regardless of specific tools and technology platforms, partnerships between business and cancer centers for software development seem to be a trend that is likely to grow.

3.7.5 Social and Organizational Implications

Interestingly, many centers I visited were working toward similar strategic goals, which usually took the form of leading the field in certain diagnostic or treatment approaches or growing disease-specific programs. However, at most sites, culture seems to eat strategy.(67)

Some notable and innovative strategies in informatics and IT involved the linking of systems and people into data integration and collaboration networks that cross departmental and organizational barriers. Several CTSA informatics groups were using EDW and data integration efforts to break down internal silos, interacting with their CTSA partner sites and other institutions in their region, and leveraging i2b2, caBIG, and other information technology to build collaboration networks for sharing research data. The individuals and institutions that integrated across silos and lowered barriers to extramural collaborations seemed most likely to lead the field. Marked differences are likely to emerge between leading, average, and trailing centers in terms of informatics that facilitate integration and connection of data and people across organizational boundaries. Sites with strong and growing regional informatics networks (e.g., Boston, Chicago, Texas, Pennsylvania, Minnesota, and the Pacific Northwest) appear likely to lead. The best competitive strategy for informatics may be to become an organization that coordinates and plays well with others both internally and externally. Individuals and groups that are self-funded or largely funded by donors may choose to go it alone, but it probably pays off in the long run to make systems as open, accessible, and widely applicable as possible, and to build collaborative opportunities by helping others catch up.

3.7.6 Rise of Design and Visualization

Last but not least, increasing value is being placed on visual and interaction design. Data visualization and making things look good is important for informatics and IT efforts. One of the reasons the Caisis system has been successful is its visual appeal. However, most user interfaces and reports from enterprise clinical and research systems are not visually appealing. Even minor changes in visual design, such as adjustments to layout, fonts, and color palettes can have a large effect on user satisfaction.(68) There is a vast need and role for visual artists and designers in informatics and IT projects.

Systems thinking, workflow design, usability, and interaction design are still major issues in IT and informatics implementations in practice. Many people I talked with reported complaints and annoyances with current systems. Many systems may work as specified but are just not visually appealing, particularly complex enterprise-level clinical systems such as EMRs. In centers that applied a "best of breed" approach to system selection, individual systems may be intuitive, but the overall informatics ecosystem that users encounter in their environment was often not intuitive.(6,29) User interface layout and behavior differences seemed to compound overall frustration and perceived complexity. I noted numerous complaints about login difficulty, password rules, colors, fonts, drop-down lists, tab and field order, excessive clicking in navigation, scrolling behavior, layout of fields on screens, and performance issues. We still have a long way to go to improve user experience in informatics and IT, and perhaps we should consider partnering more graphic artists and interaction designers with developers and informatics researchers.

3.8 Conclusion

With the significant investments in the United States on biomedical informatics and healthcare IT research, infrastructure development, and systems implementations, we need to better understand the overall trends, gaps, and opportunities among leading centers so that we spend resources wisely. Otherwise, we may squander the opportunities for progress to improve healthcare quality and accelerate collaborative research. Although there are many highly skilled and knowledgeable individuals, good informatics methods and tools, and outstanding leaders among the NCI-designated cancer centers and CTSA sites, the depth of cross-pollination between groups and centers is unfortunately still limited below the senior investigator and leader level. We are not yet making the best use of existing expertise and resources. We need to actively promote broad and deep interactions, partnerships across sites, and alignment of efforts in order to minimize duplication and missed opportunities.

Although this research was based on interviews - qualitative, but not rigorously so - it demonstrates that systematic networking and information gathering across sites is feasible and inexpensive relative to the knowledge, perspective, and contacts gained. As these 60 cancer centers move forward with systems implementations, organizational and process improvements, software and standards development, and biomedical informatics research, there will be many opportunities to connect with and learn from colleagues, leverage extramural research and tools, and help build infrastructure and relationships that improve our healthcare and research system. We should help make that happen.

3.9 Acknowledgements

This work would not have been possible without the generosity and support of many people. In particular I have to thank the following individuals: Peter T. Scardino, Wendy Perchick, Elizabeth Roby, and the MSKCC team; the NLM biomedical informatics PhD programs at Stanford, Vanderbilt, OHSU, University of Virginia, and UW; Isaac Kohane and Harvard CBMI; Bruce Trock, Steve Goodman, and the Johns Hopkins team; Bob DiLaura and Mike Kattan at CCF; Philip Kroth at University of New Mexico; Celestia Higano, Peter Tarczy-Hornoch, Tom Payne, and the Fred Hutch/UW Cancer Consortium team; and all of my friends and family who gave me a place to sleep or a home-cooked meal. The following individuals generously provided corrections and suggestions for improving this manuscript: Peter Tarczy-Hornoch and Amy Abernethy. This work was partially supported by National Cancer Institute Grant R01-CA119947.

3.10 Synthesis

This chapter described my visits to 60 cancer centers and conversations with 394 people. It identified trends in IT and informatics through the analysis of notes from those visits and the review of related literature. The lessons learned from the broad analysis in this chapter informed the strategy and design of the HIDRA system at Fred Hutch cancer center described in Chapter 4.

Sections 3.2 to 3.8 of this chapter began to address my overall question of "How can we improve access to clinical and related data about cancer patients for research?" and also informed my development of my overall hypothesis that new tools and methods from biomedical informatics could improve the availability of data for cancer research if they were applied

thoughtfully and strategically. My specific question for Chapter 3 was "What are the current opportunities for strategic application of biomedical informatics tools and methods in cancer centers?" The answer lies not in specific tools, but rather in approaches that focus on iterative process improvement rather than technology, decoupling or loosely coupling with commercial EMR, CTMS, and biorepository system implementations to minimize project risks, focusing on high security, open-source and common platforms, the skills and expertise of existing staff, and making it easy for groups to play well with other groups and centers.

The net result of my initial exploration of the question of what we can learn from the success and failures of IT and informatics in 60 cancer centers validated my decision to pursue a PhD at the University of Washington. It also led to a refinement of my question to "How can we improve access to clinical and related data about cancer patients for research?" and the generation of the overarching hypothesis that new tools and methods from biomedical informatics could improve the availability of data for cancer research if they were applied thoughtfully and strategically. The first aim of my dissertation was to apply the lessons learned from Caisis (Chapter 2) and my cancer center visits (Chapter 3) to develop and assess a modern integrated data platform to support a wide variety of cancer research, which is the focus of the next chapter.

Findings from the preliminary work in Chapter 3 might be relevant to cancer centers addressing a variety of common research challenges.

First, cancer research IT and informatics leaders should not count on clinical templates and data warehouses to solve research data problems. Perhaps a portion of cases and a proportion of clinical data elements can practically be collected through templates -which is valuable- but

the feasibility of obtaining the vast majority of clinical data elements in discrete form through clinical templates - at least in the next few years - is low.

Second, given the rise in and need to participate in research and collaborative networks, and the increasing costs of solving IT and informatics problems alone in a single cancer center, it will be important for centers to align with technology trends emerging across centers in the interest of interoperability, cross-pollination of ideas and lessons learned, and leveraging investments. Given that no system will be a perfect solution, rather than "serial dating" and looking for a shiny new CTMS or biospecimen management system or writing yet another local solution, cancer centers should when possible stick with the systems that are broadly used by other cancer centers (e.g., REDCap, i2b2, OpenSpecimen, OnCore) so that we will all wrestle with issues together.

Third, cancer centers should provide solutions that work for smaller labs and yet could have economies of scale at the institutional level. Rapid or self-service access to data by staff with limited technical skills is a common need, and most researchers favor solutions that use common, low cost, open-source technology platforms. Due to limited and variable levels of research funding in labs, most researchers cannot and will not commit to paying high, recurring software license and support fees.

Fourth, the awareness of the work of other centers and the commitment to collaborate, leverage and be interoperable with the work of other centers - to play well with others - is paramount for cancer centers. The cross-pollination of tools and methods should extend beyond just participation by leaders and researchers in professional meetings and conferences. Centers would benefit from regular IT and informatics related site visits to other centers and deeper

involvement of all levels of IT staff who are actually implementing and operating systems rather than just connecting senior IT leaders and informatics researchers.

Finally, centers should spend time and effort on overall technology strategy, organizational structure and management. Many of the barriers to progress in informatics and IT systems are due to social and organizational issues such as recruiting, retaining and managing talented staff, and integration of local staff with external technology partners and collaborating institutions.

Chapter 4 Hutch Integrated Data Repository and Archive (HIDRA): Data Platform and Clinical Research Informatics Strategy for a Consortium Cancer Center

4.1 Context

This chapter is my exploration of Aim 1, to develop and assess a modern integrated data platform to support a wide variety of cancer research. To address this aim, I answer the following questions for Chapter 4: What are the challenges and opportunities for informatics at Fred Hutch? Do these challenges and opportunities align with my previous work and lessons learned? Are there tools and methods that others or I have used before that could be successfully applied at Fred Hutch? How can we enable new types of research, faster results, and better quality of research data at Fred Hutch? How can we improve access to data with tools and methods that are scalable, extensible to new projects, sustainable, and portable to other centers?

Answering these questions helps inform the overall research question "How can we improve access to clinical and related data about cancer patients for research?" and Aim 1 is logical first step to test my hypothesis that there are new tools and methods from biomedical informatics that could improve the availability of data for cancer research if they were applied thoughtfully and strategically. The text of the chapter was shortened and submitted on date March 8, 2016, as a paper to the AMIA 2016 Annual Symposium, which explains the inclusion of the authors and title of the paper after section 4.1. For this paper, I came up with the initial strategy for HIDRA, led requirements analysis for HIDRA and Argos, and wrote the paper. The

other authors were involved in implementation of HIDRA and edited the paper. Because this is a stand-alone paper I have included a separate acknowledgements section.

Hutch Integrated Data Repository and Archive (HIDRA): Data Platform and Clinical Research Informatics Strategy for a Consortium Cancer Center

Paul A. Fearn, MBA[1,2], Sarah Ramsay, MPH[1], Emily Silgard, MS[1], Adam Rauch, BS[3], Kristin Dubrule, BA[3], Peter Tarczy-Hornoch, MD[2]

[1]Fred Hutchinson Cancer Research Center, Seattle, WA; [2]University of Washington, Department of Biomedical Informatics and Medical Education, Seattle, WA. [3]LabKey Software, Seattle, WA

4.2 Abstract

HIDRA is an initiative to develop a data platform to support current and next-generation clinical and translational research for a consortium cancer center. This paper describes the vision and strategic goals for the system, the management approach to the planning and implementation of the system, several key decisions such as the development of a high-security environment and an enterprise clinical data processing pipeline, and lessons learned. HIDRA currently contains information on more than 300,000 patients with cancer, a self-service web application and data request service for access to data, and a scalable, extensible data integration platform.

4.3 Background and Rationale

Advancement of biomedical research and improvement of patient care at cancer centers are dependent on ready access to data and the processing of that data into usable information. Many clinical trials, correlative studies, and translational research studies that require patient information acquire their own copies of these data from source systems (e.g., electronic health

records [EHRs]) through abstraction and data exports or reports, or they may implement their own customized databases for a particular disease (e.g., breast cancer), treatment modality (e.g., surgery, radiation therapy), or individual study. Often, this results in the same patient's records being abstracted repeatedly for different research efforts. This approach is expensive, duplicative, and difficult to scale. Moreover, as we increasingly explore molecular data and the expression of cancers across populations of patients, the lack of integrated data on a cancer center's patient populations across different diseases, treatment modalities, and studies locks us into siloed thinking and outdated models of disease. This legacy structure is a barrier to broad data mining, molecular diagnostic modeling and research, and precision medicine.

In the bioinformatics world, the standardization of the processing, normalization and analysis of genomics data have allowed rapid advances in analytic methods and tools. Although there are a few advances toward standardized patient phenotyping, in general the acquisition, integration, and processing of clinical data is far behind bioinformatics pipelines, and clinical research is dependent on project-specific data abstraction, physician data entry (i.e. templated notes), and data exports from clinical systems or data warehouses.

To support current and next-generation clinical, correlative and translational research in cancer, there is a need to [1] acquire patient data from disparate sources and integrate it into common data model, and [2] provide an easy-to-use front end that allows analysis and visualization of patient data while maintaining regulatory compliance.

The data acquisition goal requires developing and implementing standard and scalable methods for acquiring, integrating, and processing patient data into reusable data models and structures, and to automate this pipeline for clinical data so that centers can afford and keep pace with the increasing volume and variety of information generated. Past efforts to provide a fabric

of interoperable data and systems at a national level have had mixed results. Some, like the eMerge program,(69) have been relatively successful; others, like the caBIG program,(36) have been spectacularly unsuccessful. Most large and academic centers have implemented or are in the process of implementing common tools (e.g., REDCap(21) or an enterprise clinical trial management system [CTMS]), integrated data repositories with self-service access (e.g., i2b2(8)), and common data models (e.g., Caisis, OMOP).(3,70) Although these steps toward common data infrastructure, format, and access are important, much of patient information is still generated or transferred in narrative or nonstandardized text (e.g., pathology and radiology reports, clinical notes, treatment summaries).

Many centers have ongoing efforts to implement templated medical records with standardized fields, terminologies and text in an effort to have more useful data for downstream use. However, templated reporting has practical limits, as anyone experienced in such efforts can testify. The volume, density, and variety of information conveyed through patient care dialogues and the resulting narrative or nonstandardized text in EHRs far exceeds what can practically be communicated through standardized templates. Templates and drop-down lists are not an adequate medium for communication of rich information. It is frustrating, painful, and limiting for physicians to communicate and document in this way, and with the recent maturation of NLP and machine-learning tools and methods, this strategy is no longer necessary. Templates and standardization are great approaches when used judiciously and for information with relatively little variation, volume, or time sensitivity.

The analysis goal requires the development of easy-to-use tools for narrowing the patient population to specific cohorts of interest and analyzing the characteristics of these cohorts via domain-specific visualizations and reports. Ideally, these are provided as self-service tools,

usable by scientists in a secure environment that protects confidentiality. The system must restrict each researcher's access to the patients and fields required for their research.

Patient privacy and information security are increasingly important concerns for processing cancer center clinical data. The number of healthcare data hacks and breaches is rising, and to remain competitive for grants and contracts that involve the hosting of sensitive information, as well as mitigating local risks of breaches of patients' health information, cancer centers must have a high-security environment strategy that can evolve with the risk landscape.

Fred Hutch is a private research institute that leads the Fred Hutch/UW Cancer Consortium, an NCI-designated comprehensive cancer center.(71) As a research institute, Fred Hutch does not provide patient care or own medical records systems. It is dependent on its clinical partners (UW, SCCA, and SCH) for access to data from medical records for the Consortium patient population.

Historically, like most centers, Fred Hutch had a disparate collection of study-specific databases, clinical research data systems, and data feeds from clinical systems. The bone marrow transplant program at Fred Hutch had a robust data management system developed over 30 years, but the center had a growing need for adequate data and systems to support research for solid tumors and hematologic malignancies that were not associated with transplants.

This paper describes a five-year strategic initiative of the people, processes, and technology applied to raise the level of data and informatics capabilities of the center and to make patient and related data more readily available to Consortium investigators in a way that minimizes long-term costs and risks and provides a scalable, extensible platform for current and future research.

4.4 Methods

In 2011, Fred Hutch leadership initiated a project to develop a cancer research data warehouse to provide a competitive data and informatics platform for its Cancer Consortium with UW, Seattle Cancer Care Alliance (SCCA), and Seattle Children's Hospital (SCH). At that time, through its CTSA informatics core, the UW had recently implemented a clinical data repository (CDR) based on the Microsoft Amalga system (now Caradigm CIP, Bellevue, WA). Fred Hutch had the opportunity to partner with and leverage the expertise and data integration work of UW.

In early 2012, strategic and high-level requirements and use cases were gathered from a representative sample of Fred Hutch and Consortium investigators involved with clinical, correlative, and translational research for the Consortium cancer patient population from UW, SCCA, and SCH. These requirements and use cases were augmented through literature review, information gathered from visits to more than 60 other cancer centers in 2008–2009, and searches for publicly available information on web sites of NCI-designated cancer centers.

The HIDRA platform was envisioned as a system to integrate data about [1] patients and research subjects, [2] biospecimens, [3] clinical trials and other studies, and [4] molecular assay results or other associated datasets. The initial scope was the Consortium patient population, but the infrastructure was intended to scale to other uses.

The initial strategic goals of HIDRA were to:

- Enable the Consortium to learn from every patient who comes through the door, and integrate that knowledge back into the clinical care. This goal includes use of clinical data for research, research operations, healthcare operations, quality improvement, and public health reporting purposes.

- Provide an integrated approach to data, with easy access for clinicians and investigators. This goal includes integration of data and systems across all cancer types, linkage of specimen and molecular data with clinical data, and integration with Consortium study and biospecimen management systems.
- Automate or facilitate human information processing. This goal includes scalable manual data abstraction, EDC, data feeds and clinical data processing from medical records. It also includes the acquisition and linkage of outcomes data (e.g., from tumor registry, long-term follow-up, patient-reported data)
- Provide a strong competitive platform that is ready to meet HIPAA, FISMA, or FDA regulatory reviews or audits.

The IT and informatics goals for this resource were to:

- Provide a reliable and scalable hosting environment from storage through database and application servers
- Provide a FISMA-ready security environment that would meet or exceed HIPAA business associate requirements
- Consolidate data feeds from UW clinical systems through the UW CDR
- Provide a self-service interface for Consortium investigators to find, filter, and acquire data
- Provide the ability to integrate biospecimen data
- Provide the ability to integrate with an enterprise CTMS
- Provide the ability to link to and query a variety of molecular or other related datasets

- Develop an enterprise clinical data processing pipeline to extract and process information from narrative reports and text fields
- Integrate clinical data processing with human abstraction and review workflows, to enable a gradual transition to automated techniques and increase the reliability of these algorithms
- Provide the ability for this system to play well with other systems and collaborative partners so that is not just a Fred Hutch solution
- Provide an affordable and sustainable solution through use of lower cost or open-source options if viable

In spring 2012, the HIDRA steering group issued an RFI to five potential technology partners to assist with requirements analysis and development of the platform. In summer 2012, LabKey Software (Seattle, WA) was selected as the technology partner based on their clear understanding of the requirements and demonstrated ability to develop large-scale biomedical data management systems. Project governance was established, including a HIDRA steering committee and a vision committee to align IT and informatics with the goals of investigators and institutional initiatives. Detailed requirements working groups were initiated in fall 2012 with Fred Hutch IT and informatics staff, technical partners from LabKey Software, as well as counterparts from UW, SCCA, and SCH. This team engaged in a 90-day planning effort for the HIDRA core infrastructure and a data transport layer from the UW CDR system. The detailed requirements analysis for HIDRA began with a map of existing systems, such as the UW CDR, the Gateway bone marrow transplant database, a variety of disease- and project-specific solid tumor and hematologic malignancy databases, the Cancer Surveillance System (CSS) SEER registry,(72)

survivorship program databases, and other research data sources and applications associated with the Consortium patient population. Requirements analysis covered data feeds and connections between these systems, as well as data, relevant terminology standards, input/output, site-specific integration, data access (e.g., user interfaces, application program interfaces), information security, regulatory compliance, operations, communications, training, and documentation. The current state of data and informatics at Fred Hutch, high-level Consortium data requirements, and an initial vision and strategy for an integrated data repository were presented in October 2012 to the Consortium external advisory board. Detailed requirements and an implementation plan for the HIDRA core platform were completed by December 2012, and the core platform and data feeds from UW CDR were started in 2013. It was estimated as a 3- to 5-year project.

To facilitate the development of HIDRA, the Caisis database system was used as an initial data model that could accommodate data from all cancer types and as an existing set of user interfaces for data entry and management. Fred Hutch had used the Caisis system for prostate cancer data management since 2003, and the informatics staff was very familiar with operating and configuring it. Moreover, the data model and metadata-driven user interface in Caisis were remarkably compatible with the LabKey Server platform.(73)

Two key efforts that provided the legal and human subjects protection foundation for HIDRA implementation were [1] the establishment of a legal memorandum of understanding (MOU) for data sharing among the Consortium partner institutions and [2] the consolidation and rationalization of IRB files for data repositories (see Figure 4.1).

The MOU document contains a very clear description of the intended scope of HIDRA and principles of data sharing for Fred Hutch, UW, SCCA, and SCH. The document formalizes allowable activities (e.g., research and research operations, healthcare operations, quality

improvement, public health reporting), a detailed list of clinical data feeds (e.g., labs, clinical notes, demographics, radiology, pathology), the purpose for each of these data types in HIDRA, a data security agreement, and a HIPAA-compliant business associate agreement (BAA). It took one full year to develop this MOU, which was approved in December 2013.

Once the MOU was completed, a HIDRA IRB file for the data repository was created. This file described the scope, operations, and uses of HIDRA, including a summary of data access and use procedures and a description of the high-security environment. The HIDRA IRB file does not have its own foundational informed consent form. Rather, it relies on the MOU terms, BAA, waiver of consent, and waiver of HIPAA authorization to allow the transfer of records for the entire Consortium patient population into its high-security environment (see left side of Figure 4.1). However, any access to or provision of these data must meet requirements of both the MOU and IRB. Many of the patients in the HIDRA database have signed foundational consents to use their health information for research. The HIDRA IRB file references these foundational consents. Any data request from HIDRA, through an analyst or through self-service access, must either be for completely coded data with no protected health information (PHI), or must have IRB approval (see right side of Figure 4.1). After establishing the HIDRA IRB file, at least nine other data repository IRB files were modified to allow the consolidation of their data into the HIDRA system and to allow access to their data through the HIDRA gatekeeping process for data access and use. It took another full year to develop and approve all of these IRB files and modifications, which were completed in December 2014.

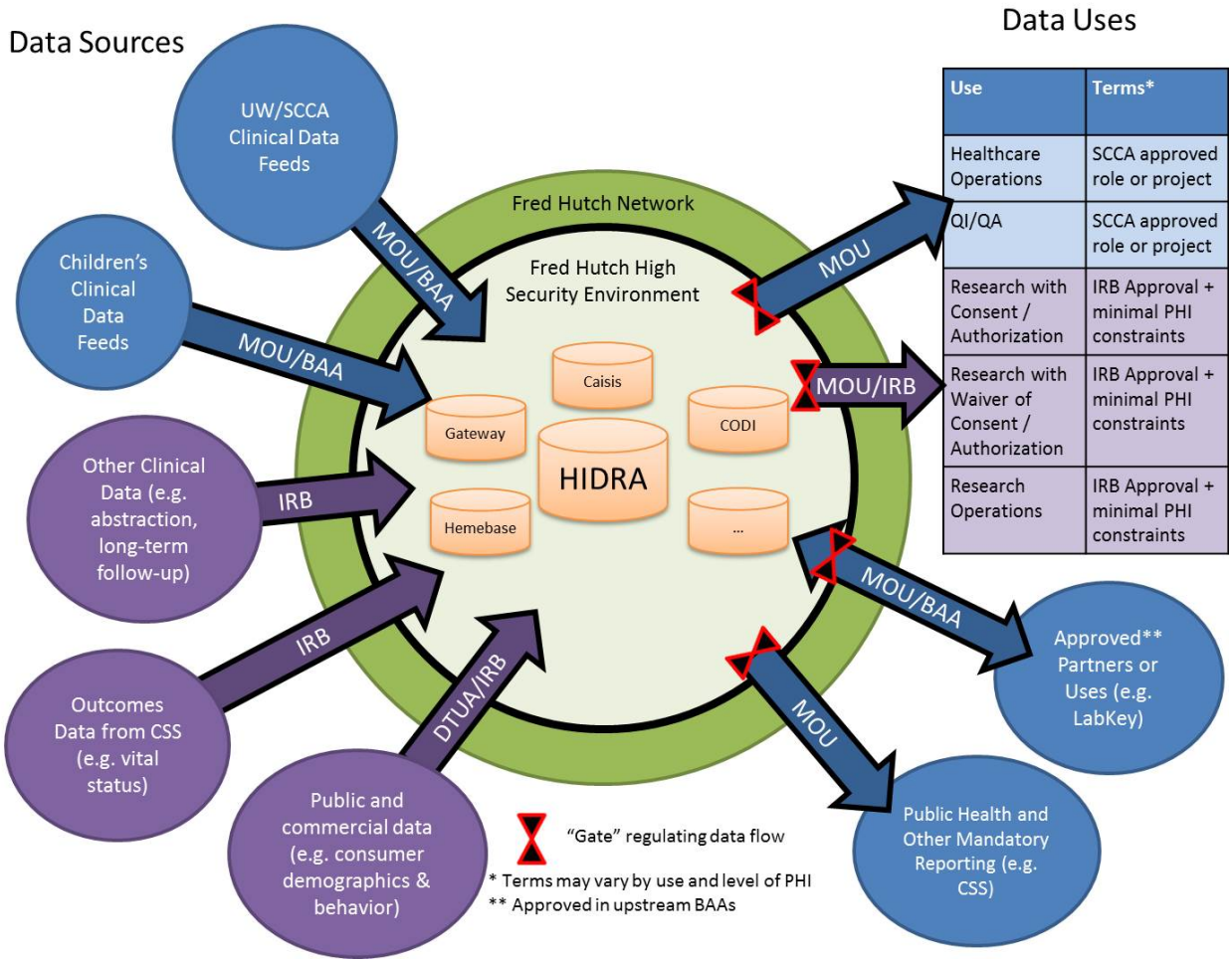

Figure 4.1. Conceptual diagram of the legal and IRB framework for HIDRA. A legal memorandum of understanding (MOU) and business associate agreements (BAA) between Consortium partners created the basis for data sharing for a variety of uses, the terms of use, and the requirements for high-security environment. The HIDRA IRB file and additional data transfer and use agreements (DTUA) are aligned with the MOU and describe the operation of HIDRA as a data source for other IRB-approved studies.

As a result of the MOU agreement, and the combined security and regulatory goals of the project, while the technical teams were advancing on implementation, a document policy and training program were developed, which resulted in 66 people completing 790 online trainings on more than 76 policies and procedures. An exhaustive Data Access Training was developed in conjunction with all Consortium regulatory partners to ensure that all users entering the system and requesting data had a thorough understanding of the proper handling and use of the data.

For self-service access to data, detailed requirements analysis, design, and implementation planning began in summer 2013. The planning team consisted of informatics personnel, IT project management, and senior LabKey Software staff. Over a six-month period ending December 2013, the planning team met with 13 disease groups, 28 subgroups, 133 participants distributed over 66 meetings. From the original interviews and use cases, similar projects at other cancer centers, and suggestions about features that would be most desired, the team developed a survey to guide discussion in meetings to solicit priorities from participants. The requirements analysis and design proceeded in three waves: detailed gathering of requirements with the first wave of meetings, refinement of requirements and priorities with the second, and validation of approach with the third wave. At each meeting, the design team collected examples of use cases, queries and questions potential users would like from the system, requested data dictionaries or lists of desired data elements, and presented emerging design concepts for feedback and refinement.

Also in summer 2013, the team evaluated 14,220 data elements from the 18 data dictionaries and wish lists compiled from meetings with disease groups as described above.(13) Each data element was normalized to a standard concept and then traced back to its source (e.g., laboratory, pathology report, patient reported in clinical visit). Per their source, the data elements were classified as being unstructured (e.g., path report) or structured (e.g., lab value, race, CPT code), known or brought in by the patient (patient reportable), or computable from other data elements. The purpose of this analysis was to determine where to focus automation efforts in the HIDRA design. We condensed more than 14,000 existing and desired fields from 13 different disease groups into just under 4,000 individual elements. Conservatively, 65% of data elements came from unstructured sources, 15% of data elements could be patient reported (about 75% of

these elements were currently coming from unstructured sources), and 15% of data elements were computed from other elements. Based on this analysis, the team prioritized the development of a clinical data processing pipeline for extracting and processing information from unstructured notes into usable data elements over other possible automation and facilitation functions such as acquiring patient-reported data.

Midway through this requirements analysis in September 2013, the HIDRA steering committee arranged an external review with informatics experts from other cancer centers. From the feedback on this external review, the team continued the development of the core HIDRA infrastructure, but limited the user interface pilot to one disease.

In January 2014, implementation of the self-service user interface for HIDRA, called Argos, was initiated, focusing on brain cancer. The Argos application was built on the widely used open-source LabKey Server data management platform and the Caisis data model. This approach allowed simultaneous development of core HIDRA infrastructure and data feeds, ongoing data entry and operations using the Caisis system, and implementation and testing of the Argos pilot for data exploration. The brain cancer pilot of Argos was completed in Summer 2014. From fall 2014 to spring 2016, the Argos application was extended to support all disease groups, and a clinical data processing pipeline and workflow were developed by LabKey Software and Fred Hutch clinical data processing experts to support both automated and human data abstraction.

The HIDRA core infrastructure and data feeds project, which was initially started as a Waterfall method project, was re-baselined in 2014. Its core functionality was completed in 2015 after shifting the Fred Hutch internal team to an Agile methodology, and its performance and

features have been improved through 2016. The Argos application was developed at LabKey Software using an Agile methodology.

The HIDRA core and Argos projects have involved approximately seven FTEs distributed over approximately 15 IT staff, not counting LabKey staff. In addition, numerous research, clinical, and administrative staff across all Consortium partner institutions have been involved in this project.

4.5 Results

As of December 2015, the UW/SCCA population selected for inclusion into HIDRA included 309,239 patients, approximately 10% of the overall population in the UW CDR database. Each of these patients was selected for inclusion in HIDRA because of a cancer-related encounter (e.g., by ICD code), so this population includes some screening and consult patients as well as those diagnosed or treated for cancer. Table 4.1 shows that a significant number of patients had records associated with at least two types of cancer.

Table 4.1. As of December 2015, count of patients for each cancer for which they had related events in the medical record (in bold on diagonal), and counts of patients associated with at least two cancer types. The total number of patients in HIDRA was 309,239. Some patients may have three or more cancer types associated with events.

	Brain	Breast	GI	GU	Gyn	H&N	Heme	Renal	Sarcoma	Skin	Thoracic
Brain	**16832**										
Breast	1960	**86508**									
GIl	2537	16639	**73397**								
GU	815	2152	6349	**12868**							
Gyn	1011	19282	12188	2480	**47025**						
H&N	2622	2629	4312	991	1551	**23474**					
Heme	3016	12093	15163	5177	8461	4669	**82013**				
Renal	796	716	1768	1635	492	505	1669	**6785**			
Sarcoma	2859	3269	5655	2327	2138	2551	5169	1033	**23023**		
Skin	2481	12371	12827	3450	8077	4128	10724	1056	5402	**72104**	
Thoracic	2275	2439	4397	1333	1402	2504	5249	1116	3460	2576	**18710**

GI, gastrointestinal; GU, genitourinary; H&N, head and neck

HIDRA has more than 800 million rows of data, 380 GB of total storage space for clinical data, 150 million lab chemistry results from 26 million orders, 48 million encounter events, 3 million diagnostic imaging orders and their results.

In line with the original strategic goals, these data are available for research and research operations, healthcare operations, quality improvement and public health reporting. These data span all cancer-related disease groups, and are integrated through data feeds and into the Caisis data model to facilitate easy access by clinicians and investigators. The data and infrastructure are ready for integration with enterprise CTMS, biospecimen management, and assay data management, and planning for those systems is underway.

The Caisis database was used for its unified database model for all of these different cancer types, as well as to support new and ongoing data abstraction and data quality management activities for different disease groups.

The basic clinical data processing pipeline for automated and facilitated information processing was presented in October 2015.(74,75) A fully functional enterprise clinical data

processing pipeline will be complete by summer 2016.(76) In addition, outcomes data from the CSS tumor registry has been integrated with the HIDRA platform.

HIDRA and Argos are currently running in an initial high-security environment with documented technical and operational controls that map to the NIST 800-53 standards for FISMA compliance. A next-generation high-security environment that includes federated authentication with Consortium partners is underway. Through the Argos requirements analysis, the team determined that FDA regulatory compliance (21 CFR Part 11) was not a high priority requirement for the entire system, but rather a potential requirement for specific trials that would be better served through sponsor systems or dedicated processes.

In terms of the IT and informatics goals, the HIDRA and Argos system are hosted on a reliable and scalable environment, and investigation of cloud hosting is underway. The high-security environment and operations (including training and documentation) have been implemented and have met HIPAA BAA requirements.

At this point, Consortium investigators can more readily access data through a data request service. More than 500 faculty and key personnel Consortium members who may use this resource. A standardized HIDRA data request form has been created and requests are processed by a team of analysts. Since the service opened for business, more than 50 data requests have been processed and delivered to researchers.

The Argos application for self-service access to data was launched in March 2015.(77) It was initially available to Fred Hutch network users. Federated authentication to allow access to users with UW, SCCA, and SCH credentials is underway and anticipated in 2016. Argos is populated with clinical data from all cancer-related disease groups, and allows search by clinical, specimen, and study parameters (see Figure 4.2). Discussions are ongoing regarding the

expansion of self-service data access through Argos or other LabKey tools, as well as access through programmers and analysts.

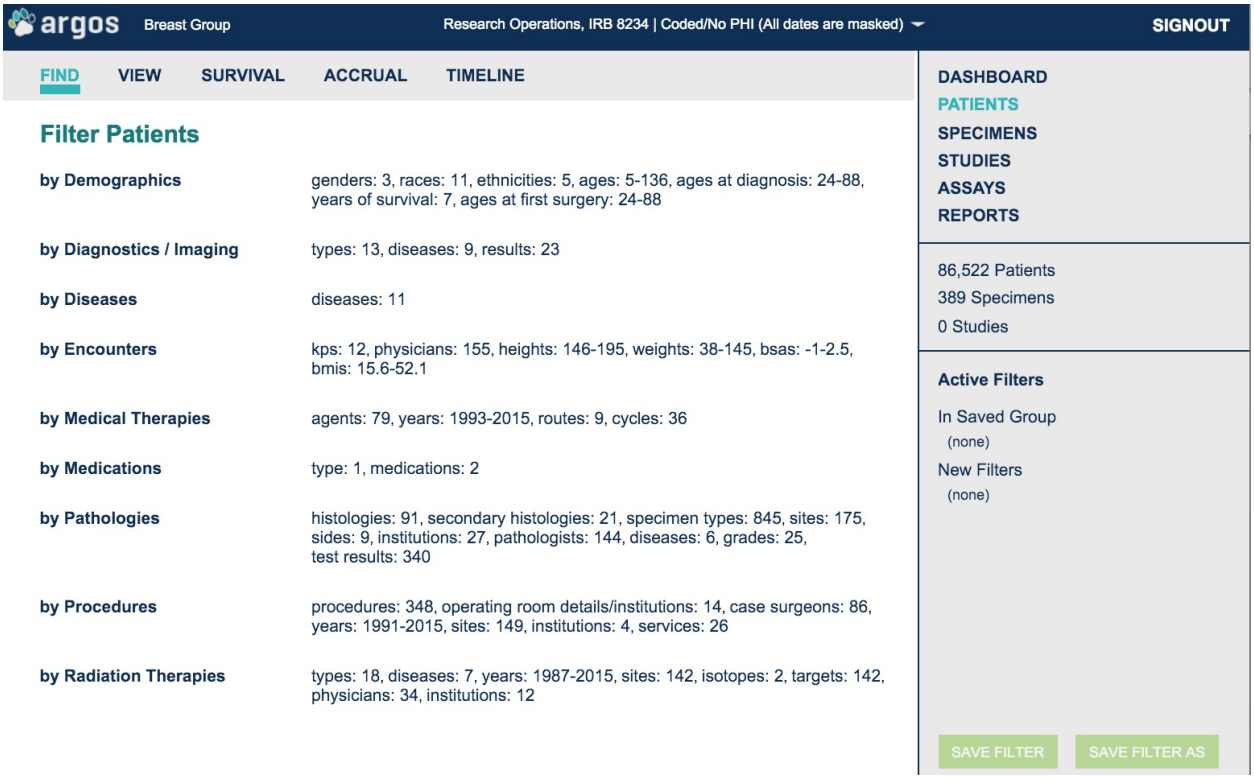

Figure 4.2. Ability for user to search by clinical parameters in Argos self-service application

For the Argos application, the top functional requirements identified and developed were the following:

- Identify subsets of patients by different parameters (e.g., clinical tumor markers and mutations)
- Search across combined data from multiple data sources
- Query and extract combined data as spreadsheets or datasets
- Provide automated abstraction (i.e. NLP) of data from notes
- Generate summary statistics for selected subsets
- Find available specimens based on patient and sample criteria

- Query based on event sequences (e.g., chemo within 30 days of death)
- Evaluate expected trial enrollment
- Explore or run queries on de-identified view of full database
- Generate survival curves for selected subsets

Argos also addressed critical nonfunctional requirements, focused on protecting patient data consistent with HIPAA and FISMA requirements:

- Data portals that restrict access to subsets of patients based on disease group, institution, study protocol, and other categories
- Access controls that limit users to the data portals, roles, and PHI levels appropriate to their organizational roles
- User declarations of role, IRB, and PHI level required for each analysis session, and agreement to appropriate terms of use
- Adaptive schema that restricts PHI column visibility based on each user's access controls and PHI level declarations
- Restrictions on extraction of patient data from the system via export or API
- Logging of all important user activities in a form easily queried by reviewers and auditors; logged activities include all:
    - Logins to the application
    - Declarations of required role, IRB, and PHI level plus agreed terms of use
    - SQL queries executed against patient data
    - Patient IDs and PHI columns viewed with each query
    - Sharing of data grids, filters, and reports with other users

Implementation of enterprise biospecimen management and CTMS for the Cancer Consortium is underway, and the HIDRA platform and Argos application are ready to integrate with these systems as described in their original vision.

For the five-year period of the project, Fred Hutch has engaged LabKey Software as a technical partner and primary software developer for Argos and the clinical data processing pipeline. A portion of internal IT and informatics staff have been dedicated to this project. As the development phase of this project comes to a close, the Consortium personnel are transitioning to operations and more incremental or specific improvements.

4.6 Discussion

Throughout this initiative, there has been value in partnering a research center (Fred Hutch) with a technology company (LabKey) for scope management, scalability, and reliable support. Developing HIDRA with LabKey rather than developing it with only internal resources was a strategic decision to maximize the ability to support the system over the long term and to make this solution portable to other centers. Both internal Fred Hutch staff and external technology partner staff from LabKey developed a healthy dynamic while implementing HIDRA. This dynamic has helped the team control project scope and prioritize development, and both sides have learned from each other. Some of the core security components - the Argos tool and operational preparations for LabKey to sign a HIPAA BAA - have proven useful for other LabKey clients, who in turn are extending the platform and applications. Moreover, partnering with all Consortium members has been increasingly rewarding. For example, the co-development of algorithms using ICD codes and other variables for selecting and tagging the

cancer patient population and attributing patients to different disease teams has been an effective and valuable Consortium collaboration.

One of the typically challenging and time-consuming tasks of building a useful data repository is the data model. A data warehouse approach that leaves data close to the format of their original source systems may allow for simpler extract-transform-load (ETL) processes to acquire data, but to query effectively, this generally requires detailed domain knowledge of sources. Alternatively, developing different data models or marts for each disease, treatment modality, or study may limit the ability to mine clinical data broadly. Adopting and adapting the Caisis data model has proven a useful strategy and allowed the team to focus on other aspects of design and implementation. In addition to the Caisis data model, the team adopted and reused multiple modules from LabKey Server that were developed for its other clients. For example, much of the user interface design for Argos came from or was inspired by the Collaborative DataSpace project, an HIV vaccine research data portal funded by the Gates Foundation and developed by LabKey.

Implementing Agile methodology using short, iterative development cycles (sprints) is critical even for large IT projects. This approach motivates the team to make steady progress in the face of daunting challenges, to control scope and project focus, and to adapt to challenges without analysis paralysis. Over the five years of this project, the team has increasingly formalized the development, testing, and production environments, as well as its project management and development processes.

An integrated data repository for cancer research relies on information, especially the tracking of consents from clinical trials and data or specimen repositories. To facilitate translational research it is also dependent on biospecimen management information, and the

management of molecular data for its patient population. Like EHRs and other clinical systems, enterprise CTMS and biospecimen management systems are intertwined with their respective complex workflows and may be out of scope for a data repository initiative. However, an integrated repository for current and next-generation research cannot be complete without information from these systems.

Although numerous commercial, academic, and open-source tools for clinical data processing are currently available, the rate-limiting factor for enterprise deployment of clinical data processing is the ongoing development of training and validation data. Rethinking the clinical data supply chain and developing the tooling for an enterprise approach to clinical data processing is a significant challenge. The Fred Hutch and LabKey approach has been to integrate existing manual abstraction workflows into the development of these training and validation data and this approach will be described and published separately.

The decision to develop an enterprise approach to high security, including implementing technology, training, operations, and documentation for a HIPAA- and FISMA-ready environment was initially daunting. However, after several iterations, the entire team has become more adept at thinking through the NIST 800-53 security requirements and operating in a high-security environment. Simply, the upfront investment and learning curve has been worth the effort. However, providing federated authentication and controls to allow users at each Consortium site to access the Argos self-service application using their home institution credentials (username and password) has proven more difficult and more resource- and time-consuming than anticipated, and is still underway.

The upfront decision to develop an MOU for data sharing and matching IRB files for the HIDRA repository that involved all Consortium partners was also worthwhile. These documents are referenced frequently and have guided numerous design and implementation decisions.

Managing the history and documentation of this project so that the team, especially new members and project managers, can recall or find detailed requirements and decisions, has been an ongoing struggle. Most of the documentation is currently managed in an institutional instance of Microsoft SharePoint.

Performance of HIDRA data feeds and the Argos self-service application has been more of a struggle than anticipated. Over time, the ETL processes have been refactored to improve performance, and although most researchers do not yet require real-time data, much room for improvement remains. Developing a de-identified dataset that is equivalent in volume and variety to the production dataset has been key to scaling the system. The Argos developers load these large de-identified datasets into their local deployments, allowing them to mimic real-world performance as they develop and test the system, with no risk of exposing real patient information.

Further work on the HIDRA platform will involve scaling the patient population and associated clinical data elements, as well as using this infrastructure for other projects. As enterprise biospecimen management, clinical trial management, and experimental results data systems are implemented, they will need to be integrated with HIDRA and Argos to facilitate ready access to data for investigators. The adoption and extension of this technology by other LabKey clients is also anticipated.

4.7 Conclusion

The development of the integrated data repository for a consortium cancer center can be facilitated through the approach used for HIDRA. Creating a legal and an IRB foundation for the repository, a high-security environment, and operations to support the repository are important to be able to scale up quickly. Having a reliable and competent technical partner may ease the pressures on local staff and may lead to creative solutions and learning opportunities. However, local staff must be deeply engaged for a successful and sustainable solution. Healthy dynamics between project leadership, internal staff, and an external technical partner helps to prioritize development and control scope. Also, adopting and adapting tools from other centers for key functions can speed up and lower total development cost. The development or adoption of a self-service data access interface is important to investigators and staff. However, there is a great need for customization for each disease group.

4.8 Acknowledgements

This project was partially supported by the FHCRC/UW Cancer Consortium Cancer Center Support Grant of the NIH under Award Number P30 CA015704, and the National Center For Advancing Translational Sciences of the NIH under Award Number UL1TR000423. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

4.9 Synthesis

This chapter described the strategic goals, design and implementation of the HIDRA system at Fred Hutch and lessons learned from this work. The rationale and design of a pipeline for clinical

data processing was briefly described in this chapter, however because of the strategic importance of the pipeline and the amount of research and development involved, it is described separately and more fully in Chapter 6. Chapters 4–6 can be considered two parts of the overall story of developing a next-generation research data repository at Fred Hutch cancer center.

Sections 4.2 to 4.7 addressed my Chapter 4 questions: What are the challenges and opportunities for informatics at Fred Hutch? Do these challenges and opportunities align with my previous work and lessons learned? Are there tools and methods that others or I have used before that could be successfully applied at Fred Hutch? How can we enable new types of research, faster results, and better quality of research data at Fred Hutch? How can we improve access to data with tools and methods that are scalable, extensible to new projects, sustainable, and portable to other centers? Together, these questions tie back to the overarching research question for this dissertation: how can we improve access to clinical and related data about cancer patients for research? They also tie back to the overall hypothesis (that there are new tools and methods from biomedical informatics that could improve the availability of data for cancer research if they were applied thoughtfully and strategically) and to Aim 1 from Chapter 4 (to develop and assess a modern integrated data platform to support a wide variety of cancer research). This work on HIDRA logically leads to Aim 2 in Chapter 6. Aim 2 was to develop and characterize a clinical data processing pipeline that is scalable to the cancer center enterprise level, is well-supported and sustainable, and can complement or streamline existing manual data abstraction and information-processing activities. As mentioned above, Chapters 4 and 5 can be considered two parts of the same overall story of developing a next-generation research data repository at Fred Hutch cancer center.

Findings from Chapter 4 are relevant to other cancer centers addressing the need to integrate a variety of clinical, specimen, study and molecular data to improve the ease and quality of research.

First, the legal, IRB and security framework from HIDRA is relevant to other centers. Already, this framework has been used to rewrite IRB files for the Hutchinson Institute for Cancer Outcomes Research (HICOR) and the SCH Research Informatics Platform. I have already received questions and requests from other cancer centers about using the HIDRA IRB files as a model and framework for their own database and data integration initiatives. The way that the regulations and policies were implemented in Argos (the self service data access tool) are also relevant to other centers. In the Argos application, we integrated and included all constraints, controls, and terms of use for all included data sources (e.g., medical records, cancer registry data, consumer data) in one framework that makes it easy for researchers to do the right thing and comply with pertinent legal, IRB and IT security regulations and policies.

Second, for consortium or matrix cancer centers that are part of a broader academic or community medical center, HIDRA provides an example of leveraging the data warehouse and clinical data repository work of a broader medical center for the purposes of the cancer center.

Third, HIDRA provides an example of adopting and extending the work from various other groups to speed up and reduce costs of implementing and integrated data repository. We used the database model and user interface of Caisis to provide data collection, processing and storage functionality without having to reinvent those components within HIDRA. We also adopted and adapted LabKey Server functionality developed for infectious diseases populations to provide the core features of the Argos self service data access application.

Fourth, the HIDRA project provides an example of a system not just for manual data entry, or data feeds, or clinical templates, but an overall strategy and pipeline for clinical data processing that can include and integrate different manual and automated forms of clinical data acquisition and processing to support research.

Fifth, a lesson learned from the HIDRA work that can be applied at other centers is the need for a system performance evaluation and improvement strategy, including providing a realistic and de-identified testing dataset for software developers and system operations staff.

Sixth, the HIDRA work identified federated security (the ability for users in a consortium or matrix cancer center to log into a system using any of their institutional username and password credentials) as a critical point of potential failure. Identifying the preferred credentials of the targeted user base is an key requirement for any similar system implemented at another center.

Finally, the Agile approach to software engineering and system implementation was challenging to implement in teams who were unaccustomed to it, but critical for project momentum and success. Breaking projects down to very small tasks that can be implemented in weeks rather than months or years allow the developers and implementers to keep up momentum, reduce over-analysis, and stay focused.

Chapter 5 Evaluation of HIDRA

5.1 Context

This chapter continues my exploration of Aim 1, adding further assessment to the HIDRA data platform covered in Chapter 4. To address the evaluation aspect of Aim 1, I answer the following question: What is the impact of the data platform developed at Fred Hutch?

For this chapter, I guided and conducted much of the original HIDRA and Argos requirements analysis, and as Director of the developing Biomedical Informatics resource for the Fred Hutch/UW Cancer Consortium, I was responsible for assuring that the system implemented was aligned with the overall vision and would support the biomedical informatics strategy for the consortium. Because this chapter is an extension to Chapter 4 rather than a freestanding work, the additional background and rationale of this chapter is minimal.

5.2 Background and Rationale

One important aspect of biomedical informatics research such as the HIDRA platform described in Chapter 4 is the impact of this work on research, and ultimately on improving the health of patients with cancer. Since the HIDRA platform was developed to support research, ultimately the impact of this system could be evaluated according to its effects on publications and presentations, clinical trials and other studies initiated or in progress, as well as grants, contracts and philanthropic funding. Because the HIDRA platform and its self-service data access tool (Argos) have not been fully deployed due to delays in implementing federated security access to the system, performance issues, and changes to the Fred Hutch IT and informatics organization and priorities, the assessment possible at this stage is limited.

Ideally, to measure impact, we would be able to attribute datasets produced from HIDRA to publications, presentations, grants, contracts, clinical trials and other studies or to philanthropic gifts. However, with limited outcomes to measure at this point, we can look at intermediate outcomes or measures of impact related to the development and implementation process. The measures of impact are described under the methods section, and the findings of this evaluation are described under the results and findings section.

5.3 Methods

5.3.1 Evaluation of Features

To determine the use cases and features for Argos that would provide the greatest value across all disease groups, we asked each disease group to rank a list of proposed features, and to validate that these were indeed the features desired. This list (shown in Figure 5.1) was then sorted to prioritize development and implementation efforts for a brain cancer pilot of Argos, and for the full implementation. To evaluate that the system developed reflected the requirements and priorities of researchers, we went back and reviewed each feature and each disease group to confirm that the functionality implemented in Argos matched the user requirements.

In addition to the requirements from users within disease groups, we reviewed the system features against the NIST 800-53 security controls to check that it conformed to appropriate HIPAA- and FISMA-level security requirements.

Scoring Legend:
3 = Critical
2= Very Important
1 = Somewhat Important
0 = Not Important

ID	FEATURE	TOTAL POSSIBLE	SUM TOTAL	AVERAGE	MODE	TOTAL COUNT	3-COUNT (critical)	3-%	2-COUNT (very imp.)	2-%	1-COUNT (somewhat imp.)	1-%	0-COUNT (not imp.)	0-%	Breast-SPORE	GI-Hepatobiliary	GI-Lower	GI-Pancreas	Gyn	Heme-AML	Heme-Inherited MDS	Heme-Myeloma	Infection Surv. #1 - Pergam	Lung	Neuro	Peds/SCH	Sarcoma - Med Onc	Sarcoma - Surg / Rad
01	Eval expected trial enrollment	42	30	2.1	3	14	6	43%	5	36%	2	14%	1	7%	3	2	2	1	2	2	0	2	1	3	3	3	3	3
02	ID subsets of Pts by	42	40	2.9	3	14	12	86%	2	14%	0	0%	0	0%	3	3	3	2	3	3	3	3	3	3	3	2	3	3
03	Generate survival curves	42	28	2.0	3	14	6	43%	4	29%	2	14%	2	14%	3	3	3	2	3	3	1	1	0	2	3	2	2	0
04	Generate summary stats	39	32	2.5	3	13	7	54%	5	38%	1	8%	0	0%	3	3	3	2	3	3	3	2	2	3	2	1		2
05	Find available specimens	42	31	2.2	3	14	8	57%	3	21%	1	7%	2	14%	3	3	3	3	1	0	3	0	2	2	3	3	3	2
06	Collect Pt-reported data	42	26	1.9	1	14	5	36%	3	21%	5	36%	1	7%	2	1	3	3	2	1	1	0	1	2	1	3	3	3
07	Provide automated abstraction	39	33	2.5	3	13	8	62%	4	31%	1	8%	0	0%	3	2	3	3	3	3	3	2	3	1		2	2	3
08	ID Pts based on Tx	42	39	2.8	3	14	11	79%	3	21%	0	0%	0	0%	2	3	3	3	2	3	2	3	3	3	3	3	3	3
09	ID Pts by tumor markers	42	35	2.5	3	14	11	79%	0	0%	2	14%	1	7%	3	3	3	3	1	3	3	3	3	3	3	1	0	3
10	Manage LTFU / data collection	36	20	1.7	2	12	3	25%	4	33%	3	25%	2	17%	2	0	2		2	3	3	1	1	1	0	3		2
11	View timelines - Individual Pt	39	21	1.6	1	13	4	31%	2	15%	5	38%	2	15%	2	1		1	1	3	3	0	3	1	2	1	3	0
12	View timelines - Multiple Pts	39	17	1.3	1	13	1	8%	4	31%	6	46%	2	15%	2	1		2	0	1	0	2	1	1	1	1	3	2
13	Add my data to HIDRA	39	22	1.7	2	13	3	23%	5	38%	3	23%	2	15%	2	2		1	0	3	1	3	3	1	0	2	2	2
14	Query/extract combined data	39	34	2.6	3	13	10	77%	1	8%	2	15%	0	0%	3	3		1	2	3	3	3	3	3	3	3	3	1
15	Query on event sequences	36	26	2.2	3	12	6	50%	2	17%	4	33%	0	0%	2	3		2	3	3	1	1	3	3	1	3		1
16	Search combined data	36	33	2.8	3	12	9	75%	3	25%	0	0%	0	0%	3	3		2	2	3	3	2	3	3	3	3		3
17	Share subsets w/ select people	39	23	1.8	3	13	4	31%	4	31%	3	23%	2	15%	3	3		2	0	3	3	1	2	1	2	1	2	0
18	Enable use for FDA-reg. studies	33	7	0.6	0	11	1	9%	0	0%	4	36%	6	55%	0	0		1	0		1	0	3	1	1	0		0
19	Integrate Pt & specimen data mgmt	36	22	1.8	3	12	5	42%	3	25%	1	8%	3	25%	3	3		0	2	0	3	0	2	1	3	3		2
20	Explore de-ID'd view of all HIDRA data	33	24	2.2	3	11	5	45%	4	36%	1	9%	1	9%	3	3			3	2	2	1	3	3	2	2		0

Figure 5.1. Ranked user requirements and features for the Argos self-service data access tool

5.3.2 Evaluation of Performance

One of the critical outcomes of a system such as HIDRA and data access tool such as Argos is the performance of data feeds and the user interface. Users are likely to become frustrated with or form a poor perception of a system that has significant lag times when browsing for data. To measure performance of the data feeds in HIDRA, we asked all potential users in the disease groups what would be an acceptable lag time between generation of information in the EMR and the delivery of that information for research in HIDRA. The researchers we interviewed unanimously agreed that a 24 hour to 1 week lag time for data feeds from the EMR into HIDRA would be acceptable for their purposes.

To evaluate the performance of the Argos self-service data access tool, we named each of the measurable components of the web-based user interface, and created a matrix of expected performance for each component, measured in seconds of lag time. The minimal, acceptable and desired performance expectations for each component were documented. A sample of the

performance expectation matrix is in Table 5.1. For each version of Argos released from LabKey, we installed the software in a staging environment identical to the production system, and assessed performance against the performance expectations.

Table 5.1. Sample of performance criteria matrix for the Argos self-service data access tool (times in seconds)

Feature	Performance Criteria (times in seconds)		
Page/Feature	Desired Performance	Acceptable Performance	Minimum Performance
Portals Page	1	2	3
Activity Dialog	1	1.1	2
Terms of Use screen	0.5	1	1.1
Dashboard Page (all components)	4	8	10
Dashboard Chart	1	3	4
Dashboard Table	1	2.5	3
Overall Counts	2	2.5	3.1
Filter Summary Page (Patients)	2	3.1	4
Individual Explorer Page (Patients)	1	2	3
Column Chooser (Patients)	1.5	3	4
View Page (Patients)	2	5	10.1
Data Grid Sorting (Patients)	1	1.1	2
Column Chooser Filter Dialog on Data Grid (Patients)	0.5	1	1.1
Save Data Grid Dialog (Patients)	0.5	1	1.1
Action of Saving Data Grid (Patients)	1	2.1	3

5.3.3 Evaluation of Usability

In addition to the evaluation of features relative to requirements and the evaluation of performance, a system like HIDRA should be assessed for usability. Unfortunately, due to the delays in implementing a federated security environment across the Fred Hutch/UW Cancer Consortium and the performance issues discovered after quickly scaling up the data feeds into HIDRA, the actual user base has remained small. Except for the IT and informatics staff developing and implementing HIDRA and Argos, there are only six end users. These end users are from the brain cancer group, because that was the original small pilot of Argos and HIDRA. The system performance issues have been resolved by engineers, and these fixes should soon be promoted to the HIDRA and Argos production environments. However, the federated security work is still in progress.

Given the delays in making the system available to more end users, we conducted a preliminary usability evaluation with a set of potential new users at SCCA that were able to obtain Fred Hutch security credentials. In this evaluation, we briefly described HIDRA and the Argos self-service data access tool, and then presented the users with a set of tasks to attempt to accomplish using Argos. These tasks included finding subsets of patients that meet certain criteria, and other tasks that mapped closely to the original requirements and features requested by disease groups (described under Evaluation of Features and in Figure 5.1). During this usability evaluation, users were encouraged to “think out loud”(78) when encountering a challenge with the system. Each user was paired with a LabKey software engineer. Findings from that evaluation are still being compiled so will not be covered in this chapter, and similar follow-up usability evaluation sessions are currently being scheduled.

5.3.4 Evaluation of Outcomes and Impact

The outcomes or impact of HIDRA could eventually be evaluated in terms of its contribution to publications, presentations, and grants. In addition to support of academic output, HIDRA could be evaluated by its perception by peers and the public and its adoption by other groups.

Given the current stage of HIDRA deployment, the academic output is limited. However, in the result and findings section I will describe the 62 HIDRA data requests that have been submitted since the system went live. I will also describe to the best of my ability to the current peer and public perception and adoption of this system by other groups.

5.4 Results/Findings

5.4.1 Results of Requirements and Feature Evaluation

The requirements that were identified and prioritized by researchers were used to guide all development of HIDRA and Argos. The first phase of development focused on a subset of the requirements and a subset of data that characterize patients with brain cancer. All of the features identified for the brain cancer pilot were completed by summer 2014. The remaining features are described in Chapter 4 and were completed in 2015 and 2016. We reviewed all of the original requirements gathering documentation from meetings with individual disease groups to confirm that what has been built includes the features and requirements that were specified.

5.4.2 Results of Performance Evaluation

The performance of HIDRA data feeds and the Argos self-service tool for data access are evaluated after each new release is completed in our staging environment, before it is released into the production environment.

The HIDRA data feeds initially performed well (within 24 hours) for the brain cancer pilot completed in summer 2014. However, when the entire cancer patient population was added to HIDRA in 2015 (over 300,000 patients and associated clinical data), the performance lagged considerably. Much of the HIDRA data feeds infrastructure has been enhanced or rewritten to get performance of data feeds for the entire population to an acceptable lag time, which is back under 24 hours.

The performance of the Argos tool has been evaluated by Fred Hutch business analysts with each monthly software release from LabKey Software. When the underlying population was expanded from the brain cancer pilot to all disease groups, performance of Argos suffered considerably. Results of one of the Argos performance tests is shown in Table 5.2.

In order to bring the Argos user interface performance down to acceptable lag times, we had to create a testing dataset with realistic volume and variety of data and that was completely de-identified so that it could be safely used by software engineers at LabKey for development and testing.

The performance of the most recent release of Argos is acceptable according to our performance matrix. This version is not yet in production but will be promoted soon to the production HIDRA and Argos environment.

Table 5.2. Example of results from performance evaluation (times in seconds)

	How Long it Took to Run (in seconds)				
Page/Feature	All Patients	Brain	GI	Thoracic	Head and Neck
Portals Page	2.39	2.39	2.39	2.39	2.39
Activity Dialog	0.349	0.031	0.445	0.046	0.048
Terms of Use screen	0.759	0.024	0.043	0.083	0.028
Dashboard Page (all components)	4.54	4.77	5.09	5.24	4.05
Dashboard Chart	3.02	3.62	3.7	2.78	2.62
Dashboard Table	4.54	4.07	5.09	5.24	4.05
Overall Counts	3.02	3.02	1.15	2.23	1.91
Filter Summary Page (Patients)	5.87	3.3	6.23	5.27	3.081
Individual Explorer Page (Patients)	0.251	0.189	0.245	0.314	0.191
Column Chooser (Patients)	7.516	6.03	7.122	6.431	5.019
View Page (Patients)	0.883	0.223	0.51	0.407	0.455
Data Grid Sorting (Patients)	0.533	0.208	0.381	0.274	0.272
Column Chooser Filter Dialog on Data Grid (Patients)	1.2	1.82	2.07	1.78	1.8
Save Data Grid Dialog (Patients)	0.04	0.04	0.04	0.04	0.04
Action of Saving Data Grid (Patients)	1.188	0.329	0.462	0.411	0.689
Patients - Survival Page (Patients)	6.7	1.66	2.788	2.272	1.642
Patients - Survival Page Apply Filters (Patients)	3.413	1.74	2.39	1.73	1.633

5.4.3 Results of Usability Evaluation

As the number of current end users is minimal (only 6 users from the brain cancer disease group), and the results from the recent usability evaluation are still being compiled, the usability findings are minimal.

However, I helped to design and conduct the usability evaluation, and in asking users to think out loud while attempting to perform selected tasks in Argos, I found that there is some confusion between the patient filters (Figure 4.2) and a feature that allows patients deeper access into the underlying database. Also, the labels used (e.g., medical therapy "agents", "histologies", "KPS") are not understood by all staff.

There is a tendency by many software engineers to overreact to user feedback, so I have cautioned the engineers and IT staff involved to rethink the usability evaluation process and repeat the evaluation with another set of potential users before making changes to the application. Also, some usability issues can be addressed through training rather than modifications to the application. In the next round of usability evaluation, LabKey will be providing a short training session, and more realistic scenarios.

5.4.4 Results of Outcomes and Impact Evaluation

Of the 62 data requests for the HIDRA system, 44 were to support research and 18 were to support healthcare operations, quality improvement, and clinical care/treatment activities.

The data delivered from HIDRA for clinical care/treatment activities included delivery of reports to support testing of ICD-10 reports, statistics on patients with head, neck and prostate cancer. Two of the requests for data to support clinical and treatment activities were removed from the HIDRA data request queue and routed to the SCCA clinical data analytics team. Data delivered for for healthcare operations included lists of referring physicians for solid tumor

cancers aand long-term follow-up reports for patients with head and neck cancers. For quality improvement, the HIDRA team provided data to compute survival for patients with stage IV non small cell lung cancer, data to determine consent status of the genitourinary cancer patient populatio, and data to support internal auditing efforts.

For research, the HIDRA system and team have provided data to support a a study of docetaxel medical therapy for genitourinary cancers, a study of temozolamide medical therapy for glioblastoma cancer, identified rare lymphoma cases and acute myeloid leukemia cases. Overall, to date most of the HIDRA data requests have been to identify populations of patients in preparation for research or to aid in cleanup of disease specific databases.

5.5 Discussion

Now that the basic functionality of HIDRA and Argos is completed, future requirements and feature development will focus on advancing individual disease groups (e.g., the hematologic malignancy research group) and be prioritized by Fred Hutch senior leadership as well as through direct collaboration of IT and informatics staff with researchers. As of summer 2016, the basic functionality of HIDRA and Argos will be considered complete.

The previous lack of de-identified testing data with realistic volume and variety was a barrier to performance evaluation. However, this lack has been remedied. All developers and testers have access to realistic but safe testing data, and performance evaluation is ongoing.

The usability evaluation of HIDRA and Argos has been limited due to the lack of federated security and performance issues. HIDRA and Argos performance issues have largely been resolved, and a solution for federated security is underway to allow users from UW, SCCA and SCH to access the systems hosted at Fred Hutch. This federated security solution is anticipated to be finished in the next 4 to 6 months.

It is still too early to evaluate the impact of HIDRA on research (e.g., publications, presentations, trials, studies, grants, contracts). However, the data requests from HIDRA have begun to support research and related activities at Fred Hutch and its consortium partners.

Anecdotally, public perception seems to be that HIDRA is up and running and is an example of a successful integrated research data repository for cancer. I have spoken with several researchers at other institutions who have heard about and requested additional information about HIDRA. Moreover, the HIDRA and Argos tools that have been implemented in LabKey Server software are being adopted and adapted by other groups, including Genomic England, the NCI, and a group of other cancer research centers.

5.6 Acknowledgements

I would like to thank Evan Carl for his assistance in the analysis of HIDRA data requests, David Sharp for his work on the HIDRA and Argos requirements analysis and prioritization, and Kristin Dubrule and Jessi Murray from LabKey Software for their leadership in designing and conducting a usability evaluation of Argos.

5.7 Synthesis

This chapter described the assessment of the HIDRA system at Fred Hutch. It is an extension of the work of Chapter 4, and is limited due to the lack of users pending resolution of security and performance issues as well as restructuring and reprioritization within Fred Hutch IT department. Sections 5.2 to 5.5 address my Chapter 5 question: What is the impact of the data platform developed at Fred Hutch? As in Chapter 4, this question ties back to the overarching research question for this dissertation: how can we improve access to clinical and related data about cancer patients for research? It also ties back to the overall hypothesis (that there are new tools

and methods from biomedical informatics that could improve the availability of data for cancer research if they were applied thoughtfully and strategically) and to Aim 1 from Chapter 4 (to develop and assess a modern integrated data platform to support a wide variety of cancer research).

Findings from Chapter 5 are relevant to other cancer centers implementing similar integrated data repositories to support research. It underscores the need to resolve federated security issues for consortium or matrix cancer centers. It identifies the need for a realistic development and testing dataset to find and resolve performance issues. It supports the importance of usability testing, and integrating usability testing and enhancements with user training. Finally, it describes the desired impact of efforts like HIDRA and how that impact can be evaluated.

Chapter 6 Scalable Clinical Data Pipeline for a Cancer Center

6.1 Context

Chapter 6 is my exploration of Aim 2, to develop and characterize a clinical data processing pipeline that is scalable to the cancer center enterprise level, is well-supported and sustainable, and can complement or streamline existing manual data abstraction and information-processing activities. To explore this aim, I answer the following questions for Chapter 6: How can we improve the quality, speed, and economics of the acquisition, processing, and delivery of clinical data to support cancer researchers? How can we make clinical data processing a core competency of Fred Hutch at an enterprise scale?

Answering these questions helps inform the overall research question for this dissertation (How can we improve access to clinical and related data about cancer patients for research?) because much of the clinical data needed for cancer research comes from narrative medical reports and because improving access for research requires the extraction and coding of that information into more distinct data elements.

Aim 2 is also a logical next step to test my overall hypothesis that there are new tools and methods from biomedical informatics that could improve the availability of data for cancer research if they were applied thoughtfully and strategically. Specifically, this chapter addresses the strategic application of clinical data processing and software engineering tools and methods to improve the availability of data for research.

Aim 2 flows from Aim 1 (to develop and assess a modern integrated data platform to support a wide variety of cancer research). The pipeline for clinical data processing (Aim 2) is part of an overall integrated data platform, specifically for the acquisition and processing of data

from medical record documents, and the pipeline is foretold in Chapter 4 (the chapter on HIDRA that addresses Aim 1).

I intend to submit this paper for publication by September 2016 after the clinical data processing pipeline is fully implemented and tested. Thus, the co-authors are included as well as the paper title after section 6.1. For this work, I initiated and led the clinical data processing strategy at Fred Hutch and wrote the first draft of the paper. The co-authors contributed to design and implementation of the clinical data processing pipeline and edited the paper. Because this is a standalone paper, I have included a separate acknowledgements section.

The pipeline for clinical data processing is my contribution. None of the NLP algorithms developed at Fred Hutch are my contribution or part of this dissertation, however they are described briefly in the Approach section at the requests of my reading committee. The pipeline is currently nearing completion and will be in production at Fred Hutch by summer 2016. Our intention is that the NLP interns for summer 2016 will implement algorithms into the clinical data processing pipeline in order to test it and provide input for further enhancements.

Scalable Clinical Data Pipeline for a Cancer Center

Paul A. Fearn, MBA[1,2], Emily Silgard, MS[1], Tony Galuhn[3], Sarah Ramsay, MPH[1]

[1]Fred Hutchinson Cancer Research Center, Seattle, WA; [2]University of Washington, Department of Biomedical Informatics and Medical Education, Seattle, WA; [3]LabKey Software, Seattle, WA

6.2 Abstract

Fred Hutch and LabKey Software have designed a scalable pipeline for processing clinical documents into usable data elements through both clinical data processing. This pipeline was developed initially for the Hutch Integrated Data Repository and Archive (HIDRA), a collaborative effort across the Fred Hutch/UW Cancer Consortium to create an integrated database that would enable scientists and physicians to learn from new and long-term patients across the consortium, where historically siloed data management has led to the duplication of abstraction and information-processing efforts and has hindered access to valued information. The proposed enterprise clinical data processing pipeline will serve to expedite access to usable data that is needed for improving healthcare operations and advancing cancer research. By reducing redundancy in abstraction and information-processing efforts, tracking and reducing variation/bias of interpretation, and making patient history data as up-to-date and complete as possible, this research aims to ease the burden of manual abstraction and improve the timeliness and quality of clinical data.

6.3 Background and Rationale

Scaling and acceleration of clinical and translational research is dependent on the acquisition and processing of information from EHRs and other clinical systems. However, even with recent advances in EHRs, text processing, and machine learning, obtaining clinical data for research is still largely dependent on manual data abstraction. Clinical trials still depend on manual data abstraction of clinical data from EHRs into EDC systems, often with duplicate data entry for quality assurance. Correlative research still relies on manual data entry from clinical systems into project-specific forms or questionnaires and project-specific data management systems. Personnel in outcomes research, cancer survivorship, quality improvement, care pathways, and cancer registration may all abstract similar data from EHRs into dedicated systems. For hospitals, some of these programs (e.g., cancer registries) are typically cost centers. With ongoing pressure from consumers, payers, and policymakers to decrease costs and increase the value of care, the current management of clinical data is unsustainable.

Most hospitals and research centers have recognized the issue of duplication of clinical data abstraction efforts across different projects, and there is an emerging consensus that the majority of information needed for operational and research activities is currently documented in narrative or unstandardized text in clinical systems.(13,79–81)

In summer 2013, the Fred Hutch NLP staff and interns evaluated 14,220 data elements from the 18 data dictionaries and wish lists compiled from meetings with disease groups as described above.(72) Each data element was normalized to a standard concept and then traced back to its source (e.g., laboratory, pathology report, patient reported in clinical visit). Per their source, the data elements were classified as being unstructured (e.g., path report) or structured (e.g., lab value, race, CPT code), known or brought in by the patient (patient reportable), or

computable from other data elements. The purpose of this analysis was to determine where to focus automation efforts in the HIDRA design. We condensed more than 14,000 existing and desired fields from 13 different disease groups into just under 4,000 individual elements. Conservatively, 65% of data elements came from unstructured sources, 15% of data elements could be patient reported (about 75% of these elements were currently coming from unstructured sources), and 15% of data elements were computed from other elements. Based on this analysis, we prioritized the development of a clinical data processing pipeline for extracting and processing information from unstructured notes into usable data elements over other possible automation and facilitation functions such as acquiring patient-reported data.

Another reason for prioritizing development of a clinical data processing pipeline is that other approaches (e.g., templates, manual data entry) do not scale. Based on my experience of implementing clinical templates at MSKCC to collect both research and clinical data during the clinical workflow, broad use of templates requires drastic reengineering of clinical workflows so that data collection is distributed across a variety of staff in a comprehensively designed data supply chain. Although some clinical data such as medication orders may be collected effectively using templates, it is frustrating for clinical staff to communicate the majority of clinical information in templated formats. Templated communication is mind-numbing in a high-volume clinic. It is difficult to distinguish between different cases when they are templated, whereas in narrative text a patient's story is more unique and memorable. Also, getting templates rolled out for all staff and all patients is problematic, so for the foreseeable future narrative text is likely to remain in the medical record. Based on findings from my visits to other cancer centers (described in Chapter 3), every cancer center has the same struggles with scaling the implementation of templates. Many centers have ongoing efforts to implement templated medical records with

standardized fields, terminologies and text in an effort to have more useful data for downstream use. However, templated reporting has practical limits, as anyone experienced in such efforts can testify. The volume, density, and variety of information conveyed through patient care dialogues and the resulting narrative or nonstandardized text in EHRs far exceeds what can practically be communicated through standardized templates. Templates and drop-down lists are not an adequate medium for communication of rich information. It is frustrating, painful, and limiting for physicians to communicate and document in this way, and with the recent maturation of NLP and machine-learning tools and methods, this strategy is no longer necessary. Templates and standardization are great approaches when used judiciously and for information with relatively little variation, volume, or time sensitivity.

Manual data abstraction does not scale well because of the variable reliability of humans for information processing, the costs of manual data abstraction and data quality assurance, and the intense training required for data entry staffs. Manual data abstractors regularly experience fatigue, and their perceptions of information may change over time as well as be affected by external factors. These factors can introduce shifts and biases in data interpretation that are exacerbated as volume their work increases. With limited and variable funding for research, it is often difficult to hire and retain data entry staff. We can and do take advantage of the lower costs of offshore data abstraction by outsourcing some data abstraction and data quality assurance work to dedicated data abstraction companies. However, this is only a temporary fix, and incurs the costs of setting up and overseeing contract data entry work. Because cancer clinical data can be complex to interpret, training data entry staff can take quite a long time and become expensive. The training requirements make the costs of acquiring (and the costs of losing) personnel pretty steep.

Importantly, the volume and variety of information communicated during the care of patients either pushes the boundaries or exceeds most methods of communication. Medical language contains numerous information-dense terms and expressions. Narrative communication between healthcare providers and in medical records is often difficult to unpack and is represented in lay language or in discrete, standardized vocabulary and data fields. Physicians in clinic often communicate at a rate and manner that far exceeds the ability of patients to fully comprehend, and medical records also reflect this information-dense communication.(82)

In healthcare IT and informatics, there have been a number of common approaches to more efficiently obtain and process the information from EHRs into desired systems and structures such as databases to support research, clinical analytics, quality improvement, and public health reporting (e.g., cancer registries). These include implementing [1] templated notes or synoptic reports to capture information in structured form as it is generated, [2] data feeds into a data warehouse, data repository, or destination system, [3] low-cost outsourced or offshore data abstraction, and [4] clinical data processing. Each of these approaches has tradeoffs and practical limits.

From a data perspective, forcing physicians to shift to templated or synoptic medical records seems attractive. However, because standard terminologies and templates restrict the expressiveness of communication, they can severely hinder medical practice. Moreover, templated notes place the burden of translating clinical stories into a standard language on the physicians and other clinical staff, and may limit their ability to distinguish between and remember many different patient stories seen in clinics. There may be other practical limits to implementing templated, standardized forms for medical communication and documentation.

Data feeds from source systems into a data warehouse, data repository, or other desired downstream system and format may work quite well for data that are natively well structured and standardized (e.g., clinical laboratory results, demographics, billing diagnosis and procedure codes, medication orders). However, unless healthcare providers use templates for systems like pathology, radiology, surgery, radiation oncology, and clinical encounter notes, there are limits to how much data acquisition and information processing can be addressed using data feeds.

Outsourcing or offshoring data abstraction to a lower cost operation can save money in the short term, but this approach just postpones the problem and may not address the issue of duplication of abstraction efforts across multiple activities (e.g., research, quality improvement, cancer registry). Also, outsourcing and transfer-of-information processing to another group does not address the variations in human information processing, or the biases introduced by each project or task. Moreover, outsourcing data abstraction is ultimately not scalable for larger programs that require extraction and processing of patient information from many healthcare providers and facilities, like the CDC's National Program of Cancer Registries (NPCR),(83) the Commission on Cancer National Cancer Data Base (NCDB),(84) NCI's SEER program,(85) and ASCO CancerLinQ.(86)

For more than a decade, NLP and machine-learning methods and tools to extract clinical information have been developed and evaluated, and with sufficient training data and methods, the performance of these tools has approached the performance of human data abstractors and domain experts. The application of clinical data processing tools and methods at an enterprise scale in a hospital or research center is attractive in that it may reduce the current trend of shifting the burden of generating structured data onto healthcare providers through templates.

Also, clinical data processing algorithms may generate more reliable, reproducible, and measurable results.

In the world of sequence and other assay data, bioinformatics pipelines, data standards, and toolkits have been developed to transform data from their raw format into standard tables of aligned and normalized data that are amenable to data mining. However, pipelines for clinical data have not yet evolved to this point. Clinical NLP research projects tend to focus on either information retrieval and extraction tools or on the performance of specific rule-based and machine-learning algorithms. End-to-end substitutes for manual data abstraction and complementary systems to facilitate manual and automated information extraction and processing from text medical records into desired end formats are lacking, although the current generation of tools and approaches are getting close.(87,88)

The current barriers to implementing clinical data processing and machine-learning methods that scale out to cover an entire healthcare enterprise or network will likely require workflows and features that are analogous to the activities of manual data abstraction and data quality assurance, as well as activities to train and evaluate clinical data processing algorithms. A complete enterprise pipeline to facilitate or automate data abstraction using clinical data processing would need to include tooling for the annotation of documents, performing quality assurance, monitoring, and comparing the performance of human and algorithm abstraction, search and information extraction, information consolidation, and processing into ultimately desired systems and formats.

6.4 Objective

The objective of this research was to design and implement an end-to-end pipeline of tools and methods to facilitate shared information processing by both algorithms and human staff, shifting the burdens of manual to automated clinical data processing over time in order to achieve scalability, consistency, and sustainability. This solution should support ongoing abstraction, ongoing generation of training and validation data, ongoing monitoring, and the opportunity to tune performance of both people and algorithms. For initial affordability, sustainability and broad applicability, such a solution should be based on an open-source platform so that algorithms and all components of the workflow are interchangeable, portable, and extensible. Finally, it should be scalable to an ever-increasing volume and variety of documents while complying with the high security requirements of clinical information processing. The intention of this work is to provide a framework for developing and deploying any kind of clinical data processing algorithm: commercial, open-source, or homegrown.

The Fred Hutch and LabKey clinical data processing pipeline is meant to be a platform not only for the deployment of clinical data processing algorithms to automatically extract data elements from clinical narratives, but also for the assignment, tracking, and completion of manual abstraction tasks, ongoing and systematic quality assurance of both manually abstracted and automatically extracted data, and the iterative creation of training data and clinical corpora through the linking of structured and unstructured data. The pipeline is designed to gradually take the place of a wide variety of EDC systems and databases currently used for manual abstraction.

6.5 Approach

The concept of an enterprise clinical data processing pipeline at Fred Hutch was initiated in 2012 as part of the Hutch Integrated Data Repository and Archive (HIDRA) program, a multiyear project to develop a data platform to support clinical and translational research. LabKey Software was selected as a technical and software development partner based on their clear understanding of the requirements and demonstrated ability to develop large-scale biomedical data management systems. At the time that HIDRA was initiated, staff at Fred Hutch had minimal practical knowledge and expertise around the practical application of NLP tools and methods. To gain expertise and build practical understanding of NLP through the center, the HIDRA leaders began presenting the basic concepts of NLP and how it could be applied to disease research teams and IT staff.(89) Fred Hutch began an NLP summer internship program that would bring in at least one graduate-level computational linguistics expert per summer and match them with research groups to explore a specific problem or opportunity that seemed amenable to an clinical data processing solution. These projects either leveraged existing datasets that had been previously abstracted for research or developed training data as part of the project. To date, two of these NLP interns have been employed fulltime at the cancer center.

Clinical data processing tools and methods were not useful without staff who had the knowledge and expertise to apply them locally, so developing a talent pipeline for NLP experts was a critical first step. Fred Hutch is part of a consortium NCI-designated comprehensive cancer center with its partners UW, SCCA, and SCH. To advance clinical data processing expertise at Fred Hutch, the NLP leadership team collaborated with the UW Computational Linguistics program, which has a practical and well-regarded Master's level training program for NLP tools and methods. Obstacles in clinical data acquisition and opportunities to advance

cancer research at Fred Hutch using clinical NLP were presented annually to UW Computational Linguistics students.

A series of projects was used both to explore the application of clinical data processing tools and methods across different disease groups (e.g., specialists in cancers of breast, brain, pancreas, lung, sarcoma, blood) and to give cancer center faculty and staff practical exposure to NLP concepts, tools, and methods. These projects included extraction and computation of pancreatic and lung cancer staging information,(90,91) extracting clinical timelines of patients with brain cancer(92), an assessment of inter-abstractor agreement for breast cancer pathology data elements from the same patient population abstracted for different research databases, determining chemotherapy administered within 30 days of death for patients with acute myeloid leukemia (AML),(93) developing a method to automatically classify patients with sarcoma from pathology report text,(94) parsing karyotypes from cytogenetic reports and classifying risk levels in patients with AML,(95) and determining the smoking history of patients with cancer from clinical notes. In addition, a project for extracting lung cancer biomarker results (e.g., ALK, EGFR) from pathology reports is underway.

The algorithms developed to determine chemotherapy regimens administered within 30 days prior to death in AML patients was trained with documents from 24 patients with AML who came to SCCA between January 2010 and December 2012, were at least 18 years old, and received chemotherapeutic agents within 30 days of death. The algorithms were tested using records from an additional 30 patients. Rules-based NLP algorithms were developed for date of death and chemotherapeutic agents. The recall for date of death and chemotherapeutic agent algorithms was 92% in the training sample. In the testing sample, accuracy was 96% and recall was 73% for both algorithms.

The algorithms developed to extract clinical events to facilitate patient timeline creation were trained with clinic notes from 330 patients with brain cancer from SCCA. These notes were divided into training and testing datasets of 165 patients. A statistical and rules-based NLP algorithm was developed to extract key words. As this work was unfinished, I do not have performance measures. However, NLP algorithm development is not intended to be part of this dissertation or this chapter anyway.

The algorithms developed to supplement service line classification of patients with sarcoma was a very simple rule-based program that employed a gazetter of sarcoma histologies and pattern matching to identify non-negated mentions of sarcoma in the final diagnosis section of pathology reports. The algorithms were developed using 42 pathology reports, 19 of which came from known sarcoma patients, and 23 of which came from known non-sarcoma patients. The algorithm had 100% precision, but only 89% recall.

The algorithms for pancreatic cancer diagnosis and staging were developed using medical oncology and procedure notes from UW and SCCA selected using ICD-9 codes 157-157.9. 63 notes were selected for training, and 26 were set aside for testing. The algorithms were rules-based, and developed using the pyConTextNLP toolkit. The algorithms has 61.5% recall, and 95.5% precision.

The algorithms for extracting lung cancer stages from free-text clinical notes were developed using a corpus of 21,535 clinic notes from 485 lung cancer patients from SCCA. 60% of the patients were used for training, and 40% were set aside for testing. The algorithms were all rules-based and has a recall of 96% and precision of 90%.

As the HIDRA planning team reached out to every disease group at the cancer center for requirements analysis, they collected data dictionaries from existing databases. In cases where no

databases existed in a disease group, the team collected lists of data elements that would be desired to support their respective clinical and translational research goals. In total, the HIDRA team collected 14,220 data elements from 13 major disease groups, including 28 subgroups. Over a 10-week period and 600 person-hours of effort, NLP experts and interns normalized the 14,220 field names into concepts and then condensed them to a single list of 4,000 elements that covered all disease groups. Each of these data elements was then traced back to its source system to determine if it was natively structured (e.g., from clinical labs, demographics, procedure and billing codes) or unstructured (e.g., from pathology reports, radiology notes, surgery notes, clinical visit notes). The research goal was to determine the potential impact of various forms of automation, including developing clinical data processing for narrative medical records, providing a mechanism for patient-reported data, and developing algorithms to derive data elements from other source data (e.g., age at surgery derived from date of birth and date of surgery). With a conservative estimate, the team found that 65% of desired data elements to support cancer research come from unstructured sources, 15% could be patient reported, and 15% could be derived from other data elements. These categories were not mutually exclusive. In terms of document sources, 50% of the data elements came from clinical visit notes, 20% from pathology reports, 15% from surgery notes, 5% from radiology reports, and 10% from other sources. This analysis provided validation that investment in an enterprise clinical data processing pipeline could yield value for research across all disease groups and was in line with similar analyses at other institutions that found the number of clinical data elements from unstructured notes to be anywhere from 45% to 80%.(76,79–81) Based on this analysis of data elements and data sources, the theoretical maximum proportion of desired clinical data elements

that could be obtained in discrete form from electronic clinical systems (including templated notes) was around 35%.

The clinical data processing team also conducted a preliminary comparison of existing commercial, academic, and open-source tools. The team evaluated and continues to track existing open-source and commercial technologies for the development of clinical data processing pipelines. There were numerous existing clinical data processing platforms for processing unstructured clinical text, such as cTAKES(88) and OpenDMAP.(96) However, these did not include parallel manual annotation and quality assurance work streams to facilitate the creation of training data for clinical data processing algorithms or auditing of automatically or manually extracted data. While these tools and other existing open-source biomedical language processing algorithms may be used for developing automated information extraction at numerous points within the clinical data processing pipeline, Fred Hutch also required the ability to incorporate manual annotation, quality assurance, and task management in one high-throughput system.

In addition, the team evaluated many well-known and widely used annotation toolkits, such as Brat(97) and Knowtator,(98) but these tools did not have any built-in clinical data processing algorithms or tools. GATE(99) is a platform that can be used for both the automated processing of raw text and the manual annotation of corpora. However, not all of its components were free and open-source, and while its flexibility was extremely helpful in the development process, it may not be user-friendly enough to provide an efficient tool for large-scale manual annotation and abstraction tasks for a variety of users. Commercial NLP packages like Linguamatics(100) offered some out-of-the-box solutions for linguistic queries and information retrieval, but the limited output and overall flexibility was not conducive to the growth and

change of a portfolio of algorithms over time, and it did not provide a platform for the creation of high-quality labeled data for training and validation of clinical data processing algorithms.

The initial model for structuring and storing information from the Fred Hutch and LabKey clinical data processing pipeline was based on the Caisis(4) system, which is an open-source, web-based cancer data management system and clinical data management system widely used by data abstractors at Fred Hutch.

Although there were numerous commercial and open-source options for components of a clinical data processing pipeline, the goals for the pipeline at Fred Hutch were more comprehensive and integrated than any existing platform could provide and required that the solution scale to support growing usage of NLP at an enterprise level. To facilitate current data abstraction and support a transition to a more facilitated and automated approach to information processing, the platform would need to support the following activities:

- Intake and management of a variety of documents and related clinical information
- Simultaneous EDC and annotation of source documents to provide tools for abstractors as well as high-volume training and validation data
- Ability to configure, manage, and evaluate the performance of a variety of ongoing data abstraction, annotation, and clinical data processing algorithm validation tasks within an integrated platform
- Scalability to support high volume (5,000 - 6,000 new cases diagnosed or treated at the Fred Hutch/UW Cancer Consortium per year, and more than double that amount of follow-up encounters, consults and other cancer-related visits to process)
- Extensibility of the platform with costs that are affordable and sustainable for centers with varying and unpredictable levels of funding

- End-to-end security appropriate for HIPAA-compliant and FISMA-ready environments, including access controls, and the ability to train and run algorithms within a high-performance computing environment that also provides robust access controls, encryption, de-identification, and audit logging
- Ease of use for a small IT and clinical data processing team to manage

In fall 2014, after the HIDRA core data platform, clinical data feeds, high-security environment, and a pilot self-service data access tool (Argos) were developed, the Fred Hutch informatics team began work on the clinical data processing pipeline in collaboration with LabKey Software. The clinical data processing pipeline vision and high-level requirements were articulated so that it would be a distinct module within the LabKey Server data integration platform,(73) and also would meet all the high-security, performance, ease-of-use, and integration requirements for the HIDRA enterprise data platform. The initial design for the pipeline leveraged LabKey experience and existing tools for bioinformatics pipelines (e.g., proteomics, genomics, flow cytometry) so that users would have a common approach within LabKey Server to develop and configure both clinical data processing and bioinformatics pipelines.

Detailed planning and development of the clinical data processing pipeline began in fall 2014. The clinical data processing pipeline development team consisted of an NLP research engineer, LabKey Software developers and project managers, and informatics leaders from Fred Hutch. It followed LabKey's Agile software development methodology that had been employed successfully for the development of the Argos self-service data exploration tool for HIDRA. Throughout the development of the clinical data processing pipeline, this team engaged with

Fred Hutch IT staff and experts to understand data sources, input and output requirements, application program interfaces (APIs), and security requirements. The team also engaged with researchers and data entry staff in disease groups to understand workflow and usability requirements for data abstraction, annotation, and clinical data processing algorithm validation.

6.6 Results

Fred Hutch and LabKey Software have designed an enterprise clinical data pipeline to serve as an integrated platform for:

- Automated information extraction
- Manual abstraction and verification
- Workflow and task management
- Generating training data for algorithm development

Conceptually (see Figure 6.1), source documents such as pathology reports are retrieved and fed from a clinical source system or data repository - in this case HIDRA - into a staging area. Clinical data processing pipeline jobs configured in LabKey Server pick up the source data and provide the tracking and tools to process the incoming clinical documents. The clinical data processing engine, as depicted in Figure 6.1, provides for interchangeable algorithms for information extraction and processing, and LabKey Server provides user interfaces for manual information extraction and processing or human verification of algorithm outputs. The clinical data processing engine is modular so that over time components or even the entire engine can be upgraded or replaced with commercial or open-source engines and algorithms. The output from

the clinical data processing engine is shepherded by LabKey Server to a downstream staging area for subsequent ETL processes to move data into other systems.

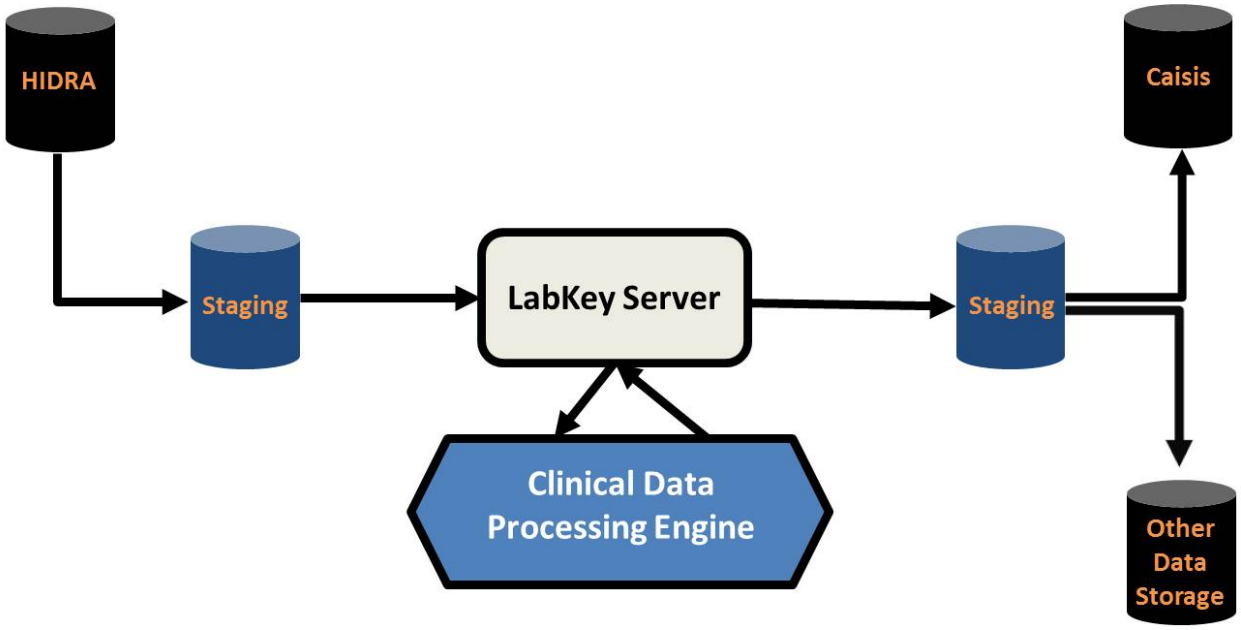

Figure 6.1. Conceptual diagram of a clinical data processing pipeline as deployed at Fred Hutch cancer center

The clinical data processing engine, shown conceptually in Figure 6.1, is intended to be interchangeable with different commercial, open-source, and academic engines to allow for innovation. The clinical data processing engine used in the Fred Hutch deployment was developed in Python; its architecture is depicted in further detail in Figure 6.2. The engine can be called as a Python script with command line arguments, which are configured in a LabKey Server clinical data processing pipeline task. Internally, the engine has a hierarchical design where each level of the pipeline represents a separate physical directory that contains the Python scripts and modules, text file gazetteers, dictionaries, and other metadata needed to extract the targeted set of data elements. The engine is controlled through a high-level script, and consists of multiple component scripts that are organized by document type (e.g., clinic notes, radiology

notes, pathology notes) and disease group (e.g., lung cancer, brain cancer, breast cancer). Because the formats and some of the language of different documents is specialized, document parsers may be specialized to document type.

Furthermore, some of the fields to be extracted may be common to a document type and not vary widely by disease group. For example, in a pathology note, algorithms to extract the accession number, date of pathology evaluation, and pathologist could exist at the document level. This organization allows for document- or disease-specific terminologies or gazetteers, data dictionaries of elements to be extracted, and other specific information-processing components. Final logic to integrate results from disease-specific modules works at the document level. This organization of Python scripts has proven to be easy for various clinical data processing engineers, NLP interns, IT operations staff, and software developers to understand and extend. Moreover, the vast majority of bioinformatics pipeline and scientific computing jobs running on the Fred Hutch high-performance computing cluster are Python or R scripts. Having a common Java-based platform (LabKey) for data integration and security, and information-processing programs written in a common language (Python) across both clinical data processing and bioinformatics jobs, allows the center to find synergies and leverage data science expertise broadly.

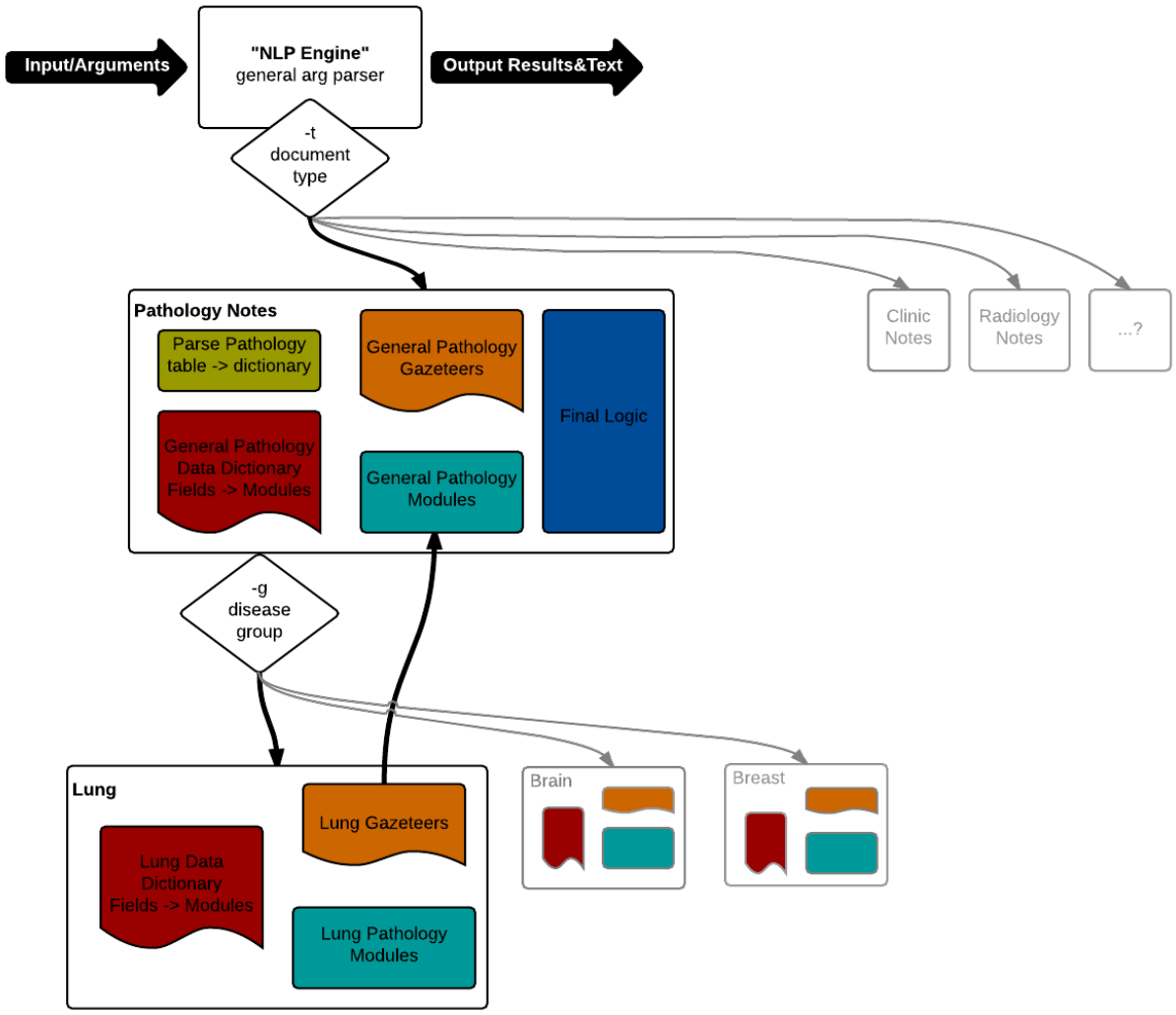

Figure 6.2. General architecture of the clinical data processing engine

The clinical data processing pipeline is engineered to support all of the interactions and typical combinations between human data abstractors, reviewers, and algorithms (Figure 6.3). The overall pipeline includes setup for abstraction, annotation, automated information processing, and review tasks through a configurable workflow module. The user interface for manual data entry (Figure 6.4) can also be used for general abstraction, quality assurance of automatically extracted data elements, audits of manually abstracted data, and creating NLP

training corpora through multiple parallel abstraction/annotation resources. This multipurpose user interface is part of LabKey Server, an open-source platform that helps translational research teams integrate, analyze, and share clinical data.

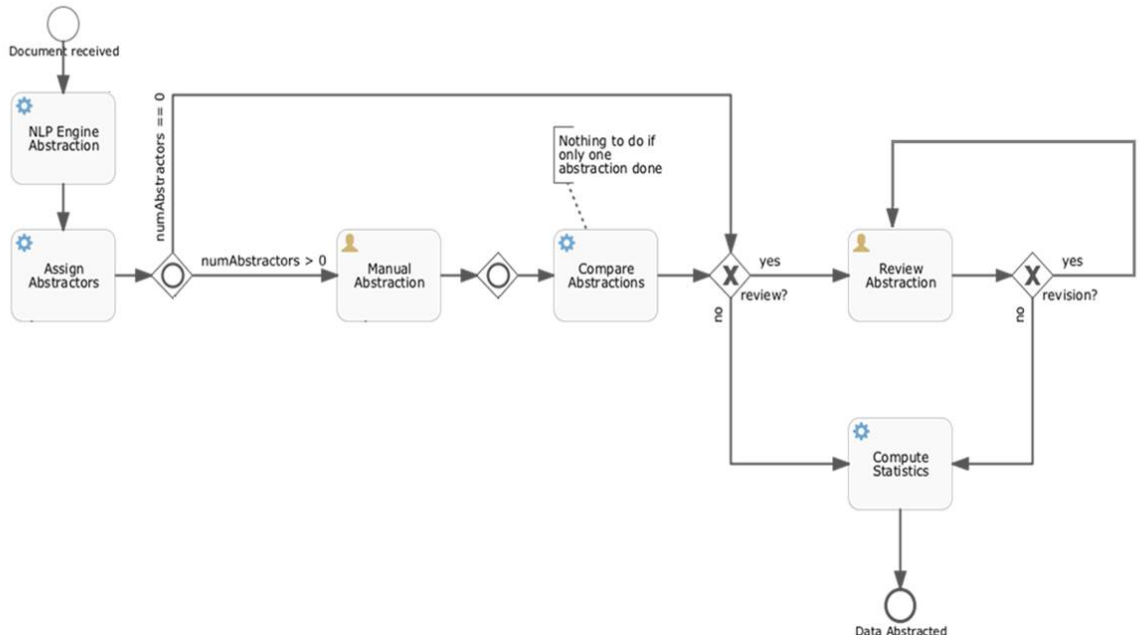

Figure 6.3. Detailed workflow diagram of the clinical data processing pipeline mediated through LabKey Server

One of the central and most important user interfaces for the clinical data processing pipeline is the data abstraction, annotation, and review interface (Figure 6.4). This interface provides a way to view and search clinical documents (left side of Figure 6.4), as well as to highlight features in the text such as keywords or phrases. It also provides an EDC tool (right side of Figure 6.4) to enter or review the abstracted or algorithm-populated output from the clinical source documents. The data elements for this EDC tool are driven by importable metadata rather than hard coded in LabKey Server. The character offsets (e.g., exact location of each keyword or phrase) and the associations between each highlighted feature in the source document with the entered or computed output values are all stored and accessible to NLP

engineers for training, validation, and quality improvement. Over time, the current EDC interfaces that are used by data abstractors and their workflows are anticipated to migrate over to this framework.

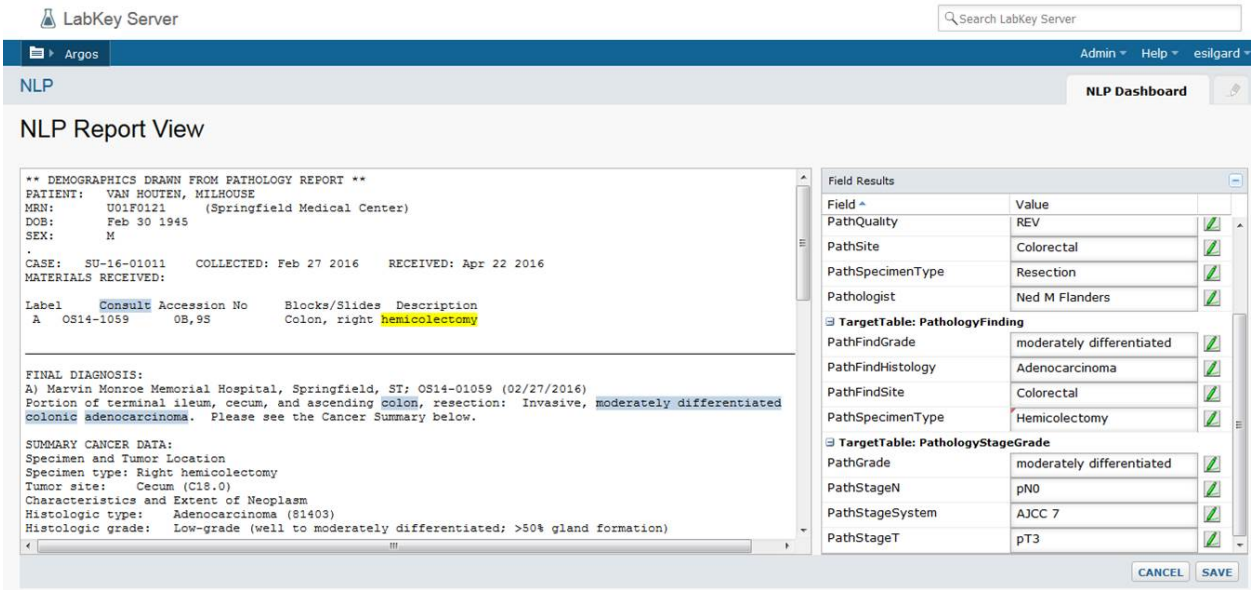

Figure 6.4. Data abstraction, annotation, and review interface. The source documents appear on the left, and a web form for the data elements to extract or review is on the right. Features in the source document that correspond to each data element are highlighted and stored so that links and locations can be used for algorithm training and tuning.

For a phased approach to development, a pilot demonstrating at least the minimal functions of the pipeline for a subset of reports was developed first. In October 2015, the use of algorithms for extracting lung cancer pathology data elements from pathology reports was demonstrated at the LabKey User Conference,(75) presented at Cancer Informatics for Cancer Centers (CI4CC),(74) and presented as a poster at the 2015 AMIA Annual Conference.(76) Subsequent work from Fall 2015 through Spring 2016 has been to implement all of the major pipeline functions in preparation for production implementation. The fully functional clinical data processing pipeline work includes [1] upstream integration to pull documents from clinical system sources and feed them to the pipeline, [2] configurable tasks for abstraction and

annotation and for running clinical data processing algorithms to populate data elements, [3] the ability to produce statistics to evaluate inter-abstractor agreement and performance of algorithms, and [4] downstream system integration to push results from staging area to other databases (see Figure 6.1).

Early clinical data processing algorithms to be deployed in the version 1.0 in production include:

- Assigning reports to disease groups
- Pathology report data element extraction for a variety of disease groups
- Karyotype parsing and risk classification in patients with AML
- Prognostic staging for lung cancer

6.7 Discussion

The understanding and application of clinical data processing has evolved at Fred Hutch and the Cancer Consortium, from early debates over what it actually is and what to even call it (e.g., “NLP”, “text mining”, “text processing”) and whether it would actually be useful at the cancer center to an increasing volume of collaborative research around NLP and requests for clinical data processing to facilitate a variety of data collection and data retrieval efforts. Although for the first couple years the intentional and consistent communications and ongoing collaborations around clinical data processing were challenging to setup, they have been worthwhile and are resulting in increasing demand from and value delivered to researchers.

The project required professional software engineers (at LabKey Software) to learn more about clinical data processing and NLP experts (at Fred Hutch) to adapt to Agile software development methodology and an enterprise data integration platform, which was a healthy

dynamic. Both groups have gained valuable experience, and the resulting module for the LabKey Server platform is now sustainable and portable to other centers. The clinical data processing engine developers do not have to maintain the system and then figure out if and how to commercialize and sustain it, which is a common problem with many tools developed at academic centers.

Many existing clinical data processing tools and cancer center efforts have tended to focus on NLP algorithm development or information retrieval (e.g., ability to search full text of reports). There do not appear to be other groups focused on the entire integrated data supply chain and framework to facilitate ongoing advancements in both clinical data processing algorithms and their implementation at the enterprise level. Also, annotation and EDC tool efforts at other centers seem to be disconnected from an overall clinical data processing pipeline. The collaborative work of Fred Hutch and LabKey Software appears to fill a gap in the field of biomedical informatics and in the landscape of commercial or open-source clinical data processing tools that are widely available, well-supported, and readily portable to other centers.

The approach of modeling the clinical data processing pipeline on bioinformatics pipelines and working with a technical partner and leveraging that expertise has accelerated mutual understanding of what and how to develop. Implementing a robust and configurable workflow engine within LabKey Server was deemed necessary given the complexity of possible workflows and the need for regular adjustments to configurations over time. The requirements for information security have proven increasingly important. Because the bioinformatics pipelines on which this clinical data processing pipeline was based did not have HIPAA- and FISMA (NIST 800-53)-level security controls, these critical features would have been difficult to build in if they had not been articulated upfront and if the team had not been able to leverage

parallel security work in LabKey Server to create a high-security platform for the HIDRA system. Because both teams were intimately familiar with HIPAA- and FISMA-level security requirements, numerous security issues were identified upfront and addressed in the pipeline implementation.

An interesting issue encountered was the storage of language models for machine-learning algorithms in a high-security environment. It was impractical to eliminate potentially sensitive information from the algorithms themselves, so the high-security platform controls and data hygiene practices for potentially confidential information had to be extended to both the clinical data processing engine and the computing environment that runs the engine. Clinical data processing algorithms would potentially require high-performance computing resources, however the clusters previously developed for bioinformatics pipelines did not require or provide HIPAA- and FISMA-level security controls. Moreover, version control of algorithms for the clinical data processing engine (see Figure 6.1) would need to be secure, and clinical data processing pipeline components in their own version control software would need to be regularly integrated and tested with the clinical data processing engine to produce versioned releases of the entire integrated platform.

As part of the development of the HIDRA database and its Argos self-service data access tool built on LabKey Server, the team found the need for realistic datasets for development and testing that could be considered completely de-identified. Although the clinical data processing pipeline is developed with security controls for PHI, such as the identifiers found in clinical documents, the team found use cases for de-identified clinical documents. Those use cases include the ability to provide full-text search of clinical documents without exposing identifiers, the ability to provide realistic de-identified training and validation data for clinical data

processing algorithm development, and the ability to provide a corpus of de-identified documents with realistic format, volume, and variety of content that can be used for further development, testing, and performance tuning for the entire pipeline.

The clinical data processing engine (Figure 6.1) and the visual abstraction and review tool (Figure 6.4) required the development of flexible data structures and user interface components to handle multiple specimens in a pathology report (or multiple findings in a radiology report). The output of clinical data processing algorithms is JavaScript Object Notation (JSON) structures, and the LabKey user interface for EDC (right side of Figure 6.4) was adapted to interpret JSON objects from the clinical data processing engine. The data elements are driven by metadata rather than being hard-coded into LabKey Server.

Several known problems with existing upstream parsers for pathology reports had a negative impact on clinical data processing engine performance; these problems are currently being investigated and remedied. For instance, an upstream parser may duplicate the records or lines in a pathology report. If the parser maintains the order of the duplicate records or lines, it poses little problem for the clinical data processing algorithm. Since the context is essentially the same, it just looks like the same thing was said twice. However, if the parser inserts a duplicate record or line in a different part of the report, the context for that line (its preceding and subsequent lines) may have changed, which could change its meaning. For example, if a line reads “1. High grade adenocarcinoma present”, and that line gets duplicated and inserted in a different location in the pathology report, it may appear that there was adenocarcinoma in two different specimens, rather than just one that the original report stated.

There are several next steps planned for this work:

- Full deployment and operation of the pipeline in the Fred Hutch high-security environment
- Gradual transition of the work of abstractors. Having outsourced data entry under a centralized service-level agreement (SLA) and centralized data quality management may facilitate this transition because contracts and SLAs can be modified incrementally and consistently over time. Transition may be more difficult in the situation of decentralized data abstraction and information-processing work.
- Ongoing clinical data processing algorithm development for potentially hundreds or thousands of algorithms to cover the range of desired data elements for cancer research
- Extension of the clinical data processing pipeline to multiple disease groups. Ideally, many of the algorithms will be common across diseases to minimize the development of disease-specific algorithms and dictionaries, however the clinical data processing engine architecture (Figure 6.2) also allows for disease- and document-specific extensions. The algorithms will need to be extended to cover multiple document types (e.g., cytogenetic reports, clinic notes, surgery and radiation oncology treatment notes, microbiology and virology lab reports). Some report types are inherently more homogeneous across all patient populations and will therefore require less disease-specific algorithm development. For cancer, several types of reports differ among diseases because of anatomic descriptions. In the case of pathology reports, there are varying histologies and related findings. In contrast, some of the anticipated data elements to extract from clinic visit notes involve social and family histories that may differ greatly from person to person, but are not substantially different across disease groups.

Because of the open and agnostic approach to the engine within the clinical data processing pipeline, LabKey Server could be used with a variety of clinical data processing engines, while keeping overall workflows and framework for annotation, review, quality assurance, and task management the same. This will allow the platform to support more rapid advancements in the practical application of clinical data processing in cancer centers or other organizations. The clinical data processing pipeline is portable to other centers, scalable to the enterprise level, and extensible both through the substitution of different clinical data processing engines and through the addition or configuration of different modules within or interoperable with the LabKey Server data integration platform. This platform and approach is also amenable to cloud-based implementation. Because of the requirements for implementing HIPAA- and FISMA-level security controls and the need to scale the storage and computational capacity to run a clinical data processing pipeline across the enterprise, a cloud-based implementation may be the most affordable and sustainable strategy for many centers.

The intent is that this platform would be relatively portable to any group or institution. The initial parsing and preprocessing of clinical document may have to be adapted to match organization-specific sources and document formats, however the data elements needed for cancer research should remain relatively similar.

6.8 Conclusion

Fred Hutch and LabKey Software have developed a platform that provides a secure and open-source pipeline for the acquisition and processing of clinical documents and performs automated information extraction and case data consolidation through the use of a clinical data processing engine (e.g., locally developed algorithms or other commercial or open-source tools). This

pipeline also provides a framework for manual information-processing tasks such as abstraction from raw reports for ongoing annotation projects as well as verification and performance evaluation of clinical data processing algorithms. The LabKey Server clinical data processing pipeline can be used to integrate multiple automated processing solutions within a single open-source but well-supported architecture. Integrating the LabKey clinical data processing pipeline with existing data abstraction workflows has the potential to incrementally decrease staff time and efforts through automations and increase the amount of annotated training data available for future clinical data processing algorithm development.

6.9 Acknowledgements

The authors wish to thank Kristin Dubrule, Adam Rauch, Susan Hert, Ian Sigmond, and Ryan Standley, LabKey employees who contributed to the design, development, and testing of the clinical data processing pipeline and workflow.

6.10 Synthesis

Chapter 6 described the pipeline for clinical data processing and is closely related to Chapter 4 about the HIDRA system. Rather than designing a new database model for HIDRA and the clinical data processing pipeline, we initially relied on the Caisis database model as a target. However, some performance and usability issues have become increasingly evident with the Caisis database model, both through the self-service data access tool, Argos, and through the development of the information-processing pipeline for clinical data. These previously known and emerging database model issues as well as potential solutions and avenues for further research are explored in Chapter 7.

Sections 6.2 to 6.8 addressed my Chapter 6 questions: How can we improve the quality, speed, and economics of the acquisition, processing, and delivery of clinical data to support cancer researchers? How can we make clinical data processing a core competency of Fred Hutch at an enterprise scale? This work ties back to the overarching question for this dissertation (how can we improve access to clinical and related data about cancer patients for research?). Getting clinical data from source documents into formats usable for most research requires extraction and transformation from narrative medical records into discrete, coded data elements, which is the purpose of the pipeline for clinical data processing. Chapter 6 also ties back to the overall hypothesis that there were new tools and methods from biomedical informatics that could improve the availability of data for cancer research if they were applied thoughtfully and strategically, and to Aim 2 to develop and characterize a clinical data processing pipeline that is scalable to the cancer center enterprise level, is well-supported and sustainable, and can complement or streamline existing manual data abstraction and information-processing activities.

All of this work logically leads to Aim 3 (Chapter 7) - to develop, model, and assess database frameworks for cancer - because the data integration and self-service data access from Caisis (Chapter 2), HIDRA (Aim 1, Chapters 4 and 5), and the document-oriented clinical data processing pipeline (Aim 2, Chapter 6) have run into potential performance and usability issues related to their underlying data models and database technologies.

Findings from this chapter are relevant to other cancer centers addressing the need to transition from manual data entry of information from medical records into database to automated approaches to data processing.

First, one of the greatest barriers to advancing automation of clinical data processing is the limited availability of training and validation datasets that have been annotated for use by

statistical algorithms, NLP or machine learning tools and methods. The work in Chapter 6 provides a solution to this problem that is agnostic to the automation tools.

Second, the work in this chapter is generalizable to other centers because the clinical data processing pipeline was developed with a technology partner, LabKey Software. Because of LabKey's—and our—desire to make this solution portable to other centers, it was not tightly coupled to the IT and informatics infrastructure or data elements. Rather, it was designed and built from the ground up a portable, scalable, extensible pipeline. The work in this chapter is already being adopted by the NCI SEER program for developing training and validation data for collaborative NLP and predictive modeling research with the Department of Energy laboratories, and may be adopted by a large research network of over a dozen cancer centers. Other cancer centers and LabKey customers have expressed interest in adopting this pipeline tool.

Chapter 7 Comparison of Database Models for Cancer Research

7.1 Context

This chapter is my exploration of Aim 3, to develop, model, and assess database frameworks for cancer. To explore this aim, I answer the following question: How can we best characterize and compare database models and big-data technologies to inform a sensible strategy for cancer research database design and technology? Because database models, design, and technology can both enhance and hinder ready access to data, answering these questions helps inform the question "How can we improve access to clinical and related data about cancer patients for research?" Aim 3 is a logical next step to test my overall hypothesis that there are new tools and methods from biomedical informatics that could improve the availability of data for cancer research if they were applied thoughtfully and strategically. There are numerous database designs promoted in the biomedical informatics community (e.g., i2b2, OMOP, BRIDG) and numerous big-data companies and technologies that promise solutions for healthcare and biomedical research data. However, these solutions are not necessarily applied thoughtfully and strategically. This chapter was my attempt to get my head around these options in order to guide the discussion and evaluation of database models and big-data technologies.

Aim 3 flows from Aim 2 (to develop and characterize a clinical data processing pipeline that is scalable to the cancer center enterprise level, is well-supported and sustainable, and can complement or streamline existing manual data abstraction and information-processing activities), from Aim 1 (to develop and assess a modern integrated data platform to support a wide variety of cancer research), and from Chapters 2 and 3 as well.

I wrote this chapter as an exploratory paper to expand on observations of the Caisis database model in Chapter 2 and emerging alternative models (e.g., i2b2) introduced in Chapter 3, and to guide strategy for data models to support the expansion of HIDRA and Argos from Chapter 4 and the clinical data processing pipeline in Chapter 6. However, after finishing this work, I do not think it is publishable yet as a distinct journal article. Rather, I plan to use some of this material in papers about related work and seek to publish this work after doing further database model experiments outlined in next steps. Because this chapter was originally written as a standalone paper, I have included a separate acknowledgements section.

7.2 Abstract

Database models selected or designed to support biomedical research can result in challenges for system implementation, usability, and sustainability. With a steady increase in volume and variety of incoming data for cancer research, and the explosion of big-data tools and database platforms, many centers struggle to understand and articulate issues and tradeoffs between approaches and to develop attainable, sustainable strategies for data management to support research. This chapter describes database and data quality concepts, common issues with database models in cancer research, and lessons learned about the tradeoffs of different modeling approaches. It recommends a general approach to thinking about and evaluating database model options.

7.3 Background and Rationale

Current and next-generation cancer research relies on ready-access to and use of data about patients from a variety of sources. The story of each cancer patient as they go through life and

their various encounters with the healthcare delivery system is expressed in data from medical records that document episodes of care (e.g., clinic visits, diagnostic imaging and lab results, procedures and medications, outcomes or complications of medical interventions) and in consumer and environmental data that may indicate exposures, behaviors, and other factors that could affect the development and progression of cancer or the outcomes of treatment.(1)

The bulk of these clinical, consumer, and environmental data are not collected for the purpose of research. Clinical data from patients' encounters with healthcare providers are typically documented in EMRs and related clinical systems, and the purpose of those systems is to record patients' histories from the perspective of physicians and other providers, to provide medical documentation for billing and legal purposes, and to operate a healthcare facility. Secondary use of the information from clinical systems for research has required manual data abstraction (chart review) and information processing (filling out paper or web-based forms for specified data elements with coded values) for each research project. Cancer diagnosis and treatment information from medical records is also abstracted and processed manually into designated systems for hospital registries and state-mandated cancer reporting.

The majority of clinical data needed for research and public health reporting is still recorded in narrative form (e.g., in pathology and radiology reports, clinical visit notes, treatment summaries) and frequently even the structured, discrete data elements available from treatment and payment systems (e.g., demographics, clinical lab results, medication orders, claims), require processing and consolidation for secondary use.(13) For example, a research project may require documenting the highest abnormal lab value that is nearest to but before primary cancer treatment rather than just documenting all lab values for a given patient by date. Using human data abstractors and data managers to review information in medical records, extract key

elements, determine which elements to record and how to code them into a desired format for research or reporting is a foundational activity for both research and cancer surveillance.

Many groups are experimenting with project-specific or enterprise-level application of clinical data processing and machine learning to automate or facilitate manual abstraction of coded data elements from raw and narrative medical records data sources. However, the difficulty in implementing these technology solutions for automation and their ultimate usefulness may partly depend on underlying database models that provide target data formats, relational structures, constraints, coding of information, and query access.

Currently, there is tremendous duplication of data abstraction and information-processing efforts to support cancer research because each investigator and each study may have a different desired database model according to their research question, perspective, and analysis approach. Efforts to integrate project-specific databases or build reusable databases that can support a variety of applications often create as many issues as they solve and can frustrate investigators. The payoff from implementing reusable clinical data repositories often takes time for the systems and operations to mature, and many investigators and research projects may not wait for the payoff. Also, each approach to implementing reusable database models comes with issues and tradeoffs that are often difficult to explain and justify to investigators, especially those without training in database models. This work is intended to provide some concepts, language, and examples to explain different database model approaches and to facilitate discussions or decisions about how to collect and store data for various investigators and studies.

In the world of cancer research, and biomedical research in general, the constraints of different database models may become usability issues. Conceptual and structural database issues manifest across a variety of IT and informatics endeavors, including the following:

- Design and usability of paper and electronic forms for collecting clinical data for a variety of studies, from clinical trials to epidemiologic and correlative research
- Design and usability of study-specific databases versus disease-specific registries and integrated data repositories
- Issues or conflicts around data structures, coding, and formats between statisticians and clinical trials experts versus relational database developers and IT/informatics staff
- Convenient duplication and fragmentation of data collection for disparate research projects versus inconvenient but unified data collection across a variety of projects
- Profusion of requirements analyses, data element mapping and recoding exercises, and competing approaches to information modeling and storage that may be more fueled by marketing and popularity of technology platforms (e.g., Hadoop and NoSQL databases) rather than thoughtful consideration of the tradeoffs of different database approaches(101,102)

Common questions and proposed solutions arise around the struggles with database models. Should we just use REDCap,(21) or have someone build a relational database and web application? Should we build an enterprise data warehouse and create data marts for each unique project to meet data requirements? Should we put everything in a Hadoop data lake on Amazon Web Services (AWS)?(103) Should we make all of our research data available in i2b2?(8) Should we build research data elements into clinical templates and checklists and then use the EMR for research data? Questions like these are examples of common symptoms of database model frustrations and notional solutions. However, each of these questions can be unpacked into underlying assumptions, tradeoffs, and design implications. There is a need for a framework

and recommendations to inform discussion of database models and to guide the application of different database designs and technologies.

7.4 Objective

The goals of this work are to [1] characterize common issues with database models in cancer research and surveillance, [2] trace common issues back to underlying assumptions and concepts, [3] describe tradeoffs between different modeling approaches based on the application of underlying assumptions and concepts, and [4] provide examples of how to recognize and apply this knowledge for cancer data. Overall, the purpose of this research is to facilitate the discussion of new technologies, to inform the application of different database approaches and technologies, and to advance the current and next generation of database design for cancer research.

7.5 Methods

For this research, common issues with database models in cancer research centers and cancer surveillance are identified and described. The purpose of this paper was to document and explore the database model concepts and technologies that I have encountered and to compare them in a manner that would inform strategy for application in cancer centers. Therefore, the methods are aligned with the goals stated in the objective above. In this work, I [1] characterize common issues with database models in cancer research and surveillance, [2] trace common issues back to underlying assumptions and concepts, [3] describe the tradeoffs between different modeling approaches based on the application of underlying assumptions and concepts, and [4] provide examples of how to recognize and apply this knowledge for cancer data.

Common assumptions dominate for this comparison of database models. First, that investigators and users of biomedical databases are concerned with data quality. They want to be able to rely on databases to accurately answer research and operational questions. In my experience, databases are frequently criticized in terms of quality, but definitions of quality vary. Therefore, it is useful to begin the exploration of database models with a definition of data quality.

A helpful definition of data quality for this work was published by the United Kingdom National Health Service (NHS).(104) NHS defined high-quality data as:

- Accurate
- Up-to-date
- Quick and easy to find
- Free from duplication
- Free from fragmentation
- Complete (also mentioned by NHS in the same source and added to this attributes list)

"Accurate" in this case will be defined either as *precision* as it is used in evaluating the performance of information retrieval and clinical data processing algorithms, or as *positive predictive value* as used in diagnostic and epidemiologic studies.(105) Precision at the data element or field level can be computed as the total number of correct values that can be verified in the source documentation (usually the EMR), over the total number of values populated in given data element or field in the database, as depicted in Figure 7.1 and Figure 7.2. The rationale for using precision as a measure of accuracy is described in section 7.6.2.

"Up-to-date" refers to the latency of information in a database relative to the point in time that the information is generated or recorded in a source system. An up-to-date database would include the most recent values from source systems. For cancer research, an up-to-date database would generally include any information in a source system that had been documented for at least 1 week before a data search, query, or export request. However, for clinical decision support tools, the latency requirements would be much stricter. Inconsistencies in data due to latency would not be acceptable when using a database model to make treatment decisions for individual patients.

Data that are "quick" and easy to find refers to both the ability to search for the data of interest and find or retrieve them quickly. This access could be either through a self-service tool or through a data request service staffed by analysts or data brokers. Quick and easy to find implies reliable naming conventions and metadata that can be searched and referenced so that less expert staff can explore the data with self-service tools with less assistance from domain and database experts.

Data "duplication" can become a problem over time in terms of costs and variations in data acquisition and handling, storage and management, and the confusion of potential database users. In addition to the expense of duplicate data abstraction and information processing, when data are duplicated across multiple databases, there is a risk of update anomalies where different copies can become out of sync. This inconsistency presents a problem for users who may query from different copies of supposedly the same data and get different results.

"Fragmentation" refers to having different aspects of a patient story in disparate databases and systems, so that queries or dataset requests involve linkages across databases, reformatting different data elements to a common format, and potentially inheriting different

assumptions and biases that affect the meaning of data. Fragmentation of sensitive data may also increase security and privacy risks. Data integration into a common database model and format on a common technology platform aim to reduce fragmentation.

Data "completeness" will be defined as recall (from information retrieval and NLP) or sensitivity (from diagnostic and epidemiologic studies).(105) As depicted in Figures 6.1 and 6.2, recall is the number of values in a data element or field with correct values over the total number of desired values for that data element or field found in the relevant source documentation.

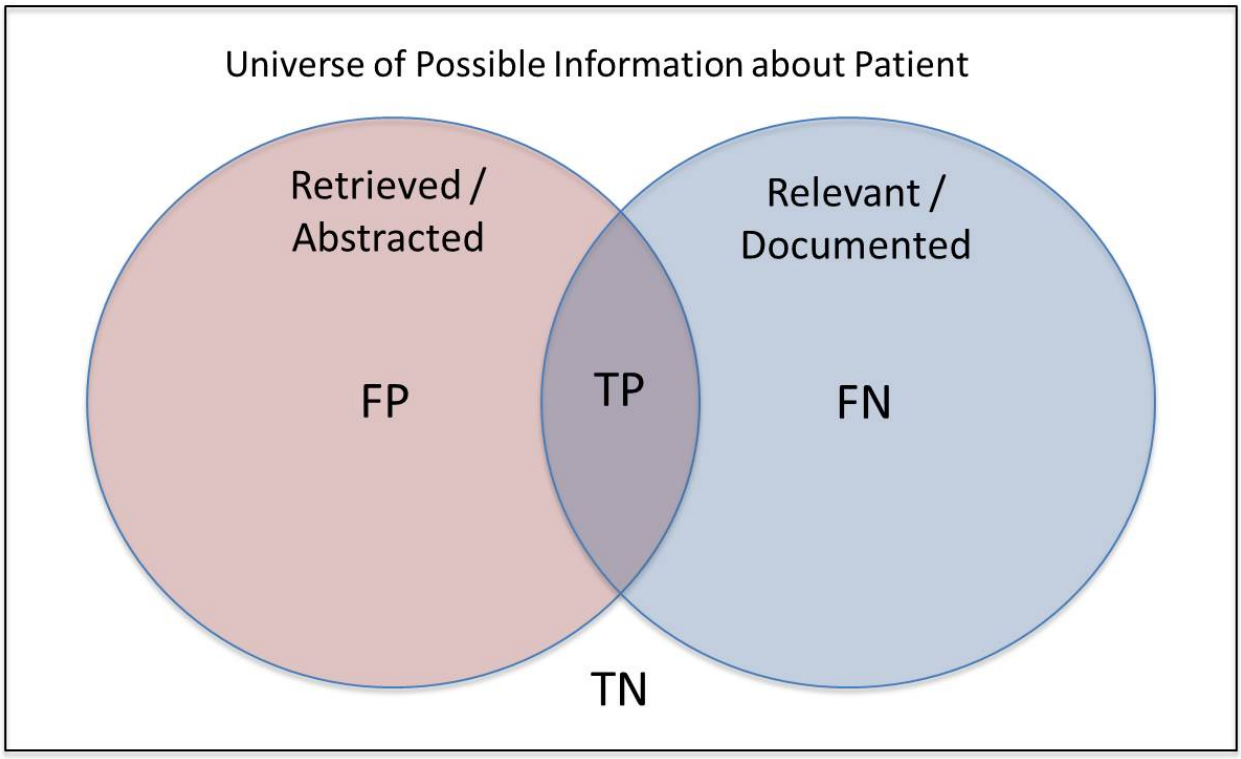

Figure 7.1. Venn diagram of relevant/documented and retrieved/abstracted information. In information retrieval, natural language processing and machine-learning methods, in place of "accuracy", performance is generally measured using precision (or positive predictive value), the number of True Positives (TP) over the total of documents retrieved or values abstracted from documents (TP+FP). Completeness is measured as recall, the number of TPs over the total number of relevant documents or documented values (TP+FN). The true negatives (TN) are not evaluated because they are neither retrieved nor relevant/documented.

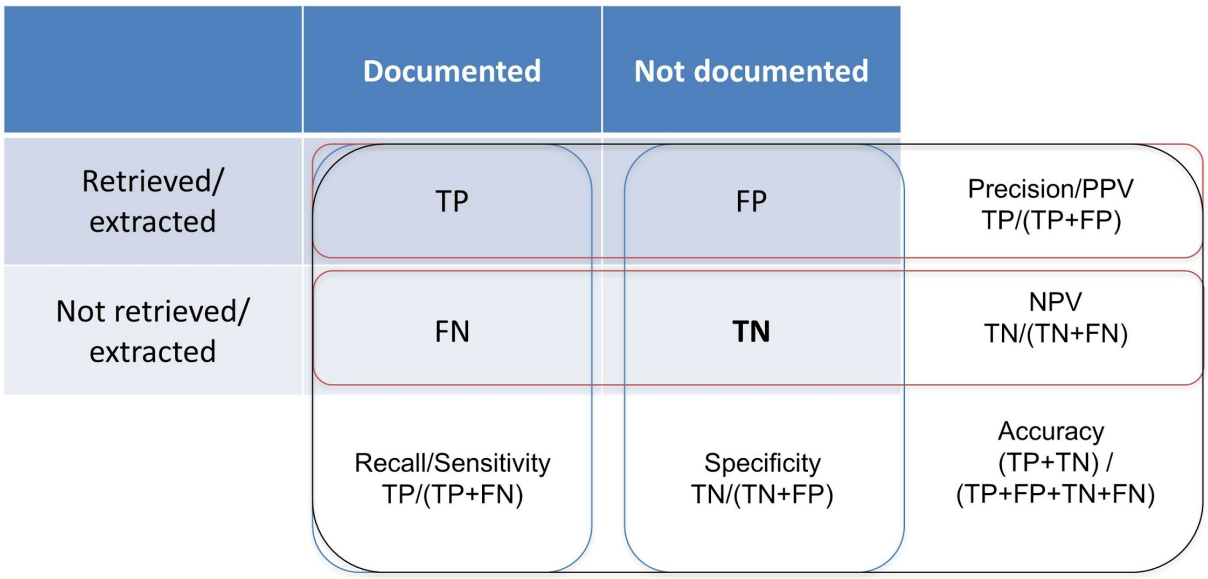

Figure 7.2. Measures of accuracy and completeness. Specificity, negative predictive value (NPV) and accuracy that include true negative (TN) values work for evaluating controlled experiments like trials, diagnostic studies, or epidemiologic cohort study surveys where the "universe" is the tightly defined, but not for clinical data repositories and cancer surveillance, where prevalence of data documented in medical records may vary greatly.

These concepts and measures of data quality will be used in the results and discussion below.

7.6 Results/Findings

7.6.1 Common Issues with Database Models in Cancer Research

Many of the issues found at the intersection of research and database development appear to come from the different viewpoints and constraints entailed by different modeling approaches and dominant technologies. A variety of common approaches address the problems of translating research and cancer registry needs into database models and system specifications. At a high level, these issues can be categorized into three areas of design: [1] how data are modeled in user interfaces for data acquisition and editing, [2] how data are modeled for storage, and [3] how data are modeled for retrieval, reporting, and analysis (Figure 7.3). The methods for structuring

data into information systems that cover these three functions tend toward either keeping the user interface, storage, and query/reporting models equivalent (e.g., in spreadsheets), or allowing the user interface, storage, and analytic models to diverge.

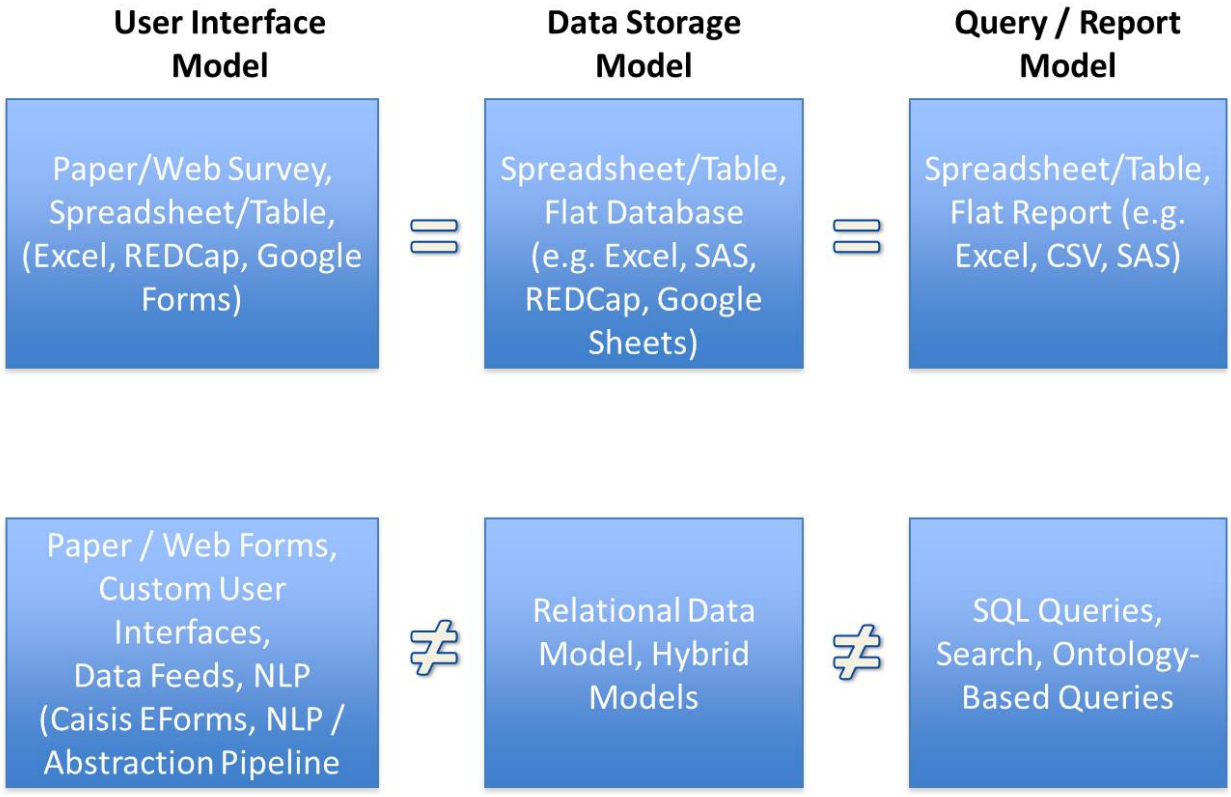

Figure 7.3. Some database models are equivalent for user interface, data storage and reporting; others compartmentalize and translate between these functions.

The database models where the three functions are closely aligned tend to make most sense from a narrower, single research project viewpoint. However, a major problem with a single, flat, project-specific database model to support all three functions is that data collected in this format is rarely reusable. Each patient record abstracted into flat formats requires data abstraction and data transformations that are project specific. Also, if these data were to be reused, the biases in data abstraction and transformation for one project may affect other

projects, especially with turnover in data management staff or students who abstract and transform information from medical records into a flat, project-specific database model.

The second set of approaches where the user interfaces, data storage, and analytic models may vary allows the development of common data entry forms and reusable tools, common data feeds or clinical data processing pipelines from source systems to populate databases, centralized database models that can store a variety of incoming data, and analytic models that can provide data output to a variety of projects. On the positive side, each of the functions may become reusable across projects, and staff can obtain platform expertise that allows them to provide cross-coverage and economies of scale across multiple projects. The downside of separating these three functions is that mapping and translation work is required both between the user interfaces and the data storage models and between the data storage and analysis functions. Mapping and translation work requires both biomedical research domain expertise and technical skills (e.g., writing ETL processes in SQL).

Through my experience of migrating numerous users to the Caisis database,(4) I have found that most data entry staff can adapt to different user interface approaches over time. In Caisis database implementations, after staff have entered all data for approximately ten patients using a general patient data user interface that closely maps to the Caisis relational database model, they are comfortable using that user interface. The issues that linger after the initial learning curve include the amount of drilling down and clicking to find the appropriate place for each data element, compared to the use of flat paper or web forms that are good for “heads down” data entry and virtually eliminate drilling down, but may lose some of the context of individual data points.

Flatter user interfaces such as custom REDCap forms, typical Access database forms, web applications that mimic case report forms or epidemiologic surveys, and Excel spreadsheets make sense intuitively to most users and investigators who do not have training and expertise in database design. However, flat data models do not generally map well to source systems (e.g., EMR, other clinical systems) and they force users to perform much more information processing to abstract, compute, recode, reformat, and otherwise transform findings from source documents into values in the desired format.

The primary advantage of flatter data storage models is that they tend to be easily understood by investigators, data managers, statisticians, or other analytics staff. These data models are often either single tables or a few tables that match the desired output format for an individual project or spreadsheets that grow over time with additional rows, columns, and sheets added to support multiple projects with a common disease (e.g., breast cancer) or a common intervention (e.g., bone marrow transplant).

Flat data storage models accumulate data integrity problems over time, such as update anomalies, repeating rows, repeating columns, and multivalued fields. Update anomalies occur when a primary or computed data point exists in multiple cells in a spreadsheet or table, and a user updates one cell but does not update the related cells. In cancer research databases, repeating rows commonly occur in spreadsheets that are organized around cancer cases or tumors (e.g., left and right primary breast cancers) rather than around patients. The repeated patient information across multiple rows is at risk for update anomalies. A common example of repeating columns in cancer research databases includes the dates and results of multiple lab tests, diagnostic images, or treatments over time for the same case or patient. Repeating columns of the same information become increasingly sparse and more difficult to query as the database

grows. Spreadsheet users may also enter multiple values in one field, separated by some kind of delimiter. These multivalued fields then become more difficult to query and update. All of these data quality issues can be considered symptoms of data potentially not fitting well in a flat database model, and there are common solutions to these issues obtainable through either migration to normalized relational database models or through transition to NoSQL models such as a column or document stores.

In cancer research, reusable databases are generally implemented as relational database models or hybrid database models involving one or more relational models as well as complementary document, file, column, or graph data stores. Most EMRs and related clinical systems either store data in a relational model or can make data available in a relational model. A generalizable approach to reusable data collection and processing to support a variety of research projects would likely entail feeding data from EMRs and other clinical systems into a data warehouse or data repository. A data warehouse would import source data from a variety of systems into a common technology platform (e.g., SQL Server database), link data by patient identifiers, classify data by disease or other high-level organizing themes, and potentially further link and organize data into individual medical encounters, episodes of care, or other temporal structures.

Most biomedical research data warehouses do not substantially transform data from source systems into a standard database model. Rather, they link and reformat data, but leave them in the data models corresponding to the source systems and rely on experienced analytics staff to have sufficient domain and source system knowledge to query and transform data from the warehouse for various projects or load it into other database models. This complexity, reliance on domain and source system expertise, and requirements for a high level of technical

expertise to transform warehouse data for individual research projects makes data warehouses of limited immediate use to most investigators and their staff. There are economies of scale in developing data feeds to data warehouses, however warehouses may not scale down well in a timely fashion to meet the long tail of desired data formats and outputs for individual research projects.

Data marts can be created from data warehouses for larger disease registries and research projects where the investment and time required for implementation and ongoing maintenance of a data mart are feasible. A data mart would perform common transformations to bring data from a warehouse into an analytic database structure that is probably closer to the desired query and reporting needs of a research group.

Biomedical data warehouses typically implement feeds, linkages, and some transformation of data from source systems but do not typically have user interfaces for manual data abstraction and information processing. The requirement for complex transformations, including both automated and manual information processing, is where a generalizable cancer research data management system like Caisis for cancer research or SEER*DMS(106) for cancer surveillance come in. These data models and systems are designed to integrate information that is manually abstracted from medical records documents with data that are fed directly from a data warehouse or source system. Similar generalizable data models for clinical trials and translational research have been developed (e.g., CDISC ODM,(107) BRIDG,(108) LS-DAM(109)), and they tend to be relatively complex in terms of the number of tables and fields and the relationships between them.

A reusable relational database can make functions available to enforce data quality that may be more difficult to implement in a spreadsheet or warehouse. These include checks for [1]

field level integrity, [2] record integrity, and [3] referential integrity. Field-level integrity means that the values or codes in an individual field match some set of allowable values. Record integrity means that various fields in one record are consistent (e.g., death date is greater than birth date in a patient record). Referential integrity means that records are correctly linked so that there are no orphan records (e.g., no lab values that are not associated with a patient).

In addition to the basic integrity constraints of a relational database, any database model and system that is reusable across multiple diseases and projects is likely to require customized data access rules and security controls, tracking of the provenance of data points (e.g., with users, timestamps, updates, deletes), customized user interfaces for data entry, editing and review of records, and customized data exports to meet analysis and reporting requirements. Implementing each of these customizations in a complex, generalizable system like Caisis requires technical expertise and system-specific knowledge to host, configure, and program, which is often beyond the skills and understanding of study staff in individual research groups. Individuals who could manage a spreadsheet or REDCap database on their own would become dependent on specialized IT staff if they adopted a data mart or generalized research database like Caisis.

To be useful to researchers, any database or system should make data readily available for queries and reports. In a flat model like a spreadsheet or REDCap, the data are already "export ready", though because of the lack of built-in controls for data integrity, they may require cleanup and re-coding before being used for analysis. Without built-in audit logs, it may be difficult to track data provenance on the editing and data wrangling performed to clean up a spreadsheet before analysis.

Reusable data models and tools to support analytics have also been developed. Two of the most prominent analytic models and systems are i2b2(8) and OMOP.(70) Compared to data

warehouses or clinical data management systems (e.g., Caisis, SEER*DMS), i2b2 and OMOP database models reduce the overall number and complexity of tables and fields, and embed complexity in ontologies or terminologies as well as in metadata. The i2b2 data model is based on the star schema, with a central fact table that can be filtered by different dimension tables and the concepts driven with an Entity-Attribute-Value (EAV) structure.(110) In i2b2, browsing, searching, traversing, and managing the ontologies associated with a repository can be complex and the single fact table, even though it is indexed, may have performance issues, high data volume, and variety. Like the i2b2 data model, the OMOP data model has relatively few tables and fields. However, OMOP is not a star schema, and rather than relying on an EAV structure to define concept dimensions, it depends on linked vocabularies for querying and representing concepts. The relative simplicity, stability and popularity of their information models make i2b2 and OMOP attractive for the development of associated tools for data exploration, self-service queries, and reporting. However, more complex queries and reports against these models typically require custom tools and ontologies to be developed, and the mapping and transformations necessary to fit a broad set of data elements into i2b2 or OMOP models can become daunting. These analytic models are typically best used for exploring relatively stable sets of data with limited sets of fields or dimensions that conform well to standard terminologies.

In addition, there have been continual advances in tools for distributed information storage, processing, and retrieval for massive consumer datasets (e.g., Google, Amazon, Facebook). Big-data companies and biomedical research groups have started exploring these tools for making data available to investigators and data scientists. big-data technologies are developed and widely used out of necessity in internet search and consumer data businesses. Platforms for distributed storage (e.g., S3, HDFS), data management (e.g., DynamoDB, HBase,

Cassandra), processing (e.g., Elastic MapReduce, Spark), search (e.g., Lucene, Solr) and machine learning (e.g., TensorFlow, Mahout) may seem like potential silver bullets for exploring and analyzing biomedical data, but at the moment these platforms still require considerable technical skills and domain expertise to apply to cancer research, and they may end up performing poorly in production if applied without a strategy and design.

A relational database model (e.g., Caisis, SEER*DMS, OMOP, i2b2) with greater integrity constraints than many NoSQL data stores requires ETL processes written by SQL and domain experts to populate. The transformations must happen before data can be loaded into the relational data model. A NoSQL database may allow some data transformation to be postponed, the so-called Extract-Load-Transform process. However, data will eventually need to be transformed in a timely manner to a database model that is understandable by end users and accessible to their preferred tools for statistical analysis, reporting, and data visualization.

Populating a data model for research typically requires access to or export from source systems, data feeds, manual data abstraction, and transformations through manual or automated coding and consolidating records into usable formats. Although it may be economically attractive to implement data feeds, templated notes in clinical systems, and centralized databases to support a variety of projects and to eliminate some duplication of effort, fragmentation of storage and inconsistencies in the accuracy and completeness of data extraction and processing, centralization does not eliminate project-specific data-processing requirements. Rather, it makes research teams dependent on centralized services for data acquisition, processing and delivery.

For a flat and project-specific database model (e.g., Excel, REDCap), each record is abstracted and transformed upfront into the desired format, and the burden for timeliness and updates falls on the study staff. If stored in same format as desired for analysis, there is little

need to transform the data at the time of analysis. In a generalizable research database, some of the extraction and translation of data to desired formats may be performed upfront, but further transformations will be needed for each project that draws from the repository as a data source.

7.6.2 Underlying Assumptions and Concepts

To understand the assumptions and concepts behind cancer database models, it is useful to consider the underlying diseases and how information about cancer fits within biomedical database models. Figure 7.4 shows an underlying cancer disease progression model (in the blue line), and how medical data relate to that underlying construct. Cancer presence and progression are often invisible to the clinician. Various instruments (e.g., physical exams, diagnostic images, blood tests, biopsies) are used to detect and estimate the extent, type, and aggressiveness of cancers, and various interventions (e.g., surgery, chemotherapy, radiation) are aimed to stop or slow progression of the cancer. Clinical data in medical records reflects not the underlying disease, but rather the documentation of diagnostics and interventions applied to the patient during their clinic and hospital visits. In addition to clinical data, there may be public and commercially available demographic, environmental, and consumer datasets that document potential exposures and behaviors related to the development of cancer and the outcomes or side effects of interventions. However, part of the underlying disease and progression is inherently unknown.

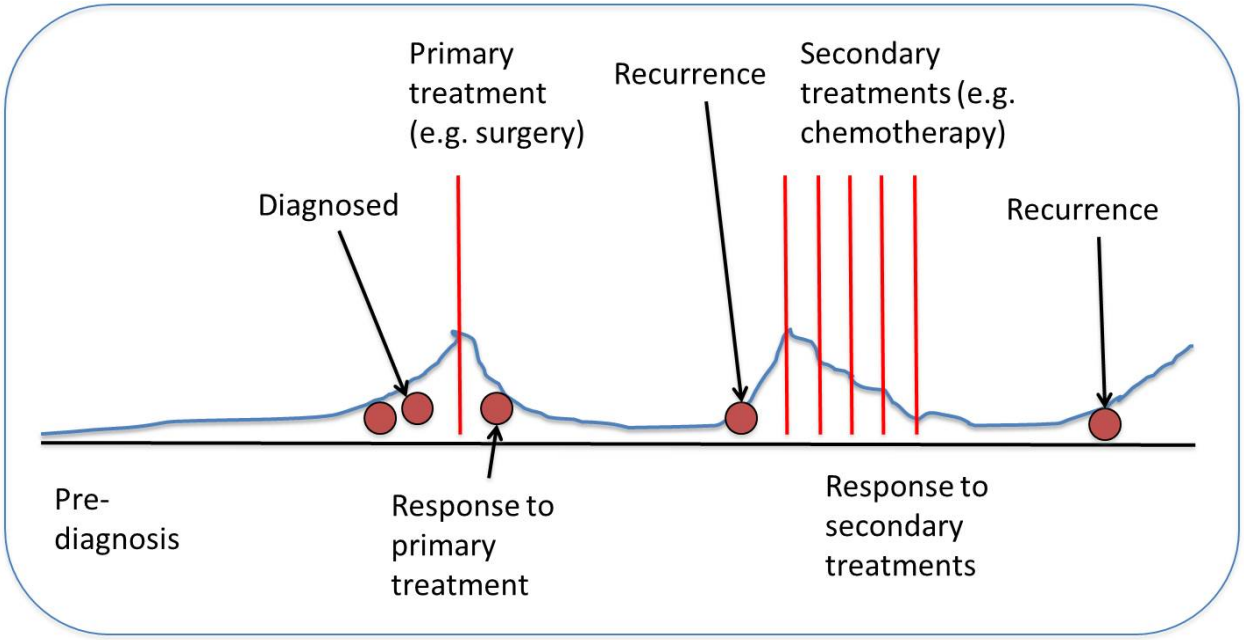

Figure 7.4. Clinical data in relation to underlying model of cancer as a chronic, progressive disease.

Most cancer database models only pertain to the documented universe (see Figure 7.1), typically abstracted or imported from medical records, and potentially augmented by public and commercially available demographic, consumer, and environmental data. This documented patient history is never complete and is always somewhat subjective, related to the instrument used to measure it and how the resulting data are provided.

Database models used in cancer research could be classified according to their assumption of an "open world" versus a "closed world".(111–113) Under a closed-world assumption, all information that is not true is considered to be false. Moreover, under a further "domain-closure" assumption, no other events are considered except those in the available medical record. For example, under the closed-world assumption, if no medical record of prostate cancer surgery is found for a given patient, it is assumed that the patient has not had prostate cancer surgery. The domain-closure assumption in this example implies that the

universe of information about the patient is contained within the medical records, so once those records have been reviewed, a data abstractor can determine whether the procedure (the prostate cancer surgery in this case) has been performed on this patient. Many case report forms, spreadsheets, clinical trials EDC systems, and other flat data models used for clinical trials, diagnostic studies, and epidemiologic cohort studies have an implied closed-world assumption. An indicator of this assumption is a question like "was surgery performed, yes or no?" or "not done" as an allowable value for a response. A closed-world assumption would allow abstractors and information processors to determine true negative (TN) values at the point in time that the medical records are reviewed, as depicted and described in Figures 6.1 and 6.2.

An open-world assumption implies that the underlying truth may not be known. Again referring to Figures 6.1 and 6.2, an abstractor or algorithm could report on what is documented or whether the value in a field was supported in source documentation or not. However, if an abstractor or algorithm could not find evidence of a value in source documentation, the truth would be "unknown" rather than "not done", "not present", etc.

A typical relational database design for a clinical system with a one-to-many relationship between patients and procedures would be more in line with the open-world assumption because there would be no record reporting when a procedure was not performed. NLP and information retrieval methods are generally trained and evaluated from an open- rather than a closed-world assumption. The truth of the underlying disease is considered unknown, and absence of evidence in medical documentation is not evidence of absence for a diagnosis or intervention. However, individual documents or assessments in a medical record (e.g., a review of systems or pathology report) may be dictated from a closed-world perspective, and it may be valuable for data entry staff to abstract from a closed-world perspective to track completion of their work.

The Caisis clinical data model and similar reusable relational data models are generally built with an open-world assumption, and when a study team seeks to migrate information from a paper form, REDCap database or spreadsheet to one of these relational models, the closed-world questions from those database models do not map cleanly to Caisis and are often a source of frustration and misunderstanding. In the Caisis model, we implemented an AbsentEvents table to explicitly note the assumed TN values that were captured in forms or flat databases under a closed-world assumption. For example, AbsentEvents could record that a lab test was not performed on a particular date or that a medication was not given. Because AbsentEvents is just another related table in Caisis, to regenerate a closed-world spreadsheet from Caisis would require a query or report to join records in the regular table with the associated records in the AbsentEvents table.

Another issue that has come up frequently in Caisis and similar cancer database models is the granularity or hierarchical level of a reported result. For example, the pathology findings from a prostate biopsy could be reported as “positive” or “negative”, or to be more granular they could be reported as “well”, “moderately”, or “poorly” differentiated cells, they could be reported as “Gleason sum” of 2 to 10, or the combination of primary and secondary “Gleason grade” of 1 to 5 each. Each of these results could be derived from the most granular result - the primary and secondary Gleason grades. The difference between these results becomes the granularity or level in the hierarchy of cancer grade classification. In the Caisis database model, there are different fields for each level of granularity. Other database models (e.g., OMOP, i2b2) may take advantage of the hierarchical features of terminologies or ontologies and may only store one level of granularity in the database - the most specific - and be able to compute the other levels of granularity based by traversing the hierarchy of an associated terminology or

ontology. The use of terminologies allows those data models to have fewer fields, but in turn they rely more heavily on their associated vocabularies.

The final underlying concept that I wanted to explain is the concept of data abstraction and its relation to natural language processing and computation. The activity of abstracting data from medical records into a database can be broken down into a number of distinct mental activities: [1] framing the task, [2] information extraction, [3] consolidation or computation to transform multiple inputs into one output, and [4] coding and formatting data to desired fields and values.

Framing an abstraction or clinical data processing task entails defining the domain of information to be considered, for example, all available pathology documents for patients with a breast cancer diagnosis. That scope would be the relevant/documented data as depicted in Figure 7.1. Framing the task also requires specifying the target values to be retrieved or abstracted from the specified relevant source documents. In practice, that would involve defining a data dictionary of which fields to abstract, how they are defined, and how they should be coded in an output database or form.

The second activity for both data abstraction and clinical data processing tasks is information extraction. This step involves the review of source documents, and the identification of words, formats, positioning and other features in those documents that are associated with the desired data points to extract or abstract. In NLP, this step is called feature engineering. Features are used in manual processes or algorithms to extract information from the source document. In clinical data processing, the extracted elements may be returned as multiple hits or in some other intermediate data structure. In manual data abstraction, these elements are held in memory or possibly as notes on a paper form or screen.

The third activity is consolidation and computation from one or more extracted features into the desired data elements. For example, if multiple values are found in source documents, which one is correct? Which one is from the desired point in time? Which one has the highest score, maximum dimension, or worst result? Computed data points may involve functions of two or more other data points. Examples of computed data points are “Age at Diagnosis”, “Date of First Contact”, “Primary Treatment”, “Cancer Stage”, and “Date of First Recurrence”.

Finally, the extracted and consolidated or computed results are coded and formatted to the desired vocabulary for the fields and values specified in the data dictionary. The consolidation of data from sources into desired values and formats may be lossy (i.e. original information may be eliminated) or lossless (i.e. all of the original information is retained after data consolidation and/or computation) depending on the scope, structure, and vocabulary of the destination database model.

7.6.3 Describe Tradeoffs between Different Modeling Approaches

Given the increasing volume of biomedical data and the limited resources of most organizations, it is impractical to pursue many different approaches to database modeling. Each set of standards and structures was designed with a certain perspective and strengths in mind. As with any organizational design and management challenge, decisions about foundational structures mitigate some issues but generate other issues, which require management. An organization should consider their strengths in terms of staff knowledge and experience and their goals and sustainable resources before establishing enterprise database structures and incurring costs of managing their operations within those structures. Most cancer research organizations will end up with a great variety of database models, however they may not be structured or managed for

the greatest value to the organization unless they are coordinated under an overall data and informatics strategy.

Figure 7.3 shows examples of how some data storage models are closely aligned with user interfaces and analysis or reporting functions. Other data models are designed to align more closely with either data entry and management (e.g., Caisis, SEER*DMS) or desired data output (e.g., OMOP, i2b2). The Caisis database model was originally designed with a temporal and stackable structure to facilitate predictive modeling and algorithm development.(3) The Caisis tables are also closely aligned with the organization of data in EMRs and other clinical systems such that each clinical system (e.g., pathology, radiology, clinical laboratory, surgery) has a corresponding set of related tables in Caisis. A key result of this design is that it maps easily to source systems, and thus facilitates the development of data feeds from clinical systems. For example, all of the information that would be extracted from a pathology report would go into the Caisis pathology tables, and all of the information from radiology systems would go into the Caisis diagnostic imaging tables.

The Caisis data model and others like it (e.g., SEER*DMS, BRIDG) have less overall complexity than data warehouse models, which typically amalgamate the schemas of multiple source systems. However, these models are more complex than database models designed for self-service query and analytics tools (e.g., OMOP, i2b2, CDISC ADaM(114)) and study data reporting tools (e.g., CDISC ODM). The Caisis database model makes a good intermediate format for clinical data storage and editing, but it is difficult to use and does not perform well as an analytics tool.

There are a plethora of new tools and approaches for big-data management, and it is often difficult to assess their value and fit for biomedical research. Most of these tools grew out of the

distributed hosting and processing needs of massive data-driven companies like Google, Amazon, and Facebook. Relational databases were optimized for transaction processing and accounting systems, however their structures are difficult to change quickly or dynamically and the processing required to meet all relational database integrity constraints often exceeds the value of the processed data. For many big-data companies, a lazier approach to inserts, updates, and constraint checking as well as enabling a variety of search and processing functions without all of the joins required in a complex relational database model is necessary.

Big database platforms (e.g., from Amazon Web Services, Cloudera and Hortonworks Hadoop, Google, Microsoft Azure) typically include distributed storage, scalable SQL and NoSQL databases, and tools for data acquisition, processing, search, and machine learning. For clinical and other sensitive data, these companies have begun to offer high-security cloud hosting that can allow them to sign HIPAA BAAs and meet HIPAA- and FISMA-level security requirements.

Key-value stores are probably the simplest NoSQL database model. They provide distributed storage of data indexed by a key field. This concept of indexed storage is already used frequently in clinical systems and biomedical research databases. For example, the Caisis database allows users to set pointers (file pathnames) to files associated with any record that are then stored on a file server. EAV models, which are similar to key-value stores but without distributed storage, can also be implemented within a relational database. Big-data key-value stores are optimized for distributed storage and distributed processing platforms like Hadoop HDFS or Amazon S3. Amazon's Dynamo key-value store design was on the leading edge of development of NoSQL database platforms.(102,115)

Beyond relational database stores, document stores allow for further data complexity. For example, the Javascript Object Notation (JSON) documents in a MongoDB database can store nested and complex data structures.(116) They can also be indexed and queried by attributes within those JSON document structures. However, this allowance for complexity comes at a cost in terms of query performance and burden on the knowledge and skill of application or middleware programmers to implement queries and maintain database integrity.

Column store databases add complexity to key-value stores, and many are derived from Google BigTable.(117) Basically, column stores can provide collections of key-value stores (columns) that point back to rows. Unlike a relational database, each row does not need the same sets or numbers of columns, and the "primary keys" are the values in columns. Column store databases are efficient in terms of writes, can handle sparse and irregular records, and are optimized for queries. They could be considered an alternative to analytic database models and data marts (e.g., OLAP cubes) and are potentially well suited for managing and querying high-dimensional data like genomic assays.(118)

Graph stores like Neo4j are specialized tools for managing graphs or triple stores. They are useful for applications that are network-graph centric, like social networks or predictive modeling, but may be less useful as a generalized database model for cancer research.

In addition to the NoSQL database systems, many relational database vendors have built NoSQL features into their platforms. For example, some vendors have implemented table indexing that is more in line with a column store approach and have used JSON or XML data types for the storage of documents as values in text fields.

Overall, these NoSQL database models and systems offer options for storing massive amounts of data with flexible or unknown schemas that would not be practical to load into

relational databases. They also offer alternatives to the solutions that have previously been hacked together in SQL databases, such as pointers to files, EAV databases, document storage in text fields, OLAP cubes, and star schema models.

7.7 Discussion

7.7.1 Lessons Learned about Data Modeling Approaches

In context, the rise of NoSQL and database models can be seen as a natural evolution of technology rather than a big-data revolution (see Figure 7.5).(119) With the pace and demands of cancer research, investigators may have little tolerance for database models that impede their work. The lure of the flat, simple spreadsheet, REDCap, or custom database model is always present.

The depth and complexity of highly normalized relational data models with many tables such as Caisis can make navigation and entry of detailed data tedious. However, there are benefits to the Caisis relational database structure in that it maps well to clinical source systems and has a functional and customizable web-based data entry, review, and editing interface. The Caisis database model becomes challenging to use when you have to write a query that crosses many details nested in subtables (e.g., determining extent of disease, which could have findings under pathology, diagnostic imaging, and potentially other tables). As shown in Figure 7.6, queries across findings from different clinical source systems and stored in subtables may require multiple joins. Facilitating these types of queries is one reason that some cancer research databases create high-level tables to organize data around cases, tumor sites, treatments, or studies. Or, they may flatten the database altogether. However, these structural approaches reduce the reusability of the database models across projects and make data feeds from clinical

source systems more complicated to manage. A significant shortcoming of relational database models like Caisis is the amount of knowledge represented in their tables and their relationships. Querying and integrating that knowledge requires considerable technical and domain expertise.

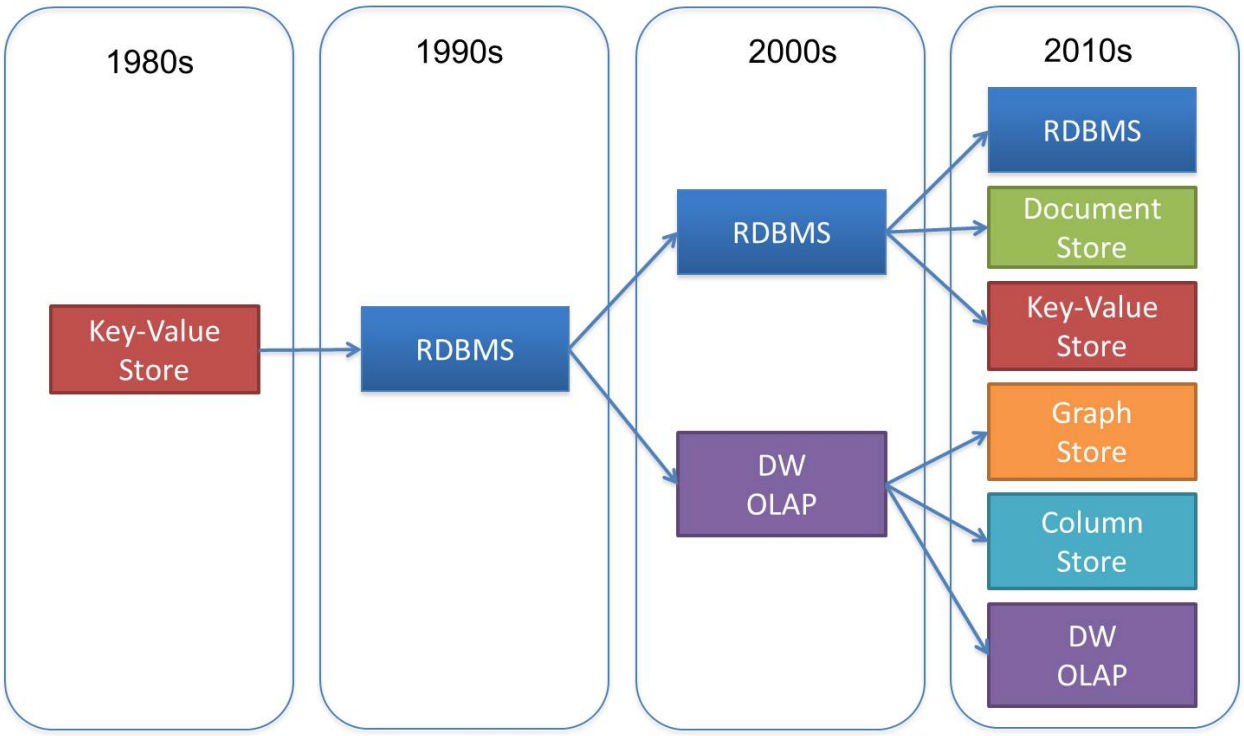

Figure 7.5. Natural evolution of database technology, adapted from Kalakota (119)

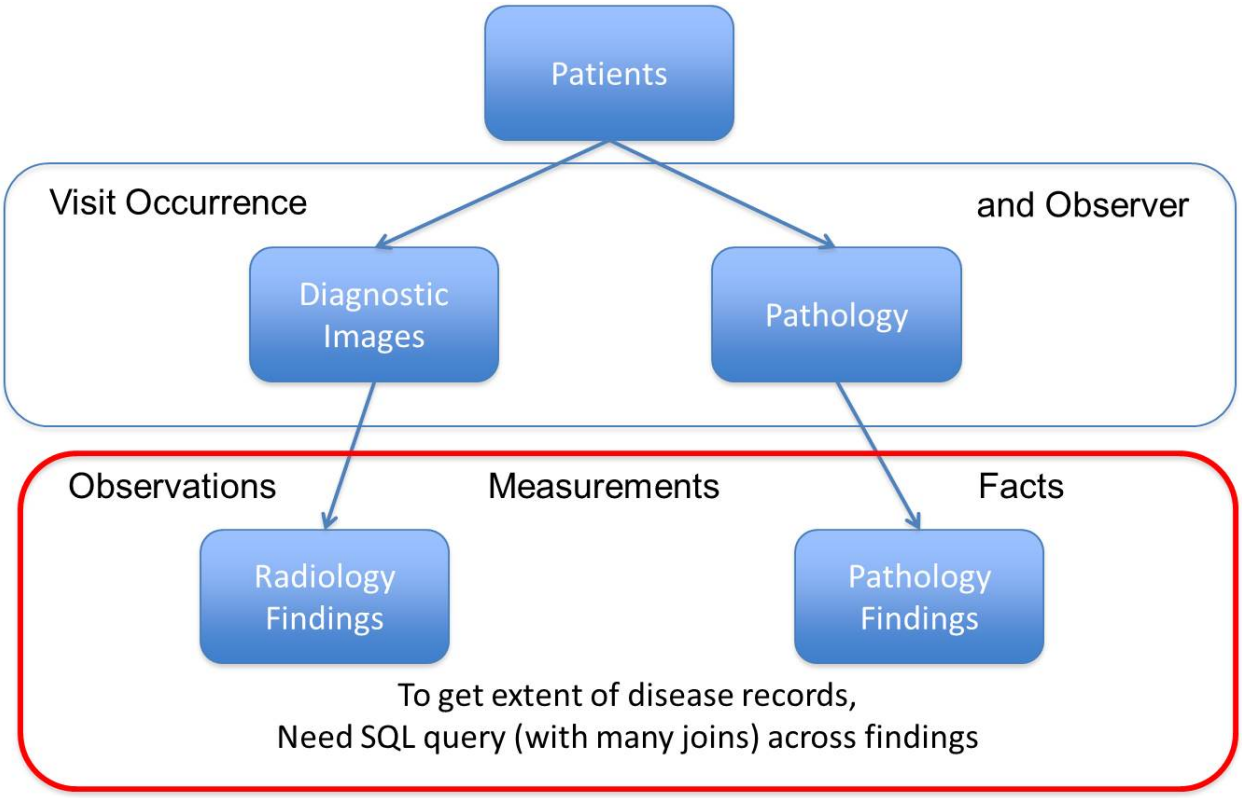

Figure 7.6. Normalized relational database model may be difficult to query and require multiple joins to integrate data across subtables.

One approach to the difficulty in querying a relational model like Caisis is to develop a complementary analytic data model, such as a star schema or column store. Many examples of analytic data models for biomedical research have emerged over the last decade, including i2b2 and OMOP. These models tend to have fewer tables and relationships than models like Caisis, and they embed much of the complexity of the data in terminologies or ontologies rather than in multiple tables, fields, and relationships. Figures 6.7 show the simplification of the analytic data model and the shifting of managing complexity to a terminology or ontology (as used by i2b2) compared to the normalized relational database model depicted in Figure 7.6. Figure 7.8 shows an alternative approach to a simplified analysis data model used by OMOP. Importing data into analytic models often requires extensive mapping and transformation, and querying these models

requires navigating the associated terminologies. They have proven quite useful for facilitating self-service cohort selection queries using a relatively narrow set of parameters, but they do not appear to scale well for data of high volume or variety.

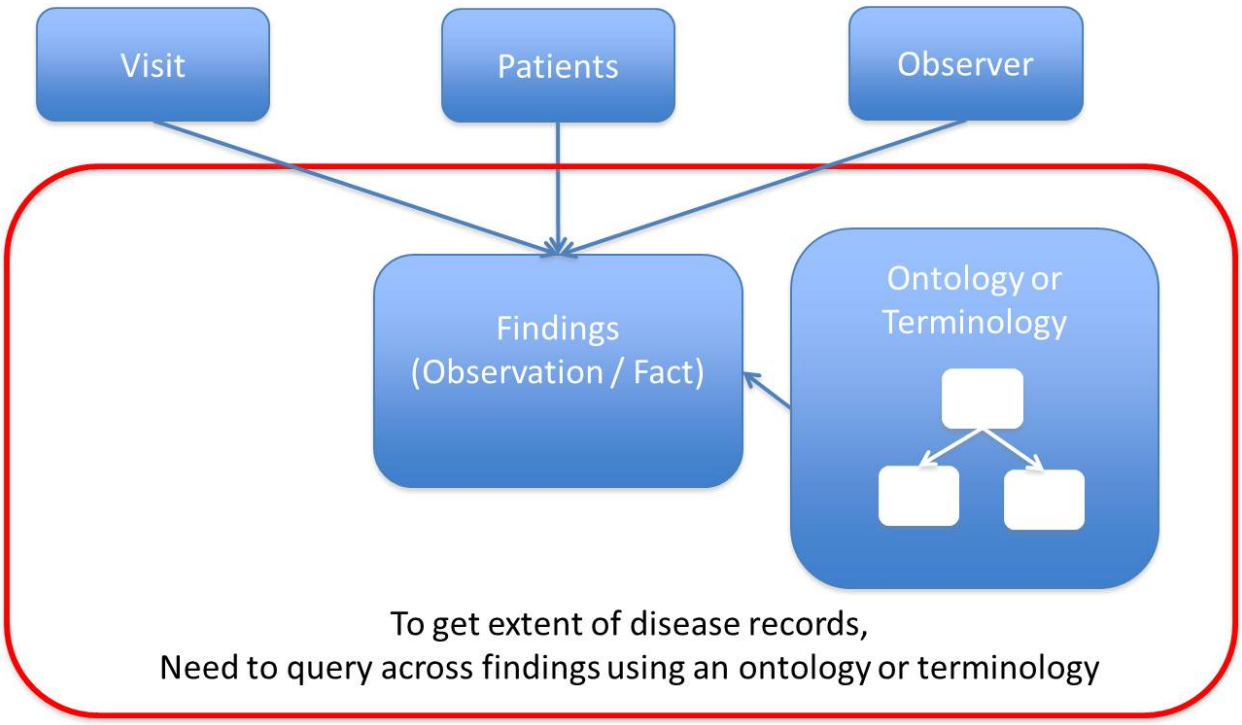

Figure 7.7. Data complexity in ontology or terminology and start schema (e.g., i2b2).

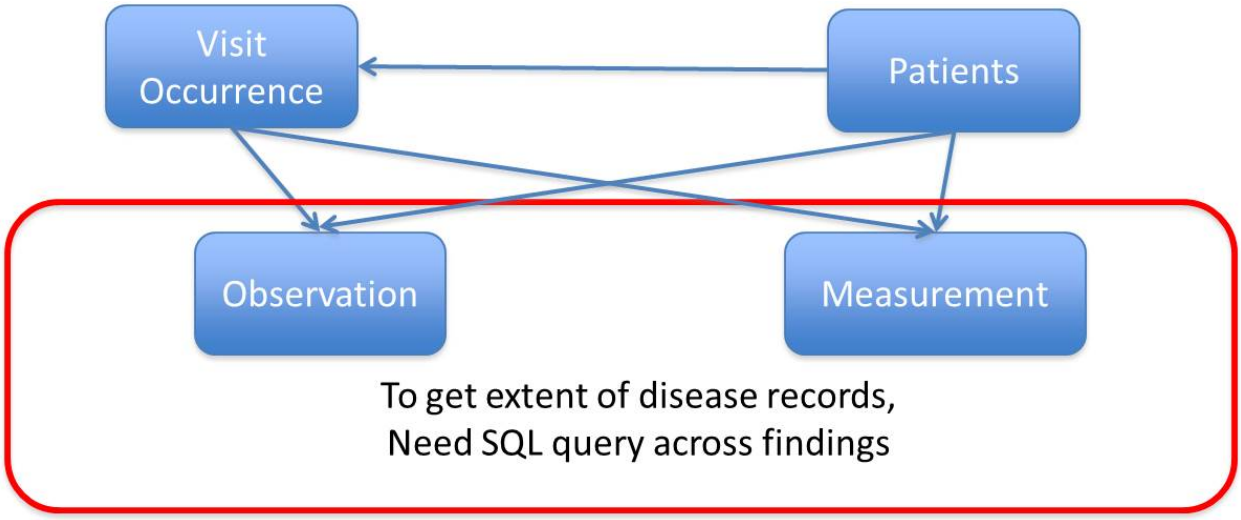

Figure 7.8. Data complexity in the field names and relationships between tables (e.g., OMOP).

The majority of clinical data still comes from text documents, and long text fields or notes tend to be distributed broadly and nested deeply in relational database designs. The ability to conduct full-text searching across both documents and databases details remains attractive. Full-text searching by querying across the lowest level of details in a relational model like Caisis could involve numerous joins - and could be mind-numbingly difficult to write in SQL. However, there are several options to facilitate search through documents or text fields. Most mature relational database management systems offer built-in features for indexing and full-text search. Another more powerful option for full-text search could be to employ a full-text indexing and search tool like Lucene (a search engine) or Solr (an application that uses and extends the Lucene search engine) to crawl documents stored in a relational database and facilitate search and retrieval.

Figure 7.9 shows an alternative approach to modeling complex clinical data from documents like pathology reports, still maintaining the alignment with clinical source systems, but allowing the flexibility for variations in data structure and handling storage, query, and full-

text search of documents. A model like Caisis could be simplified by allowing the high-level entities like diagnostic imaging and pathology to include JSON documents rather than modeling out anticipated fields and relationships in a normalized relational database approach. A database model like this could accommodate data structure changes without redesigning the database and could allow for full-text indexing and search functionality.

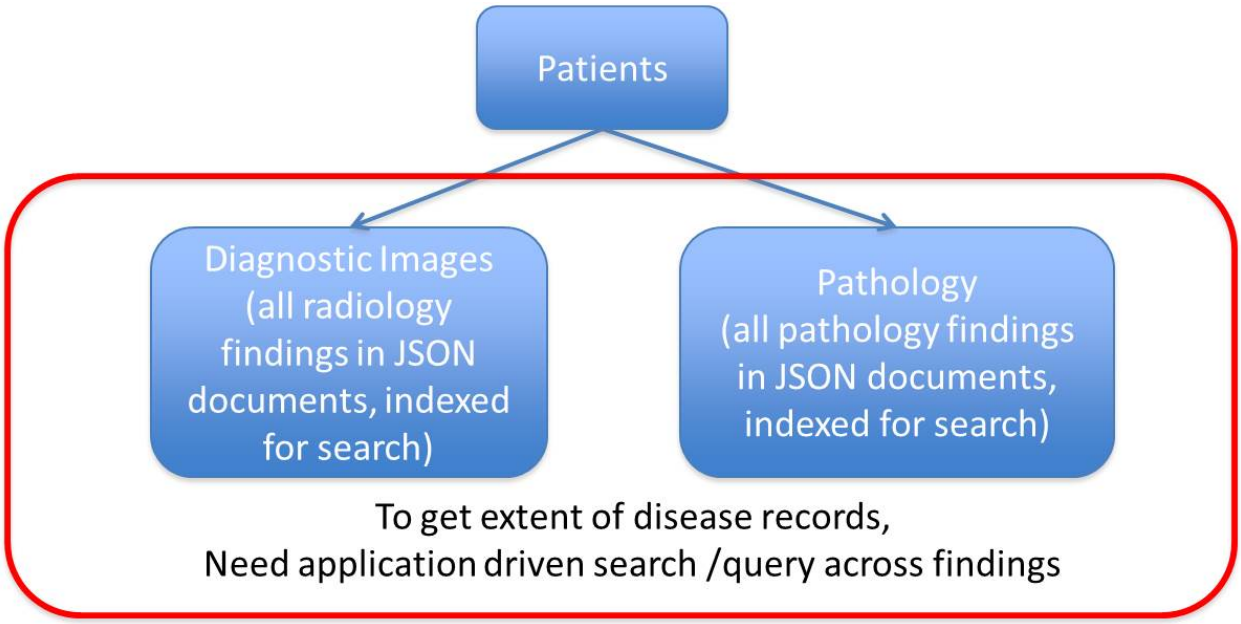

Figure 7.9. Data complexity in structured documents (e.g., MongoDB or JSON features in relational databases).

One of the downsides of a document store oriented database as depicted in Figure 7.9 is the potential lack of a mature, domain-specific user interface for data entry, review, and editing. A software engineer could write custom data entry, review, and editing tools, but this could become expensive to develop and to support over time. The user interface tools, configuration tools, and ease of customization of mature systems like Caisis or SEER*DMS makes them an attractive stepping stone in the overall evolution of database models and systems. The Caisis or

SEER*DMS databases have stable and extensible user interfaces for viewing, adding, and editing records in their respective data models. So, although querying these systems may be difficult because of the join complexity depicted in Figure 7.6, having a user interface for data entry and editing make these solutions attractive for manual data acquisition and processing. If a group decided to adopt a similar relational database model, like BRIDG, or to develop a custom database model for their disease or project, they would also need to develop and manage the corresponding user interfaces. If a configurable tool for data abstraction and information processing could be developed and plugged into any database model (e.g., relational, column, store, document store, hybrid), it would likely be more affordable and less risky for research groups and centers to choose more optimal database models.

EAV models and key-value stores (e.g., pointers to files on a network file server) have been implemented in relational databases. However, since relational databases were not designed to efficiently store and query EAV and key-value data stores, they may perform poorly or be difficult to manage at scale. With EAV and key-value models, it may be challenging to represent and query different hierarchical groupings or nestings of data (e.g., records in fields or multivalued arrays in fields). EAV models or key-value stores may be useful as an extension to an existing relational database model for the case where there could be many sparsely populated fields. EAV models may also be useful in generating audit logs for specific fields, as in SEER*DMS. In the Caisis model, an EAV model is used to allow virtual fields to be configured within the system admin tool as additions to any existing table. This feature reduces the need for structural changes to relational database tables, but has proven cumbersome to manage when querying the data or performing structural upgrades to the database. It is also easy to overlook EAV extensions to tables when users go directly to the database to write queries. Transforming

data into pivoted and denormalized views or an analytic data model for querying could help with the usability of EAV models used for data storage. Also, the performance of EAV views could become slow because of the limited ability to add table- and field-specific indexing. Field and record integrity checking would have to be handled at the application level for EAV structures and key-value stores since they cannot rely on native relational database constraints, indexes, and data types. For distributed and flexible storage of individual values or objects, a key-value store database could be a useful adjunct to a relational database model, but probably not as a complete cancer database solution.

Document stores may be most useful when the data element structures are complex and variable, with nesting or relational characteristics rather than more array-like structures. A primary example of these structures in cancer research is clinical data from narrative documents. Patient histories do not fit well into high-dimensional arrays. A document store could be suitable to facilitate data abstraction, document annotation, clinical data processing, and other information-processing tasks. Document storage could be managed in specialized databases such as MongoDB or by using emerging JSON or XML document store extensions to SQL databases. When storing JSON in a relational database field or in a document store, values within the document structure may need to be indexed for reasonable search and query performance. Otherwise, finding data points within documents would require a developer to implement nested looping in the query processor or the application user interfaces.

Column store database models may be useful for query optimization for high-dimensional but roughly tabular or cubed data. In cancer research, any of the high-dimensional array data from molecular profiling would fit well into a column store. Because of the indexing and simplicity of querying values within each column, a column store could be optimized for

querying for specific tumor mutations. A column store could also be useful for storing and querying time series sensor data (e.g., from wearable devices). Column stores tend to have lower overhead and latency for writing than relational databases. Finally, a column store could be a useful addition to or substitute for a data mart or an analytical data model, particularly those that take the form of cubes or star schemas (e.g., OLAP, i2b2).

Graph databases and triple stores could be useful for storing and querying ontologies, vocabularies, or for other types of knowledge representation where the graph structure is particularly useful to facilitate computation. For example, developing predictive models for cancer progression or treatment outcomes may be facilitated through the development and population of graph databases. Public and commercially available demographic, environmental, and consumer data, which may contain different constellations of data points related to each patient, may also fit well within a graph database. Because relatively few staff are comfortable with graphical models, graph databases, and the tools to facilitate graph-based queries, these database models are probably not well suited to the generalized storage of cancer data.

The integration and management of vocabulary with database models is fraught with challenges, particularly for the models that rely heavily on terminologies or ontologies for knowledge representation, navigation and querying (e.g., i2b2, OMOP). The numeric coding of values (e.g., 1=”male”, 2=”female”) and storage of the codes in database models is still quite prevalent in clinical and epidemiologic studies. However, since each study may have different coding rules, it is difficult to develop a reusable database around coded values. Vocabulary management systems such as the one in Caisis allow for the configuration and assignment of synonyms, parent or child records, and other attributes associated with terms. So, for example, a code of “1” could be assigned to the “male” term for one study, and a code of “2” could be

assigned to “male” for another study. Queries or data access tools would have to map the correct codes to be reusable across studies. This may be difficult to implement in practice, so many reusable databases like Caisis have been designed with minimal use of numeric codes with the burden of recoding managed by study staff or centralized data delivery staff.

Sometimes, terminologies may have restrictions on their distribution and use (e.g., SNOMED-CT and AJCC have licensing requirements, but LOINC and UICC are less restrictive). The distribution and licensing terms of vocabularies may become an issue in environments where many cancer centers participate in data sharing for national or international collaborations. Even a highly curated and suitable terminology may lose to a terminology that is open-source and freely distributed, or is at least available in terms that match the anticipated funding, scope, scalability, and portability goals of a database.

Notions of data quality and underlying open- or closed-world assumptions find their way into database implementations and management. Some awareness of and agreement on a definition of data quality between study and IT staff may reduce - or at least provide a framework for - discussion when these issues arise.

Huser and Cimino(120) recommend the following desiderata for an integrated data repository:

- Single patient identifier or master patient index
- An information model that is sufficiently generic but extensible so that it incorporates new data sources but remain stable over time
- Support for the nesting of facts (e.g., collection of data from each specimen in a pathology report, or each finding in a radiology report)
- Semantic integration and terminology model

- Documentation and metadata (e.g., data dictionaries)
- Provenance tracking (or capturing the historical evolution of data in the repository)
- Management of protected health information

Tracking and mapping to a single patient identifier may be handled in a relational database system because of the importance of consistency. The remainder of the desiderata above implies that there may be no ideal single platform solution. Rather, hybrid database models and systems to support hybrid models will most likely be required to advance cancer research and precision medicine initiatives.

A core generic and standardized database model may be best implemented in a relational database, however to be extensible and support the nesting of facts, it may call for the application of NoSQL models such as a column stores for high-dimensional genomic data and a document stores for complex clinical data derived from medical records. Further, a terminology model or database to facilitate predictive analytics may be best managed in a graph database for the optimization of navigation and queries using the knowledge represented in trees and graphs.

No matter which database models and platforms are used, the careful design of security boundaries, constraints to PHI, and logging to document provenance of data points will be important foundational requirements.

7.7.2 Future Research/Next Steps

As big-data requirements and use cases for hybrid database models to support research are growing, it will be important to conduct pilot studies to evaluate the performance of various data stores and hybrid database models.

The performance and usability of document stores (e.g., MongoDB) versus embedding documents (e.g., JSON) in columns in a relational database will need to be evaluated for clinical data management. The performance evaluation should include write and read performance, indexing and optimization requirements, full-text search capabilities, and technical skills and work required for implementation and use.

Analytic database models (e.g., i2b2, OMOP) should be evaluated compared to column stores for query performance, ease of use, transformations required to load, and knowledge and technical skills required for implementation.

Because the lack of configurable and easily customized tools for data abstraction and editing is a likely barrier to the adoption of some database models, there is a need to advance data abstraction and clinical data processing pipeline tools as flexible, extensible, and portable interfaces for hybrid database models.

7.8 Conclusion

No single database model clearly meets the requirements of managing clinical and molecular data for biomedical research. Many organizations will need conceptual and practical guidance to better understand and make decisions about system requirements and design.

This chapter and the following recommendations should be useful guides for thinking about, discussing, and planning database models for cancer:

1. Identify each incoming data type and estimate its volume, variety, velocity, veracity, and security requirements. As big-data technologies are evolving rapidly, it may not make sense to invest or plan more than a couple years ahead of projected growth.

2. Identify and estimate the ways in which these data will need to be manipulated, queried, or visualized to support research and operations.
3. Determine how database functions and contents will be evaluated in terms of performance and data quality.
4. Identify candidate database model approaches, most of which are likely to be hybrid models. High-dimensional array data like genomics assays and sensor data may perform well in column store like Cassandra. Clinical data may perform well in a combination relational database and document store for acquisition, processing, and storage. For high-performance data exploration and visualization, a column store may offer the best performance for both clinical and high-dimensional data. However, if the database must be used to support clinical decision-making, the latency and consistency of returned results must be factored into the platform selection, database design, and configuration.
5. Conduct write and read simulations using different approaches to confirm hypotheses regarding performance and assess the time required to implement and manage each structural approach. Each database model implemented will require staff time and resources for mapping and recoding.
6. Consider implementing a data landing area that is within a high-security environment, but allows a variety of incoming data formats with minimal upfront mapping and transformation, and allows for the linkage and classification of incoming data. Most organizations will need to acquire and process an increasing volume and variety of data sources over time.

7. Consider implementing a flexible and extensible pipeline for manual data abstraction and clinical data processing to transform data from a landing area or primary database model into increasingly usable data structures and analytic models.
8. The primary database model for data entry, editing, and storage should be reusable across multiple projects in order to recoup the costs of implementation and operation. It should be lossless in terms of data and extensible so that planned and unplanned data can be reliably acquired, stored, and accessed. Also, given the prevalence and persistence of flat models with closed-world assumptions, the primary database model should provide a simple way to document absent events or absence of evidence. The primary database model could be a based around retrospective research (e.g., Caisis), cancer surveillance (e.g., SEER*DMS) or clinical trials. If organized around clinical trials, the associated EDC systems could export in CDISC ODM format that from which SDTM and then ADaM can be derived, or adhere closely to CDISC CDASH guidelines, which map to SDTM format for required reporting.(121–123)
9. Analytic database model(s) or column stores may be implemented depending on the anticipated usage or participation in research networks. Analytic database models may be somewhat lossy or narrow, may require complex mapping and transformation to populate, and may have consistency issues when querying. For research and statistical modeling, small random errors in data points caused by inconsistencies or latencies in analytic models or big NoSQL database models may be acceptable. However, if implementing precision medicine for an individual patient, for example in a clinical decision support tool based on molecular data, consistency and low latency would need to be prioritized, which would affect the performance of some database models. Analytic

database models for specific analyses (e.g., graph database for predictive modeling) or reporting requirements (e.g., CDISC SDTM) should be loaded from a primary database model. However, because they are generally not reusable for a variety of different projects, they should not be chosen as primary database models.

10. Assess the current and attainable domain expertise and technical skills of staff. The current database knowledge and technical skills of staff, and the ability to hire, train, and retain talent, should always be factors in the planning of database models. There is a tremendous learning curve and responsibility on programmers to implement reliable and sustainable hybrid database models.

7.9 Acknowledgements

For this work, I would like to thank Emily Silgard and Dirk Petersen from Fred Hutch, Marina Matatova from Stanford, Matt Bellew from LabKey Software, and Mark Danese from Outcomes Insights for sharing their opinions and insights on database models and big-data technologies.

7.10 Synthesis

Chapter 7 explored some of the issues with database models to support cancer research, compared traditional database model approaches that have been implemented, and introduced new big-data tools that may offer performance and usability improvements to cancer research data systems. Sections 7.2 to 7.8 addressed my Chapter 7 question: how can we best characterize and compare database models and big-data technologies to inform a sensible strategy for cancer research database design and technology? Various common database models were explored and

compared, from traditional relational models like Caisis to analytic models like i2b2 and OMOP and to big-data technologies like column and document stores.

This work ties back to the overarching question for this dissertation (how can we improve access to clinical and related data about cancer patients for research?) because access to data may be facilitated or hindered by the database models and database technologies selected for implementation. Without a thoughtful strategy and design of underlying data models, processing of and access to clinical data through informatics tools can be impaired.

This work also ties back to the overall hypothesis that there are new tools and methods from biomedical informatics that could improve the availability of data for cancer research if they were applied thoughtfully and strategically. A number of big-data technologies could be valuable augmentations of existing databases that support cancer research.

The aim of this chapter was to develop, model, and assess database frameworks for cancer. This data modeling chapter is informed by all of Chapters 2–6. The Caisis relational database model described in Chapter 2 was useful for prospective data entry, but becomes complex for queries and supporting analysis tools without transformation. The rise of i2b2 and REDCap described in Chapter 3 is related to the need for a database model designed to support analytics. Although the Caisis model was selected as a starting point for HIDRA described in Chapter 4, it may have limited ability to support the Argos self-service query tool. The clinical data processing and abstraction pipeline described in Chapter 6 would be tedious to expand if it continued to rely on the Caisis database model and other traditional relational models.

Although the scalability and flexibility needs at a single cancer center may justify the use of new big-data tools and hybrid database models described in Chapter 7, for clinical data, the big-data challenges of cancer registries and cancer surveillance are orders of magnitude greater

than the experience at any single cancer center. Chapter 8 describes the cancer surveillance data and information-processing challenges and aims to determine what informatics tools and methods have been applied in that space and what the opportunities might be to apply lessons learned from the previous chapters in this dissertation.

Chapter 8 Informatics and Cancer Surveillance: Literature Review and Vision

8.1 Context

This chapter is my exploration of Aim 4, to characterize the big-data needs and informatics opportunities for cancer surveillance at the national level. To explore this aim I answer the questions for Chapter 8: What informatics tools and methods have already been applied successfully in the cancer surveillance domain? What are the current and coming opportunities for applying or developing new tools and methods for advancing biomedical informatics in this domain?

Answering these questions helps inform the overall research question: how can we improve access to clinical and related data about cancer patients for research? Before this research, I had encountered only anecdotal reports of the lack of informatics research applications in cancer surveillance. This literature review provides at least a baseline account of the informatics tools and methods applied in this domain. The literature reviewed provides some indication of the current and coming opportunities for applying and advancing biomedical informatics, NLP, and machine learning within cancer registries and cancer surveillance.

Aim 4 is also a logical next step to test my hypothesis that there are new tools and methods from biomedical informatics that could improve the availability of data for cancer research if they were applied thoughtfully and strategically. Many research projects and databases at cancer centers rely on cancer registries for basic diagnosis, treatment, and outcomes data about their patient population, and they may integrate cancer registry data into larger data repositories like HIDRA described in Chapter 4. Improving the registry databases with

informatics tools and methods could have far-reaching effects for the cancer research community.

Aim 4 (to characterize the big-data needs and informatics opportunities for cancer surveillance at the national level) flows from Aim 3 (to develop, model, and assess database frameworks for cancer) because of the scale issue. The big-data scale challenges in cancer surveillance are orders of magnitude greater than those of any single cancer center. Database models that may work well in a single center are likely to fail at the scale of national cancer surveillance, and implementing big-data technologies like distributed column and document stores becomes more urgent and valuable in this context.

Aim 4 also flows from Aim 1 (to develop and assess a modern integrated data platform to support a wide variety of cancer research) and Aim 2 (to develop and characterize a clinical data processing pipeline that is scalable to the cancer center enterprise level, is well-supported and sustainable, and can complement or streamline existing manual data abstraction and information-processing activities) because all of the legal and IRB issues related to a data sharing and governance framework as well as self-service data access described for Aim 1 (Chapters 4–5, development and assessment of HIDRA) and the clinical data processing pipeline described in Chapter 6 (Aim 2) are not only applicable but essential for the acquisition, processing, and dissemination of cancer registry data described here in Chapter 8 (Aim 4).

I performed the literature review and wrote this chapter. The co-authors provided feedback on Figure 8.1 and agreed to be co-authors when this paper is edited and submitted for publication. This paper will likely be submitted to the *Journal of the National Cancer Institute*, *Journal or Registry Management*, or similar publication, thus the title is repeated and the authors are listed

below. Because this was written as a standalone paper I have included a separate acknowledgements section.

Informatics and Cancer Surveillance: Literature Review and Vision

Paul A. Fearn, MBA[1,2], Eric Durbin, DrPH[3], Lynne Penberthy, MD, MPH[4]

[1]Fred Hutchinson Cancer Research Center, Seattle, WA; [2]University of Washington, Department of Biomedical Informatics and Medical Education, Seattle, WA; [3]University of Kentucky, Lexington, KY. [4]National Cancer Institute, Surveillance Research Program, Rockville, MD.

8.2 Abstract

This chapter describes cancer surveillance, its current data collection and information-processing activities, and its big-data challenges. This introduction is followed by a literature review of informatics tools and methods that have been applied to cancer registries and concludes with the identification of opportunities for further research and the application of existing adjacent research to the challenges of cancer registries and surveillance. A total of 683 articles were retrieved through searches of PubMed, Google Scholar, and a Google internet search. Of these articles retrieved, 66 articles were relevant and were categorized into common functional themes for cancer registration. Applicable NLP and machine-learning research was found regarding the identification of reportable cancer cases and a handful of other examples of applied informatics for cancer surveillance. As expected and anecdotally reported, there is a dearth of and opportunity for informatics, NLP, and machine-learning research in the domain of cancer surveillance and cancer registries.

8.3 Background and Rationale

Cancer surveillance is the systematic collection and reporting of information about the diagnosis and treatment of cancer cases in a population. Data that are collected for cancer surveillance are used for research, public health planning, and the assessment of the effects of policies and improvements in healthcare.

The reporting of cases is mandated at the state level, collected at hospitals and a variety of other healthcare-related facilities, and is consolidated at state or regional central cancer registries, illustrated in Figure 8.1. In 1956, the American College of Surgeons (ACS) mandated the operation of hospital registries for its Commission on Cancer (CoC) accredited cancer programs.(124) Before the National Cancer Act of 1971, only a few central cancer registries were operating: Connecticut (1935), New York (1940), Hawaii (1960), New Mexico (1966), Wyoming (1967), Kansas and Colorado (1968), Idaho (1969), and Los Angeles and Virginia (1970). Most of the current central registries became operational from the 1970s through the 1990s.(2)

Cancer surveillance data are aggregated at the national level by three programs. The Surveillance, Epidemiology and End Results (SEER) program of the NCI was initiated in 1973 to advance cancer surveillance. The SEER program funds population-based data collection and reporting for 30% of the US population, and it makes these data available for a wide variety of research.(125) In the United States, approximately 1500 CoC-accredited hospitals (out of the more than 5,000 registered hospitals in the United States) currently contribute data to the National Cancer Data Base (NCDB), which was initiated in 1989 and is funded by both ACS and the American Cancer Society.(126) In 1992, the Center for Disease Control (CDC) initiated the National Program of Cancer Registries (NPCR) to improve the data quality and population

coverage of data collection and reporting for state cancer registries and to establish registries in 10 states that previously had no state registry.

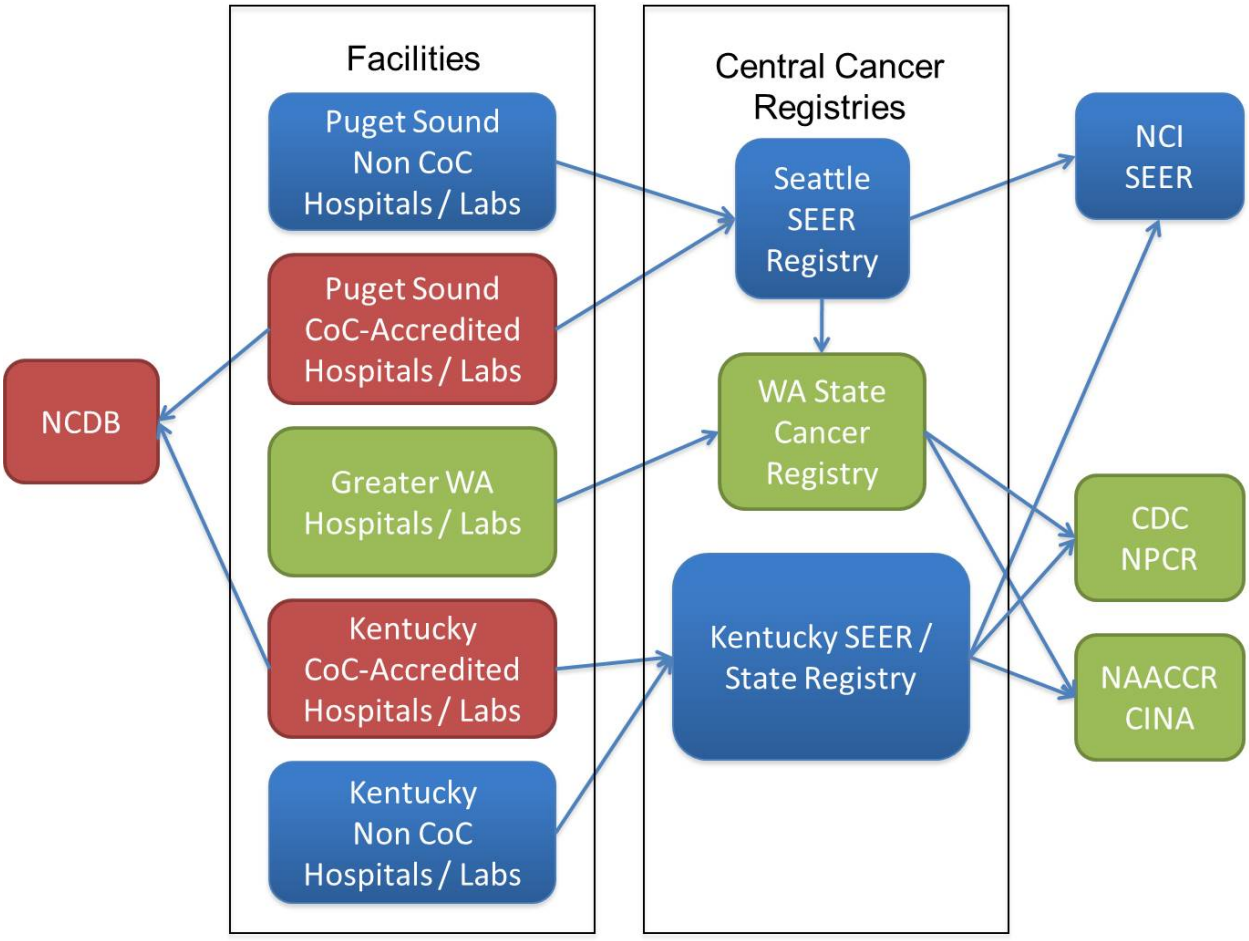

Figure 8.1. Flow of cancer case reporting, from hospitals, labs, and other healthcare facilities, to central cancer registries and to aggregated databases. Only Commission on Cancer (CoC) accredited hospitals report to NCDB (e.g., UW, SCCA, and University of Kentucky Healthcare). Many other community hospitals and facilities report to central registries. Some SEER-funded registries (e.g., Seattle, Detroit) cover regional or special populations rather than entire state populations like the Kentucky SEER registry. De-identified data are provided to NCI from SEER-funded registries, to NPCR from CDC-funded registries, and to NAACCR Cancer in North America (CINA) from NAACCR-certified registries.

The primary data collection activities for cancer surveillance typically happen within hospital cancer registries or through cancer cases that are reported directly from diagnostic and treatment facilities to central cancer registries. Data about patients with cancer and their respective cancer diagnoses, treatments, and survival are generally abstracted by registry staff

who access hospital medical records and enter information into commercial or open-source tumor registry software (e.g., Oncolog, Abstract Plus, C/NET, SEER*Abs)(83). The data are abstracted from medical records and entered into software that can output NAACCR abstract record format.(127) Internally these data abstraction software tools use a variety of different database models.

Central cancer registries use additional software and operations to consolidate reports from multiple hospital registries, cancer diagnostic or treatment facilities, and laboratories. Many of the NCI SEER-funded central registries use the SEER*DMS software(106), and many of the CDC NPCR-funded registries use the Registry Plus suite of tools.(128) These tools for central registries aim to facilitate quality control of incoming data, editing and consolidation activities, and reporting.

Overall, the current process for cancer surveillance has a long history and relies heavily on manual data abstraction from medical records, manual consolidation and error checking, and manual data quality control. Information processing relies on expert cancer registrars, who are often highly trained, experienced, and certified.(129) With increased funding from SEER and CDC, central registries have worked to improve data quality with a focus on accuracy, de-duplication of reports, and improved population coverage.

Because central cancer registries' mandates for reporting cases and authority to access patient identifiers lies at the state level, finding and resolving duplicate reports of cancer cases across different states remains an ongoing challenge. Florida and New York state registries recently piloted the use of the National Death Index (NDI) to find and resolve duplicate reports.(130)

With current funding, practices and infrastructure, hospital registries, central registries, and national programs face tremendous challenges to sustain high-quality and up-to-date operations.(131,132) As cancer prevention, diagnosis, and treatments become more successful, the rates of cancer-related deaths are declining, and patients with cancer may have more complicated medical histories that involve multiple cycles of diagnosis, treatment, response to treatment, remission, and recurrence. Each of these episodes of healthcare incurs costs to patients and payers, and may be associated with various side effects and complications. The increasing financial and quality-of-life impact of cancer as a chronic disease can be a significant burden on patients and society.

Originally, cancer registries focused data collection on primary diagnosis, primary treatment, and survival outcomes. However, to be relevant for the current and next generation of cancer research, the planning of healthcare policy, and funding to support precision medicine, cancer registries must expand the volume and variety of data collection to include detailed patient timelines with potentially multiple rounds of care. In addition to new types of data (e.g., molecular biomarker results) and more granular data (e.g., details of medical therapies) that are becoming available in medical records, researchers (and thus central cancer registries) should have access to claims data to study costs of care, to pharmacy data, and to public and commercially available demographic, environmental, and consumer data that expand the view of cancer research to include factors that affect the development of cancer in populations and the factors that affect the outcomes of healthcare interventions, including financial outcomes.

The data challenges facing cancer registries and the SEER program meet common definitions of big data: increasing volume, variety, velocity, and veracity of data.(133,134) There are increases in the volume and variety of data needed to support research. Improving the

timeliness (velocity) of data collection, information processing, and reporting of or access to data is incredibly important to support timely research and public health planning. The quality (veracity) of data will always be scrutinized by investigators and will affect the perceived validity and usefulness of cancer registries. Finally, assembling big data about patients with cancer incurs some risks to patient privacy and requires investment in constantly improving technical and operational security controls. To meet the big-data challenges and mitigate the associated risks with finite resources will require rethinking and reinventing the cancer registry data supply chain.

8.4 Objective

Informatics tools and methods such as clinical NLP, machine learning, workflow and interface design, security, and database design may provide valuable approaches to the development of the next generation of cancer surveillance. The goals of this work were to [1] search the recent literature of informatics for cancer registries, [2] to identify key insights and ideas for application of informatics in this domain, and [3] to articulate a vision for further informatics research and development to advance cancer registries and cancer surveillance.

8.5 Methods

To develop a search methodology for informatics and cancer surveillance, we explored MeSH terms and the Medline database for terms to identify cancer registries and relevant informatics topics. Because in our experience we have not seen much published research on applied informatics for cancer registries, we prioritized recall to retrieve the broadest set of potentially useful articles. Also, because of the tremendous changes in biomedical informatics tools and

methods over time, especially the dramatic rise of NLP and machine learning in the biomedical literature over the last decade, we limited the retrieval of articles over the last ten years, from 2005 - 2016. We also limited the search to English language articles.

A search term strategy was developed using both MeSH headings and other terms. For the retrieval of PubMed-indexed articles related to cancer registries, the ("Registries"[mh] AND "Neoplasms"[mh]) AND "SEER Program"[mh] searches with MeSH subject headings missed numerous pertinent articles and returned many irrelevant articles when compared to retrieval using cancer registry specific terms listed in Table 8.1.

Table 8.1. Cancer registry search terms explored in PubMed. From Feb 7, 2016.

Search Term	Term by Itself	Additional Items	Total Items
"cancer registry"	13,065	12,983	12,983
"SEER"	6,639	5,962	18,945
"cancer registries"	3,656	2,268	21,213
"tumor registry"	2,370	2,027	23,240
"cancer register"	713	651	23,891
"National Cancer Data Base"	494	393	24,284
"tumor registries"	405	271	24,555
"National Cancer Database"	269	186	24,741
"tumor register"	47	37	24778
"tumour registries"	41	33	24811
"tumour register"	19	15	24826
"NCDB"	253	14	24840

The “SEER” search term by itself did not miss any of the 4,681 “SEER Program”[mh] tagged articles. However, the combined queries in Table 8.1 missed 8,932 of the 20,759 articles tagged with "Neoplasms"[mh] AND "Registries"[mh] MeSH headings. “NPCR” returned mostly false-positive hits, and “National Program of Cancer Registries” did not yield additional results because the “Cancer Registries” term was already searched. Most of the articles returned from the “register” term came from the Swedish National Prostate Cancer Register (NPCR), but were included for completeness.

The most comprehensive search for cancer registry related articles in Medline was:

(("cancer" OR "tumor" OR "tumour" or “Neoplasms”[mh]) AND ("registry" OR "registries" OR "register" or “Registries”[mh]) OR "SEER" OR "National Cancer Data Base" OR "National Cancer Database" OR "NCDB") AND 2005:2016[dp] AND English[la]

This search returned 28,074 articles on Feb 7, 2016.

For informatics research applicable to cancer registries, the MeSH database was searched to identify indexed concepts, which were tested individually and in combination (see Table 8.2).

Table 8.2. Summary of informatics articles retrieved by individual terms and combined.

Search Term	Term by Itself	Additional Items	Total Items
"Algorithms"[mh]	233,006	354	354
“Software”[mh]	120,141	121	475
“Computer Systems”[mh]	146,734	113	588
“Systems Analysis”[mh]	38,427	44	632
"Informatics"[tiab]	9,435	135	667
“Natural Language Processing”[mh]	2,847	0	667
"Computer Security"[mh]	5,805	6	673
"Informatics"[ot]	986	0	673
“Automatic Data Processing”[mh]	13,310	1 false positive	673
"big data"	1,525	10	683

Because “Algorithms” is a parent term of “Natural Language Processing” in MeSH and returned all of the NLP-related articles, “natural language processing” was not used in the query.

The combined search for informatics research applicable to cancer registries was: ("big data" OR "Computer Security"[mh] OR "Informatics"[tiab] OR “Systems Analysis”[mh] OR “Computer Systems”[mh] OR “Software”[mh] OR "Algorithms"[mh]))

When combined with the cancer registry query above, this search returned 683 Medline articles on February 7, 2016.

To supplement the automated search of Medline, additional manual search was performed for cited papers, papers and authors known by reputation, and papers indexed on the internet. The manual search was not restricted to the medical domain or to a specified time period. To search the internet we used the same terms as for the literature search. The following query was created for the supplementary internet search using Google: ("big data" OR "Computer Security" OR "Informatics" OR “Systems Analysis” OR “Computer Systems” OR “Software” OR "Algorithms") (("cancer" OR "tumor" OR "tumour") ("registry" OR "registries" OR "register") OR "SEER" OR "National Cancer Data Base" OR "National Cancer Database" OR "NCDB")). Google search filters were set to the date range 2005–2016 (after:2005/01/01) and English language. For Google Scholar, the query had to be shortened to: ("informatics" OR "natural language processing" OR "automation") (("cancer" OR "tumor" OR "tumour") ("registry" OR "registries" OR "register") OR "SEER" OR “NCDB”). The 2005–2016 date range was used with Google Scholar as well, and patents were excluded.

The top 250 hits from both internet (Google) and Google Scholar searches were reviewed to identify additional pertinent articles. Only a couple additional relevant articles were found.

Papers were considered relevant if they described the application of NLP or other informatics tools and methods to improve the acquisition and processing of cancer registry data or to automate operations of cancer registries. Articles that dealt with analysis of registry data or

tools and methods for data analysis were not considered relevant for this review. From anecdotal reports and our sense of this domain, we hypothesized that relative to the informatics problems and opportunities for clinical data processing and machine learning to improve cancer registries, there would be few publications of modern informatics tools and methods to registry operations. The search process was therefore designed and anticipated to have the highest recall (sensitivity) possible, but most likely very low precision (positive predictive value).

After filtering out nonrelevant papers, the selected literature was grouped into common themes that would likely be at least somewhat familiar to informatics researchers and the cancer surveillance community. Then, each theme-oriented set of papers was reviewed. I took notes related to the informatics tools or methods applied and the results. These notes were then summarized into one results section per theme (the subsections of 8.6 below).

Finally, based on the background that described the challenges in cancer surveillance and based on the findings described in section 8.6 below, I describe a vision for further informatics research and development to advance cancer registries and cancer surveillance.

8.6 Results

Overall, this review confirmed anecdotal reports and our sense that there would be a dearth of applied informatics, NLP, and machine-learning research for cancer registries and surveillance.

8.6.1 False Positives and Exclusions

Of the 685 articles retrieved from PubMed and Google, 220 false positives dealt with the analysis of registry data. These were mostly retrieved because of the inclusion of the "Algorithms" MeSH term. Although not the topic of this paper, the number of published analyses based on registry data indicates the importance of cancer registries to research. A total of 86

articles dealt with image registration for radiologic and radiation oncology studies. Although imaging-related articles could have been filtered out from the queries, they were included in order to potentially identify any articles about the use of radiology reports in cancer registries (e.g., to assess progression and recurrence of cancer).

A total of 57 articles were related to disease-specific registries, including rare diseases and conditions or exposures that may be risk factors related to the development of cancer. A total of 54 articles focused on registries to facilitate the enrollment of patients to studies or clinical trials. There were 24 articles focused on procedure- or treatment-specific registries (e.g., bone marrow transplant, gynecologic surgery quality, mammography, robotic surgery). The search identified 21 articles that used cancer registries for identification of participants for a study. A total of 13 articles focused on registries of screening for cancer or follow-up of cancer survivors. Ten articles were from developing disease-focused repositories that incorporated cancer registry data. Two articles were about biologics registries, and 1 described an immunization registry. There were 5 registry overview papers that merely mentioned informatics, 4 articles on the value of registry data and guidelines for use for research, and 13 articles on metadata registries. Ten of the informatics methods articles retrieved by the "Algorithms" search term dealt with quality improvement or the imputation of missing values. A total of 69 articles were otherwise not relevant because of misclassification in Medline or only incidental mention of registries and informatics search terms. On review of their abstracts, 40 additional articles were excluded for being off topic or not broadly applicable to cancer registries. In sum, 66 articles were categorized and reviewed below.

8.6.2 Identifying Reportable Cancer Cases

The largest proportion of articles about informatics, NLP, and machine learning applied to cancer registries dealt with automating the identification of reportable cancer cases to improve the efficiency and accuracy of surveillance. Deterministic algorithms for case-finding using coded data sources such as claims and hospital discharge summaries were implemented and validated by a number of groups.(135–141). Tognazzo et al (142) from the Venetian Tumour Registry reported the acquisition of 55% of cancer cases through algorithms and automation using ICD-9 codes from hospital discharge records and death certificates, and SNOMED codes from pathology reports. In a follow-up study, Tognazzo et al(143) acknowledged the difficulty of implementing deterministic rules to automate cancer reporting and have explored the use of probabilistic classifiers to determine reportability based on coded data from different sources of evidence. At Vanderbilt, Naser et al(144) described efficiency gains of approximately 80 person-hours per month through automated matching algorithms of patients and cases before manual review to determine if they were new analytic (reportable) cases. Similarly, to both reduce manual chart reviews and augment registry data, Chubak et al(145) developed algorithms with hospital administrative data to identify breast cancer recurrence or second primaries. Haque et al(146) at Kaiser Permanente also developed algorithms and a semi-automated approach to identify recurrent or second primary breast cancers from pathology reports and administrative data, and they hypothesized that NLP could further improve the review of pathology reports if machine-readable electronic reports were available. Gold and Do(147) evaluated three published algorithms to identify breast cancer data using codes from Medicare claims data and found substantial variation in the performance of the algorithms by geographic region. They questioned how good an algorithm's performance needs to be for staff to rely on its application to new data.

Beyond the deterministic algorithms above, some groups trained statistical algorithms on coded data. Fenton et al(148) developed statistical algorithms (classification and regression trees) using claims data to distinguish between screening mammograms versus diagnostic mammograms. Mahnken et al(149) developed statistical models to identify cases of oral and pharyngeal cancers using claims data.

Some groups have shifted from deterministic and statistical algorithms to keyword searches and simple clinical data processing to facilitate case-finding. Asgari et al(150) from Kaiser Permanente developed text string (keyword) searches of pathology reports to improve case-finding for basal cell carcinomas. Cogle et al implemented both claims-based algorithms(151) and keyword algorithms(152) on pathology reports to improve the reporting of myelodysplastic syndromes (MDS) that are frequently diagnosed in outpatient clinics and may be underreported in cancer registries, especially with ICD-9 codes where MDS was not classified as a neoplasm. In Detroit, Eide et al(153) compared claims-based algorithms and NLP of pathology reports to identify non-melanoma skin cancers, and concluded that NLP methods performed notably better than claims-based algorithms. Hanauer et al(154) developed an open-source registry case-finding engine using rules-based clinical data processing that searches for keywords, phrases, and negation.

March et al(155) extended existing commercial software(156) for pathology report mining to use with radiology reports for the detection of reportable cases of central nervous system (CNS) tumors. This extension involved developing a lexicon for CNS neoplasm-related terms in CT and MRI reports, heuristic testing, and validation against a reference dataset prepared for their study. All potentially reportable cases were still reviewed by tumor registrars before inclusion in the registry.

In France, Jouhet et al(157) applied supervised machine learning (SVMs and Naive Bayes classifiers) to classify pathology reports for reportable cancer cases and assign ICD-O-3 topography and morphology codes. The classifiers performed quite well with pathology reports treated as a "bag of words" and using off-the-shelf machine-learning toolkits, which the authors noted leaves much room for improvement. In addition, Jouhet et al(158) hypothesized that selecting, prioritizing and reformatting relevant records for processing could reduce noise for manual validation by registry staff.

In Australia, Patrick et al(159) have applied more sophisticated NLP to radiology reports to identify cancer cases for the registries. Initially, Nguyen et al(160) evaluated the performance of classifying free-text pathology reports for reportable cancer cases using SNOMED CT codes and related concepts as well as context from pathology report free text. Subsequently, Nguyen and Patrick(161) incorporated machine learning and active learning models and a rich feature set from source documents including key terms, linguistic context, and negation.

At University of Alabama at Birmingham, Osborne et al(162) have successfully applied NLP and machine-learning methods using the Unstructured Information Management Architecture (UIMA) framework, which is also used by both cTAKES and IBM Watson tools, making their solution potentially portable to other centers and systems.

8.6.3 Natural Language Processing

In addition to case-finding, a few registries have applied NLP tools and methods to extract reportable data elements from narrative medical documents. McCowan et al(163) developed machine-learning approaches for cancer staging from pathology reports. Nguyen et al(164) developed rules-based clinical data processing algorithms for registries to compute lung cancer stage from free-text pathology reports using the open-source Generalized Architecture for Text

Engineering (GATE)(165) platform. At the Kentucky cancer registry, Kavuluru et al(166) applied statistical and machine-learning methods for text classification to extract cancer topology and morphology data elements from pathology reports. In Taiwan, Liang et al(167) developed a simple automated system for extracting data elements from free-text pathology reports to facilitate the work of registrars.

Although not specifically applied to cancer registries, Luo et al(168) from MIT, Massachusetts General, and Harvard explored the automation of lymphoma classification from pathology reports using graph-based concept tagging and machine learning. This and other papers indicate that there are likely existing clinical data processing tools and methods already applicable in the cancer domain that could be applied to cancer registry information-processing tasks. Moreover, clinical data processing beyond cancer may be applicable to the cancer registry domain. For example, training data and algorithms developed for the i2b2 smoking status(169) and obesity and comorbidity(170) challenges are certainly relevant to enriching the cancer registry databases for prevention research.

8.6.4 Data Linkages

Although a thorough exploration of the opportunities of data linkages for cancer registries is beyond the scope and was not the intention of this review, several articles described linkages. There were some innovative uses of claims and administrative data to augment information collected in cancer registries. Hassett et al(171) used administrative and claims data to develop algorithms to detect recurrence for lung and colorectal cancers. Bikov et al(172) developed an algorithm to identify medical therapies for colon cancer using SEER-Medicare data. Warren et al(173), Pezzi(174), and Haejin et al(175) all remarked on the importance of recurrence information for cancer registries to support research, the lack of recurrence information in

registry data, and that heuristic algorithms developed for limited populations using claims data may not scale well or would at least need to be validated across multiple cancer registries before broad implementation. In the United Kingdom, Ashley et al(176–178) explored linkage of patient-reported outcome (PRO) systems with cancer registries to enable broad research on cancer outcomes, quality of life, and survivorship.

8.6.5 Integrated Platforms

Several articles described integrated data platforms where cancer registry data were a core component. Weber et al(179) determined that the cancer registries make an ideal foundation for comparative effectiveness research and based their data dictionaries around SEER standards where possible, augmenting it with EMR and other data sources. Other groups such as the ORIEN(180), SPIN(53), and CRN(181) networks have gravitated to cancer registry data and standards as a core of clinical information for an integrated data platform to support cancer research and augmented the registry data through linkage to other data sources.

Beyond the data itself, cancer registry staff and systems have been leveraged for informatics platforms. Dhir, Patel et al(182) describe the employment of cancer registry staff as honest brokers for linkage and queries of identified data in the operation of an integrated data and specimen repository to support research. Hurdle et al(183) describe a data integration and query system allowing investigators to search cancer registry, genealogy, vital records, and EMR data to find cohorts for trials or retrospective studies. Houser et al(184) describe the challenges of using a cancer registry system (C/NET Solutions(185)) to support clinical trials eligibility screening.

8.6.6 Automation

One of the anticipated benefits of informatics applied to cancer registries was the automation of registry operations. Levin et al (186) developed heuristic algorithms for automating the consolidation of data from multiple sources. Zhang et al(187) developed similar algorithms to consolidate dates of cancer diagnosis. Zhang et al suggested that standardizing data-consolidation algorithms and applying approaches from computer science (specifically information retrieval) could help develop more scalable and robust approaches to automated consolidation of information reported to registries.

Some groups re-engineered their registry management software to support automation. In Japan, Shiki et al(188) developed a Unified Modeling Language (UML) of the cancer registration process and were able to thereby identify opportunities to improve efficiency. In Italy, Contiero el al(189) reported on the development of software for automating and managing the Varese Cancer Registry. These automation approaches involved the structure of systems and entailed handcrafted rules for data processing rather than machine learning.

8.6.7 Race and Ethnicity Algorithms

A few registries have reported on use of algorithms to facilitate race and ethnicity determination. Schwartz et al(190) in Detroit developed a database and algorithms to help identify Arab Americans. Hsieh et al(191) in Louisiana evaluated the NAACCR Asian/Pacific Islander Identification Algorithm to improve coding of Asian ethnicities in cancer registries. Boscoe et al(192) in New York compared statistical and heuristic algorithms for determining Hispanic ethnicities in cancer registries, without finding significant differences between the results of

heuristic versus statistical approaches. Maringe et al(193) used a South Asian Names and Group Recognition Algorithm to study cancer survival differences in England.

8.6.8 Security

Given the privileged access of registries to sensitive information, privacy and security have been key concerns of registry staff. Registry staffs are typically proficient in sequestering HIPAA-specified identifiers (e.g., name, address, social security number) from their records before delivering datasets to researchers or aggregated databases (e.g., NCI, NPCR, CINA in Figure 8.1). Howe et al(194) described a software program developed for registries that looks for potential confidentiality breaches based on a statistical model of combinations of key variables that could determine small numbers of individual patients (e.g., sex, race, age group, year of diagnosis, cancer site, county). In Utah, Hurdle et al(183) minimized potential re-identification of individuals through removing the results of aggregated data where queries returned fewer than 5 individuals.

The authority and responsibility of cancer registries to collect and protect identifiers and sensitive information is sometimes more impacted by policy than security measures. Anderson and Storm(195) remark that recent potentially excessive data protection rules in Europe could lead to missing data from registries and that there may be an inherent conflict between data confidentiality and public health needs. Hakulinen et al(196) emphasize the importance of the use of identifiers for data linkages that enable public health research. Kerr(197) and Casali(198) also express concern that new European regulation on data protection that would potentially require consent for the collection of cancer registry data would undermine cancer research. Rahu and McKee(199) report a case of the Estonian Cancer Registry being impaired through data

protection regulation to the point of providing misleading information. The inability to link with death certificates led to artificially high survival rates reported from registry data.

8.6.9 Software

Like the operations of central cancer registries that are highly complex with big-data scale challenges, the systems used to manage registry operations are also highly complex and challenged by scale issues. We found very little research has been published about central registry software. The New Jersey State Cancer Registry and Rutgers Cancer Institute of NJ(200) reported on their implementation of the CDC's Registry Plus web-based system for direct electronic reporting of cases by physician offices and ambulatory care centers. This kind of software to support direct reporting by facilities helps to reduce the work of central registries, many of which are running on minimal staffing. Other than web-based documentation, we did not find any publications on the SEER*DMS system, which is a widely used program for central cancer registries to manage their operations and reporting.

8.6.10 Other Findings

A decade ago, the implementation of standard checklists and synoptic pathology reports was anticipated as a solution to improve the quality and timeliness of pathology data for registries.(201,202) However, these structured reporting efforts have been slow to materialize, so registries must still rely heavily on conventional narrative text pathology reports for the majority of their case-finding and abstraction workflows.

There was little information on underlying data models for cancer registries. Esteban-Gil et al(203) developed OWL ontologies and semantic data repository to transform a cancer registry database for new types of analysis and data visualization. However, we have not found other

reports of innovative application of knowledge representation or big-data tools and methods to cancer registries.

From England, Rashbass and Peake(204) described the value of cancer registries for providing high-quality population-level data for research and public health, and the current limits of registries due to incomplete data and latency in reporting. They remarked that with the addition of claims and administrative data, registries could provide a good starting point for a wide variety of cancer research, including the annotation of tissue from repositories and clinical trials recruitment. Because of the scale of national registry operations, they are inherently difficult to change quickly, but the evolution of registries is critical and possible.

8.7 Discussion

8.7.1 Implications of Findings

As expected, there appears to be a dearth of informatics, clinical data processing, and machine-learning research applied to cancer registries, particularly in the United States. Although this literature search was intentionally broad, the difficulty in conducting searches with high recall and the number of false-positive results may reflect the scattered nature and inconsistent classification of informatics research for cancer registries. Refining these queries for monitoring published literature as well as publishing and presenting overviews and opinion pieces around informatics, clinical data processing, and machine learning for cancer registries at national meetings could help to increase awareness in the informatics community of the problems and opportunities for applied research in this area.

Numerous articles retrieved for this review described domains adjacent to cancer registries. For disease-, procedure-, or treatment-specific registries at cancer centers, perhaps the

lack of local access to cancer registry data and the inadequacy of clinical details therein have motivated researchers to develop their own dedicated systems. The historic purpose and scope of cancer registries has been to collect data on confirmed diagnoses and treatments of cancer. However, the abundance of adjacent databases discovered in this review (e.g., for screening, survivorship, and follow-up, conditions that may be precursors to cancer, rare and underreported diseases) may indicate opportunities for collaboration with and linkage to cancer registries. Perhaps there are opportunities for registry linkages that can support closely related work in biorepositories, rare or orphan diseases, new treatments and procedures, registries of conditions and exposures that may be precursors to cancer, survivorship, and clinical trial networks.

There are challenges of data volume, variability, and security to apply clinical data processing to cancer registries. Because of these current challenges, research into the application of locally developed informatics tools and methods to central cancer registries could help to advance the translation of novel tools and methods to into enterprise and interinstitutional scale. Although case-finding automation may create a quality control burden for registry staff in the short term, as the process and algorithms are smoothed out, there should be substantial savings of staff time and effort through automation.(205)

8.7.2 Next Steps and Opportunities

Although some progress has been made in the use of NLP and machine learning for case-finding and abstraction of data elements from pathology reports, there appears to be a tremendous gap and need to validate, apply, and improve these methods for central cancer registries. The research and application of clinical data processing, and machine-learning tools and methods could be extended to support other kinds of automation for registries, such as the consolidation of dates of diagnosis or multiple reports. NLP and machine learning are not even mentioned in

the Medical Informatics Basics for Cancer Registry guide from NAACCR.(206) Registries must be better informed about the opportunities of informatics, clinical data processing and machine learning, and funded to help advance the application of these technologies to improve the efficiency and quality of their operations.

To keep the information security and data confidentiality protections in central registry systems up-to-date with an evolving understanding of risks and controls, we need much more research and development of robust information security tools and methods applicable to cancer registries.

With the scale and complexity of systems that support central cancer registries, there is likely value in publishing the evaluation and lessons learned from the implementation and operation of these systems (e.g., the SEER*DMS system(106)). Also, the enhancement of these systems to support higher volumes and greater variety of data, greater automations, new data linkages, and improved security should be shared both for peer review and for increasing awareness of cancer registry challenges for the biomedical informatics community.

Finally, an extensive literature review should be conducted of modern clinical data processing tools and methods that could be applied to cancer registries.

8.8 Conclusion

Cancer surveillance is a broad and well-established operation for the acquisition and processing of clinical data and making those data available to advance research and public health. Although cancer registries are experiencing big-data problems that are orders of magnitude greater than individual hospitals and research centers, their operations still rely largely on manual information processing and simple heuristic algorithms for automation. To expand the data they collect and

remain relevant to current and next-generation cancer research, cancer registries and the organizations that fund them will need to adopt big-data tools and methods such as NLP and machine learning, and to actively engage with the biomedical informatics research community.

This work was a review of the literature [1] to validate our assumption that there has been a lack of applied informatics research for cancer registries, [2] to identify areas where some progress has been made, [3] to discuss opportunities to advance registry operations, and [4] to rethink the role and opportunity for registries to accelerate cancer research.

8.9 Acknowledgements

I would like to thank Lynne Penberthy, Steve Schwarz, Eric Durbin, and NCI Surveillance Research Program staff for their feedback on Figure 8.1 and orientation to cancer registries and cancer surveillance. Also, thanks to Linda Coyle and the team at IMS for explanations of the SEER*DMS system and information-processing workflows for cancer registration.

8.10 Synthesis

Sections 8.2 to 8.8 addressed my Chapter 8 questions: What informatics tools and methods have already been applied successfully in the cancer surveillance domain? What are the current and coming opportunities for applying or developing new tools and methods for advancing biomedical informatics in this domain? The literature review of informatics applied to cancer research directly addresses the first question. The tools and methods applied in this domain are summarized in section 8.6, and the opportunities for applying existing informatics tools and methods or developing new ones for cancer surveillance are directly addressed in the discussion (section 8.7), next steps and opportunities (also section 8.7), and conclusion (section 8.8) above.

The work in this chapter ties back to the overarching question for this dissertation: how can we improve access to clinical and related data about cancer patients for research? Cancer registries and national cancer surveillance provide both a massive data source and a challenge for research and public health. Their data are used as a foundation or an augmentation for databases and integrated data repositories across many cancer centers. Because of the economies of scale and the mechanisms used to mandate, fund, and operate cancer registries, it appears feasible to significantly augment and automate them, and in turn make high-volume cancer surveillance data available for every cancer center and the entire cancer research community.

This work also ties back my overall hypothesis that there are new tools and methods from biomedical informatics that could improve the availability of data for cancer research if they were applied thoughtfully and strategically. Chapter 8 identified the themes of informatics tools and methods that appear to improve data availability for cancer registries and described the background and opportunities for biomedical informatics in cancer surveillance in order to stimulate strategic thinking and new research in this area.

Finally, this work ties back to Aim 4 (to characterize the big-data needs and informatics opportunities for cancer surveillance at the national level). Chapter 8 described the big-data problems in cancer surveillance, reviewed of the literature at the intersection of informatics and cancer registries, and assessed the opportunities for research and applications of biomedical informatics and IT in this domain.

Chapter 8 is informed by Chapter 2 (Caisis) because the database model, challenges, and lessons learned about acquisition of data for the Caisis database are all present in cancer surveillance—however in greater scale by orders of magnitude. Chapter 8 is informed by Chapter 3 (Recurring Themes in Informatics and IT from 60 Cancer Centers) in that the

challenges of data acquisition, processing, and dissemination to researchers is present across all cancer centers and difficult for any one center to completely address because of the inefficiencies of their relatively small scale. Chapter 8 is informed by Chapter 4 (HIDRA: Data Platform and Clinical Research Informatics Strategy for a Consortium Cancer Center) in that that the legal and IRB frameworks, types of data, and needs of researchers covered in Chapter 4 are all applicable at the national cancer surveillance level, again at a greater scale. Chapter 8 is informed by Chapter 6 (Scalable Clinical Data Pipeline for a Cancer Center) in that automated approaches to information extraction and processing become not only feasible but imperative at a national scale. Also, the integration of manual annotation and abstraction with clinical data processing and machine learning has strategic importance and is feasible at the national cancer surveillance level, so the the workflows and tooling developed in Chapter 6 are applicable to the domain described in Chapter 8. Chapter 7 (Comparison of Database Models for Cancer Research) also informs Chapter 8 because many of the database model approaches characterized in Chapter 7 have been proposed as solutions for cancer surveillance, and the underlying data quality and database model concepts described in Chapter 7 show up in discussions of cancer surveillance technology. The database framework and lessons learned from the exploration of database models in Chapter 7 can be immediately applied in an overall strategy for research and the application of informatics tools and methods for cancer surveillance.

Findings from this chapter are relevant to other cancer centers addressing the costs of operating a cancer registry or the potential integration and use of cancer registry data to support research.

First, cancer registries are typically cost centers. Hospitals or other healthcare facilities that are accredited by the American College of Surgeons Commission on Cancer (CoC) are

required to abstract and report clinical data elements to the National Cancer Data Base (NCDB). This requirement is an unfunded mandate, but necessary to maintain their credentials. This chapter describes research from several groups showing that automated clinical data processing, especially using NLP and machine learning methods, can enhance cancer registry data processing. These findings indicate that cancer registry tasks, such as case finding (identifying reportable cancer cases) and extracting required data fields from pathology reports for cancer reporting, are candidates for automation. Automation of case finding and manual data abstraction could save hospitals money, and could speed up the data acquisition and processing timeline, making registry data more up-to-date, reliable, and thus useful as a data source for research data repositories.

Second, central cancer registries and national cancer surveillance programs typically invest millions of dollars on to support manual data abstraction efforts at healthcare facilities, data quality control and data consolidation efforts at central registries and further consolidation and reporting to national databases. The work in this chapter indicates that automation of clinical data processing (e.g., through NLP and machine learning methods) has performed well in several cases and has potential to facilitate the work that these cancer surveillance programs fund.

Third, this work identifies clinical data processing automation that is directly applicable to cancer registries and cancer surveillance as an area of potential research for biomedical informatics that would have value at the local and national level.

Chapter 9 Conclusions

Each section in this chapter explains the findings from prior chapters. The order of the chapters and conclusions below is based on the natural progression of this work. The first two chapters (Chapters 2 and 3) cover the experiments and issues identified through the preliminary work on the Caisis database at a single cancer center and the exploration of IT and informatics issues across multiple cancer centers. The next three chapters (Chapters 4–6) cover the implementation of the Hutch Integrated Data Repository and Archive (HIDRA) and a critical strategic part of it, an enterprise pipeline for clinical data processing. The last two chapters (Chapters 7 and 8) cover issues of scalability and the potential application of big-data tools and methods. The conclusions include extrapolation beyond the individual chapters. At the end, limitations and avenues for future research are discussed.

9.1 Conclusions from Chapter 2. Preliminary Work: Caisis

The two questions related to this chapter were [1] What aspects of Caisis are applicable today to current and future informatics platforms for cancer research? and [2] What are the limits of Caisis that would need to be addressed with new tools and methods?

I started this research with a background in developing relational database models and operations that could scale well and support predictive modeling. Over the past decade, the need to scale data systems and the growth of predictive modeling and personalized medicine initiatives has only increased the demand for this type of biomedical informatics research and solution development.

The Caisis project, summarized in Chapter 2, was aimed to develop a reusable database to support predictive modeling and a variety of cancer research projects. From the preliminary

work on Caisis, it was evident that a database model that is temporally organized and aligned with clinical source systems could support high-volume data abstraction, as well as the reuse of clinical data across multiple projects in a cancer center. Having clear naming conventions in the database model, built-in vocabulary management, and a metadata-driven web application made this system relatively easy to configure without programming and was critical for extensibility and sustainability. However, to make data readily available to researchers, analytic tools and staff with outstanding database skills were needed to denormalize and query the data. For extensibility of the relational database model behind Caisis, neither structural changes to tables and fields nor EAV virtual fields turned out to be ideal solutions. The biomedical domain knowledge and technical skills of staff were barriers to optimal implementation. Moreover, there were limits to the amount of data acquisition that could be provided through data feeds, structured data entry in clinical source systems and manual data abstraction.

Although the Caisis system was adopted by several other cancer centers in the United States and abroad, the experience and lessons learned were somewhat biased to one large standalone cancer center, Memorial Sloan-Kettering (MSKCC). In the visits to cancer centers in the following chapter, I sought to determine if the findings from the Caisis work were generalizable and if there were other common findings and trends that could contribute to a more successful strategy for cancer research data management.

In answer to the first question for this chapter (What aspects of Caisis are applicable today to current and future informatics platforms for cancer research?), the Caisis database model and its existing configurable and metadata-driven web interface for data entry and editing are still applicable today, at least in the short term until an alternative is developed. Future informatics platforms will need to maintain the flexibility of Caisis, but will need to go much

further in terms of database models and interfaces to support data science and self-service data exploration. The answer to the second question for this chapter (What are the limits of Caisis that would need to be addressed with new tools and methods?) is that the current database model is too deep with too many tables and requires too many joins to query easily, and the design of Caisis to integrate with clinical practice and acquire all data in a structured format is likely not an attainable solution. There may be real practical limits to the acquisition of structured data in clinics, and with the maturation of NLP and machine-learning technology, there may no longer be such a need to capture structured data at the point of service. Systems like Caisis may now be able to step back from highly normalized relational database models and explore hybrid database approaches (described in Chapter 7 on database models) and implement a data acquisition and processing pipeline that is more document-based, automated and amenable to clinical data processing and machine learning (described in Chapter 6 on the pipeline for clinical data processing).

The generalizable contributions of Chapter 2 are a working a comprehensive database model that is temporally organized and has the ability to stack into an analytic structure for predictive modeling and an associated web-based tool for data abstraction. This system is freely distributed under an open-source license, meets common requirements for IT security, extensibility and supportability, and it has already been adopted and extended by numerous other cancer centers in the United States and internationally.

9.2 Conclusions from Chapter 3. Preliminary Work: Recurring Themes in Informatics and IT from 60 Cancer Centers

The question addressed in this chapter is "What are the current opportunities for strategic application of biomedical informatics tools and methods in cancer centers?"

Chapter 3 reviewed the findings from visits to more than 60 cancer centers to identify trends in IT and informatics. Through these site visits and subsequent analysis, I sought to reason from the specific case of Caisis at MSKCC, to more general biomedical informatics problems in cancer centers and ultimately to be able to summarize and apply general findings from multiple centers back to a single cancer center.

First, I found that expecting the majority of usable research data to be collected in discrete form through EMRs appeared to be unrealistic in the near term across all centers. There appeared to be practical limits to structured data collection through templated clinical documentation.

Second, clinical trial and biospecimen management systems tended to be extremely complex, expensive, and time consuming to implement across a cancer center, and simple systems like REDCap were likely to win out over more complex systems. Investigators and staff were likely to vote with their feet by adopting the simplest solutions. Although there could be benefits to integrating data from clinical trial and biospecimen management systems to search for cohorts and specimens for further study, those system implementations were often resource drains and should probably be managed separately from an integrated data repository effort.

Third, ready access to data for research in familiar flat formats (e.g., Excel) was critical to researchers. If research systems did not provide this lowest common denominator of data formats without excessive barriers to access, researchers and staff were not likely to use them.

Fourth, there appeared to be a trend toward the use of lower-cost and more open-source systems across all cancer centers. Most centers have limited funding and resources for informatics initiatives. Few centers have the resources to purchase expensive technology or to completely develop their own solutions for data acquisition, processing, and access. In the current and projected future environments of limited research funding, difficulty in building up local talent, and big-data pressures for scalability and security, a good strategy for most cancer centers would likely be to partner with other centers in collaborative networks and to leverage open-source and existing technology wherever possible.

Finally, the domain knowledge and technical skills of staff, as well as organizational structure and management at a center, could either enable or frustrate the goals of research IT and informatics leaders. To enable the meaningful and sustainable adoption of tools and methods from other cancer centers, the cross-pollination of ideas and collaboration would need to be deeper than just the involvement of senior leadership and individually funded research collaborations.

From this exploration and the summarization of IT and informatics issues across US cancer centers, these findings informed the strategy for the development of the Hutch Integrated Data Repository and Archive (HIDRA, Chapter 4) and its pipeline for clinical data processing (Chapter 6).

The answer to the question for this chapter (What are the current opportunities for strategic application of biomedical informatics tools and methods in cancer centers?) is that there are some key opportunities for all cancer centers. The first is actively seeking out and adopting or adapting work from other centers. The Cancer Informatics for Cancer Centers (CI4CC) community has now evolved into a mechanism for cross-pollination of ideas across centers.

Low-cost, open-source tools and the implementation of high-security environments are strategic opportunities for any cancer center.

The generalizable contributions of Chapter 3 are the following: First, the volume and variety of data elements that can practically be collected through clinical templates is limites. Second, given the importance of research and collaboration networks, cancer centers should adopt or at least be interoperable with common platforms like REDCap, i2b2, OpenSpecimen and OnCore so that we can wrestle with common issues as a community. Third, due to limited and variable funding for research, solutions need to scale down to affordable levels for individual researchers and labs. Fourth, site visits and active cross-pollination of tools and methods across center must extend deeper into all levels of IT and informatics staff rather than just connecting senior IT leaders and informatics researchers. Finally, centers should spend time and effort resolving social and organizational barriers to progress in informatics and IT.

9.3 Conclusions from Chapter 4. Hutch Integrated Data Repository and Archive (HIDRA): Data Platform and Clinical Research Informatics Strategy for a Consortium Cancer Center

The questions answered in this chapter are the following: What are the challenges and opportunities for informatics at Fred Hutch? Do these challenges and opportunities align with my previous work and lessons learned? Are there tools and methods that others or I have used before that could be successfully applied at Fred Hutch? How can we enable new types of research, faster results, and better quality of research data at Fred Hutch? How can we improve access to data with tools and methods that are scalable, extensible to new projects, sustainable, and portable to other centers?

The HIDRA project in Chapters 4–5 was aimed to develop a data platform to improve data access and support a variety of research efforts at the Fred Hutch/UW Cancer Consortium. This aim was achieved and is summarized in Chapters 4–5. For the initial HIDRA strategy, I used previous experience from the Caisis database at MSKCC (Chapter 2) and findings and trends identified across multiple cancer centers (Chapter 3). Near the start of the HIDRA project, I interviewed around thirty Fred Hutch and Cancer Consortium investigators to assess current IT and informatics challenges and to confirm my hypothesis that the findings from Caisis and other cancer centers would apply at Fred Hutch. I also identified unique requirements and opportunities at Fred Hutch.

Overall, my strategy was to implement a reusable data platform like Caisis, but I realized that [1] data acquisition through manual data abstraction, data feeds, and templated notes in clinics would be insufficient, [2] the legacy legal and IRB framework for data sharing across the Fred Hutch/UW Cancer Consortium was causing an awkward technical architecture and an integrated repository would be impossible without reforming that framework, [3] grant and contract opportunities for a center that mastered FISMA security requirements and risks for centers that were lagging in security made a high-security environment a critical requirement, [4] we did not have the technical skills and mature development processes in-house to build HIDRA ourselves, so we needed a technical partner, and [5] we should adopt and build on technology from other groups to enable HIDRA to be sustainable and to play well with other centers. I also knew that we had to address the biggest weakness of Caisis and the focus of tools like i2b2: the ability for researchers to query and get access to the data themselves through a self-service user interface.

In the HIDRA project, we found that creating a legal and IRB foundation for the repository, a high-security environment, and operations to support the repository were important factors to be able to avoid wasted time and scale up quickly. Throughout all of this work, information security and attentiveness to regulations was paramount. There was no responsible way to create scalable and integrated data systems without implementing a high-security environment, including operations and documentation to support it.

We also found that the lack of federated authentication that allows users to log in with their preferred organization credentials can cutoff big-data integration projects at the knees if it is intended to support a consortium cancer center or a network of centers. A high-security environment should include provisions for relatively easy access across the intended base of users, using their own organizational authenticated credentials.

We found that having a reliable and competent technical partner may ease the pressures on local technical staff and may lead to creative solutions and learning opportunities. However, local staff must be deeply engaged for a successful and sustainable solution developed with an external technical partner. Healthy dynamics between project leadership, internal staff, and an external technical partner help to prioritize development and control scope. Also, adopting and adapting informatics tools from other centers for key functions can speed up and lower total development cost. The development or adoption of a self-service data access interface is important to investigators and staff. However, there is still a great need to customize user interfaces, data export formats and reports for each disease group.

The pipeline for clinical data processing for HIDRA was a complex and critical set of functionality, and it is described in more detail in Chapter 6. HIDRA has initially relied on the Caisis database model for abstraction, data feeds and storage of cancer data, and the Caisis data

has been processed into cubes to optimize queries for the Argos self-service data exploration and access tool for HIDRA. The limits of the Caisis database model and potential future database model options are characterized in Chapter 7.

Again, the questions answered in this chapter are the following: What are the challenges and opportunities for informatics at Fred Hutch? Do these challenges and opportunities align with my previous work and lessons learned? Are there tools and methods that others or I have used before that could be successfully applied at Fred Hutch? How can we enable new types of research, faster results, and better quality of research data at Fred Hutch? How can we improve access to data with tools and methods that are scalable, extensible to new projects, sustainable and portable to other centers? The assessment of impact of HIDRA is covered in Chapter 5.

The answer to the first question (What are the challenges and opportunities for informatics at Fred Hutch?) is that perhaps the greatest challenges have been the skills and expertise of the current team and the management of a large, complex project. Bringing in new talent and working to keep everyone involved on a common roadmap with common understanding of what we were trying to accomplish has been extremely challenging. Establishing a legal and IRB framework for a consortium cancer center was an unanticipated challenge that turned into an opportunity after it was solved. Implementing a high-security environment and high-security approach to the development and implementation of HIDRA has turned from a challenge to an opportunity for leadership and a competitive advantage.

The answer to the second question (Do these challenges and opportunities align with my previous work and lessons learned?) is both yes and no. The challenges of consolidating multiple disparate database efforts into one integrated strategy for HIDRA aligns closely with my experience in developing Caisis at MSKCC. However, performing this work in a consortium

center across multiple teams rather than in a hierarchical organization like MSKCC has proven very difficult. In both situations, it appears that strong and sustained leadership is critical to success.

The answer to the third question (Are there tools and methods that others or I have used before that could be successfully applied at Fred Hutch?) is yes. The Caisis database model and system that I developed at MSKCC (described in Chapter 2) was applicable to HIDRA and allowed us to temporarily bypass the design of a new database model, with all of the challenges therein (e.g., time, domain and technical resources, resolution of conflicting approaches). The adoption and adaptation of the LabKey Server platform and the technical partnership with LabKey Software resulted in another successful application of previously developed tools and methods.

The answer to the fourth question (How can we enable new types of research, faster results, and better quality of research data at Fred Hutch?) is that HIDRA may address all of these issues. By integrating multiple disparate databases into one platform and database model, we are now able to more easily research cancers that cross traditional boundaries of disease types (e.g., breast cancer, leukemia, sarcoma) and medical specialties (e.g., surgery, radiology, radiation oncology, medical oncology). By implementing both self-service tools like Argos and a HIDRA dataset request service staffed by analysts, we have enabled faster access to data for research. Also, integrating data into the Caisis database model so that data entry staff and analytics are coming from the same tables and fields, implementing direct data feeds from clinical source systems, and working through issues with each feed or database consolidated into HIDRA have uncovered data quality issues and forced us to resolve them, set standards, and continuously improve the performance of the system and the quality of its contents.

The answer to the fifth question (How can we improve access to data with tools and methods that are scalable, extensible to new projects, sustainable, and portable to other centers?) is that leveraging and extending existing open-source technology like LabKey Server and sticking to our goals of a single database model has allowed us to improve access to data. Having a technology partner (LabKey Software) that has a vested interest in making their software portable to other centers and their business sustainable has kept us aware of these issues and allowed us to control the scope and focus of software engineering to balance local enhancements with broader goals of extensibility, portability, and sustainability. Creating a large but safely de-identified testing dataset of realistic data that could be freely used by LabKey and Fred Hutch developers, operations staff, and testers was useful to improve scalability and tune performance.

The aim of this chapter was to develop and assess a modern integrated data platform to support a wide variety of cancer research. This aim was achieved through the development and implementation of the HIDRA system at Fred Hutch. Open questions remain around the role of the Caisis database model and web-based data-entry system as HIDRA is opened up to new users and groups. Future work on HIDRA will likely involve redesign of the database model (which was the motivation for Chapter 7 on database models) and the implementation of the clinical data pipeline for clinical data processing (which is described in Chapter 6). The database model evolution and implementation of the clinical data pipeline are likely intertwined. Each depends somewhat on the other for successful and scalable deployment.

The generalizable contributions of Chapter 4 are the following: First, the legal, IRB, and security framework for HIDRA is relevant to other centers and has been applied to at least two similar efforts. Second, HIDRA provides an example of leveraging a clinical data repository at a broader academic medical center to support a cancer-specific data repository. Third, HIDRA

provides an example of adopting and extending the IT and informatics work of other groups to solve local issues economically. Fourth, HIDRA provides an example of an overall strategy for clinical data acquisition, processing, storage and self service data access. Fifth, HIDRA identified the need for a realistic and de-identified testing dataset to facilitate software development and system implementation. Sixth, the HIDRA work found that lack of federated security for a consortium or matrix cancer center is a critical barrier to progress on an integrated data repository. Finally, the HIDRA project found that the Agile approach to software engineering and system implementation was critical for project momentum and success.

9.4 Conclusions from Chapter 5. Evaluation of HIDRA

The following question is addressed in this chapter: What is the impact of the data platform developed at Fred Hutch?

9.5 Conclusions from Chapter 6. Scalable Clinical Data Pipeline for a Cancer Center

The questions answered in this chapter are the following: How can we improve the quality, speed, and economics of the acquisition, processing, and delivery of clinical data to support cancer researchers? How can we make clinical data processing a core competency of Fred Hutch at an enterprise scale?

Developing a scalable pipeline for clinical data processing was considered a critical part of HIDRA. It addressed the findings and conclusions from preliminary work in Caisis and findings from other cancer centers that there were practical limits to what could be achieved through templated clinic notes, data feeds from clinical systems, and typical manual data

abstraction. It also targeted one of the greatest barriers to developing enterprise-level clinical data processing solutions: the lack of ongoing training and validation data. The clinical data processing pipeline would allow us to adjust existing data entry work across the center to generate data suitable for clinical data processing algorithm development and to provide ongoing quality assurance data to validate and improve clinical data processing algorithms.

This pipeline provides a secure and open-source tool for the acquisition and processing of clinical documents, performs automated information extraction and case data consolidation through the use of an clinical data processing engine (e.g., locally developed algorithms or other commercial or open-source tools), and can feed the resulting data into an integrated data repository. The clinical data pipeline provides a framework for manual information-processing tasks such as abstraction from clinical reports for ongoing research or new annotation projects as well as the verification and performance evaluation of clinical data processing algorithms. The clinical data processing pipeline was developed by Fred Hutch and LabKey and can be used to integrate multiple automated processing solutions within a single open-source but well-supported architecture. Integrating the LabKey clinical data processing pipeline with existing data abstraction workflows has the potential to incrementally decrease staff time and efforts through automations and increase the amount of annotated training data available for future clinical data processing algorithm development.

Although there are many clinical data processing algorithms developed for cancer research, this pipeline fills an unmet need to help us transition from individual projects and algorithms to employing clinical data processing at an enterprise scale as part of everyday processes at a cancer center.

Through the development of the clinical data processing pipeline, we encountered challenges with the Caisis database model and the need to store structured documents and extracted data elements as they are processed. These requirements have naturally led to the investigation of document stores (e.g., MongoDB) and ways to store structured documents within relational databases. These database model issues are discussed further in Chapter 7.

The answer to the first question for this chapter (How can we improve the quality, speed, and economics of clinical data acquisition, processing, and delivery to support cancer researchers?) is that I still think that integration of clinical data processing, machine learning, and manual data abstraction into a single flexible clinical data pipeline is a viable strategy for improving quality, speed and economics of data acquisition, processing and delivery. Manual data abstraction and data processing are just not scalable or sustainable, given the cost constraints and variable funding in cancer centers and the ever-increasing volume and variety of clinical data as cancer treatments become more successful and cancer becomes more of a chronic disease with multiple treatments and ongoing monitoring. In our experience, clinical data processing and machine-learning algorithms have proven to perform well across a variety of clinical data applications at a cancer center.

The answer to the second question (How can we make clinical data processing a core competency of Fred Hutch at an enterprise scale?) is that it takes a sustained and concerted effort to introduce and spread deeper and deeper understanding of clinical data processing across an enterprise. After clinical data processing competency was established as a core part of the strategy for HIDRA and Fred Hutch, we have given multiple presentations and engaged in research collaborations with different groups within the cancer center. These presentations and collaborations continue and have been steadily growing. Hiring core NLP research engineer staff

that are technically competent and mentored to present and collaborate well with a variety of groups is so critical that we established a summer internship program in NLP. I think the combination of telling the clinical data processing story repeatedly in multiple venues and engaging in collaborative clinical data processing research while implementing an enterprise pipeline is the road to establishing clinical data processing as a core competency of Fred Hutch and other centers.

The aim of this chapter was to develop and characterize a clinical data processing pipeline that is scalable to the cancer center enterprise level, is well-supported and sustainable, and can complement or streamline existing manual data abstraction and information-processing activities. This aim has been largely achieved as described in this chapter. A preliminary version of this pipeline was presented in October 2015 at the LabKey User Conference and CI4CC meeting, and a fully functional version of the pipeline is on track for completion in summer 2016. The open questions include how to successfully transition existing and new manual data abstraction and clinical data processing work to the pipeline and what modifications will be needed in terms of the pipeline functionality, performance, and usability to achieve user acceptance and broad deployment. During the upcoming summer, the 2016 NLP interns will be working with the clinical data processing pipeline for their research projects in order to start addressing these open questions and informing future work on the pipeline.

The generalizable contribution of Chapter 6 is a tool for shifting the work of manual data abstraction so that it generates training and validation data suitable for development of automated clinical data processing algorithms (e.g., statistical algorithms, NLP, machine learning). This tool was built with a technology partner, LabKey Software, so that it can be

portable to other clients, scalable, and extensible to different clinical data sources and databases. This tool is already being adopted by other groups.

9.6 Conclusions from Chapter 7. Comparison of Database Models for Cancer Research

The question addressed in this chapter was How can we best characterize and compare database models and big-data technologies to inform a sensible strategy for cancer research database design and technology?

From my preliminary work on the Caisis database model in Chapter 2, findings from other cancer centers in Chapter 3, and through the challenges of implementing database models to support HIDRA and the clinical data processing pipeline in Chapters 4 and 6, I decided it would be worthwhile to explore and compare database models and their potential applications to cancer research systems. Furthermore, with the current number and variety of big-data companies marketing various tools and database models in the biomedical domain, there was a need to think through and be able to articulate the essence and tradeoffs of the various database models.

There is clearly no single database model to meet the requirements of managing clinical and molecular data for cancer research. Many organizations will need conceptual and practical guidance to better understand and make decisions about system requirements and design. Each database model implemented will require staff time and resources for mapping, formatting, and recoding of data. Most organizations will need to acquire and process an increasing volume and variety of data sources over time.

As big-data technologies are evolving and disseminating rapidly, it may not make sense to invest or plan more than a couple years ahead of projected database growth. High-dimensional data like genomics assays and sensor data may perform well in a column store like Cassandra. Clinical data may perform well in a combination relational database and document store for acquisition, processing, and storage. For high-performance data exploration and visualization, a column store may offer the best performance for both clinical and high-dimensional data. However, if the database must be used to support clinical decision-making, the latency and consistency of returned results must be factored into the design and platform selection.

To achieve economies of scale in a cancer center, consortium, or network, the primary database model for data entry, editing, and storage should be reusable across multiple projects in order to recoup the costs of implementation and operation. It should be lossless in terms of data and extensible so that planned and unplanned data can be reliably acquired, stored, and accessed. Also, given the prevalence and persistence of flat database models with closed-world assumptions to support trials and epidemiologic studies, the primary database model should provide a simple way to document absent events or absence of evidence at a point in time.

Star schemas, analytic database models or column stores may be implemented depending on the anticipated usage or participation in research networks. Analytic database models may be somewhat lossy or narrow, and may require complex mapping and transformation to populate. Also, column stores may have consistency issues when querying. All of these issues need to be researched and tested before broad implementation.

The current database knowledge and technical skills of staff and the ability to hire, train, and retain relevant talent should always be factors in planning to adopt different database

models. There is a tremendous learning curve and responsibility on today's programmers to implement reliable and sustainable hybrid database models and technology for big data.

Based on this chapter, I would advocate a hybrid strategy of different database models and technology based on the center's strategic needs and the technical skills of its staff. I would definitely try to adopt and adapt models from other groups rather than reinventing new database models or trying to make old technology and models do new tricks. Database models and big-data technologies appear to be fruitful areas for further research at cancer centers.

The answer to the question addressed in this chapter (How can we best characterize and compare database models and big-data technologies to inform a sensible strategy for cancer research database design and technology?) is that cancer database models may be characterized by their suitability to facilitate different attributes of data quality (e.g., accurate, complete, up-to-date, freedom from duplication, freedom from fragmentation), different notions of datasets prevalent in clinical and research systems (open- and closed-world assumptions), and different types of data (e.g., clinical documents, high-dimensional array data). After this work, I see that database models and technologies can be characterized as evolving into an increasingly specialized collection of approaches that each have strengths and tradeoffs.

The aim of this chapter was to develop, model, and assess database frameworks for cancer. This aim was achieved in that I now have a better understanding of how each of these database models and technologies may perform and how they relate to each other. Open questions remain around the performance of column stores versus star schemas and analytic database models for queries and data exploration tools, and document stores versus relational models to support information-processing workflows such as the pipeline for clinical data

processing described in Chapter 6. Based on the work in this chapter, I can now design and guide strategic exploration of these different models and big-data technologies.

9.7 Conclusions from Chapter 8. Informatics and Cancer Surveillance: Literature Review and Vision

This chapter addresses the following questions: What informatics tools and methods have already been applied successfully in the cancer surveillance domain? What are the current and coming opportunities for applying or developing new tools and methods for advancing biomedical informatics in this domain?

Chapter 8 explores the extension of all of the previous chapters to cancer registries and cancer surveillance. Cancer surveillance is a broad and well-established operation for the acquisition and processing of clinical data and making those data available to advance research and public health. Although cancer registries are experiencing big-data problems that are orders of magnitude greater than individual hospitals and research centers, their operations still rely largely on manual information processing and simple heuristic algorithms for automation. To expand the data that they collect and remain relevant to current and next-generation cancer research, cancer registries and the organizations that fund them will need to adopt big-data tools and methods such as NLP and machine learning and to actively engage with the biomedical informatics research community to evaluate or develop new tools and methods that can scale appropriately.

Because cancer registries and broader cancer surveillance efforts, such as the NCI SEER(125) program, have been operating at a scale orders of magnitude greater than any single cancer center and most research networks, they may provide the opportunity to develop robust

tools and methods for data acquisition, clinical data processing, machine learning and database models that are relevant and applicable for individual centers.

One of the patterns seen across all of this research is staff taking an engineering approach to each information-processing task, breaking it down into parts, and writing "simple" data processing rules. While rules-based approaches to data processing make sense to most software engineers, at the scale of consortia, networks of centers, or cancer surveillance, engineering our way through increasingly complex information-processing workflows is just painful and likely untenable. This rules-engineering approach to facilitating and automating information processing in workflows may be playing out in clinics with EMRs, structured clinical templates, alerts, etc.

To develop solutions that scale, my thinking has changed about how to approach information processing—from an engineering perspective to an education perspective. How do we train people to abstract or process medical information? We provide examples of the inputs (e.g., documents) and desired outputs (e.g., list of data elements and desired coding standards), and then we walk through a supervised training process. Initially, staffs learn to recognize features in documents, what to extract, and how to process the features into desired outputs through a set of heuristics and repetition. Some of the examples used for educational purposes are routine cases; others are exceptional cases that may require differential information processing. The training examples may refer to and require the use of heuristic rules or they may require more complex parsing of linguistic structures and reasoning. Through repetition, staffs begin to rely mostly on pattern recognition for information processing. When the pattern is not obvious, staff may turn to heuristic rules and guidelines for abstraction and data processing. When these heuristics are insufficient, staff turn to deeper language parsing and reasoning, often bringing in more experienced staff to help. I think we need to take this approach to train and

apply clinical data processing and machine-learning algorithms for scalable information processing.

Rather than thinking of algorithm development as an engineering task, my conclusion from this work is that we should think of algorithm development as an educational task. We should use the same kinds of training examples with machine-learning algorithms that we use with human staff, and create a learning environment that is a hybrid for human and artificial intelligence (AI) systems. This approach would involve rethinking and standardizing inputs, guidelines (e.g., training documents), outputs, and performance evaluation so that they can be used interchangeably for human and AI systems. Also, rather than trying to start over with this approach, we could adapt existing systems and processes to capture input and output data continuously in a format that it can be used to train and evaluate algorithms, people, or both. Rather than over-engineering workflows and making life painful for clinical and research staff, we should be thinking of and training hybrid systems where people and AI algorithms are collaborating to complete information-processing tasks. Ultimately, this is the kind of data environment that would be required to scale and support cancer surveillance, cancer research across networks of centers, and precision medicine initiatives.

The answer to the first question (What informatics tools and methods have already been applied successfully in the cancer surveillance domain?) is that numerous informatics tools and methods have been applied to cancer surveillance with purported success. These tools and methods include NLP, machine-learning, statistical and heuristic data processing algorithms, data linkage methods, integrated platforms, workflow engineering, and automation, information security, software design and engineering and even ontologies. Relatively few of these methods have been applied in the United States or applied broadly across multiple cancer registries.

The answer to the second question (What are the current and coming opportunities for applying or developing new tools and methods for advancing biomedical informatics in this domain?) is that given the seeming success of informatics tools and methods in the cancer registry domain where they have been applied and the huge and increasing challenges of scalable and affordable data acquisition and processing for cancer surveillance, informatics, clinical data processing, and machine-learning tools and methods appear to be highly applicable in this domain. Moreover, cancer registries have a tremendous amount of existing data for developing and validating new tools and methods, as well as an opportunity to implement those tools and methods broadly to have a large and sustainable effect on improving data for cancer research.

The aim of this chapter was to characterize the big-data needs and informatics opportunities for cancer surveillance at the national level. This aim has been achieved through this chapter's overview of cancer surveillance and its big-data issues, literature review of informatics applied in this domain, summarization of findings, and description of opportunities for research and development. Future work will involve the collection and characterization of clinical NLP and machine-learning work applied in the cancer domain that is likely applicable to registries and cancer surveillance.

The generalizable contributions of Chapter 8 are the following. First, it provides a description of cancer registries and cancer surveillance from an informatics perspective, including the case for automation. Second, it contributes a review of informatics tools and methods applied to cancer registries that indicates potential for automation of clinical data processing. Third, it identifies cancer registries and cancer surveillance as an area for funding and advancing biomedical informatics research.

9.8 Limitations

There are several limitations to this work. Aside from a panel of clinically actionable mutations from tumor samples(207) and biomarker results reported in pathology reports, HIDRA (covered in Chapter 4) and cancer registries (covered in Chapter 8) have not integrated much genomic data. At this point in time, many cancer centers still wrestle with the cost of acquiring large numbers of molecular assays to associate with patients' clinical data for data mining and translational research. The Argos self-service user interface for HIDRA has not been widely used, and the broad rollout of this tool to users is pending federated authentication of user credentials (logins) from all of the cancer consortium partners as well as the resolution of performance issues.

A fully functional pipeline for clinical data processing (covered in Chapter 6) is nearing completion, so this functionality has not been evaluated in practice. We anticipate that testing of the pipeline and changes to existing data abstraction workflows will be a gradual process with numerous challenges.

The comparison of database models (covered in Chapter 7) summarized my experience and a targeted review of literature only. It was not a comprehensive review of big-data technologies and did not include testing of the various database models and tools. However, this chapter was only intended to provide frameworks for further discussion and research.

The review of informatics and cancer surveillance literature (Chapter 8) was high level and strategy focused, so it did not get into the details of algorithms or approaches used. It also focused exclusively on cancer registries and did not cover clinical NLP and machine learning from adjacent biomedical domains.

9.9 Overall Conclusions and Avenues for Future Research

My overarching research question has been: how can we improve access to clinical and related data about cancer patients for research? I think I have answered this question through the characterization of strategies and systems developed and implemented at individual cancer centers (Caisis at MSKCC described in Chapter 2, and HIDRA at Fred Hutch described in Chapter 4). Creating legal, IRB, security, and technology foundations and integrating disparate databases into systems that have self-service data access and data request services are my answer. The development and implementation of these foundations rely heavily on local talent, a strategy and plan for work, and ongoing management. I have also answered the overall research question through broad looks across multiple cancer centers (described in Chapter 3) and across the cancer registry and cancer surveillance domain (described in Chapter 8). Some advancements in data platform development, clinical data processing, and machine learning may depend on training data and may be driven by pressures to achieve scale and economies of scale. Most individual cancer centers do not have and can hardly afford the resources to tackle all of these issues and develop their own solutions. Also, most centers do not have the patient volumes to justify expensive systems or software development efforts. To improve access to clinical and related data will require partnering across networks of centers, so a strategy of adopting flexible, portable, and extensible systems locally and funding their advancement nationally is important. Finally, a deeper understanding and application of database models and big-data technologies at local cancer centers and across networks of cancer surveillance can potentially help improve the speed and current difficulties of sharing data about cancer patients for research.

My hypothesis for this dissertation was that new tools and methods from biomedical informatics could improve the availability of data for cancer research if they were applied

thoughtfully and strategically. I addressed this hypothesis through multiple angles, from the actual implementation and assessment of the Caisis (Chapter 2) and HIDRA (Chapter 4) systems at individual centers, a broad set of site visits to 60 cancer centers to characterize trends in IT and informatics (Chapter 3), and a broad literature review of cancer surveillance (Chapter 8). In particular, advancing pipelines for clinical data processing (Chapter 6) and database models (Chapter 7) are areas within biomedical informatics that I think are most urgent and readily applicable to improve the availability of data for cancer research.

Based on this work, future research could be conducted around the following areas:

- Performance of various big database models (e.g., star schema analytical data models, column stores, documents stores, hybrid models) to support information acquisition and processing workflows and to support data access (e.g., self-service exploration, searches and queries, reports and data visualizations)
- Developing and evaluating tools and methods for high-security environments. A key aspect of this research should be exploration of the usability and scalability implications of any security approach.
- Approaches to portable application and scaling of clinical data processing and machine-learning algorithms developed by different research groups.
- Developing and evaluating education-based approaches for both human and machine learning of information extraction and processing tasks. This work involves developing strategies and tools for efficient and ongoing annotation and abstraction, human review of algorithms outputs, algorithm review of human outputs, and how to most effectively combine human and AI in hybrid systems that perform well and are sustainable.

This chapter summarized the findings and conclusions from each of the previous chapters, from the preliminary work on Caisis and cancer center site visits, through the development of HIDRA and the pipeline for clinical data processing, to the explorations of both database models to support cancer research and issues of scale and extensibility to cancer registries.

Bibliography

1. Weber GM, Mandl KD, Kohane IS. Finding the missing link for big biomedical data. JAMA [Internet]. 2014 Jun 25;311(24):2479–80. Available from: http://dx.doi.org/10.1001/jama.2014.4228

2. Board of Scientific Advisors Ad Hoc Working Group. An assessment of the impact of the NCI cancer biomedical informatics grid (caBIG®) [Internet]. 2011. Available from: http://deainfo.nci.nih.gov/advisory/bsa/archive/bsa0311/caBIGfinalReport.pdf

3. Fearn P, Sculli F. The CAISIS Research Data System. In: Biomedical Informatics for Cancer Research [Internet]. Springer US; 2010 [cited 2016 Feb 21]. p. 215–25. Available from: http://link.springer.com/chapter/10.1007/978-1-4419-5714-6_11

4. Caisis [Internet]. 2013 [cited 2016 Feb 19]. Available from: http://caisis.org/

5. Fearn PA, Regan K, Sculli F, Katz J, Kattan MW. A chronological database as backbone for clinical practice and research data management. In: Computer-Based Medical Systems, 2003 Proceedings 16th IEEE Symposium [Internet]. 2003. p. 9–15. Available from: http://dx.doi.org/10.1109/CBMS.2003.1212759

6. Fearn P, Regan K, Sculli F, Fajardo J, Smith B, Alli P. Lessons Learned from Caisis: An Open Source, Web-Based System for Integrating Clinical Practice and Research. In: Twentieth IEEE International Symposium on Computer-Based Medical Systems (CBMS'07) [Internet]. IEEE; p. 633–8. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4262719

7. OMOP Common Data Model | OHDSI [Internet]. [cited 2016 Feb 21]. Available from: http://www.ohdsi.org/data-standardization/the-common-data-model/

8. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc [Internet]. 2010 Mar;17(2):124–30. Available from: http://dx.doi.org/10.1136/jamia.2009.000893

9. Fearn PA, Lafferty HJ, Bauer MJ, Kattan MW. A clinical research database solution for HIPAA privacy and security requirements. MedInfo, San Francisco, CA. 2004;

10. Dinu V, Nadkarni P. Guidelines for the effective use of entity-attribute-value modeling for biomedical databases. Int J Med Inform [Internet]. 2007 Nov;76(11-12):769–79. Available from: http://dx.doi.org/10.1016/j.ijmedinf.2006.09.023

11. Nadkarni PM, Brandt C. Data Extraction and Ad Hoc Query of an Entity—Attribute—Value Database. J Am Med Inform Assoc [Internet]. The Oxford University Press; 1998 Nov 1 [cited 2016 Feb 21];5(6):511–27. Available from: http://jamia.oxfordjournals.org/content/5/6/511.short

12. Cohort Jigsaw [Internet]. [cited 2016 Feb 21]. Available from: http://jigsawanalytics.com/

13. Silgard E, Fearn PA, Nichols K, Tran J, Omaiye A, Velagapudi N. Characterization of clinical data elements for secondary use in a comprehensive cancer center. AMIA Joint Summits on Translational Science; San Francisco, CA.

14. Caisis. Caisis Collaborators [Internet]. Caisis; [cited 2016 Apr 23]. Available from:

http://caisis.org/collaboration.html

15. Welcome to CTSA Central [Internet]. [cited 2016 Feb 21]. Available from: https://ctsacentral.org/consortium/institutions/

16. CancerCenters - OCCWebApp 2.1.0 [Internet]. [cited 2016 Feb 21]. Available from: http://cancercenters.cancer.gov/Center/CancerCenters

17. Informatics and Stimulus Funding [Internet]. Bioinformatics@Becker. 2009 [cited 2016 Feb 21]. Available from: http://beckerinfo.net/bioinformatics/informatics-and-stimulus-funding/

18. DiLaura R, Turisco F, McGrew C, Reel S, Glaser J, Crowley WF Jr. Use of informatics and information technologies in the clinical research enterprise within US academic medical centers: progress and challenges from 2005 to 2007. J Investig Med [Internet]. 2008 Jun;56(5):770–9. Available from: http://dx.doi.org/10.231/JIM.0b013e3175d7b4

19. Wallace PJ. Reshaping cancer learning through the use of health information technology. Health Aff [Internet]. 2007 Mar;26(2):w169–77. Available from: http://dx.doi.org/10.1377/hlthaff.26.2.w169

20. McIntosh LD, Sharma MK, Mulvihill D, Gupta S, Juehne A, George B, et al. caTissue suite to OpenSpecimen: Developing an extensible, open source, web-based biobanking management system. J Biomed Inform [Internet]. 2015 Aug 29; Available from: http://dx.doi.org/10.1016/j.jbi.2015.08.020

21. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform [Internet]. 2009 Apr;42(2):377–81. Available from: http://www.sciencedirect.com/science/article/pii/S1532046408001226

22. TheBrain :: Mind Mapping Software, Brainstorming, GTD and Knowledgebase Software [Internet]. [cited 2016 Feb 21]. Available from: http://www.thebrain.com/

23. Epic [Internet]. 2016 [cited 2016 Feb 21]. Available from: https://www.epic.com/

24. Cerner [Internet]. 2016 [cited 2016 Feb 21]. Available from: http://www.cerner.com/

25. GE Centricity [Internet]. 2016 [cited 2016 Feb 21]. Available from: http://www3.gehealthcare.com/en/Products/Categories/Healthcare_IT/Electronic_Medical_Records/Centricity_EMR

26. Allscripts [Internet]. 2016 [cited 2016 Feb 21]. Available from: http://www.eclipsys.com/

27. McKesson [Internet]. 2016 [cited 2016 Feb 21]. Available from: http://www.mckesson.com/

28. Berg M. Implementing information systems in health care organizations: myths and challenges. Int J Med Inform [Internet]. 2001 Dec;64(2-3):143–56. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11734382

29. Stead WW. Rethinking electronic health records to better achieve quality and safety goals. Annu Rev Med [Internet]. 2007;58:35–47. Available from: http://dx.doi.org/10.1146/annurev.med.58.061705.144942

30. Berner ES, Detmer DE, Simborg D. Will the wave finally break? A brief view of the adoption of

electronic medical records in the United States. J Am Med Inform Assoc [Internet]. 2005 Jan;12(1):3–7. Available from: http://dx.doi.org/10.1197/jamia.M1664

31. Chute CG, Beck SA, Fisk TB, Mohr DN. The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. J Am Med Inform Assoc [Internet]. 2010 Mar;17(2):131–5. Available from: http://dx.doi.org/10.1136/jamia.2009.002691

32. Vogel. Oncology IT: A glimpse of where medicine and IT are headed. ADVANCE for Health Information Executives. 2009 Oct;22–4.

33. Simborg DW. Promoting electronic health record adoption. Is it the correct focus? J Am Med Inform Assoc [Internet]. 2008 Mar;15(2):127–9. Available from: http://dx.doi.org/10.1197/jamia.M2573

34. DiLaura RP. Clinical and translational science sustainability: overcoming integration issues between electronic health records (EHR) and clinical research data management systems "separate but equal." Stud Health Technol Inform [Internet]. 2007;129(Pt 1):137–41. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17911694

35. Bernstam EV, Hersh WR, Johnson SB, Chute CG, Nguyen H, Sim I, et al. Synergies and distinctions between computational disciplines in biomedical research: perspective from the Clinical andTranslational Science Award programs. Acad Med [Internet]. 2009 Jul;84(7):964–70. Available from: http://dx.doi.org/10.1097/ACM.0b013e3181a8144d

36. Masys DR, Harris PA, Fearn PA, Kohane IS. Designing a public square for research computing. Sci Transl Med [Internet]. 2012 Aug 29;4(149):149fs32. Available from: http://dx.doi.org/10.1126/scitranslmed.3004032

37. Velos [Internet]. 2016 [cited 2016 Feb 21]. Available from: http://velos.com/

38. Forte OnCore [Internet]. 2016 [cited 2016 Feb 21]. Available from: http://forteresearch.com/enterprise-research-oncore/

39. mdlogix [Internet]. 2016 [cited 2016 Feb 21]. Available from: http://www.mdlogix.com/

40. medidata [Internet]. 2016 [cited 2016 Feb 21]. Available from: https://www.mdsol.com/en

41. PhaseForward (now owned by Oracle) [Internet]. 2010 [cited 2016 Feb 21]. Available from: http://www.oracle.com/us/corporate/Acquisitions/phaseforward/index.html

42. OpenClinica [Internet]. 2016 [cited 2016 Feb 21]. Available from: https://www.openclinica.com/

43. Sim I, Tu SW, Carini S, Lehmann HP, Pollock BH, Peleg M, et al. The Ontology of Clinical Research (OCRe): an informatics foundation for the science of clinical research. J Biomed Inform [Internet]. 2014 Dec;52:78–91. Available from: http://dx.doi.org/10.1016/j.jbi.2013.11.002

44. Weng C, Tu SW, Sim I, Richesson R. Formal representation of eligibility criteria: a literature review. J Biomed Inform [Internet]. 2010 Jun;43(3):451–67. Available from: http://dx.doi.org/10.1016/j.jbi.2009.12.004

45. Morris MJ, Basch EM, Wilding G, Hussain M, Carducci MA, Higano C, et al. Department of Defense prostate cancer clinical trials consortium: a new instrument for prostate cancer clinical research. Clin Genitourin Cancer [Internet]. 2009 Jan;7(1):51–7. Available from: http://dx.doi.org/10.3816/CGC.2009.n.009

46. Speakman J. The caBIG ,Clinical Trials Suite. In: Biomedical Informatics for Cancer Research [Internet]. Springer; 2010. p. 203–13. Available from: http://link.springer.com/chapter/10.1007/978-1-4419-5714-6_10

47. The International Society of Biological and Environmental Repositories presents Abstracts from their Annual Meeting. Biopreserv Biobank. 2009;7(1):51–88.

48. Kroth PJ, Schaffner V, Lipscomb M. Technological and administrative factors implementing a virtual human biospecimen repository. AMIA Annu Symp Proc [Internet]. 2005;1011. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16779298

49. Lowe HJ, Ferris TA, Hernandez PM, Weber SC. STRIDE--An integrated standards-based translational research informatics platform. AMIA Annu Symp Proc [Internet]. 2009 Nov 14;2009:391–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20351886

50. 5AM Solutions [Internet]. 2016 [cited 2016 Feb 21]. Available from: https://www.5amsolutions.com/

51. eMERGE network [Internet]. 2014 [cited 2016 Feb 21]. Available from: https://emerge.mc.vanderbilt.edu/

52. Prostate SPORE NBN Pilot - Docs and Files [Internet]. 2014 [cited 2016 Feb 21]. Available from: https://wiki.nci.nih.gov/display/GFORGEARCHIVES/Prostate+SPORE+NBN+Pilot+Project+-+Docs+and+Files

53. Drake TA, Braun J, Marchevsky A, Kohane IS, Fletcher C, Chueh H, et al. A system for sharing routine surgical pathology specimens across institutions: the Shared Pathology Informatics Network. Hum Pathol [Internet]. 2007 Aug;38(8):1212–25. Available from: http://dx.doi.org/10.1016/j.humpath.2007.01.007

54. Patel AA, Gilbertson JR, Parwani AV, Dhir R, Datta MW, Gupta R, et al. An informatics model for tissue banks--lessons learned from the Cooperative Prostate Cancer Tissue Resource. BMC Cancer [Internet]. 2006 May 5;6:120. Available from: http://dx.doi.org/10.1186/1471-2407-6-120

55. Patel AA, Gilbertson JR, Showe LC, London JW, Ross E, Ochs MF, et al. A novel cross-disciplinary multi-institute approach to translational cancer research: lessons learned from Pennsylvania Cancer Alliance Bioinformatics Consortium (PCABC). Cancer Inform [Internet]. 2007 Jun 8;3:255–74. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19455246

56. Murray MD, Smith FE, Fox J, Teal EY, Kesterson JG, Stiffler TA, et al. Structure, functions, and activities of a research support informatics section. J Am Med Inform Assoc [Internet]. 2003 Jul;10(4):389–98. Available from: http://dx.doi.org/10.1197/jamia.M1252

57. SAS [Internet]. 2016 [cited 2016 Feb 21]. Available from: https://www.sas.com/en_us/home.html

58. Nahm M, Zhang J. Operationalization of the UFuRT methodology for usability analysis in the clinical research data management domain. J Biomed Inform [Internet]. 2009 Apr;42(2):327–33. Available from: http://dx.doi.org/10.1016/j.jbi.2008.10.004

59. Rayner TF, Rocca-Serra P, Spellman PT, Causton HC, Farne A, Holloway E, et al. A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. BMC Bioinformatics [Internet]. 2006 Nov 6;7:489. Available from: http://dx.doi.org/10.1186/1471-2105-

7-489

60. Sansone S-A, Rocca-Serra P, Brandizi M, Brazma A, Field D, Fostel J, et al. The first RSBI (ISA-TAB) workshop: “can a simple format work for complex studies?” OMICS [Internet]. 2008 Jun;12(2):143–9. Available from: http://dx.doi.org/10.1089/omi.2008.0019

61. Dudley JT, Butte AJ. A quick guide for developing effective bioinformatics programming skills. PLoS Comput Biol [Internet]. 2009 Dec;5(12):e1000589. Available from: http://dx.doi.org/10.1371/journal.pcbi.1000589

62. Fleming V, Garland J, Morgan F, Bolger R, Baum LF, Langley N, et al. The wizard of Oz [Internet]. Turner Entertainment Company; 2009. Available from: http://www.devoir-de-philosophie.com/pdf_free/245107.pdf

63. SOCRA The Society of Clinical Research Associates [Internet]. 2016 [cited 2016 Feb 21]. Available from: http://www.socra.org/

64. AMIA 10x10 Courses [Internet]. 2016 [cited 2016 Feb 21]. Available from: https://www.amia.org/education/10x10-courses

65. ConvergeHEALTH [Internet]. 2016 [cited 2016 Feb 21]. Available from: http://www.converge-health.com/

66. LabKey Software [Internet]. 2016 [cited 2016 Feb 21]. Available from: https://www.labkey.com/

67. Weeks J. On management: Culture eats strategy. Management Today. 2006;

68. Rose AF, Schnipper JL, Park ER, Poon EG, Li Q, Middleton B. Using qualitative studies to improve the usability of an EMR. J Biomed Inform [Internet]. 2005 Feb;38(1):51–60. Available from: http://dx.doi.org/10.1016/j.jbi.2004.11.006

69. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet Med [Internet]. 2013 Oct;15(10):761–71. Available from: http://dx.doi.org/10.1038/gim.2013.72

70. Voss EA, Makadia R, Matcho A, Ma Q, Knoll C, Schuemie M, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. J Am Med Inform Assoc [Internet]. 2015 May;22(3):553–64. Available from: http://dx.doi.org/10.1093/jamia/ocu023

71. Fred Hutch / University of Washington Cancer Consortium [Internet]. 2016 [cited 2016 Jan 17]. Available from: http://cancerconsortium.org/en.html

72. Cancer Surveillance System [Internet]. 2016 [cited 2016 Jan 17]. Available from: https://www.fredhutch.org/en/labs/phs/projects/cancer-surveillance-system.html

73. Nelson EK, Piehler B, Eckels J, Rauch A, Bellew M, Hussey P, et al. LabKey Server: an open source platform for scientific data integration, analysis and collaboration. BMC Bioinformatics [Internet]. 2011 Mar 9;12:71. Available from: http://dx.doi.org/10.1186/1471-2105-12-71

74. Ramsay S. Unlocking Medical Records with Natural Language Processing [Internet]. CI4CC; 2015 Oct 21; La Jolla, CA. Available from: http://www.ci4cc.org/

75. Ramsay S, Silgard E, Rauch A. Unlocking Medical Records with Natural Language Processing [Internet]. LabKey User Conference; 2015 Oct 1; Seattle, WA. Available from: https://www.labkey.com/conference

76. Silgard E, Galuhn T, Egan K, Rauch A, Fearn P. An Enterprise Clinical Data Pipeline for a Cancer Center. AMIA 2015 Annual Symposium; San Francisco, CA.

77. HIDRA [Internet]. 2016 [cited 2016 Jan 17]. Available from: https://www.fredhutch.org/en/labs/hidra.html

78. Holzinger A. Usability Engineering Methods for Software Developers. Commun ACM [Internet]. New York, NY, USA: ACM; 2005 Jan;48(1):71–4. Available from: http://doi.acm.org/10.1145/1039539.1039541

79. Hwang S. The Duke Breast Data Repository [Internet]. Duke Joint Health Informatics Seminar; 2014 Aug 27; Durham, NC. Available from: http://www.dukeinformatics.org/8-27-14-seminar-breast-cancer-data-mining-shelley-hwang-md-mph-1/

80. Amorosano D. Unstructured data a common hurdle to achieving guidelines: Healthcare organizations are increasingly looking for solutions to transform paper-based processes into more efficient electronic workflows. Health Manag Technol [Internet]. 2012;33(6):28–9. Available from: http://europepmc.org/abstract/med/22787953

81. Pennic J. Healthline Launches HealthData Engine to Harness Unstructured Data [Internet]. HIT Consultant. 2014 [cited 2016 Feb 23]. Available from: http://hitconsultant.net/2014/09/22/healthline-launches-healthdata-engine/

82. Back A. Patient-physician communication in oncology: what does the evidence show? Oncology [Internet]. 2006 Jan;20(1):67–74; discussion 77–8, 83. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16572594

83. CDC - National Program of Cancer Registries (NPCR) [Internet]. [cited 2016 Feb 7]. Available from: http://www.cdc.gov/cancer/npcr/

84. National Cancer Data Base [Internet]. 2016 [cited 2016 Jan 22]. Available from: https://www.facs.org/quality%20programs/cancer/ncdb

85. Surveillance, Epidemiology and End Results Program [Internet]. 2016 [cited 2016 Jan 22]. Available from: http://seer.cancer.gov/

86. ASCO CancerLINQ [Internet]. 2016 [cited 2016 Jan 22]. Available from: http://cancerlinq.org/

87. IBM Watson Healthcare [Internet]. 2016 [cited 2016 Jan 22]. Available from: http://www.ibm.com/smarterplanet/us/en/ibmwatson/health/

88. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc [Internet]. 2010 Sep;17(5):507–13. Available from: http://dx.doi.org/10.1136/jamia.2009.001560

89. I speak computer: Making medical information Big Data-ready [Internet]. 2014 [cited 2016 Jan 22]. Available from: https://www.fredhutch.org/en/news/center-news/2014/11/natural-language-processing-of-medical-information.html

90. Nichols K, Silgard E, Fearn P, Yahne J, Pillarisetty V. Extracting pancreatic cancer diagnosis and stage from clinical text. AMIA 5th Annual Summit on Clinical Research Informatics; San Francisco, CA.

91. Aldrichl R, Silgard E. Rule-Based Extraction of Lung Cancer Stages from Free-Text Clinical Notes. AMIA 2015 Joint Summits in Translational Research; San Francisco, CA.

92. Kahn A, Silgard E, Fearn P, McFerrin L. Automated discovery of keywords for clinical event extraction to facilitate patient timeline creation from clinic notes. AMIA 2015 Join Summits in Translational Research; San Francisco, CA.

93. Park HM, Sandhu V, Fearn P, Dorcy KS, Estey EH, Silgard E. Using Natural Language Processing to Determine Chemotherapeutic Regimens Administered within 30 Days Prior to Death in Acute Myelogenous Leukemia Patients. Blood [Internet]. Am Soc Hematology; 2014;124(21):1267–1267. Available from: http://www.bloodjournal.org/content/124/21/1267.abstract

94. Silgard E, Fearn P, Mohedano A, Hammond S. Supplementing service line classification with natural language processing. AMIA 2015 Joint Summits in Translational Research; San Francisco, CA.

95. Silgard E, Sandhu V, Estey E, Herman D. An Automated System for Parsing and Risk Classifying Karyotype Nomenclature for Acute Myeloid Leukemia. ASH Annual Symposium; Orlando, FL.

96. Hunter L, Lu Z, Firby J, Baumgartner WA Jr, Johnson HL, Ogren PV, et al. OpenDMAP: an open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression. BMC Bioinformatics [Internet]. 2008 Jan 31;9:78. Available from: http://dx.doi.org/10.1186/1471-2105-9-78

97. Brat [Internet]. 2016 [cited 2016 Jan 22]. Available from: http://brat.nlplab.org/index.html

98. Knowtator [Internet]. 2016 [cited 2016 Jan 22]. Available from: http://knowtator.sourceforge.net/

99. GATE [Internet]. 2016 [cited 2016 Jan 22]. Available from: https://gate.ac.uk/projects.html

100. Linguamatics [Internet]. 2016 [cited 2016 Jan 22]. Available from: http://www.linguamatics.com/

101. Apache Hadoop [Internet]. 2014 [cited 2016 Feb 2]. Available from: https://hadoop.apache.org/

102. Leavitt N. Will NoSQL Databases Live Up to Their Promise? Computer [Internet]. 2010 Feb;43(2):12–4. Available from: http://dx.doi.org/10.1109/MC.2010.58

103. Amazon Web Services (AWS) [Internet]. 2016 [cited 2016 Feb 2]. Available from: https://aws.amazon.com/

104. Data Quality FAQs. NHS [Internet]. 2010 [cited 2016 Jan 24]. Available from: http://webarchive.nationalarchives.gov.uk/20130502102046/http://www.connectingforhealth.nhs.uk/factsandfiction/systemsfaqs/dataquality

105. Powers DM. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Bioinfo Publications; 2011; Available from: http://dspace2.flinders.edu.au/xmlui/handle/2328/27165

106. SEER Data Management System [Internet]. [cited 2016 Feb 7]. Available from:

http://seer.cancer.gov/seerdms/

107. CDISC Operational Data Model [Internet]. 2016 [cited 2016 Feb 2]. Available from: http://www.cdisc.org/odm

108. Biomedical Research Integrated Domain Group (BRIDG) Model [Internet]. 2015 [cited 2016 Feb 2]. Available from: http://www.bridgmodel.org/

109. Life Sciences Domain Analysis Model (LS-DAM) [Internet]. 2015 [cited 2016 Feb 2]. Available from: https://wiki.nci.nih.gov/pages/viewpage.action?pageId=23401587

110. Murphy S. Ontology services for the i2b2 query platform [Internet]. 2011 [cited 2016 Feb 1]. Available from: http://wiki.siframework.org/file/view/Shawn+Murphy+-+Query+Health+Concept+Working+Group+12-20-2011.pdf

111. Maria Keet C. Open World Assumption. In: Encyclopedia of Systems Biology [Internet]. Springer New York; 2013 [cited 2016 Feb 23]. p. 1567–1567. Available from: http://link.springer.com/referenceworkentry/10.1007/978-1-4419-9863-7_734

112. Hustadt U, Others. Do we need the closed world assumption in knowledge representation? In: KRDB [Internet]. 1994. Available from: http://ceur-ws.org/Vol-1/hustadt-long.pdf

113. Drummond N, Shearer R. The open world assumption. In: eSI Workshop: The Closed World of Databases meets the Open World of the Semantic Web [Internet]. 2006. Available from: http://www.nesc.ac.uk/talks/701/OWA_NDrummond.pdf

114. DePuy V. SDTM What? ADaM Who? A Programmer's Introduction to CDISC [Internet]. Bowden Analytics. 2014 [cited 2016 Jan 31]. Available from: http://analytics.ncsu.edu/sesug/2014/PH-11.pdf

115. DeCandia G, Hastorun D, Jampani M, Kakulapati G, Lakshman A, Pilchin A, et al. Dynamo: Amazon's Highly Available Key-value Store. Oper Syst Rev [Internet]. New York, NY, USA: ACM; 2007 Oct;41(6):205–20. Available from: http://doi.acm.org/10.1145/1323293.1294281

116. Trivedi A. Mapping Relational Databases and SQL to MongoDB [Internet]. 2014 [cited 2016 Feb 1]. Available from: http://code.tutsplus.com/articles/mapping-relational-databases-and-sql-to-mongodb--net-35650

117. Chang F, Dean J, Ghemawat S, Hsieh WC, Wallach DA, Burrows M, et al. Bigtable: A Distributed Storage System for Structured Data. ACM Trans Comput Syst [Internet]. New York, NY, USA: ACM; 2008 Jun;26(2):4:1–4:26. Available from: http://doi.acm.org/10.1145/1365815.1365816

118. Wang S, Pandis I, Wu C, He S, Johnson D, Emam I, et al. High dimensional biological data retrieval optimization with NoSQL technology. BMC Genomics [Internet]. 2014 Nov 13;15 Suppl 8:S3. Available from: http://dx.doi.org/10.1186/1471-2164-15-S8-S3

119. Kalakota R. The NoSQL and Spark Ecosystem: A C-Level Guide [Internet]. 2015 [cited 2016 Feb 2]. Available from: http://practicalanalytics.co/2015/06/02/the-maturing-nosql-ecoystem-a-c-level-guide/

120. Huser V, Cimino JJ. Desiderata for healthcare integrated data repositories based on architectural comparison of three public repositories. AMIA Annu Symp Proc [Internet]. 2013 Nov 16;2013:648–

56. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24551366

121. Kuchinke W, Aerts J, Semler SC, Ohmann C. CDISC standard-based electronic archiving of clinical trials. Methods Inf Med [Internet]. 2009 Jul 20;48(5):408–13. Available from: http://dx.doi.org/10.3414/ME9236

122. Meineke FA, Stäubert S, Löbe M, Winter A. A comprehensive clinical research database based on CDISC ODM and i2b2. Stud Health Technol Inform [Internet]. 2014;205:1115–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25160362

123. Verplancke P. Advantages of a real end-to-end approach with CDISC standards [Internet]. 2016 Feb 1; North Bethesda. Available from: http://www.cdisc.org/system/files/all/generic/application/pdf/6c3_verplancke.us_interchange2013.advantages_of_cdiscend_to_end.20131107.pdf

124. History - National Cancer Registrars Association [Internet]. [cited 2016 Feb 7]. Available from: http://www.ncra-usa.org/i4a/pages/index.cfm?pageID=3873

125. SEER Registries - About SEER [Internet]. [cited 2016 Feb 7]. Available from: http://seer.cancer.gov/registries/

126. Cancer Surveillance Programs and Registries in the United States [Internet]. [cited 2016 Feb 7]. Available from: http://www.cancer.org/cancer/cancerbasics/cancer-surveillance-programs-and-registries-in-the-united-states

127. Naaccr I. Volume II, Data Standards and Data Dictionary [Internet]. [cited 2016 Feb 7]. Available from: http://www.naaccr.org/StandardsandRegistryOperations/VolumeII.aspx

128. CDC - Cancer - NPCR - Software and Tools for Cancer Registries and Surveillance [Internet]. [cited 2016 Feb 7]. Available from: http://www.cdc.gov/cancer/npcr/tools/index.htm

129. Certification - National Cancer Registrars Association [Internet]. [cited 2016 Feb 7]. Available from: https://www.ncra-usa.org/i4a/pages/index.cfm?pageid=3864

130. Wohler B, Qiao B, Weir HK, MacKinnon JA, Schymura MJ. Using the National Death Index to identify duplicate cancer incident cases in Florida and New York, 1996-2005. Prev Chronic Dis [Internet]. 2014 Sep 25;11:E167. Available from: http://dx.doi.org/10.5888/pcd11.140200

131. Subramanian S, Tangka FKL, Beebe MC, Trebino D, Weir HK, Babcock F. The cost of cancer registry operations: Impact of volume on cost per case for core and enhanced registry activities. Eval Program Plann [Internet]. 2015 Nov 30;55:1–8. Available from: http://dx.doi.org/10.1016/j.evalprogplan.2015.11.005

132. Tangka FKL, Subramanian S, Beebe MC, Weir HK, Trebino D, Babcock F, et al. Cost of Operating Central Cancer Registries and Factors That Affect Cost: Findings From an Economic Evaluation of Centers for Disease Control and Prevention National Program of Cancer Registries. J Public Health Manag Pract [Internet]. 2015 Dec 3; Available from: http://dx.doi.org/10.1097/PHH.0000000000000349

133. Laney D. 3D data management: Controlling data volume, velocity and variety. META Group Research Note. 2001;6:70.

134. The Four V's of Big Data [Internet]. IBM Big Data & Analytics Hub. [cited 2016 Feb 7].

Available from: http://www.ibmbigdatahub.com/infographic/four-vs-big-data

135. Baldi I, Vicari P, Di Cuonzo D, Zanetti R, Pagano E, Rosato R, et al. A high positive predictive value algorithm using hospital administrative data identified incident cancer cases. J Clin Epidemiol [Internet]. 2008 Apr;61(4):373–9. Available from: http://dx.doi.org/10.1016/j.jclinepi.2007.05.017

136. Larsen MB, Jensen H, Hansen RP, Olesen F, Vedsted P. Identification of patients with incident cancers using administrative registry data. Dan Med J [Internet]. 2014 Feb;61(2):A4777. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24495885

137. Lai S-M, Jungk J, Garimella S. Colorectal Cancer Identification Methods Among Kansas Medicare Beneficiaries, 2008-2010. Prev Chronic Dis [Internet]. 2015 Jul 9;12:E107. Available from: http://dx.doi.org/10.5888/pcd12.140543

138. Eide MJ, Krajenta R, Johnson D, Long JJ, Jacobsen G, Asgari MM, et al. Identification of patients with nonmelanoma skin cancer using health maintenance organization claims data. Am J Epidemiol [Internet]. 2010 Jan 1;171(1):123–8. Available from: http://dx.doi.org/10.1093/aje/kwp352

139. Craig BM, Rollison DE, List AF, Cogle CR. Underreporting of myeloid malignancies by United States cancer registries. Cancer Epidemiol Biomarkers Prev [Internet]. 2012 Mar;21(3):474–81. Available from: http://dx.doi.org/10.1158/1055-9965.EPI-11-1087

140. Couris CM, Polazzi S, Olive F, Remontet L, Bossard N, Gomez F, et al. Breast cancer incidence using administrative data: correction with sensitivity and specificity. J Clin Epidemiol [Internet]. 2009 Jun;62(6):660–6. Available from: http://dx.doi.org/10.1016/j.jclinepi.2008.07.013

141. Penberthy LT, McClish D, Agovino P. Impact of automated data collection from urology offices: improving incidence and treatment reporting in urologic cancers. J Registry Manag [Internet]. 2010 Winter;37(4):141–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21688743

142. Tognazzo S, Andolfo A, Bovo E, Fiore AR, Greco A, Guzzinati S, et al. Quality control of automatically defined cancer cases by the automated registration system of the Venetian Tumour Registry. Quality control of cancer cases automatically registered. Eur J Public Health [Internet]. 2005 Dec;15(6):657–64. Available from: http://dx.doi.org/10.1093/eurpub/cki035

143. Tognazzo S, Emanuela B, Rita FA, Stefano G, Daniele M, Fiorella SC, et al. Probabilistic classifiers and automated cancer registration: an exploratory application. J Biomed Inform [Internet]. 2009 Feb;42(1):1–10. Available from: http://dx.doi.org/10.1016/j.jbi.2008.06.002

144. Naser R, Roberts J, Salter T, Warner JL, Levy M. An informatics-enabled approach for detection of new tumor registry cases. J Registry Manag [Internet]. 2014 Spring;41(1):19–23. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24893184

145. Chubak J, Yu O, Pocobelli G, Lamerato L, Webster J, Prout MN, et al. Administrative data algorithms to identify second breast cancer events following early-stage invasive breast cancer. J Natl Cancer Inst [Internet]. 2012 Jun 20;104(12):931–40. Available from: http://dx.doi.org/10.1093/jnci/djs233

146. Haque R, Shi J, Schottinger JE, Ahmed SA, Chung J, Avila C, et al. A hybrid approach to identify subsequent breast cancer using pathology and automated health information data. Med Care [Internet]. 2015 Apr;53(4):380–5. Available from:

http://dx.doi.org/10.1097/MLR.0000000000000327

147. Gold HT, Do HT. Evaluation of three algorithms to identify incident breast cancer in Medicare claims data. Health Serv Res [Internet]. 2007 Oct;42(5):2056–69. Available from: http://dx.doi.org/10.1111/j.1475-6773.2007.00705.x

148. Fenton JJ, Zhu W, Balch S, Smith-Bindman R, Fishman P, Hubbard RA. Distinguishing screening from diagnostic mammograms using Medicare claims data. Med Care [Internet]. 2014 Jul;52(7):e44–51. Available from: http://dx.doi.org/10.1097/MLR.0b013e318269e0f5

149. Mahnken JD, Keighley JD, Girod DA, Chen X, Mayo MS. Identifying incident oral and pharyngeal cancer cases using Medicare claims. BMC Oral Health [Internet]. 2013 Jan 1;13:1. Available from: http://dx.doi.org/10.1186/1472-6831-13-1

150. Asgari MM, Eide MJ, Warton EM, Fletcher SW. Validation of a large basal cell carcinoma registry. J Registry Manag [Internet]. 2013 Summer;40(2):65–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24002130

151. Cogle CR, Craig BM, Rollison DE, List AF. Incidence of the myelodysplastic syndromes using a novel claims-based algorithm: high number of uncaptured cases by cancer registries. Blood [Internet]. 2011 Jun 30;117(26):7121–5. Available from: http://dx.doi.org/10.1182/blood-2011-02-337964

152. Cogle CR, Iannacone MR, Yu D, Cole AL, Imanirad I, Yan L, et al. High rate of uncaptured myelodysplastic syndrome cases and an improved method of case ascertainment. Leuk Res [Internet]. 2014 Jan;38(1):71–5. Available from: http://dx.doi.org/10.1016/j.leukres.2013.10.023

153. Eide MJ, Tuthill JM, Krajenta RJ, Jacobsen GR, Levine M, Johnson CC. Validation of claims data algorithms to identify nonmelanoma skin cancer. J Invest Dermatol [Internet]. 2012 Aug;132(8):2005–9. Available from: http://dx.doi.org/10.1038/jid.2012.98

154. Hanauer DA, Miela G, Chinnaiyan AM, Chang AE, Blayney DW. The registry case finding engine: an automated tool to identify cancer cases from unstructured, free-text pathology reports and clinical notes. J Am Coll Surg [Internet]. 2007 Nov;205(5):690–7. Available from: http://dx.doi.org/10.1016/j.jamcollsurg.2007.05.014

155. March S, Cernile G, West K, Borhani D, Fritz A, Brueckner P. Application of automated pathology reporting concepts to radiology reports. J Registry Manag [Internet]. 2012 Autumn;39(3):95–100. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23443452

156. HOME - Artificial Intelligence In Medicine (AIM) [Internet]. Artificial Intelligence In Medicine (AIM). [cited 2016 Feb 27]. Available from: http://www.aim.on.ca/

157. Jouhet V, Defossez G, Burgun A, le Beux P, Levillain P, Ingrand P, et al. Automated classification of free-text pathology reports for registration of incident cases of cancer. Methods Inf Med [Internet]. 2012;51(3):242–51. Available from: http://dx.doi.org/10.3414/ME11-01-0005

158. Jouhet V, Defossez G, CRISAP, CoRIM, Ingrand P. Automated selection of relevant information for notification of incident cancer cases within a multisource cancer registry. Methods Inf Med [Internet]. 2013 Apr 24;52(5):411–21. Available from: http://dx.doi.org/10.3414/ME12-01-0101

159. Patrick J, Asgari P, Li M, Nguyen D. Using NLP to identify cancer cases in imaging reports

drawn from radiology information systems. Stud Health Technol Inform [Internet]. 2013;188:91–4. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23823294

160. Nguyen A, Moore J, Zuccon G, Lawley M, Colquist S. Classification of pathology reports for cancer registry notifications. Stud Health Technol Inform [Internet]. 2012;178:150–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22797034

161. Nguyen DHM, Patrick JD. Supervised machine learning and active learning in classification of radiology reports. J Am Med Inform Assoc [Internet]. 2014 Sep;21(5):893–901. Available from: http://dx.doi.org/10.1136/amiajnl-2013-002516

162. Osborne JD, Wyatt M, Westfall AO, Willig J, Bethard S, Gordon G. Efficient identification of nationally mandated reportable cancer cases using natural language processing and machine learning. J Am Med Inform Assoc [Internet]. 2016 Mar 28; Available from: http://dx.doi.org/10.1093/jamia/ocw006

163. McCowan IA, Moore DC, Nguyen AN, Bowman RV, Clarke BE, Duhig EE, et al. Collection of cancer stage data by classifying free-text medical reports. J Am Med Inform Assoc [Internet]. 2007 Nov;14(6):736–45. Available from: http://dx.doi.org/10.1197/jamia.M2130

164. Nguyen AN, Lawley MJ, Hansen DP, Bowman RV, Clarke BE, Duhig EE, et al. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. J Am Med Inform Assoc [Internet]. 2010 Jul;17(4):440–5. Available from: http://dx.doi.org/10.1136/jamia.2010.003707

165. Cunningham H, Maynard D, Bontcheva K, Tablan V. GATE: A framework and graphical development environment for robust NLP tools and applications, 2002. In: Proc 40th Anniversary Meeting of the Association for Computational Linguistics (ACL).

166. Kavuluru R, Hands I, Durbin EB, Witt L. Automatic Extraction of ICD-O-3 Primary Sites from Cancer Pathology Reports. AMIA Jt Summits Transl Sci Proc [Internet]. 2013 Mar 18;2013:112–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24303247

167. Liang Y-F, Chu P-Y, Chang C-S, Wang C-H, Chang P. Developing and evaluating a simple, spreadsheet-based pathology report extraction system for cancer registrars. AMIA Annu Symp Proc [Internet]. 2006;1008. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17238627

168. Luo Y, Sohani AR, Hochberg EP, Szolovits P. Automatic lymphoma classification with sentence subgraph mining from pathology reports. J Am Med Inform Assoc [Internet]. 2014 Sep;21(5):824–32. Available from: http://dx.doi.org/10.1136/amiajnl-2013-002443

169. Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. J Am Med Inform Assoc [Internet]. 2008 Jan;15(1):14–24. Available from: http://dx.doi.org/10.1197/jamia.M2408

170. Uzuner O. Recognizing obesity and comorbidities in sparse data. J Am Med Inform Assoc [Internet]. 2009 Jul;16(4):561–70. Available from: http://dx.doi.org/10.1197/jamia.M3115

171. Hassett MJ, Uno H, Cronin AM, Carroll NM, Hornbrook MC, Ritzwoller D. Detecting Lung and Colorectal Cancer Recurrence Using Structured Clinical/Administrative Data to Enable Outcomes Research and Population Health Management. Med Care [Internet]. 2015 Jul 29; Available from: http://dx.doi.org/10.1097/MLR.0000000000000404

172. Bikov KA, Mullins CD, Seal B, Onukwugha E, Hanna N. Algorithm for identifying chemotherapy/biological regimens for metastatic colon cancer in SEER-Medicare. Med Care [Internet]. 2015 Aug;53(8):e58–64. Available from: http://dx.doi.org/10.1097/MLR.0b013e31828fad9f

173. Warren JL, Yabroff KR. Challenges and opportunities in measuring cancer recurrence in the United States. J Natl Cancer Inst [Internet]. 2015 Aug;107(8). Available from: http://dx.doi.org/10.1093/jnci/djv134

174. Pezzi CM. Big data and clinical research in oncology: the good, the bad, the challenges, and the opportunities. Ann Surg Oncol [Internet]. 2014 May;21(5):1506–7. Available from: http://dx.doi.org/10.1245/s10434-014-3519-7

175. In H, Bilimoria KY, Stewart AK, Wroblewski KE, Posner MC, Talamonti MS, et al. Cancer recurrence: an important but missing variable in national cancer registries. Ann Surg Oncol [Internet]. 2014 May;21(5):1520–9. Available from: http://dx.doi.org/10.1245/s10434-014-3516-x

176. Ashley L, Jones H, Forman D, Newsham A, Brown J, Downing A, et al. Feasibility test of a UK-scalable electronic system for regular collection of patient-reported outcome measures and linkage with clinical cancer registry data: the electronic Patient-reported Outcomes from Cancer Survivors (ePOCS) system. BMC Med Inform Decis Mak [Internet]. 2011 Oct 26;11:66. Available from: http://dx.doi.org/10.1186/1472-6947-11-66

177. Ashley L, Jones H, Thomas J, Forman D, Newsham A, Morris E, et al. Integrating cancer survivors' experiences into UK cancer registries: design and development of the ePOCS system (electronic Patient-reported Outcomes from Cancer Survivors). Br J Cancer [Internet]. 2011 Nov 8;105 Suppl 1:S74–81. Available from: http://dx.doi.org/10.1038/bjc.2011.424

178. Ashley L, Jones H, Thomas J, Newsham A, Downing A, Morris E, et al. Integrating patient reported outcomes with clinical cancer registry data: a feasibility study of the electronic Patient-Reported Outcomes From Cancer Survivors (ePOCS) system. J Med Internet Res [Internet]. 2013 Oct 25;15(10):e230. Available from: http://dx.doi.org/10.2196/jmir.2764

179. Weber SC, Seto T, Olson C, Kenkare P, Kurian AW, Das AK. Oncoshare: lessons learned from building an integrated multi-institutional database for comparative effectiveness research. AMIA Annu Symp Proc [Internet]. 2012 Nov 3;2012:970–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23304372

180. ORIEN | Oncology Research Information Exchange Network [Internet]. [cited 2016 Feb 27]. Available from: http://www.oriencancer.org/

181. Hornbrook MC, Hart G, Ellis JL, Bachman DJ, Ansell G, Greene SM, et al. Building a virtual cancer research organization. J Natl Cancer Inst Monogr [Internet]. 2005;(35):12–25. Available from: http://dx.doi.org/10.1093/jncimonographs/lgi033

182. Dhir R, Patel AA, Winters S, Bisceglia M, Swanson D, Aamodt R, et al. A multidisciplinary approach to honest broker services for tissue banks and clinical data: a pragmatic and practical model. Cancer [Internet]. 2008 Oct 1;113(7):1705–15. Available from: http://dx.doi.org/10.1002/cncr.23768

183. Hurdle JF, Haroldsen SC, Hammer A, Spigle C, Fraser AM, Mineau GP, et al. Identifying clinical/translational research cohorts: ascertainment via querying an integrated multi-source

database. J Am Med Inform Assoc [Internet]. 2013 Jan 1;20(1):164–71. Available from: http://dx.doi.org/10.1136/amiajnl-2012-001050

184. Houser A, Curran D, Spadt V. Using the cancer registry to meet the Commission on Cancer clinical trials accrual standard. J Registry Manag [Internet]. 2013 Winter;40(4):188–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24625773

185. C/NET Solutions - Your Total Cancer Registry Solution [Internet]. C/NET Solutions. [cited 2016 Feb 28]. Available from: http://www.askcnet.org/

186. Levin G, Scharber W, Herna M, Stearns P, Peace S. Automated Tumor Consolidation: The Florida Algorithm. In: Proceedings, North American Association of Central Cancer Registries Annual Conference.

187. Zhang X, Kahn AR, Boscoe FP, Buckley PM. An automated algorithm for consolidating dates of diagnosis from multiple sources. J Registry Manag [Internet]. 2013 Spring;40(1):36–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23778696

188. Shiki N, Ohno Y, Fujii A, Murata T, Matsumura Y. Unified Modeling Language (UML) for hospital-based cancer registration processes. Asian Pac J Cancer Prev [Internet]. 2008 Oct;9(4):789–96. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19256778

189. Contiero P, Tittarelli A, Maghini A, Fabiano S, Frassoldi E, Costa E, et al. Comparison with manual registration reveals satisfactory completeness and efficiency of a computerized cancer registration system. J Biomed Inform [Internet]. 2008 Feb;41(1):24–32. Available from: http://dx.doi.org/10.1016/j.jbi.2007.03.003

190. Schwartz K, Beebani G, Sedki M, Tahhan M, Ruterbusch JJ. Enhancement and validation of an Arab surname database. J Registry Manag [Internet]. 2013 Winter;40(4):176–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24625771

191. Hsieh M-C, Pareti LA, Chen VW. Using NAPIIA to improve the accuracy of Asian race codes in registry data. J Registry Manag [Internet]. 2011 Winter;38(4):190–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23270092

192. Boscoe FP, Schymura MJ, Zhang X, Kramer RA. Heuristic algorithms for assigning Hispanic ethnicity. PLoS One [Internet]. 2013 Feb 6;8(2):e55689. Available from: http://dx.doi.org/10.1371/journal.pone.0055689

193. Maringe C, Li R, Mangtani P, Coleman MP, Rachet B. Cancer survival differences between South Asians and non-South Asians of England in 1986-2004, accounting for age at diagnosis and deprivation. Br J Cancer [Internet]. 2015 Jun 30;113(1):173–81. Available from: http://dx.doi.org/10.1038/bjc.2015.182

194. Howe HL, Lake AJ, Shen T. Method to assess identifiability in electronic data files. Am J Epidemiol [Internet]. 2007 Mar 1;165(5):597–601. Available from: http://dx.doi.org/10.1093/aje/kwk049

195. Andersen MR, Storm HH, Eurocourse Work Package 2 Group. Cancer registration, public health and the reform of the European data protection framework: Abandoning or improving European public health research? Eur J Cancer [Internet]. 2015 Jun;51(9):1028–38. Available from: http://dx.doi.org/10.1016/j.ejca.2013.09.005

196. Hakulinen T, Arbyn M, Brewster DH, Coebergh JW, Coleman MP, Crocetti E, et al. Harmonization may be counterproductive--at least for parts of Europe where public health research operates effectively. Eur J Public Health [Internet]. 2011 Dec;21(6):686–7. Available from: http://dx.doi.org/10.1093/eurpub/ckr149

197. Kerr DJ. Policy: EU data protection regulation--harming cancer research. Nat Rev Clin Oncol [Internet]. 2014 Oct;11(10):563–4. Available from: http://dx.doi.org/10.1038/nrclinonc.2014.148

198. Casali PG, European Society for Medical Oncology (ESMO) Switzerland. Risks of the new EU Data Protection Regulation: an ESMO position paper endorsed by the European oncology community. Ann Oncol [Internet]. 2014 Aug;25(8):1458–61. Available from: http://dx.doi.org/10.1093/annonc/mdu218

199. Rahu M, McKee M. Epidemiological research labelled as a violation of privacy: the case of Estonia. Int J Epidemiol [Internet]. 2008 Jun;37(3):678–82. Available from: http://dx.doi.org/10.1093/ije/dyn022

200. New Jersey State Cancer Registry: Implementing CDC's Registry Plus™ Web Plus for Ambulatory Centers and Physicians' Offices. J Registry Manag [Internet]. 2015 Spring;42(1):29. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26625481

201. Lusky K. Pilot points way to speedier cancer surveillance. CAP Today [Internet]. 2005 Feb;19(2):5–6, 8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15787106

202. Bjugn R, Casati B, Norstein J. Structured electronic template for histopathology reports on colorectal carcinomas: a joint project by the Cancer Registry of Norway and the Norwegian Society for Pathology. Hum Pathol [Internet]. 2008 Mar;39(3):359–67. Available from: http://dx.doi.org/10.1016/j.humpath.2007.06.019

203. Esteban-Gil A, Fernández-Breis JT, Boeker M. Analysis and Visualization of Disease Courses in a Semantic Enabled Cancer Registry. SWAT4LS [Internet]. Citeseer; 2014; Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.664.5755&rep=rep1&type=pdf

204. Rashbass J, Peake M. The evolution of cancer registration. Eur J Cancer Care [Internet]. 2014 Nov;23(6):757–9. Available from: http://dx.doi.org/10.1111/ecc.12259

205. Kulhawick H. A Case for Cancer Registry Automation [Internet]. [cited 2016 Feb 28]. Available from: http://health-information.advanceweb.com/Columns/Registry-Perspectives/A-Case-for-Cancer-Registry-Automation.aspx

206. National Cancer Registrars Association. Medical Informatics Basics for Cancer Registry [Internet]. 2008 [cited 2016 Feb 28]. Available from: http://www.ncraeducationfoundation.org/pdfs/NCRAInformaticsBrochure.pdf

207. Pritchard CC, Salipante SJ, Koehler K, Smith C, Scroggins S, Wood B, et al. Validation and implementation of targeted capture and sequencing for the detection of actionable mutation, copy number variation, and gene rearrangement in clinical cancer specimens. J Mol Diagn [Internet]. 2014 Jan;16(1):56–67. Available from: http://dx.doi.org/10.1016/j.jmoldx.2013.08.004

Appendix A: Sites Visited

Site	Date	Contacts	CC	CTSA	NLM	Caisis	SPORE	EMR
City of Hope	8/27/2008	21	Y	Y		Y	1	Eclipsys
UCLA	8/28/2008	1	Y	Y	Y		2	
UCSD	9/2/2008	1	Y	Y				Epic
USC	9/5/2008	2	Y	Y				McKesson
UC-Irvine	9/8/2008	6	Y	Y	Y	Y		
St. Jude	9/10/2008	2	Y					
Univ Kansas	9/29/2008	14		Y		Y		Epic, Cerner, McKesson
UAB	10/6/2008	10	Y	Y		Y	4	Cerner, McKesson
MSKCC	10/14/2008		Y			Y	1	Eclipsys
Stanford	10/21/2008	10	Y	Y	Y			Epic, Cerner
UCSF	10/22/2008	4	Y	Y			3	Epic
UC-Davis	10/24/2008	4	Y	Y		Y		Epic
MDACC	10/27/2008	7	Y			Y	9	Homegrown
Univ Washington	11/3/2008	13	Y	Y	Y	Y	2	Cerner
Penn State	11/13/2008	2		Y				
GWU	11/21/2008	6				Y		Allscripts
EVMS	11/24/2008	6				Y		Epic, Cerner
Univ Virginia	11/25/2008	13	Y	Y	Y	Y		Epic, GE
Baylor	12/2/2008	10	Y			Y	3	GE
Univ Arizona	12/8/2008	7	Y	Y			1	Allscripts,Eclipsys
OHSU	1/6/2009	18	Y	Y	Y	Y		
Burnham	1/15/2009	1	Y					Not applicable
Salk	1/15/2009	2	Y					Not applicable
ASU	1/20/2009	1		Y				
Univ New Mex	1/22/2009	3	Y	Y				Cerner
UTHSC-SA	1/26/2009	3	Y					Epic
Vanderbilt	2/2/2009	16	Y	Y	Y			Homegrown
Univ Florida	2/9/2009	4		Y		Y		Epic
Moffitt	2/10/2009	3	Y	Y				
MSMC	2/11/2009	3				Y		
Cedars Sinai	2/27/2009	10				Y		Epic
Wake Forest	3/9/2009	11	Y					GE
Duke	3/10/2009	12	Y	Y			2	Cerner
UNC	3/11/2009	2	Y	Y		Y	2	
Hopkins	3/17/2009	15	Y	Y	Y		7	Eclipsys, GE

Site	Date	Contacts	CC	CTSA	NLM	Caisis	SPORE	EMR
Georgetown	3/19/2009	2	Y	Y				GE
Univ Maryland	3/23/2009	1	Y	Y				Epic, Cerner, GE
Univ Penn	3/24/2009	2	Y	Y		Y		Epic
Jefferson	3/25/2009	1	Y					Allscripts, Cerner, GE
Wistar	3/25/2009	3	Y				1	Not applicable
Fox Chase	3/26/2009	19	Y			Y	1	
CINJ	3/30/2009	4	Y			Y		
Northwestern	4/21/2009	1	Y	Y			1	Epic, Cerner
Univ Chicago	4/22/2009	6	Y	Y		Y	1	Epic
Albert-Einstein	4/23/2009	2	Y					Epic, GE
Roswell Park	4/28/2009	5	Y					Eclipsys, Cerner
UPMC	4/29/2009	4	Y	Y	Y		2	Epic, Cerner
Cleveland Clinic	4/30/2009	11		Y				Epic
Case Western	4/30/2009	1	Y	Y		Y		
Univ Rochester	5/4/2009	14		Y		Y		Allscripts
Harvard	5/19/2009	7	Y	Y	Y		8	Homegrown
MIT	5/20/2009	1	Y					Not applicable
Dartmouth	5/26/2009	1	Y	Y				Homegrown
Maine	5/28/2009	2				Y		Not applicable
Jackson Lab	5/29/2009	12	Y					Not applicable
Mayo Clinic	6/11/2009	18	Y	Y			4	Cerner,GE
Wash Univ	6/16/2009	9	Y	Y				Allscripts
Karmanos	6/18/2009	9	Y			Y		Eclipsys
Univ Colorado	6/25/2009	3	Y	Y	Y			Epic
Univ Utah	6/29/2009	7	Y	Y	Y			Epic, Cerner, GE

Vita

Paul Fearn is a Director of Biomedical Informatics at Fred Hutchinson Cancer Research Center. He is focused primarily on the Hutch Integrated Data Repository and Archive (HIDRA) project for the Fred Hutch/UW Cancer Consortium, advancing applications of natural language processing, and big data cancer registry informatics initiatives with the National Cancer Institute Surveillance Research (SEER) program. Previously, he was the Informatics Manager for the Department of Surgery and the Office of Strategic Planning and Innovation at Memorial Sloan-Kettering Cancer Center (MSKCC), where he initiated and led the Caisis project, an open-source system that is currently used at multiple centers. Paul has a BA in Spanish from the University of Houston, biostatistics training from the University of Texas School of Public Health in Houston, an MBA from the New York University Stern School of Business, and is a PhD Candidate at University of Washington Department of Biomedical Informatics and Medical Education. He has more than 20 years of experience in cancer research informatics at Baylor College of Medicine, MSKCC and Fred Hutch.