Title
The Health Research Exchange: A collaborative model for improving participation in health research

Permalink
https://escholarship.org/uc/item/13w3b315

Author
O’Connor, Laurie Jean

Publication Date
2013

Peer reviewed|Thesis/dissertation
The Health Research Exchange:
A Collaborative Model for
Improving Participation in Health Research

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Computer Science

by

Laurie Jean O’Connor

2013
© Copyright by

Laurie Jean O’Connor

2013
ABSTRACT OF THE DISSERTATION

The Health Research Exchange:
A Collaborative Model for Improving Participation in Health Research

by

Laurie Jean O’Connor

Doctor of Philosophy in Computer Science

University of California, Los Angeles, 2013

Professor D. Stott Parker, Chair

Advancement in medical therapies requires clinical research to test a therapy’s efficacy and safety in humans, yet this process can be slow, inefficient, and error prone. Two key problems resulting in long delays are: 1) the translation of research ideas into relevant, testable hypotheses, and 2) the enrollment of subjects in clinical research that test these hypotheses. A crucial need for computational support is indicated by several challenges – the complex, multi-level nature of the medical field, the large number of possible health hypotheses, and the variety, and size of clinical research data.

Strong social and economic forces are now changing the way healthcare works. Clinical research is evolving to emphasize increased efficiencies, productivity, and inclusion of health-consumer-centered outcomes. In this period of change, this research seeks to address these challenges through the design of a Health Research Exchange (HRE) that is a development and communication hub for health research hypotheses. The HRE improves upon existing health research databases, such as clinicaltrials.gov, in that it addresses the process of hypothesis
formation. It seeks to facilitate health hypothesis development for non-medical professionals, enabling health research to become more relevant to health consumers.

The HRE is centered on the innovative Molecular Hypothesis Model (MoHM), featuring structured representation and reasoning about health hypotheses. The MoHM – based upon the principles of connectedness, modularity, and functional specificity – decomposes hypotheses into reusable, function-specific domains connected to each other and to facts, relationships, and other hypotheses. Implemented in a Hybrid Hypergraph Description Logic (HDL), integrating SNOMED-CT medical terminology with hypergraph and Boolean formula representations, the MoHM overlays functional domains – specialized for efficient, scalable reasoning – with hypergraph connectivity of concepts.

The Molecular Hypothesis Model provides the foundation for three essential HRE capabilities: 1) an incremental hypothesis development methodology enabling the reuse and recombination of hypothesis domain elements to form new hypotheses; 2) a hypothesis management capability, leveraging MoHM connectivity and modularity to provide support for construction and querying of hypotheses, and 3) a hypothesis collaboration protocol enabling progressive engagement in health research collaboration, and efficient matching of health profiles to health research. These innovations are illustrated in the context of several health research examples.
The dissertation of Laurie Jean O’Connor is approved.

Robert M. Bilder

Junghoo (John) Cho

Wesley W. Chu

D. Stott Parker, Committee Chair

University of California, Los Angeles
2013
DEDICATION

This dissertation is dedicated to my loving husband, Donald,
whose support, encouragement, and infinite patience made this possible,
to my tenacious father, Laurence,
who instilled in me at an early age a love for mathematics and constructing things,
and
to my brilliant mother, Carol,
who always valued and exemplified life-long learning.
| CHAPTER 1 | Introduction .................................................................................................................. 1 |
| 1.1 Motivation .............................................................................................................. 2 |
| 1.2 Contributions ........................................................................................................ 6 |
| 1.3 Roadmap .................................................................................................................. 7 |
| CHAPTER 2 | Foundations .............................................................................................................. 10 |
| 2.1 Clinical Research Background ............................................................................... 10 |
| 2.2 Medical Informatics Background .......................................................................... 23 |
| 2.3 Query support for medical systems ....................................................................... 32 |
| CHAPTER 3 | The Health Research Exchange (HRE) ........................................................................ 38 |
| 3.1 Stakeholders and Requirements ............................................................................ 38 |
| 3.2 Current best practice - NIH database .................................................................. 46 |
| 3.3 The HRE system architecture ............................................................................... 53 |
| 3.4 HRE knowledge representation foundations ............................................................ 59 |
| 3.5 HRE knowledge representation approach - HDL ...................................................... 67 |
| 3.6 Conclusions ............................................................................................................ 75 |
| CHAPTER 4 | Defining Health Hypotheses .................................................................................... 77 |
| 4.1 Hypothesis models ............................................................................................... 78 |
| 4.2 The HRE Molecular Hypothesis Model (MoHM) ....................................................... 91 |
| 4.3 Representing Health Hypotheses with Domains ....................................................... 103 |
| 4.4 Summary and Conclusions ................................................................................... 122 |
| CHAPTER 5 | Leveraging Health Hypotheses ................................................................................... 124 |
5.1 Hypothesis Formation with Case-Network Domain Reasoning ........................................... 126
5.2 Managing Health Hypotheses with HyRE ........................................................................ 140
5.3 Active PHRs with Personal Health Hypotheses ................................................................. 146
5.4 Contributions .................................................................................................................... 153

CHAPTER 6 Health research collaboration in the HRE .............................................................. 157

6.1 Models for Collaboration .................................................................................................... 158
6.2 HRE Hypothesize/Engage for Collaboration .................................................................... 163
6.3 Design for Scale – From MoHM to OBDD's ...................................................................... 171
6.4 Conclusion ........................................................................................................................ 179

CHAPTER 7 Summary and Contributions ................................................................................ 181

7.1 Challenges .......................................................................................................................... 182
7.2 Contributions ...................................................................................................................... 184
7.3 Looking Back ...................................................................................................................... 189
7.4 Going Forward .................................................................................................................... 190
7.5 Wrap-up .............................................................................................................................. 192

CHAPTER 8 References ......................................................................................................... 194
LIST OF FIGURES

Figure 1. Stakeholder Interactions................................................................. 43
Figure 2. The HRE Architecture...................................................................... 54
Figure 3. HRE HYPEG Collaboration Protocol .............................................. 58
Figure 4. Example Hypothesis Formation Process ........................................ 85
Figure 5. Conceptual structure and content of the MoHM Domains................ 102
Figure 6. Mapping Hypothesis Statements to Predictions.............................. 106
Figure 7. Plausibility requirement for hypotheses........................................... 112
Figure 8. Hypothesis Design Pattern: Parameterized Difference Hypothesis .......... 135
Figure 9. The active Personal Health Record (aPHR)................................... 149
Figure 10. Collaboration Model Diagram...................................................... 166
LIST OF TABLES

Table 1. Some commercial EHR systems and capabilities .......................................................... 25
Table 2. Example PHR System .................................................................................................. 28
Table 3. Interoperability organizations, standards, and architectures ..................................... 29
Table 4. Medical content coding standards .............................................................................. 31
Table 5. HRE stakeholders – people and organizations............................................................. 41
Table 6. HRE stakeholders – systems ...................................................................................... 42
Table 7. Use of sentence templates to construct hypotheses from the NIH database ............. 49
Table 8. Examples of correspondence between DL and FOL [Grosef] ................................ 65
Table 9. Hypothesis State scoring approach ............................................................................ 104
Table 10. Project Domain information content ....................................................................... 105
Table 11. Example H-Statements and Predictions .................................................................. 108
Table 12. Prediction Domain information content .................................................................... 111
Table 13. Example Plausibility Domain (Basis) Elements ....................................................... 113
Table 14. Example population Domain operations .................................................................. 116
Table 15. Test Domain operations .......................................................................................... 119
Table 16. Example analysis rules ............................................................................................ 121
Table 17. Falsifiability Domain operations .............................................................................. 122
Table 18. HRE Hypothesis problem and solution features ....................................................... 129
Table 19. MoHM Operations .................................................................................................. 141
Table 20. WMR and PMR concept comparison ....................................................................... 142
Table 21. Hypothesis query ..................................................................................................... 167
ACKNOWLEDGEMENTS

Many people have contributed to this dissertation. I am grateful to all who have inspired it, who have made it possible, and who have made it even better than I ever could have by working alone.

My deepest gratitude is to my advisor, Dr. D. Stott Parker. Few people are as fortunate to have an advisor so kind and sincere, and whose technical interests so closely align. When I first began working with Stott, he told me that he believes that an advisor’s primary role is as a cheerleader and coach, to provide encouragement as well as advice. That has truly been the reality. The encouragement has been offered when needed most; the coaching is not limited to technical matters, but extends to philosophical approaches to research, writing, and working with people. Stott offers superb advice; two pieces of advice that he did not let me forget are: 1) The work must be of high quality, and 2) It should be fun. Stott’s interest and current work in Hypotheses and the Hypothesis Web provided the foundation for this dissertation. Arising from that foundation, the idea of a health hypothesis as a data structure for collaboration bridged the seemingly deep chasm between my interest in clinical research and my desire to incorporate that into a Computer Science dissertation.

My committee members provided helpful critique, insightful questions and a variety of perspectives that contributed in many ways to this dissertation. Dr. Wesley Chu’s prior work in the area of medical knowledge bases, and current work in the CNP Hypothesis Web project greatly influenced my technical approach. Dr. Junghoo (John) Cho’s questions and helpful criticisms were thought provoking, leading me to modify some aspects of my design. I am grateful to Dr. Robert M. Bilder for including me in many of the activities and meetings for the Consortium for Neuropsychiatric Phenomics (CNP) LA5C project. These interactions and Bob’s leadership provided insights into the needs and challenges faced by clinical researchers working in the same medical domain with shared data, but many different hypotheses and research goals.

I would like to acknowledge Dr. Adnan Darwiche, advisor during my Masters degree program, whose work in Knowledge Compilation influenced my approach of addressing performance as an element of the system design. I would like to recognize professors, Dr. Richard E. Korf, Dr. Peter Reiner, Dr. Alfonso F. Cardenas, and Dr. Carlo Zaniolo, for their excellent courses that provided some of the background necessary to begin this work.

I would also like to thank Verra Morgan, who offered continual encouragement and regularly “pinged” me to finish this endeavor. Until her retirement, Verra advised me on how to approach the people-related challenges of the program and helped me over the administrative bumps along the way.

I am grateful to the UCLA Computer Science Department professors and staff for their
consistently excellent academic program. In addition to interesting technical content, I found that Computer Science coursework offers key life lessons and valuable practical advice, if one takes the time to read between the lines of code. My four favorite Life Lessons in Computer Science are (paraphrased). 1) If you look for problems, you will find them. If you don’t look for problems, they will find you (Computer Security, Dr. Peter Reiher). 2) The most effective way to approach a problem is that it is like another problem that has been solved or is not yet solvable (Algorithms and Complexity, Dr. Majid Sarrefzadeh). 3) When solving urgent problems, there is often a time tradeoff between planning and execution. That is provable (Problem Solving and Search, Dr. Richard Korf). And finally, 4) There is probably not enough time or space to answer all of the questions one would like to ask (Automated Reasoning, Dr. Adnan Darwiche).

Many thanks to Dr. Douglas R. Partridge, who introduced me to the field of Computer Science, hiring me as a Knowledge Engineer at Hughes Research Laboratories, and ultimately convincing me that Computer Science is much more than just programming. Doug encouraged me to take Computer Science coursework to augment my Chemical Engineering background, setting in motion the long chain of events that led to this dissertation.

I am deeply indebted to my first clinical research mentors, Hassan Movahhed, my manager, and Dr. David Green, a colleague at Amgen, who together introduced me to the fascinating world of clinical research. Hassan’s passion for improving clinical trial enrollment partially inspired the topic for this dissertation. David provided insights into the challenges of working with hundreds of North American clinical investigators with diverse medical specialties, research agendas and needs. My colleagues’ continuous requirements for timely, detailed, and comprehensive clinical operations data challenged me to thoroughly understand the costs, drivers and potential roadblocks of site initiation and patient enrollment.

I would also like to acknowledge the many Amgen study managers who participated in meetings discussing the unanticipated challenges of clinical trial enrollment, and the clinical research M.D.’s who shared with us their knowledge of patient populations, treatment patterns and patients’ practical constraints.

I am grateful to the UCLA Master’s in Clinical Research Program leadership, especially Dr. Joy Frank and Dr. Robert M. Elashoff, for allowing me to participate in their program, which provided in-depth training and hands-on experience in clinical research design and data analysis, as well as realistic insights into the challenges and ethics of designing high quality clinical research.

Finally, I would like to thank the many friends, relatives, and colleagues, throughout the years, who have expressed their desire and need for better ways to engage in learning about their medical conditions, and the options available to them, including emerging medical research. Without your thoughtful discussion and analysis this dissertation could not have been written.
1980  Bachelor of Science, Chemical Engineering  
University of Arizona, Tucson, Arizona

1980-81  Process Engineer  
Manufacturing data analysis, process analysis and improvement  
Procter & Gamble

1981-85  Engineer. Real-time, hardware-in-the-loop, dynamic systems simulation  
Rocket engine and aircraft control systems, heat transfer, and fluid flow  
Rocketdyne Division, Rockwell International

1985-88  Member of the Technical Staff  
Knowledge-based diagnostic and planning systems  
Artificial Intelligence Center, Hughes Research Laboratories

1989-99  Member of the Technical Staff and Manager, Information Technology  
Knowledge-based design and concurrent engineering systems  
Rockwell Science Center, Rockwell International

1993-98  Worldwide Project Manager, Product Data Management  
Engineering product data management and workflow systems  
Meritor Automotive, Europe, U.S., Canada, Brazil, Australia

1999-2003  Project & Program Management, Senior Manager  
HR, Finance, and Commercialization Information Systems  
Research Information and Automation Technologies  
Amgen, Inc.

2004  Master of Science, Computer Science  
University of California, Los Angeles

2003-09  Director/Executive Director, Clinical Research Services  
Clinical operations and study-start-up  
Clinical Trial Management Systems  
Amgen, Inc.

2004--08  Clinical Research Methodology Curriculum (9 courses)  
(concurrent with Computer Science Ph.D. program)  
David Geffen School of Medicine, UCLA


O’Connor, L., W. Pardee, SeeQFD specification and technical liaison for commercialization by Digital Equipment Corporation.


Knowledge Management Session moderator, Meritor Automotive Global Conference, 1996.


CHAPTER 1  Introduction

Healthcare costs are increasing due to many factors. One factor that is often overlooked is the true cost of advancing the practice of medicine, addressing unmet medical needs through new therapies. Phenomenal amounts of money are spent each year for clinical research, reflected back to the consumer in the high cost of medical drugs, tests and procedures. For example nearly twenty five billion dollars was spent to fund the clinical trials completed in the United States in 2010 alone [1]. On average, drug companies spend between $100 million and $800 million per drug candidate, most of which is spent in the clinical development phase. One major concern is the need for more emphasis on health-consumer-centered outcomes versus outcomes important to pharmaceutical company competitive strategies. The Patient-Centered Outcomes Research Institute (PCORI) has recently sponsored several initiatives addressing this need [2]. Also of concern is the slow pace of development of new therapies, which often take as long as ten years to twenty years from concept to market [3]. The clinical development phase during which drug candidates are tested in humans may range from a few months for vaccines, to five to seven years for cancer or chronic conditions such as osteoporosis. If health costs are to be contained, and ideally to be decreased, the cost, duration, and relevance of clinical research must be addressed. Consequently, there is an urgent need for increased efficiencies, productivity, and inclusion of health-consumer-centered outcomes in the field of clinical research. As in other fields, where productivity has been enhanced, information technologies are an enabling factor.

Many of the concepts and methods described in this dissertation, while inspired by clinical research needs, may also be applied to health research that is about people, but does not directly
require their participation. For that reason, the term *health research* may be used if a discussion is not specific to clinical research. This dissertation addresses an important aspect of health research – communicating, sharing, and managing the health hypotheses that are the subject of, and drivers of, health research.

### 1.1 Motivation

Advancement in medical therapies requires the implementation of clinical research studies, which test a therapy’s efficacy and safety in humans. Two problems resulting in rate-limiting delays are 1) the development of clear clinical research objectives, through the translation of research ideas into testable hypotheses, and 2) the timely recruitment and enrollment of subjects in clinical research that test these hypotheses. Related to these problems is the lack of opportunities for individual health consumers to provide input into the health research definition process, so that health research outcomes are relevant to health consumers. There are several, often interacting, reasons for these challenges.

*Lack of access to information.* A core problem is that individuals who might benefit from, or who wish to contribute to, clinical research are often not aware of ongoing research activities. Physicians, who typically refer patients to clinical trials, often are overloaded with patient workload, and may not be able to stay current with the most recent clinical research. Patients do not have full access to clinical research, and may not have the technical background to read and understand publically available resources that describe clinical research, such as [www.clinicaltrials.gov](http://www.clinicaltrials.gov) [4]. Clinical researchers, themselves, may be unaware of related ongoing research, even within their own organizations. In addition, clinical researchers have insufficient,
and often difficult, access to information about the population of patients that exists. Thus, physicians, patients, and researchers lack access to essential information.

**Lack of an adequate medium for communication.** The purpose of clinical research is to test specific hypotheses, with extensive inclusion and exclusion criteria defining eligible research participants. Hypothesis ideas may potentially originate from many sources including researchers, physicians, and patients. But hypothesis ideas may not gain traction or be tested if they cannot be clearly communicated. Hypotheses may change over time, often beginning as vague conjectures that mature to more precise, testable statements. However, the process of hypothesis development is relatively unstructured; there is no generally accepted hypothesis representation (other than natural language), no methods for hypothesis exchange, nor tools for managing a large body of related hypotheses and their associated information. There are no standard tools for linking health hypotheses to health data. The lack of structure and the lack of an adequate medium make it difficult to have efficient and meaningful communication and understanding between stakeholders in health research including researchers, physicians, and potential research participants.

**Inefficiency.** With inadequate access to information, and lack of a meaningful communication medium for research hypotheses, efficiency of the research endeavor suffers. Many clinical research efforts can be large and complex, utilizing hundreds, or even thousands, of clinical parameters, and many different types of information. The number of potential hypotheses to be explored is, consequently, extremely large. Researchers often overlap in their focus due to an inability to communicate and collaborate with regard to their research hypotheses. Enrollment of
research subjects may be slow and inefficient due to difficult assessment of hypothesis-related inclusion/exclusion criteria, and to a dearth of interested subjects. Without an efficient, reliable way to clearly, and broadly, communicate research intention, two or more ill-timed clinical research studies may even compete with each other for eligible subjects, slowing the pace of research in an entire patient population. Furthermore, collection of baseline data, such as medical and medication history, needed for determination of which subjects are eligible, may be error-prone, and labor intensive, due to the manual intervention and medical expertise needed to interpret medical files in the context of a clinical research hypothesis.

**Need for computational support.** Many of the challenges of clinical research described in the previous section can potentially be addressed with computational support. However, existing information tools and methods are not sufficient to address the shortfalls of **accessibility**, **understandability**, and **efficiency** in clinical research. The need for computational support for hypothesis-based clinical research arises from three key factors: **size**, **complexity**, and **change**.

**Size.** The size of clinical research endeavors can far exceed the ability of one individual to understand and manage. Many different areas of expertise are required, resulting in a plethora of stakeholders whose needs must be met, and who must efficiently communicate their different priorities and perspective, coming to agreement on key issues. The size of clinical research databases can be extremely large, with hundreds of variables or data items, and thousands of patients. Large clinical research endeavors may result in a very large number of potential hypotheses that must be efficiently managed and prioritized.
**Complexity.** The complexity of clinical research is due in part to the many levels of abstraction that may be spanned, from the patient symptom and behavior level, to the anatomical and physiological process level, to the molecular and genomic levels. Each of these levels has its own languages, representations, and types of data. Many of these data types, such as imaging data and genomic data, are themselves very complex, requiring computational support merely to interpret an individual data item, such as an image. Hypotheses that span multiple levels of domain abstraction must connect to many different types of data, and may implicitly participate in a complex web of domain knowledge and related hypotheses.

**Change.** Change is an important factor in clinical research. During a research study that may last several years, the medical status of participants will change; the population of subjects may change; even the clinical protocol that describes the clinical hypothesis may change. The practice of medicine is continually changing, with one common denominator in all subfields being the continued improvement in sensitivity and accuracy of medical testing, as well as the addition of new medical and genomic tests. In addition, researchers may make key discoveries that change a field’s understanding of fundamental processes. Changes in a medical area lead to changes in the information schema representing clinical data. The ever-present nature of change in clinical research results in a need to represent the evolution of hypotheses and their associated test and evaluation strategies.

**Need for a computational model.** While size, complexity, and change all point to the need for computational support for managing health hypotheses, there is a major problem. There is no computational model for representing, manipulating, and managing clinical or health hypotheses.
Some work has been done in specific domains to provide automatic support for reasoning about hypotheses, but there is no generally accepted model for hypotheses that provides the representation and reasoning support needed for clinical research.

Need for a collaboration model. In addition to, and partially because of, the lack of a computational model for clinical hypotheses, there is no general collaboration model, other than person-to-person discussions, that enables multiple people, including health consumers, to collaborate in the development and testing of a clinical hypothesis. A model for collaboration applied to hypotheses must provide for a way to suggest and exchange hypotheses, to elaborate on or specialize hypotheses, to measure the status or state of a hypothesis, to manage a set of hypotheses, to express interest in a hypothesis, and to exchange and evaluate specific data that is needed to determine whether a subject can be included in the population of a hypothesis.

1.2 Contributions

This research seeks to address some of these challenges through the design and development of a Health Research Exchange (HRE) that acts as a communication hub for health research hypotheses. The HRE is centered on the innovative Molecular Hypothesis Model (MoHM), which features structured representation and reasoning about health hypotheses. The MoHM – based upon the principles of connectedness, modularity, and functional specificity – decomposes hypotheses into reusable, function-specific domains that are connected to each other and to facts, relationships, and other hypotheses. Implemented in a Hybrid Hypergraph Description Logic (HDL), which integrates SNOMED-CT medical terminology with hypergraph and Boolean formula representations, the MoHM overlays functional domains – specialized for efficient, scalable reasoning – with hypergraph connectivity of concepts.
The MoHM provides the foundation for three important HRE capabilities: 1) an incremental hypothesis development model that enables the reuse and modular recombination of hypothesis domain elements to form new hypotheses; 2) a hypothesis management capability that leverages MoHM connectivity and modularity to provide support for construction, manipulation, and querying of hypotheses; and 3) a hypothesis collaboration protocol that facilitates health research information exchange and efficient matching of clinical research to health consumers. These innovations are illustrated and explored in the context of several health research examples.

1.3 Roadmap

Chapter 2, Foundations, provides a background in clinical research, then discusses Medical Informatics, including personal health records and medical terminology. It concludes with an overview of prior work in query support for health systems.

Chapter 3, The Health Research Exchange (HRE), begins by establishing the requirements for a collaborative development environment for health research. In the context of these requirements, this chapter evaluates the current state-of-the-art for clinical research repositories, clinicaltrials.gov, reporting on capabilities, gaps, and challenges. An overview of the HRE architecture is then provided, laying out the architectural components, their functions, and how they interact. Finally, a Hybrid Hypergraph Description Logic (HDL) representational formalism is introduced, which integrates Description Logics and the medical terminology language, SNOMED-CT, with hypergraphs and Boolean formulas. This enables the use of medical terminology within complex health hypotheses, while enabling connectivity between
concepts and reasoning, which is not possible in a traditional Description Logic.

**Chapter 4, Defining Health Hypotheses**, introduces the hypothesis as an incrementally developed, rich information structure, that requires a prediction, supported by a basis of prior findings and a method for falsifying the prediction. This structure meets the requirements for a scientific hypothesis, and opens new paths for describing relevant clinical research. A structured hypothesis representation, the Molecular Hypothesis Model (MoHM), implemented in HDL, is defined. The MoHM is specialized to a subset of hypotheses, the Population Health Hypothesis (PoH), and in particular, the Difference hypothesis.

**Chapter 5, Leveraging Health Hypotheses**, describes how the MoHM and libraries of existing hypotheses are leveraged to provide useful capabilities. Chapter 5 introduces the Case-Network Domain Reasoning (CaNDoR) methodology for constructing hypotheses by reusing components of existing, similar hypotheses. The notion of Hypothesis Design Patterns is introduced as a means to capture and reuse common patterns of hypothesis expression. Hypothesis management capabilities provided by the Hypothesis Reasoning Engine (HyRE), are then presented, including the transformation of hypotheses into sharable, transportable, XML-compatible data structures. Finally, the concept of Personal Health Hypotheses, and Health Condition Profiles are introduced as a means of integrating MoHM health hypotheses with active Personal Health Records.

**Chapter 6, Health Research Collaboration in the HRE**, demonstrates how the MoHM can be used to engage earlier and more constructively in clinical research. A Hypothesize/Engage (HyPEG) protocol for the exchange of personal and population hypotheses and Health Condition
Profile (HCP) information is introduced. In support of the communication protocol, this chapter describes how health consumer HCP’s can be efficiently matched to health hypothesis population criteria, through the transformation to and evaluation of ordered binary decision diagrams (OBDD’s).

*Chapter 7, Summary and Conclusions*, provides a synopsis of the research and its findings, as well as suggesting fertile avenues for future work.

*Chapter 8, References*, provides a comprehensive set of references, spanning hypothesis models, clinical and health informatics, as well as current systems and architectures for health information management.
CHAPTER 2  Foundations

This chapter provides a foundation for understanding key practices, challenges, and ongoing research in clinical research and health information management. Section 1 introduces clinical research practices and processes. Section 2 provides background material in medical informatics. Section 3 outlines prior work in query support for health systems.

2.1 Clinical Research Background

Clinical research is an exciting field. It offers hope for millions of people with currently unmet medical needs. However, the field of clinical research is fraught with challenges – with many logistical, ethical, financial, legal, informatics, and scientific hurdles. This section provides some background in clinical research, beginning with its definition, an overview of the NIH clinical research database, and a description of legal issues in clinical research. It then outlines some of the important processes in clinical research including development of a research protocol, the study-startup process, and the patient enrollment problem. Finally, it discusses some of the information management aspects of clinical research, specifically, clinical research management systems, data analysis and data mining.

2.1.1 What is clinical research?

Clinical research is a broad term that encompasses many types of research. The NIH definition of clinical research is that it is research conducted with human subjects or with materials of human origin, where an investigator interacts directly with human subjects [5]. Clinical research may include research into drug interventions, new technologies or devices, associations of health conditions with factors such as environment, lifestyle, or genetics, and mechanisms or progression of disease.
There are several different classifications of clinical research. Interventional Trials test an intervention, such as a drug or medical device, in a specific patient population to determine if it is effective. Randomized, interventional clinical trials are often referred to as the “gold standard”, where random assignment to treatment arms accounts for unknown factors that may influence the outcome. Observational Trials observe patients in their natural settings, seeking explainable differences in populations that have a disease condition or risk factors. Clinical research, especially observational research, may involve the development of large databases that support the investigation of many different clinical research questions.

Clinical trials are typically divided into Phases, based upon their function in the research process. Phase I clinical trials are designed to test safety in humans, and to determine initial dosing levels for later trials. Phase II trials provide a small-scale test of efficacy and provide further information on dosing and safety. Phase III trials are designed to determine efficacy of a treatment, and to understand any safety or other types of concerns. Trials in Phase IV are conducted after FDA approval of the therapeutic for various reasons, often to study long-term safety, and sometimes to extend the available knowledge to cover various subsets of the population, such as children or the elderly. A Phase 0 designation has been adopted by some researchers, referring to studies in with the purpose of determining normal disease progression characteristics, and distribution of population features. Clinical research may be sponsored or conducted by many different types of organizations, including medical institutions, pharmaceutical firms, and government agencies.
2.1.2 Patient-Centered Outcomes Research

Patient-Centered Outcomes are outcomes important to patients. Patient-Centered Outcomes Research is that research that seeks to provide evidence-based information that helps patients and health care providers to make more informed decisions. For example, a research endeavor that provides guidelines for diagnosing, predicting, and treating chronic pain may help patients and their caregivers make better decisions about dealing with pain, one that takes into account the impact on patients ability to function mentally, physically, and socially. This type of research – not directly supporting the agenda of a pharmaceutical company, or even the agenda of an organization focused on a single health condition – has been in the past largely unaddressed.

The Patient-Centered Outcomes Research Institute (PCORI) [2], an outcome of Section 6301 of the United States Patient Protection and Affordable Care Act, is a private, nonprofit organization, that in 2012 established and initiated funding for a national agenda for Patient-Centered Outcomes Research. In seeking to involve patients and their caregivers in defining a new research agenda, it gives hope for improved health care outcomes and new impetus for change.

2.1.3 NIH clinical trials database

The most comprehensive source of information for health research is the U.S. National Institute of Health (NIH) Clinical Trials Database (www.clinicaltrials.gov). In 2007, the NIH began requiring that all interventional research, or clinical trials, be entered in a common database. While there are other specialized clinical trials databases, such as that of National Cancer Institute [6] with over 19,000 clinical trials, all U.S. trials in human subjects must also be housed in the NIH database. Now, five years later, in 2012, with over 100,000 current entries for clinical research, the NIH database holds information for each human research project that has been designed, reviewed and approved [4].
The database, which is online and open to the public, offers querying capabilities. Users may search on 24 different fields. A UMLS-powered backend enables synonym matching, so, for example, a database query on a term “heart attack” may yield, as a result, studies that refer to “cardiac infarction”. Responses to queries are displayed on-line, and download options are available for 20 different fields in CSV, tab-delimited, or XML. Alternatively, an XML file detailing all elements of each study may be downloaded.

Medical information relevant to the trials is described, partially through structured database fields, and partially through free-form text fields. Example medical fields include: condition, intervention, criteria (inclusion and exclusion), and outcome measures. Several types of project-focused information are represented in each clinical project entry, including project background, project schedule, investigators and their organizations. Project-related data includes a large number of date fields, fields describing specific roles, such as investigators, and the status of the trial relative to subject recruiting. Rationale for the trials is not in a single location; it may be distributed over several different fields including background, and short and long titles. Some information related to results may be found for some studies, as well as information about the methods applied to analyze the results.

While the clinicaltrials.gov database is useful, it does have limitations. It does not hold information about health research that is in the formative state of development. The focus of the clinical trials database appears to be that of a program management tool. It is only suitable and targeted for projects that have already been approved. It cannot be used for those hypotheses
that are in earlier stages of development. A health consumer who would like to provide feedback or insights on a clinical research project has no avenue for expression, other than to contact a posted investigator directly. Trials that are not actively recruiting do not include information about clinical investigators or their contacts.

While the database is open to the public, it is not easy to understand what one finds through queries on the database, unless one has a background in medicine or clinical research. A person who is knowledgeable about a medical condition may be able, through a series of targeted searches, find research that is relevant. It is not easy to find trials for which one might be eligible, since the inclusion and exclusion criteria include much non-structured medical terminology. A person without much medical background may find the terminology and technical responses to be intimidating.

There are other examples of health hypothesis databases although none as large and broad as the NIH clinical trials database. The Cochrane database [7] features a wide range of health hypotheses that have undergone meta-analysis, combining the results of multiple studies. The Alzforum website [8] organizes hypotheses concerning Alzheimer’s disease as a web of claims supported by other claims, with links to supporting literature. The NCBI Gene database [9], represents an implicit set of hypotheses, linking genome and phenome information. However, use of each of these databases requires specialized knowledge, including specialized medical terminology or health conditions. Other than the clinical trials database, there is no general source of information addressing the many different types of conditions that the health consumer may encounter.
2.1.4 Clinical research and the law

In clinical research, there is great concern for the human element, and the risks inherent in a therapeutic being used in humans, prior to knowing the long and short-term ramifications. After World War II, through the Geneva Convention, a new set of guidelines was developed, outlining the basic human rights people have when involved in clinical research. These have evolved into the ICH-GCP (International Conference on Harmonization / World Health Organization Good Clinical Practice Guidelines, as well as the European Agency for the Evaluation of Medicinal Products (EMEA) guidelines [10]. In addition, in the United States, through a set of federal regulations known as HIPAA, there are some protections for privacy of a person’s medical records and information [11].

Ethical issues are a concern, and the basis for many controversial topics of discussion. One of the ethical guidelines is that patients should be fully informed of the clinical research, and that there should be no undue influence in obtaining their informed consent. Situations involving potential conflict of interest, in those conducting clinical trials, are closely monitored. There are also strict guidelines on language used, ensuring the informed consents are understandable at an average level of education, and are in the natural language appropriate to the population of subjects. A best practice, with respect to ethics, is to maintain a buffer between clinical researchers and the patient. This is not always possible, for example in smaller studies, or in cases where the individual’s physician is a co-investigator in clinical research. Because of the need to avoid conflict of interest, and to buffer patients from research sponsors, pharmaceutical companies generally do not interact directly with patients in clinical studies. Pharmaceutical companies generally interact with physicians who take on the role of investigator. Furthermore
pharmaceutical employees, due to HIPAA and other regulations, are not allowed to directly view patient records to determine eligibility for clinical trials. Third party organizations, that may not be as familiar with the subject of the clinical research, may sometimes be engaged to provide support to physicians in sifting through medical records to identify eligible subjects.

In order to conduct any type of clinical research, there are several types of approval that are necessary. An institutional review board (IRB) or Ethics Committee (EC) must review a research protocol that specifies the research to be done, and the specific criteria used to select the research subjects. In many cases the FDA is involved in approval of a research protocol. Since the year 2000, all clinical research investigation of serious, or life-threatening conditions, as well as any gene-transfer studies, must be entered in a FDA registry, clinicaltrials.gov. However, these are not entered until after the protocol is complete and approved by the sponsoring organization [12].

2.1.5 The research protocol and the clinical hypothesis

In order to gain approval for any type of clinical research, a detailed research protocol, a comprehensive document, is required. The research protocol should clearly state the hypotheses being tested, and the population to be tested. A well-written protocol will also provide sufficient background knowledge, and links to published research in the field, that provide the basis for the thinking that leads to the clinical research hypothesis. The research protocol is the binding document that indicates how a trial is to be conducted. It specifies which measures are to be made, and exactly how those measurements will be performed. Usually there is one primary outcome, or hypothesis, to be tested and analyzed, with a small number of secondary outcomes
stated. However, in some types of clinical research, such as Patient Registries, the protocol may specify that large volumes of a variety of data are collected for the purpose of data mining, looking for signals of associations that may later prove to be important for efficacy or safety.

Elements of the protocol include descriptions of the conditions to be treated, known treatments, mechanisms of actions, i.e., how a drug or device is supposed to work, a detailed specification of when therapeutic agents are to be administered, or tests to be performed, and what data is to be collected, in what form and when. A well-written protocol will also include details of the analysis approach to be utilized in analyzing the data. It estimates the proposed patient population size, that is calculated based upon how many subjects are needed to show a statistically significant difference, given assumptions about the natural variation and range of values for specific parameters in the selected population. While statistically significant difference is usually understood to be important, an even more important factor is clinically significant difference. For example, a statistically significant difference in blood pressure in a normal population may not be clinically significant, while it may be hugely significant for a population of cardiac arrest patients undergoing a new therapy. A trial designed to detect a large, but clinically insignificant, difference is not viewed as a good use of research resources.

It is not easy to succinctly and clearly state a research hypothesis and link it to measurable data that can be obtained from human subjects. Development of a good protocol requires incorporation of knowledge from many different fields, and often requires the interaction of many different types of experts over several months, including medical specialist M.D.’s, biostatisticians, health outcomes researchers, clinical trial operations experts, international law
experts, clinical data managers, quality management, and people in the field who monitor and interact with clinical sites. All of these roles must clearly understand the specific goals of the clinical research. Often, research protocols take months to be reviewed and approved by all the governing bodies, before research is begun. The worst outcome for a clinical trial is not the failure of an intervention or expectation, but the poor design of a protocol, leading to an outcome where human health has been put at risk, with no new knowledge gained.

2.1.6 The study start-up process

The study startup process refers to the time between when the protocol is approved by the governing bodies to the time subject enrollment begins. Each site involved in a clinical study goes through a study startup process. Sites are those locations where the subject will be treated and/or evaluated, such as a clinic, a hospital, or a center. In large clinical studies, there can be hundreds of sites and thousands of subjects. Before a site can begin enrolling patients, it must be approved by its IRB, and agreements made between the clinical trial sponsor and the research site about any monetary reimbursement there may be for expenses incurred specific to the trial. In addition, each site must provide information about potential financial conflicts of interest, and the background and credentials of the principal investigator and sub-investigators participating in the clinical research at that site. Each site must ensure that all supporting personnel are fully trained on the research protocol. During the study start-up process, a large amount of required documentation is compiled. When the study startup process at a site is complete, patient enrollment can begin.
2.1.7 The patient enrollment problem

Clinical trial enrollment is probably the single biggest contributor to delays in the clinical research process, and is one of its most significant challenges. The key challenges in patient enrollment are: 1) Evaluating the current patient population who meet enrollment criteria, as well as the expected incidence rate of new patients, 2) Assessing the patient referral flow – the flow of patients from general practitioners to specialists and through specialized testing centers, such as radiology, 3) Determining which specific patients are eligible, 4) Communicating the research to patients who may consider enrolling, 5) Collecting and assessing health data to determine eligibility, 6) Ensuring the patient has sufficient information and discussions with their physician to understand the impact of the research on the patient, 7) Obtaining the Informed Consent from the patient. Since research sponsors may not be contacting patients directly, intermediate companies are often contracted with to do advertising of clinical trials, or to assess patient populations. Sponsors of clinical trials commonly underestimate the time it will take to fully enroll a study. In addition to initial enrollment, another challenge related to enrollment is minimizing the dropout rate. Dropouts are due to a number of factors including adverse health effects, desire to try an alternative treatment, inconvenience of the assessment and visit schedule, and waning interest in the study.

A major problem, central to the patient enrollment problem, is that potential subjects often do not have access to information about what trials are being planned, so as to give an eligible person time to do the research needed to determine whether he or she wants to participate. In today’s healthcare environment, people may learn about clinical research from several sources – from newspaper or radio advertisements, from their physicians, by looking at pharmaceutical websites
by looking at the clinicaltrials.gov website, or by engaging with a specialized health research, for example, the Michael J. Fox Foundation for Parkinson’s Research [13]. However, the information from these sources is often not sufficient to make a truly informed decision about whether to enroll in a clinical trial. While physician investigators must have access to the research protocol, subjects often never see it. In addition, it may be the case that not all of the risks of the trial are easily identifiable. Taking the time to fully understand the risks involved in the trial of a specific new therapy is critically important, and not doing so can have serious repercussions for the subject. Even healthy patients involved in clinical trials may unknowingly be at risk for serious adverse events [14].

During the enrollment process, tests may be performed in order to determine health and eligibility status, and extensive baseline health data for each must be gathered. This information, required to assess enrollment eligibility, usually is manually compiled by research coordinators at the site. A time-consuming task is summarization of medication histories. The problem is that, except for laboratory data analyzed at large central laboratories, most of the clinical trial data is entered by hand. Traditionally, case report forms have been used to capture data, in triplicate (or more), with third-party contractors doing double data entry to ensure its accuracy. More recently, centrally managed Electronic Data Capture (EDC) systems have been used to capture data at the site, performed by the research coordinator, who transcribes the information from lab tests and medical records. Currently, electronic medical records are not translated directly to the clinical research data system, although this is clearly a gap, and an area receiving much attention from standards bodies and software vendors. There are currently no standards for personal health records compatibility with clinical research systems.
2.1.8 Clinical research management systems

Several types of Clinical Research Management Systems have evolved during the last decade, in response to a growing volume of clinical trials and associated clinical data. However, these are not well-integrated with each other, or even with electronic medical records.

Clinical Data Management (CDM) Systems are designed to hold the data obtained from clinical trials. Generally, a separate database is created for each clinical trial. Variants of this are databases that hold data from Patient Registries – groups of people who are registered as post-market users of specific therapies, or databases holding data from observational research. This data is typically entered, either through an Electronic Data Capture system, or as double data entry processes from carbon-copy Case Report Forms completed manually. Laboratory data from automated testing facilities may be directly transferred from testing facilities. While some industry clinical research organizations aspire to create data warehouses of clinical data, combining data from different clinical trials, this is a hard problem due to the practice of optimizing the data schema for individual clinical trials, as well as to the multiple measurements at different time points for a single indicator, that may vary by subject.

Clinical Trial Management Systems record data important for the administration and management of clinical trials. These contain information about patient visits, changes to the clinical protocol, cost payment structures, and site start-up information, focused upon study start-up requirements. There are some specialized systems for modeling and simulating patient enrollment in global clinical trials, but these do not directly address the enrollment problem, other than making problems much more apparent and trackable. Safety Monitoring and
Reporting Systems record and report on any adverse reports or safety issues.

While cut-and-paste methods with word processors may reduce the amount of time for development of a clinical protocol, these often lead to errors that may cause extensive rework and expense if not found early in the process of clinical trial design. Automated protocol writing tools have been developed and are becoming more widespread. These automated tools use structured text and database fields to organize information about a specific clinical protocol that is being developed. Some tools have been developed for use in a collaborative environment, with inputs from multiple people.

FDA systems exist that organize and manage information with clinical research. For example, the registration of clinical trials in ClinicalTrials.gov, is currently required, by law for all clinical research projects meeting certain criteria, are represented in this repository. This repository contains information for clinical research that has been approved, but not research that is in the process of being defined. In addition, there is a FDA problem-reporting database that enables reporting of concerns about therapies by health practitioners [15].

2.1.9 Data analysis and data mining

In clinical research, clinical outcomes are determined by statistical analyses appropriate to the type of data and the type of relationship being tested. The data analysis is performed according to the protocol, and usually by experts in biostatistics. Data mining methods are sometimes used post hoc to analyze clinical data, to search for signals that might inspire additional clinical studies with testable hypotheses. Even when data mining results show very strong results or connections, in the medical field, this is not sufficient to convince physicians to change the way
they practice medicine. The healthcare culture is such that the practice of medicine changes as a result of statistical hypothesis testing, not as a result of connections made in data mining.

2.2 Medical Informatics Background

Biomedical or Health Informatics is an interdisciplinary field seated in the intersection between computer science, biomedicine, and probability, statistics and decision sciences [16]. It deals with the resources, devices and methods required to optimize the acquisition, storage, retrieval, and use of information in health and biomedicine [17]. This field has evolved due to the specialized concepts, languages, and terms that are needed to represent, store and reason about biomedical information, to include medical record standards, information content coding, interoperability standards and architectures, query capabilities, security, personal digital assistants, wireless technologies, and many others.

2.2.1 Electronic Medical Records

In the United States, the health care infrastructure is diverse and fragmented across many different geographic locations and medical specialties. With the increasing proportion of medical information available electronically, it is becoming conceptually feasible to support an all-electronic medical record. An important technology is the Electronic Medical Record (EMR), also known as Electronic Health Record (EHR). The Healthcare Information and Management Systems Society defines EHR as “a secure, real-time, point-of-care, patient-centric information resource for clinicians for providing access to patient health record information when they need it, as well as evidence-based decision support [18]”. Another definition of EHRs is that of Health Level 7 which lists over 110 different functions that an EHR should have [19]. ISO defines an EHR system as a system for recording, retrieving, and manipulating
information in electronic health records [20]. EHR systems are generally aimed at the health care provider as the end user and can be integrated with other Health IT Systems. Some commercial EHR providers and their capabilities are listed in Table 1.

While some definitions refer to the EHR as patient-centric, it is generally designed as a resource that is owned and controlled by the service providing institution or physician’s office. Note the emphasis on the business functions such as billing. While patient medical records are a central element, most of these systems are designed with an eye on the business of managing healthcare. In fact, the primary justification for EHR systems in hospitals and other healthcare organizations is often that of decreasing the total cost of providing healthcare, through improved management of healthcare processes [21]. Since most EHR systems are “tethered” to the institution providing the services, a patient may not be able to take the entire electronic record with them when moving to another healthcare provider. Today’s EHRs are diverse, with a variety of architectures and information representations. Transferring the entire health record for an individual between health service providers, while conceptually feasible, is usually not practical or supported by today’s system.
### Table 1. Some commercial EHR systems and capabilities

<table>
<thead>
<tr>
<th>EHR Provider</th>
<th>Capabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allscripts</td>
<td>Solutions for physicians practices, hospital and care management, clinical education, and medication services</td>
</tr>
<tr>
<td><a href="https://www.allscripts.com">www.allscripts.com</a></td>
<td></td>
</tr>
<tr>
<td>eClinicalWorks</td>
<td>Front office management, electronic medical records, patient history, medications, labs, charts, claims submission, and portal.</td>
</tr>
<tr>
<td><a href="https://www.eclinicalworks.com">www.eclinicalworks.com</a></td>
<td></td>
</tr>
<tr>
<td>Epic Systems Corporation</td>
<td>Enterprise-wide systems for inpatient and ambulatory care, hospital and professional billing, shared, web-accessible medical records for patients and affiliates, reference lab orders and results, enterprise scheduling, bed utilization, patient registration</td>
</tr>
<tr>
<td><a href="https://www.epicsystems.com">www.epicsystems.com</a></td>
<td></td>
</tr>
<tr>
<td>GE Healthcare (Centricity)</td>
<td>Healthcare support systems including perioperative, acute care, laboratory, provider order entry, medication management, diagnostics, monitoring, and imaging and information systems</td>
</tr>
<tr>
<td><a href="https://www.gehealthcare.com">www.gehealthcare.com</a></td>
<td></td>
</tr>
<tr>
<td>MedComSys</td>
<td>Information systems framework and toolsets for acute/sub acute care, home health care, military care, and ambulatory care</td>
</tr>
<tr>
<td><a href="https://www.medcomsys.com">www.medcomsys.com</a></td>
<td></td>
</tr>
<tr>
<td>Medinformatix</td>
<td>A suite of tools including EMR, enterprise practice management, radiology information, and communications</td>
</tr>
<tr>
<td><a href="https://www.medinformatix.com">www.medinformatix.com</a></td>
<td></td>
</tr>
<tr>
<td>Microsoft Healthcare Solutions</td>
<td>System solutions for management of electronic medical records, HIPAA compliance, Medicaid claims, vital records, lab and mental health information, and disease tracking</td>
</tr>
<tr>
<td><a href="https://www.microsoft.com/health">www.microsoft.com/health</a></td>
<td></td>
</tr>
<tr>
<td>NextGen</td>
<td>Products include electronic medical record, electronic practice management, community health solutions for sharing patient data, image control system, patient portal</td>
</tr>
<tr>
<td><a href="https://www.nextgen.com">www.nextgen.com</a></td>
<td></td>
</tr>
<tr>
<td>Physician Micro Systems</td>
<td>Electronic health record, medical billing, and appointment scheduling software</td>
</tr>
<tr>
<td><a href="https://www.pmsi.com">www.pmsi.com</a></td>
<td></td>
</tr>
<tr>
<td>Synamed</td>
<td>Integrated practice management solution including scheduler, waiting room oversight, document and prescription management, diagnoses and medications, drug-drug and drug-allergy interaction warnings, patient demographics, digital signature</td>
</tr>
<tr>
<td><a href="https://www.synamed.com">www.synamed.com</a></td>
<td></td>
</tr>
</tbody>
</table>

#### 2.2.2 Personal Health Records

A Personal Health Record (PHR) is a kind of electronic medical record, but one that is centered on the needs of the person instead of on those of a particular institution. PHRs typically have the goal of integrating all health related information for that person, so that it can be used by patients and physicians to manage the care of the individual. There has been a surge of interest in
providing PHRs to the consumer, as evidenced by the recent implementation of HealthVault by Microsoft [23] and several examples of programs established by employers and insurance companies to provide on-line PHRs to individuals [24].

There are two major classifications of PHR. “Tethered” PHR systems, similar to EHR’s are linked to electronic health records that are maintained by a healthcare provider, such as a primary health care facility, a hospital, a health maintenance organization (HMO) such as Kaiser Permanente, an employer-based PHR service, or an insurance-paid PHR service. “Untethered” PHRs, intended to be independent of a specific health care provider, focus on collecting the entirety of a person’s health record in a single location, fully within the control of that individual or those with designated access. Simple, untethered PHRs may be isolated record systems or they may be systems that can download information from multiple sources. They may be hosted on personal computers, or they may be made available through internet-based service providers.

**Functions of PHRs.** Much attention has been paid to the functional requirements of a PHR. In Gearon’s Perspectives on the Future of Personal Health Records, the following requirements are proposed by Patricia Bently: [25]

- Information about a person’s health states, health practices, and use of health services
- Patient preferences for services, such as advanced directives
- Decision logics that initiate alerts, warnings, or recommendations
- An individual’s observations about his or her physical and social environments
- Rules regarding privacy and access to, and use of information
- Middleware tools that manage identity

The American Health Information Management Association (AMIA) has developed a set of basic principles for personal health records and a position statement that summarizes the basic
contents of a PHR. The AHIMA requirements include [26]:

- Personal identification, name and birth date
- People to contact in case of emergency; contact information for health care providers
- Health insurance information
- Living wills, advance directives, or medical power of attorney
- Organ donor authorization
- A list and dates of illnesses and surgeries
- Current medications and dosages
- Immunizations and their dates
- Allergies or sensitivities to drugs or materials
- Events, dates, and hereditary conditions
- Results from a recent physical examination
- Opinions of specialists
- Important tests results; eye and dental records
- Correspondence with health providers
- Current educational materials (or appropriate web links) relating to one’s health

**Ambulatory Care Functions.** In healthcare terminology, ambulatory care is any patient care that is delivered outside of the hospital setting. A large proportion of healthcare is performed away from a primary or centralized provider, so that even the most comprehensive tethered systems may not have all of the information needed for effective healthcare. Ambulatory care information that should be captured in a PHR are pharmacy records, patient-administered medications, dental treatments, home-monitoring data records, consulting visits to physicians, health-related events, such as falls or exposure to toxic chemicals, and extreme stress-inducing events. Requirements for Personal Health Records, including ambulatory care, have been developed by the Certification Commission of Healthcare Information Technology [27].

**Privacy and Security.** Privacy and security are a major challenge for consumers using a PHR. While on-line services may be convenient, many are not legally bound to protect consumer health information. HIPAA (Health Information Protection and Accountability Act), which
ensures privacy of health information held by health providers and health insurers, does not extend to other organizations, such as health information websites and PHR vendor [28].

**Advanced functions.** While many requirements have been developed by healthcare organizations, there is no commonly accepted definition or standard. Advanced functions that have been identified by researchers in the field of health information management, but are not well-covered, include the ability to quickly search for and find information, to recognize changes and trends in health status, to enable answering of temporal and spatial queries, to match patient treatments or disease progression against a standard template, and to aggregate information from multiple PHRs to determine health or treatment patterns.

<table>
<thead>
<tr>
<th>PHR Systems</th>
<th>Description</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>American Health Management Association (AHIMA) PHR</td>
<td>Website with a procedures and forms for constructing a PHR <a href="http://www.myphr.com">www.myphr.com</a></td>
<td>Standalone Free</td>
</tr>
<tr>
<td>My HealtheVet (MHV) PHR</td>
<td>PHR for veteran health benefits <a href="http://www.myhealth.va.gov">www.myhealth.va.gov</a></td>
<td>No cost to Vets</td>
</tr>
<tr>
<td>Google Health, Practice Fusion, Quest Diagnostics, &amp; Cleveland Clinic (discontinued)</td>
<td>Provided free electronic health records for physician and patient use. Ability to look at diagnostic tests online.</td>
<td>Free</td>
</tr>
<tr>
<td>iHealth Record (Medem)</td>
<td>Joint venture of AMA and healthcare organizations for sharing information between patients and health care providers. <a href="http://www.ihealthrecord.org">www.ihealthrecord.org</a></td>
<td>No cost to patients</td>
</tr>
<tr>
<td>Microsoft HealthVault <a href="http://healthvault.com/WhatIsHealthVault.htm">http://healthvault.com/WhatIsHealthVault.htm</a></td>
<td>Services and software platform, allows consumers to manage healthcare information online. Integrates with PCs, PDAs &amp; health monitoring devices.</td>
<td>Free. Integrated</td>
</tr>
<tr>
<td>PersonalMD <a href="http://www.personalmd.com/medrecro">www.personalmd.com/medrecro</a> dintro.shtml</td>
<td>Medical record, personalMD emergency card, medication reminder, drug interactions tool</td>
<td>Web-based, standalone</td>
</tr>
<tr>
<td>Vision Tree <a href="http://www.visiontree.com/index.cfm/fuseaction/products.healthcare">www.visiontree.com/index.cfm/fuseaction/products.healthcare</a></td>
<td>Patient-centered health record management and system stores personal medical records, and integrates with medical practices</td>
<td>Integrated Tethered</td>
</tr>
</tbody>
</table>

**Table 2. Example PHR System**
### 2.2.3 Inter-Operability Organizations

<table>
<thead>
<tr>
<th>Inter-Operability Organizations</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HL7 (Health Level 7 of OSI model)</strong></td>
<td>Several HL7 initiatives are ongoing. The Reference Information Model (RIM) is an object model that provides a pictorial and lifecycle representation of clinical data. The HL7 messaging standard is an application protocol for electronic data exchange. The Clinical Document Architecture is a model for exchanging clinical documents, using XML, RIM, and coded vocabularies [30].</td>
</tr>
<tr>
<td><strong>GEHR/openEHR</strong></td>
<td>OpenEHR is a set of projects developing specifications and open source tools, including the Requirements Project, the Architecture and an openEHR service model [31].</td>
</tr>
<tr>
<td><strong>European Health Telmatics</strong></td>
<td>European Health Telmatics Research Program is an organization defined standard, similar to OpenEHR, that is a multi-media data architecture for using and sharing electronic healthcare records [32].</td>
</tr>
<tr>
<td><strong>CEN</strong></td>
<td>CEN, the European committee for standardization, has developed CEN EN 13606 as a communication standard for transmitting part or all of the health record for a single person [33].</td>
</tr>
<tr>
<td><strong>DICOM</strong></td>
<td>Developed by the Digital Imaging and Communications in Medicine organization, DICOM is a standard methodology for transmission of medical images and associated information. It specifies a network protocol utilizing TCP/IP, and an application layer [34].</td>
</tr>
<tr>
<td><strong>ISO</strong></td>
<td>International Organization for Standardization, Technical Committee 215 – “Health Informatics” has created standards and requirements for EHR architectures, and interoperability standards for tele-health and tele-learning systems [35].</td>
</tr>
<tr>
<td><strong>IHE</strong></td>
<td>Integrating the Health Enterprise is an industry-sponsored consortium focused on improving the way computer systems in healthcare share information. Its uses “Technical Frameworks” to represent information in many health specialties [36].</td>
</tr>
<tr>
<td><strong>HHS</strong></td>
<td>The Department of Health and Human Services (HHS) has created a HealthIT organization to address standardization and tools [37].</td>
</tr>
<tr>
<td><strong>Microsoft HealthVault Architecture</strong></td>
<td>The HealthVault architecture includes facilities for storage of health information on the web, exchange of health information, interfaces with PC’s and devices, as well as a software integration architecture and toolkit for external party applications [23].</td>
</tr>
<tr>
<td><strong>CDISC</strong></td>
<td>CDISC is an industry-sponsored consortium focused on the representation of data in clinical research. In the CDISC model, which utilizes HL7’s CDA, data flows from the EMR system, to the physician, to the sponsor, and ultimately to the FDA [38].</td>
</tr>
</tbody>
</table>

Table 3. Interoperability organizations, standards, and architectures.

PHR’s must be able to utilize and interoperate through existing and future standards. Data from
healthcare providers must be able to be downloaded to a PHR. Data from home or outpatient care must be able to be uploaded to electronic health records at institutions. Individuals and institutions must be able to control, at a granular level, the information that is provided to others. Several standards organizations, focused on interoperability issues and architectures specific to healthcare, have developed a variety of overlapping and complementary standards. These have been analyzed in depth in the context of EHRs in [22]. In addition, extensive work is ongoing in standards alignment across standards bodies, medical disciplines, and countries. Descriptions of some important standards are provided.

### 2.2.4 Medical Information Content Coding

A large body of work has already been done to address representation of medical information. In the field of medicine, the most common approach has been to formulate important concepts as codes or classifications. Coding standards were originally used to track patient mortality. Over the years they have evolved to meet many different objectives including determining disease epidemiology, tracking medical procedure cost and outcomes, identifying patterns of problems, and determining utilization of supplies. Coding is used to represent many different types of concepts such as diseases, symptoms, body locations, therapeutics, and equipment or supplies. Medical codes used in practice have limited representational flexibility. A few of the more prevalent coding schemes are described here. A variety of other medical coding standards are used by health organizations for a variety of purposes. These are often incorporated in health information systems, and are linked to the patient medical record. Table 4 summarizes a representative set of medical information content coding standards.
### Table 4. Medical content coding standards.

Some medical terminology standards combine coding schemes with additional representational power. An example of this is SNOMED-CT, Systemized Nomenclature of Medicine - Clinical Terms. Originally developed as SNOMED by the College of American Pathologists, it was combined with the UK National Health Service Clinical Terms, in an ontological framework supporting class hierarchies and attributes [44]. SNOMED-CT is a subset of the Description Logic EL++ that it does not support negation, disjunction or numeric comparison. However, it

<table>
<thead>
<tr>
<th>Medical Information Coding Standard</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>International Classification of Diseases (ICD).</strong></td>
<td>First used in 1893 to report mortality statistics, it has evolved into a tool for disease classification. Revised every 10 years by the World Health Organization, it has a minimum of three digits for mortality reporting, with optional 4th and 5th digits [39].</td>
</tr>
<tr>
<td><strong>Diagnosis Related Group (DRG).</strong></td>
<td>Developed by Yale to facilitate Medicare payments, it classifies factors affecting the length and cost of hospitalizations [40][41].</td>
</tr>
<tr>
<td><strong>International Classification of Primary Care (ICPC)</strong></td>
<td>Published in 1992 by WHO, the IPC-2 provides 1400 diagnostic classifications partially mapped into ICD-9 [42]. It defines the coordinated use of 7 dimensions to represent a clinical situation.</td>
</tr>
<tr>
<td><strong>SNOMED</strong></td>
<td>Developed by the College of American Pathologists, SNOMED represents pathologic findings with anatomic, morphologic, etiologic (cause of origin) and functional terms [43].</td>
</tr>
<tr>
<td><strong>SNOMED-CT</strong></td>
<td>Combines best features of SNOMED and UK National Health Service Clinical terms to support a comprehensive medical terminology. Available through NLM [44].</td>
</tr>
<tr>
<td><strong>NDC – National Drug Code</strong></td>
<td>Developed by the Federal Drug Administration, it uniquely classifies drugs [45].</td>
</tr>
<tr>
<td><strong>Healthcare Common Procedure Coding System (HPCS)</strong></td>
<td>Developed by the American Medical Association and required by the Center for Medicare and Medicaid, it provides healthcare procedure codes to facilitate payment of insurance claims [46].</td>
</tr>
<tr>
<td><strong>Unified Medical Language System (UMLS)</strong></td>
<td>Begun in 1986, it provides a unified approach to integrating medical terminology, featuring a meta-thesaurus of over one million terms from 1000 sources. It represents concepts by tying terms together through semantic relationships [47].</td>
</tr>
<tr>
<td><strong>Interchange registration of coding schemes</strong></td>
<td>Proposed to register and code the coding schemes, the Interchange Registration of Coding Schemes, through CEN project team PT 004, has defined a unique six character Health Card Coding Scheme Designator (HCD) for each registered code, to ensure unambiguous meaning [48].</td>
</tr>
</tbody>
</table>
supports tractable answering of subsumption queries, which are important, since medical terms may be stated in a variety of ways. SNOMED-CT and Description Logics will be discussed further in Chapter 2.

### 2.3 Query support for medical systems

Section 2.2 described important standards for medical content coding. Some of these, such as ICD-9, are used in large medical facilities, such as hospitals, to identify eligible potential participants for clinical research. While medical coding standards provide many useful, standardized representations of medical information, these do not offer capabilities for general searching for information, or for answering specific or ad hoc questions. Practical implementation of medical queries will require the ability to answer queries across both structured, coded data and unstructured, non-coded data.

Imagine a physician/patient encounter in which the patient has an electronic health record that has gigabytes of diverse information about medical conditions, procedures, medications, reported symptoms, and so on. Consider the situation where the patient has some symptoms of dementia, and the physician wants to know if the patient has ever had a head injury, has undergone stress-related trauma, or has been treated for mental disorders. No physician will want to scan through each patient’s health record to find prior episodes or problems. Instead, the physician, will want to ask specific questions that will help to support a line of reasoning or to disprove a hypothesis. The physician will want to ask questions, and then drill down to detail that supports the answers to those questions. The physician may also want to query external information sources to look for similar situations or patterns. The following research, while not all directed at electronic or
personal health records, provide the foundation for support of many types of health queries.

### 2.3.1 Content-based text retrieval (CBTR)

Efficient retrieval of relevant documents can be achieved through indexing. A good description of general purpose indexing techniques is provided by [49]. Multi-word indexing utilizes ordered sequences of terms as indexes, and depend upon word proximity. Removing filler words, such as “in” and “the”, using stop-word lists, reduces the number of words needed, but concepts are limited to individual sentences. The use of frequent item sets makes this approach computationally feasible, but it is mostly effective for specialized query types.

Medical information presents a challenge for term-matching because many medical conditions may be represented by terms that have the same meaning but are substantially different in form. These terms may be different due to naming conventions in different medical specialties, or in differences in naming between the lay person and trained physicians. For example, a medical condition, “heart failure” might also be referred to as “coronary infarction”, “chest pains”, “tachycardia”, or “angina pectoris”, depending upon the actual situation and the person doing the reporting. Medical terminology-driven matching utilizes knowledge of medical terminologies to facilitate the matching process. An example of terminology-driven mining of biomedical literature is given in [50].

### 2.3.2 Content-based image retrieval (CBIR)

In the early 1990s, work began on a content-based approach to medical records. The Knowledge-based Medical Distributed Database System (KMed) at UCLA [51] was one of the first to address retrieval by image content and by text content. CBIR focuses on indexing and
retrieving images based upon their visual characteristics, such as color, texture, shape and spatial orientation [52]. Image management and retrieval technology has the potential to make a significant contribution in the medical area. Other work has been done in the bio-medical arena, for the indexing and retrieval of biological images, such as stained cells. A large body of CBIR research has been done in other domains that could be applied to medical images.

2.3.3 Multimodal document retrieval

A generic architecture for multimodal document retrieval is proposed by Chen [52]. The architecture does not specify the mechanisms used for indexing and retrieval, but shows the basic process flow of multimodal document retrieval. Beginning with a collection of documents, each document is indexed by content, using both text-based and image-based methods. Each query is also indexed by content, using both text-based and image-based methods, since a query may contain both text and images. Associations are made between the textual and image features in the document base. These associations are used to inform the retrieval process. A feedback loop is provided when the user interacts with the system to provide relevance feedback, which is used to refine the query. The final set of documents is evaluated to determine how well the system performed in retrieving relevant answers to the query. A more recent approach to multi-modal document retrieval is the AMORE system which provides three mechanism for finding images of interest: 1) specified keywords, 2) visual similarity, or 3) semantic similarity [54].

2.3.4 Use of domain knowledge in text-based retrieval

In biomedical applications, domain knowledge is very helpful in achieving useful results in text mining and query answering. In this approach, biomedical concepts are modeled. Semantic relationships are important, while word ordering and exact word or stem matches are used as a
supplementary method. Early work by Aronis [55] focused on the use of features, classes, and inheritance networks, linking medical data to botanical knowledge, and providing an extended semantics for use with automated discovery programs. It drew on a large set of data from telephone calls to poison centers from the American Association of Toxic Control Centers Toxic Exposure Surveillance System regarding potentially toxic exposure to plants. Aronson’s MetaMap [56] maps medical text to UMLS concepts. This is accomplished by a series of steps: parsing, variant generation, candidate retrieval, candidate evaluation, and mapping construction. The mapping files must be updated with every release of UMLS Knowledge Sources. Sarnoski mapped free-text anatomical descriptions to UMLS [57]. Chu et al [58] describe a knowledge based approach for indexing and retrieving scenario-specific medical text documents, where scenarios refer to commonly occurring tasks such as performing a diagnosis or prescribing a treatment. The approach includes use of the UMLS Metathesaurus and semantic expansion of concepts to improve matching of query concepts. Another knowledge-based approach, phrase-based indexing, combines a stem-based vector space model (VSM) with a concept-based VSM, mapping multiword concepts to UMLS concepts [59]. A third approach, called Indexfinder, [60] permutes words from text to generate candidate concepts and uses the UMLS knowledge base to match the candidates. A fourth approach is the use of Type Abstraction Hierarchies (TAH’s) [61]. TAH’s represent hierarchical levels of abstraction that can be used as an aid to approximate query answering. Swanson’s ABC (transitive closure) model of discovery [62] has been used to “discover” new relationships from publicly available on-line sources. The idea is to use bibliographic evidence of relationships between concepts from diverse and unconnected sources, and link these through the following reasoning. If A is related to B and B is related to C (in some dimension), then A may be related to C. Weeber [63] describes several scientific
discoveries that have been initiated through the use of Swanson’s model of discovery used in conjunction with UMLS, including hypothetical new applications of the drug thalidomide as an immunologic agent. Srinivasan et al [64] describe another use of domain knowledge mining of MEDLINE for discovering new knowledge from the literature, finding several implicit links between dietary substances and disease.

2.3.5 Generic Query Sets

A generic set of queries was developed by the Stanford Medical Informatics group. The idea was to enable user-friendly and efficient search of on-line medical reference information. UMLS medical terminology was used as a foundation for the query set. The generic query model included definitions, risk factors, etiologies, cause and effect relationships, distinguishing features, diagnostic and follow-up information usage for evaluation, relationship of therapies to treatments, relationship of behaviors, conditions, or treatments to prevention, performance characteristics of therapies in settings, comparisons of therapies in a certain setting, contraindications, sequelae and prognosis of manifestations or pathologies, and physical properties. The query model was applied to index medical texts for information retrieval and to support queries of medical texts [65]. This technology could be conceptually used in a personal health record to create a generic set of queries used by health professionals to query patient PHR’s about their medical history.

2.3.6 Temporal Reasoning

Several researchers have noted the importance of the temporal dimension in healthcare. Baxter [66] discusses patterns of health care in the context of data mining. An important observation is that some medical conditions, such as asthma, have an episodic event pattern, i.e., there are
sporadic, unpredictable occurrences of treatments and health state. Since events tend to be encoded as irregular time points, it is more difficult to vectorize health data so that clustering methods can be used for data mining. Another important characteristic of clinical health records is the important role of clinical narrative, where health care providers use narrative to document a patient’s health history, status, and prognosis, as well as recommendations for treatment. Hyun [67] developed a representation for marking up medical records that includes anatomical reference point, direction, number, patterns and units of time. The model was developed to contribute to temporal reasoning by providing structural information to be used in decision support, data mining and process analysis. Another characteristic noted by some researchers is the predominance of medical protocols for assessing and treating medical conditions. A major challenge is to align actual events to the planned or expected progression of events in a protocol, and to determine if there is a notable deviation to the protocol. RetroGuide was developed for retrospective analysis of lifetime electronic health record data, in the context of a defined health management protocol [68]. Lifeline was developed as a graphical, interactive tool for visualizing temporal events over multiple health categories simultaneously, and which offers methods for temporal alignment of health records [69].

These information technologies and research efforts offer promise for query support in health records management. However, while many of these methods may be used to retrieve documents and answer queries about the content of health records, they do not address the important issue of being able to determine the context in which a medical test was requested, and the reason for its importance.
CHAPTER 3  The Health Research Exchange (HRE)

This chapter develops the foundations for a health research collaboration system and environment – the Health Research Exchange (HRE). The Health Research Exchange supports a vision of making health research more accessible, useful, and interactive. Its goal is to bring together health researchers and health consumers, encouraging collaborative development of research ideas. The HRE acts as a broker of health research information, featuring a two-way exchange of information, a health hypothesis retrieval and reuse approach, and an approach for matching health hypothesis ideas to potential research subjects.

Section 1 presents the case for collaboration in health research then describes potential HRE stakeholders and their associated requirements. Section 2 evaluates the state-of-the-art clinicaltrials.gov database in the context of the requirements. Section 3 presents the HRE system architecture, its components and their interactions. Section 4 presents representational features of the HRE that address the requirements, including the use of SNOMED-CT for medical terminology, the suitability of Description Logics for health research queries, and a Description Logic extension, the Hybrid Hypergraph Description Logic (HDL).

3.1  Stakeholders and Requirements

3.1.1 The case for collaboration in health research

Currently, it is primarily clinical researchers, specialized in a particular medical area, that create hypotheses that eventually evolve into a clinical protocol that can be tested. While clinical research may involve many different scientific and legal specialties, these are engaged relatively
late in the process in order to support further refinement of a research idea that is already mostly formed. Early in the process of clinical research idea and hypothesis formation, there are fewer opportunities for collaboration and sharing, in large part due to the fact that the ideas and hypotheses are in an amorphous stage - not fully formed and, consequently, not easy to communicate and share. In addition, there are many different medical conditions, sometimes referred to as orphan diseases, which are not getting the attention they need from medical researchers, simply due to the rareness of a condition, and the consequent lack of funding available. Some individuals have successfully taken medical research into their own hands, [70]-[71] but most lack access to a community of researchers who might be able to provide ideas, context, and coaching they need in order to move the state of knowledge further.

There are people in many fields, other than clinical research, who may have relevant insights into a disease area. Physicians may have noted interesting patterns in patient treatment, for instance a perceived higher incidence of pancreatic cancer in patients taking a particular diabetes drug, reported to the FDA’s adverse events database [72]-[73]or a increasingly seen side effect, such as tendon ruptures, from taking a particular antibiotic [74]. Patients or their caregivers may have made observations and perceived connections that might be the basis for a research hypothesis. For example, caregivers of Parkinsons Disease patients who participate in Tai Chi might hypothesize that a customized Tai Chi regimen might improve balance in their patients. Patients and caregivers may also be able to identify which clinical research measures are the most meaningful to those who have to live day-to-day, year-after-year, with a condition. For example, out-of-pocket costs or impact of side effects on ability to drive, or work, may be important measures that might not otherwise be included as clinical research measures.
Improved understanding and the ability to collaborate with researchers may create more interest in participating in an evolving, future clinical research endeavor. People searching for information related to a medical condition of interest to themselves and their families may be better able to judge whether clinical research is relevant if they can participate in health research formulation, and observe its evolution. With the need to speed the clinical research and patient enrollment, and with an increasing number of people using the internet for researching and discussing health conditions, the time may be ripe for a central clearing house for the useful, structured exchange of information about health research and health hypotheses.

Currently, there are few vehicles for widely collaborating with a diversity of people in the identification, formation, and incremental development of medical hypotheses. A few web sites, such as Alzforum, for Alzheimers disease [75] provide discussion forums, and links to important research and results. Other sites, such as WebMD [76] provide a variety of medical information. But, there are no central forums for collaboratively creating clinical research questions and agendas in multiple and diverse areas. There are no general forums for patients and caregivers to suggest measures for clinical research that are meaningful to them. What is needed is a model for collaboration, which supports the discussion of health hypotheses, by providing a way to suggest, represent, and exchange hypotheses, and to elaborate on or specialize hypotheses, and related information.
3.1.2 HRE Stakeholder-focused design approach

The HRE addresses the need for a collaboration model and environment. Its design is based upon the needs of potential stakeholders in this collaboration. Stakeholders are those people, organizations, or systems that will play a key role in the use and success of the Health Research Exchange. The people and organizations most likely to benefit from the Health Research Exchange include: health research investigators, health consumers, physicians and other health providers, healthcare facilities management and health-focused organizations. These are described in Table 5. An individual may participate in the HRE in more than one stakeholder role. For instance, a physician may act as a health care provider, a researcher, or a health consumer. Alternatively, a health consumer who would like to share and develop a health hypothesis may assume the role of a health researcher to propose and design a health hypothesis.

<table>
<thead>
<tr>
<th>Stakeholder</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health research investigator</td>
<td>Person</td>
<td>Responsible for designing and conducting clinical research. Includes investigators, sub-investigators, coordinators, biostatisticians, and operations personnel for patient enrollment and site monitoring.</td>
</tr>
<tr>
<td>Health Consumer</td>
<td>Person</td>
<td>Utilizes health services to diagnose, prevent, and treat illnesses, and maintain health. Health consumers may take an active role in managing their own health or that of their friends or families.</td>
</tr>
<tr>
<td>Physician or health care provider</td>
<td>Person</td>
<td>Provides services directly, or indirectly to the health consumer. These may include diagnosis and treatment of injuries or disease, preventive measures, and health assessment.</td>
</tr>
<tr>
<td>Healthcare Facility Management</td>
<td>Person</td>
<td>Responsible for managing the business side of the healthcare facility for turnaround time, efficiency, and effectiveness. Ensures quality, and makes decisions about which health procedures are supported.</td>
</tr>
<tr>
<td>Health-focused Organization</td>
<td>Group</td>
<td>Organizations that participate in information sharing and healthcare improvement, but that do not directly render health services. Examples of these are regulatory and government agencies, such as the FDA and NIH, and research organizations, such as CHDI, for Huntington's Disease.</td>
</tr>
</tbody>
</table>

Table 5. HRE stakeholders – people and organizations.
Several system types may act as key participants in the Health Research Exchange, since there is an increasing need for health-related systems to interconnect. The system types most likely to interact with the HRE include Clinical Research Management Systems, Personal Health Records, Electronic Medical Records, Hospital Business Systems, and Health-focused websites. These are described in Table 6.

<table>
<thead>
<tr>
<th>Stakeholder Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical Research Management Systems System</td>
<td>Supports the practice of clinical research; used by research groups to store and analyze the large volume of information gathered during the research process. This includes systems for managing subject data, for managing and tracking enrollment, and for tracking adverse events.</td>
</tr>
<tr>
<td>Personal Health Record (PHR) System System</td>
<td>Contains histories of health conditions, procedures, lab measurement, diagnoses, prescriptions, and doctors’ narratives. PHRs are managed by the health consumer, are not directly tied, or tethered, to a health care provider.</td>
</tr>
<tr>
<td>Electronic Health Record (EHR) System System</td>
<td>Maintains individual health records for a healthcare organization. EHRs are designed to meet the needs of healthcare practices and providers. The EHRs of an individual are usually distributed over a large number of medical service providers.</td>
</tr>
<tr>
<td>Hospital Business System System</td>
<td>Supports healthcare facilities management in keeping track of operations, costs, quality and efficiency through coding systems, such as ICD-9, that use a set of predefined vocabularies and categories.</td>
</tr>
<tr>
<td>Health-focused Websites System</td>
<td>Supports health-focused organizations in collecting and disseminating information about a specific perspective on health. Example are the FDA’s clinicaltrials.gov, and the Alzforum website for Alzheimer’s disease.</td>
</tr>
</tbody>
</table>

**Table 6. HRE stakeholders – systems.**

This research focuses on the interaction of the HRE with the stakeholder roles most likely to interactively participate in and directly benefit from, the HRE information exchange. The key stakeholders are the Health Consumer, and the Clinical Research Investigator. In the context of the HRE, the physician may assume either of these roles. The key system stakeholder is the
PHR, which may acquire some of its information from EHR’s, but may also acquire information, in the form of Health Condition Templates, from the HRE. Figure 1 depicts the interactions between the HRE, and its people, organizations, and systems stakeholders.

Figure 1. Stakeholder Interactions

3.1.3 HRE Requirements

Requirements act as a set of constraints that the system must fulfill in order to meet system goals. There are two kinds of requirements, functional and non-functional. Functional requirements are driven by stakeholders, those who interact with or benefit from the system, while non-functional requirements are required quality characteristics of the system, for example, performance, reliability, and quality of service. Together, functional and non-functional requirements drive
the technical requirements of the system, which include system and knowledge representations and operations that must be performed. Technical requirements inspire the design of artifacts, or system components that interact through defined protocols. The following sections define the functional requirements for the HRE from the perspectives of each of the primary stakeholders, and then identify non-functional and technical requirements.

3.1.4 Functional Requirements

A set of functional requirements was developed for each of the key HRE stakeholders.

Health Research Investigator (HRI) requirements:
The system must:
• Provide the means to develop and share health research ideas for a population
• Provide methods for representing diverse types of information, such as concepts, relationships, inclusion and exclusion criteria, and analysis strategies
• Provide information about the number of subjects meeting proposed research criteria
• Provide information about potential research participants, who have expressed an interest
• Provide an environment for incremental, efficient, development of health hypotheses
• Improve communication in the area of clinical research idea development
• Be able to identify related and overlapping research ideas

Health Consumer (HC) requirements:
The system must:
• Provide convenient access to information about ongoing clinical research
• Provide the means to suggest and develop health research ideas and relevant measures
• Be able to efficiently create relevant health profiles specific to a health condition
• Be able to post candidate health profiles that enable matching to relevant clinical research
• Be able to provide input on health research ideas

Personal Health Record (PHR) requirements:
The system must:
• Provide ease of entry for health profile information relevant to clinical research
• Identify health research relevant to a health condition
• Enable sharing relevant health information at the direction of the PHR owner
• Provide privacy mechanisms to avoid inadvertent sharing of personal health data

A key challenge is the specification and development of a collaboration model and infrastructure
that will enable people, who are not medical experts, to participate in the health research formulation process. A related challenge is to leverage the information in Personal Health Records to facilitate the collaboration process.

3.1.5 Non-functional Requirements

In addition to the functional requirements, there are several non-functional requirements.

The system must:
- Handle multiple levels of abstraction
- Be able to share information where users are decoupled in time and space
- Share domain terminologies, and other domain knowledge networks that can help to inform and support research idea development
- Scale to hundreds of thousands research ideas and millions of concepts
- Efficiently match research ideas against millions of potential research participants
- Provide secure storage and exchange of information
- Ensure HIPAA-compliance, and reserve the anonymity of HRE participants, unless explicit permission is given to share identify
- Provide robustness, speed, and efficiency, and ease of use

A key challenge related to non-functional requirements is architectural scalability. In order for a solution to be usable in practice, it must support efficient matching of hundreds of thousands of research ideas to millions of potential research participants, via millions of concepts.

3.1.6 Technical Requirements

Technical requirements specify the artifacts needed to meet the requirements.

- Data types and structures for expressing health research ideas
- Methods for representing and reasoning about medical terminology
- Structures and methods for represent, reuse, and manipulate domain knowledge
- Methods for linking domain knowledge structures to research ideas
- Methods for managing the distribution of published knowledge networks
- The definition and answering of queries
- Methods for incrementally developing, measuring and tracking hypotheses
- Strategies for interacting with PHRs
- User interaction strategies and interfaces that facilitate the communication process
A key technical challenge is the development of an information abstraction for health research ideas that enables incremental specification of research ideas and that can interface with heterogeneous data and inference methods in medical domains at multiple levels of abstraction.

### 3.2 Current best practice - NIH database

The clinical trials database, clinicaltrials.gov, represents today’s best practice in describing a diverse and large number of clinical research ideas [4]. Consequently, it was utilized as a resource for determining current practice in large-scale representation of research ideas, and health hypotheses in particular. Since, by the definition of clinical research, each research project must have a hypothesis in order to be approved by an Institutional Review Board (IRB), the clinical trials database is arguably the largest repository of health hypotheses in the world.

#### 3.2.1 Hypothesis Representation

However, in clinicaltrials.gov there is not a specific field for hypotheses, in fact the word hypothesis is not consistently mentioned. In over 100,000 data base entries, the word “hypothesis” or “hypotheses” was mentioned in approximately 8% (8093) of studies, the word “hypothesize”, by less than 3% (2982) of the studies. Since there is no predefined method of specification of the hypothesis under investigation, it is not possible to consistently and automatically extract the hypothesis statement. Within the clinical trials database there is no guidance or format for specifying a health research idea, except for the guidance provided in the instructions for populating the database fields. Apparently, the clinical trials database is viewed, by its sponsor, as primarily a project repository.
The clinical trials database, as a project repository, does fulfill the important function of tracking existence, content, and progress of clinical trials. But it misses the opportunity to leverage the fact that all approved clinical research must have a stated hypothesis and a test and analysis strategy, and that people who are experts in creating these, the biostatisticians, ensure that these are of good quality. Since the work is already done, it seems reasonable to require a hypothesis statement and full description that can be extracted and reused – that states the hypothesized relationship, the basis for it, and how it will be tested. In addition, if there are conceptual relationships or research population definitions that provide the foundation of the hypothesis, it would be ideal to represent them in a way that can be reused by others who may be doing similar or related research. For example, a person developing a new hypothesis about the effects of Selective Serotonin Reuptake Inhibitors (SSRI) on ADHD might be able to reuse conceptual relationships from clinical research on autism, describing SSRI effects on dopamine pathways, as well as population definitions describing sets of people who were exposed to SSRI’s during gestation (mother’s prescribed SSRI’s).

3.2.2 Health Hypothesis Extraction

With the goal of being able to consistently extract a health hypothesis, the clinicaltrials.gov available fields were analyzed for their ability to provide information relevant to hypotheses. A subset was selected, based on the potential for matching potential subjects to health research. The fields selected were: 1) Condition, 2) Intervention, 3) Criteria, 4) Outcome Measures, and 4) Study start and end dates. The author identified three fields – condition, intervention, and outcome measures – as being useful for constructing a surrogate for a hypothesis statement. For instance, with the following sentence template, a sentence emerges that is a reasonable statement
of a hypothesis:

“For the condition {condition}, the intervention {intervention} may have an impact on {primary outcome measure}, as well as {secondary outcome measure}.

If there is no intervention specified, a different sentence template may be useful for extracting a simple hypothesis sentence.

“The condition {condition} is associated with primary outcome {primary outcome} and other outcomes {[secondary outcomes]}.

Applied to a Parkinson’s Disease Tai Chi study published recently the New England Journal of Medicine [77] and summarized in the ClinicalTrials.gov NCT00611481, this template approach results in the following hypothesis: For the condition, {Parkinson’s Disease}, the intervention, {Behavioral: Tai Chi}, will have an impact on {balance, gait}, as well as {gait, physical performance, Unified Parkinson’s Disease Rating Scale, falls, Muscle strength}.

Experiments with this method yielded a wide range of success, with some of the statements being understandable as hypotheses while others were not easy to understand or were very complex. Table 7 demonstrates how example hypothesis statements may be manually constructed from these fields, using existing clinical trial entries in the NIH database.
<table>
<thead>
<tr>
<th>Title</th>
<th>Condition</th>
<th>Intervention</th>
<th>Outcome Measure</th>
<th>Hypothesis constructed from template</th>
</tr>
</thead>
<tbody>
<tr>
<td>Impact of Exercise on Asthma Morbidity</td>
<td>Condition= {Asthma}, Intervention= {Aerobic Exercise 12 weeks of supervised exercise, 3x per week, one hour}</td>
<td>Primary Outcome= {Asthma Control Questionnaire} Secondary Outcome= {{Asthma quality of life questionnaire, Asthma control test, inflammation markers}}</td>
<td>For the condition asthma, the intervention of aerobic exercise (12 weeks supervised exercise, 3x per week, one hour), may have an impact on asthma control questionnaire, as well as asthma quality of life questionnaire, asthma control test, and inflammation markers.</td>
<td></td>
</tr>
<tr>
<td>Lenalidomide Plus Bendamustine and Rituximab for Untreated CLL/SLL</td>
<td>Condition= {CLL/SLL} Intervention= {{Lenalidomide, Bendamustine, Rituximab}}</td>
<td>Primary Outcome= {identify the maximum tolerated dose of lanalidomide when combined with bendamustine and rituximab in CLL/SLL}</td>
<td>For the condition CLL/SLL, the intervention of Lenalidomide, Bendamustine, Rutuximab may have an impact on identify the maximum tolerated dose of lanalidomide when combined with bendamustine and rituximab in CLL/SLL.</td>
<td></td>
</tr>
<tr>
<td>A study of Swine-origin A/H1N1 Influenza Vaccines in Healthy Europeans, Children 6 to 35 months</td>
<td>Condition= {{Influenza, swine-origin A/H1N1 influenza}} Intervention= {Swine A/H1N1 influenza vaccine}</td>
<td>Primary Outcome= {{immunogenicity, safety}}</td>
<td>For the condition, swine-origin A/H1N1 influenza, the intervention swine A/H1N1 influenza vaccine, may have an impact on immunogenicity and safety.</td>
<td></td>
</tr>
</tbody>
</table>

Table 7. Use of sentence templates to construct hypotheses from the NIH database

With the addition of the phrase, “This hypothesis is applicable for subjects with {criteria}”, the hypothesis may be narrowed to a specific population. Unfortunately, the Criteria field, which was expected to provide the most useful information for subject identification, is most commonly represented as a “text block”. The inclusion and exclusion criteria contained in the text block could not be automatically mapped to a structured representation without further and extensive processing.
The criteria descriptions can be quite long, unstructured, and technical in nature. For example, in the Parkinson’s Disease Tai Chi study, the inclusion and exclusion criteria were stated in ClinicalTrials.gov as:

“clinical diagnosis of Parkinson’s disease, with a disease severity rating of stage 1 to 4 on the Hoehn and Yahr scale (which ranges form 1 to 5, with higher scores representing more severe disease); an age of 40 to 85 years; at least one score of 2 or more for at least one limb for the tremor, rigidity, postural stability, or bradykinesia items in the motor section of the Unified Parkinson’s Disease rating score (UPDRS) III); stable medication use; ability to stand unaided and walk with or without an assistive device; medical clearance for participation; and willingness to be assigned to any of the three interventions. Exclusion criteria were current participation in any other behavioral or pharmacologic study or instructor-led exercise program, a Mini-Mental State examination score lower than 24 (indicating some degree of cognitive impairment), debilitating conditions or vision impairment that would impede full participation in the study, and unavailability during the study period.”

Note the representational need for complex logic as well as numerical comparison operations, and reasoning with medical terminology.

A second type of analysis was performed on clinicaltrials.gov, with the goal of finding frequent terms, other than database keywords, which might be useful in the construction of hypothesis statement. Frequent terms that were of interest, and not specific to a health condition, include:


The type of clinical trial does affect the types of information available. For example, a large proportion of the NIH database entries are interventional studies. An intervention is a treatment that intervenes in a normal course of events, or pathway, to cause a difference in outcome. The data for interventional trials include a description of an intervention, which does not appear in
non-interventional trials. In addition, interventional studies may informally describe the pathways that are hypothesized to be enhanced or disrupted by the intervention. These pathways often are described as relationships, such as: “leads to”, “caused by”, and “associated with”.

The common occurrence of research descriptions and publications that describe rationale in terms of chains of these types of relationships, indicates that there is an opportunity in health research to capture, store and share, small networks of domain knowledge. For example, health research studies that incorporate, in their chain of reasoning, the concepts “associated with bone resorption”(29), “associated with bone health” (329), and “leads to atherosclerosis” (119), might be able to share and build upon common knowledge networks in each area.

The clinicaltrials.gov database was evaluated for its ability to meet the requirements for a collaborative clinical research idea environment, identified in Section 3.1. Of the more than 30 requirements, a little over a third were met by clinicaltrials.gov, but these are only met in the context of approved clinical research, and not of clinical research in the formative stages. Even in the case of approved clinical research, clinicaltrials.gov misses opportunities to engage health consumers in providing feedback on what is meaningful to them, and to make productive use of the plethora of domain knowledge and health research approaches contained in its many descriptive information elements.

The study of the clinicaltrials.gov database yielded three conclusions: 1) the current representational framework does not meet the requirements for enabling consistent representation and ease of creation of hypotheses, and 2) the current framework misses
opportunities to engage health consumers in suggesting new ideas and important measures, and
3) the current framework does not make productive use of its diverse and comprehensive medical
knowledge. The purpose of the HRE is to address these gaps.

3.2.3 Challenges

As evidenced by the large variety of clinical research descriptions in the NIH clinical trials
database, one challenge in health hypothesis development is that it is not an exact science.
Development of a testable hypothesis often requires extensive knowledge of the terminology and
important parameters for the health condition(s) of interest, as well as knowledge of clinical
research design and statistics. Even experts in the field may require several iterations in order to
develop a health hypothesis that is testable through a clinical research project. Related to this are
the challenges of incorporating inputs from multiple people, with different backgrounds, who are
separated in time and space, of handling diverse types of health conditions, and providing
sufficient support so that the health consumer will be able to collaboratively contribute in a
health research endeavor.

A key technical challenge is the development of an information abstraction for health research
ideas, expressed as hypotheses, that can represent and interface with heterogeneous data and
inference methods, in multiple medical domains at multiple levels of abstraction. The
information abstraction must accommodate the practical need and practice of incremental
development of health research ideas, such that they can be productively shared from inception
to full specification. Another challenge is the specification and development of a collaboration
model and architecture that will enable people, who are not medical experts, to participate in the
health research formulation process. A related challenge is the leveraging of information in Personal Health Records to facilitate the collaboration process. In order for a solution to be usable in practice, the architecture must be scalable; it must ultimately support efficient matching of hundreds of thousands of research ideas to millions of potential research participants, via hundreds of thousands of concepts.

The remainder of Chapter 3 will focus on two topics: 1) the overall architecture for the HRE – which describes the high-level components of the HRE and how they work together to meet the requirements, and 2) the underlying representational formalism that will provide the foundation for the representation of health research ideas.

### 3.3 The HRE system architecture

The purpose of the HRE is to provide a forum and methodology for discussion of health research in a variety of fields. It is not limited to a single medical specialty or health condition. Instead, it seeks to leverage knowledge across many different health conditions by providing a unifying framework for discussion of health hypotheses. Consequently the HRE is focused on a general representation and methodology for health hypotheses and the concepts needed to support these. While the HRE may include or link to descriptions of health data, it does not include actual health data for the individual or for clinical research. As a system, its primary interfaces are with health researchers, health consumers and personal health records. This section describes the system architecture for the Health Research Exchange, including the structure and behavior of the HRE, the decomposition of the system into its constituent elements, and the interconnections between these. Figure 2 depicts the Health Research Exchange system architecture.
The HRE comprises four major levels of abstraction, each one providing a foundation for the level above. Level 1 features an underlying representation, HDL, for HRE concepts and connectivity. Level 2 describes conceptual and representational models for hypotheses. Its Molecular Hypothesis Model (MoHM) is the central, enabling concept for the HRE. Level 3 utilizes the MoHM to provide hypothesis management capabilities, including hypothesis reuse (CaNDor), management (HyRE), and messaging (HCP). Level 4 adds a collaboration capability with the HYPEG protocol and methods for sharing health hypotheses and health condition information. Each major architectural level and its components are outlined below.

**Figure 2. The HRE Architecture**
3.3.1 Level 1. Concepts and Connectivity

Within the HRE, concepts and connectivity between concepts are represented with an extension to Description Logics (DL), called Hybrid Hypergraph Description Logic (HDL). HDL extends the notion of pairwise DL roles, represented as graph edges, to hyperroles, represented as hyperedges. This extension accommodates the need to represent the many-to-many relationships required for representing some hypotheses. In addition, HDL integrates the DL-based medical terminology language SNOMED-CT, and incorporates Boolean formulas that enable evaluation of health research inclusion and exclusion criteria. HDL is described later in Chapter 3.

The HRE will require an intuitive HDL interface for accessing SNOMED-CT concepts and applications. However, a large body of work already exists addressing interface strategies for SNOMED-CT, including automatic SNOMED-CT coding from medical records, menu-based facilities for searching for and selecting terms, and servers that can efficiently answer SNOMED-CT queries, such as subsumption. This research focuses on HRE architectural and representational challenges, and does not address SNOMED-CT utilization or interface issues. These are left as a challenge for implementation.

3.3.2 Level 2. Hypothesis Models

A variety of conceptual models have been previously considered for hypotheses. For the HRE, the underlying, core concept for the HRE architecture is the Molecular Hypothesis Model (MoHM) – based upon the principles of connectivity, modularity, and domain specificity – and represented in HDL. The MoHM decomposes hypotheses into six modular, connected domains. Each domain has a specific function and representation. For example, the Prediction
Domain has the function of describing relationships between concepts, where the relationships are represented as hyperedges. The Plausibility Domain has the function of connecting hypotheses to existing knowledge in the form of facts and relationships, represented by Domain Concept Networks – small graphs or snippets of domain knowledge. The Population domain, represented as Boolean functions, defines the inclusion and exclusion criteria for health research populations. Chapter 4 describes historical hypothesis conceptual models, the MoHM, its properties, and the six MoHM domains.

**3.3.3 Level 3. Hypothesis Management**

Population Health Hypotheses represented as MoHMs are aggregated in Case Libraries, and stored in a form such that health hypotheses or individual domains of a health hypothesis may be retrieved and re-used. Hypothesis management functions are needed for hypothesis construction, reuse, modification, sharing, and messaging. Within the HRE, an approach called Case-Network Domain Reasoning (CaNDoR) is used for development of new health hypotheses. The purpose of the CaNDoR approach to hypothesis construction is to enable people, who are not medical experts, to develop hypotheses that will be understandable by clinical research professionals. In CaNDoR, queries yield health hypothesis cases with similar content or structure. These hypothesis cases contain information that can be reused, such as relevant sets of medical measurements, testing strategies, references to existing knowledge, statistical approaches, and research population criteria, each of which represent clinical research knowledge that can take considerable effort to develop. Using a CaNDoR approach, a person who has a health research idea, but little knowledge of how health research is specified, can search for similar ideas, or hypotheses, for similar population criteria, or for similar test strategies, and then reuse parts of
these to create a new research hypothesis that will be much closer to being complete and communicable than one developed from scratch. CaNDoR is described in Chapter 5.

The Hypothesis Reasoning Engine (HyRE) provides facilities for constructing, modifying, archiving and transforming hypotheses. For every hypothesis and associated hypothesis structure, there are two forms of representation – the Working Model Representation (WMR) and the Packaged Model Representation (PMR). HyRE works within the WMR representation to provide services to aid in hypothesis formation and reuse, and to answer queries about hypothesis content, connectedness, and similarity. HyRE also provides the capabilities to translate between WMR and PMR representations, and to export PMR hypotheses to XML for sharing. HyRE is described in Chapter 5.

A Health Condition Profile (HCP) is a Personal Health Hypothesis that states that an individual is a member of a health condition population, and provides predictions that can be tested to assess that membership. An individual can have many HCP’s, each one specific to a health condition. HCP’s can express HCP messages that export health parameter values that can be used to match individuals with proposed health research. Chapter 5 describes Health Condition Profiles and messages.

3.3.4 Level 4. Collaboration

Level 4 proposes a collaboration model that will enable people, who are not medical experts, to participate in the health research formulation process. The Hypothesize/Engage Protocol (HYPEG) is a collaboration protocol, which features an incremental, progressive interaction
between health consumer and health researcher by defining levels of interaction and by protecting anonymity until both parties have agreed to reveal their identity. Both health consumers and health researchers can propose health research ideas, critique health research ideas in formation, and suggest health measures that are meaningful to them.

Figure 3. HRE HYPEG Collaboration Protocol

Through HYPEG, any health consumer may take the additional step of anonymously sharing their Health Condition Profiles for a specific condition. In addition, health researchers may express interest in a posted Health Condition Profile, and may offer to a health consumer the opportunity to become engaged in a research process. Within the Hypothesize/Engage Protocol, the challenge of matching large numbers of health hypotheses to large numbers of health profiles is addressed. An approach is utilized where health research population criteria are compiled to
Ordered Binary Decision Diagrams (OBDD), for efficient matching of research hypothesis population criteria to large numbers of posted Health Condition Profiles. HYPEG, depicted in Figure 3, is described in detail in Chapter 6.

3.3.5 HRE architecture contributions

The HRE architecture addresses the requirements for information abstraction, incremental development and collaboration in Health Research. It does so in three ways: 1) by defining a modular representational formalism for health research ideas, 2) by defining a template-based method for managing and reusing health hypotheses and their components, facilitating the involvement of non-health professionals in health research discussions, and 3) by defining a collaborative model for hypothesis discourse, enabled by large scale matching of health research ideas to patient populations.

3.4 HRE knowledge representation foundations

This section describes the underlying knowledge representation approach for the HRE. Beginning with foundational work in the medical terminology language, SNOMED-CT, and the Description Logics (DL) family of knowledge representations, this section describes a Hybrid Hypergraph Description Logic (HDL) framework, which enables connectivity between concepts not expressible in a DL, and the evaluation of hybrid queries that include both DL and non-DL functions. HDL will be used in Chapter 4 to define the Molecular Hypothesis Model, a modular approach for representing health hypotheses.
3.4.1 Medical terminology representation and SNOMED-CT

The incorporation of medical terminology is an essential component of any computational model for health research. A representation for medical terminology must be precise enough to uniquely define a medical condition, yet expressive enough to incorporate the many ways that a medical condition can be stated. It must support a “living terminology” that can rapidly change to keep up with the pace of medical technology. In addition, it must be able to facilitate the efficient answering of queries, such as whether a person has a condition, or some subset of a condition.

The Systemized Nomenclature of Medicine – Clinical Terms (SNOMED-CT), is a medical terminology language meeting these requirements. Distributed by the International Health Terminology Standards Development Organization (IHTSDO), it has over 280,000 medical concept codes, over 700,000 active terms, and over 900,000 active defining relationships [78]. Each concept is represented both by a text string and by a unique number. An example SNOMED term is: “Open wound of hand with tendon involvement (disorder) “[78]. Concepts can be combined to create new concepts, using hierarchical, unary “isa” relationships, binary attribute relationships, and a compositional grammar Three main hierarchies – finding, disease, and procedure – are augmented by over 12 supporting hierarchies. Examples of relationships are “finding site”, “causative agent”, “associated morphology”, and “using”. SNOMED-CT has a Lisp-like syntax. An example SNOMED-CT representation for History-of (myocardial infarction) described by Spackman [79] is:
While SNOMED cannot correctly interpret negation or disjunction within SNOMED terms, it is possible to express negative finding patterns such as: clinical finding absent, clinical finding unknown, no history of, no family history of, procedure not done, and drug (or procedure) counter-indicated. Representing these “negative” concepts explicitly is quite important for the expression of health research exclusion criteria, which state clear guidelines for conditions that cannot be present in an eligible participant. Reasoning in SNOMED-CT is focused on answering many practical subsumption questions, such as: “Which patient records contain findings of heart failure,” when heart failure may be expressed in many different ways, and there may be many kinds of heart failure defined.

SNOMED-CT is itself only a terminology, but supporting technologies have been developed around it, including SNOMED-CT interpreters and terminology selection user interfaces. SNOMED-CT is continually being studied and improved. SNOMED terminology servers have been built which are optimized to handle efficient SNOMED-CT queries. SNOMED-CT is widely used in electronic health records, including in the EPIC EMR system [80] [81] and in Kaiser Permanente’s Health Connect, as part of their Convergent Medical Technology (CMT) standard [82]. At Kaiser, several recent studies have been performed on SNOMED-CT, analyzing its capabilities. For example, experimental addition of negated SNOMED-CT terms in a large existing database resulted in performance consistent with operational needs [83], but that
the use of negation resulted in incorrect term classification due to negated terms classified as positive occurrences of a condition. Additional projects with SNOMED-CT are being explored by Kaiser and others to examine the tradeoffs between SNOMED-CT expressivity, query accuracy, and tractability in medical systems.

SNOMED-CT can be used with several other terminologies. It has been integrated into UMLS, largely through mapping of concepts. It also supports a direct, one-way conversion to ICD-9 terms, which are required for billing purposes, and are used by some hospitals to identify patient populations for clinical research. While it is possible that more than one medical terminology can be used, and more than one may be needed for practical support of clinical research, the HRE will focus on the SNOMED-CT representation for medical terms. Because of its widespread usage, excellent coverage of medical terms, and formal foundations, it can be effectively leveraged for the creation of a representation for health hypotheses.

3.4.2 Description Logics

SNOMED-CT offers an important advantage over other medical terminologies. It is based upon a powerful, and well-understood logical formalism called Description Logic. Description Logic (DL) is a formal knowledge representation approach, utilizing decidable fragments of first order logic [84]. It defines families of languages based upon expressivity, i.e, what can be said and asked, and tractability, i.e, what can be efficiently computed, in each language. For instance, a base language, Attributive Language, AL, allows atomic negation, concept intersection, universal restrictions, and limited existential quantification, while an extension, ALC, adds the ability of expressing a complement of any concept. Additional properties result in additional
expressivity but less tractability. Recent work in DLs has focused more on adding expressiveness at the expense of tractability. The Web Ontology Working Group developed OWL2 for semantic web expression, based on the description language SROIQ(D), which allows several other properties, but cannot answer many queries in polynomial time [85].

Description logics, which are less expressive than propositional logic, but more decidable, are sometimes referred to as “islands of tractability”. The restrictions on the expressivity in each language result in well-defined properties with respect to query types. For instance, the language EL allows only concept intersection and existential restrictions. Reasoning in EL means computing subsumption between concepts, and checking satisfiability. Because of the restriction on operations, subsumption and satisfiability can be computed in polynomial time. An extension of EL, EL++ allows the use of numeric comparison, nominals, and other operations, while retaining the ability to answer some, but not all, types of queries in polynomial time. SNOMED_CT, which was originally based partially on the lisp-like DL KRSS, is expressible in EL with additional role properties, defined by EL++ [86]. Because of its limited expressiveness, SNOMED-CT is able to answer many important medical terminology queries efficiently, however, by itself it is not expressive enough to support the operations required by the HRE. These missing operations include numeric comparisons and negation of complex concepts required for patient population criteria. While other DL’s may provide more expressiveness, they do not have the same level of medical terminology support nor do they provide the ability to address complex queries involving more than one DL or other representations.
3.4.3 Interpreting DL concepts and roles

There are two kinds of concepts in Description Logics (DL), atomic concepts and atomic roles. DL concepts do not contain variables. A concept represents all the individuals satisfying the designated properties. Terms are built using constructors, such as union, $\cup$, intersection, $\cap$, complement, $\neg$, some, $\exists$, and all, $\forall$. A set-theoretic interpretation of DL treats “concepts as sets of individuals and roles as sets of ordered pairs of individuals,” where the roles are from a specified domain. Atomic concepts are subsets of a potentially infinite interpretation domain [87]. Constructs built from atomic concepts and roles are used to specify a subset of individuals. For instance, the value restriction construct, $\forall R.C$, requires that all individuals in $R$ belong to the class $C$. The construct $\text{Subject} \cap \forall \text{gender.} \text{Female} \cap \exists \text{child.} \text{ADHD}$ denotes the set of individuals that are instances of the concept “Subject” and are connected through the role “gender” only to instances of the concept “Female” and, through the role “child”, to some instance of the concept “ADHD”. In other words, the construct describes the concept that is the set of subjects who are female, having at least one child that has ADHD. The basic operation in DL is subsumption, denoted as $A \subseteq B$, which determines whether the concept $B$ is more general than the concept $A$. Satisfiability in DL determines whether a concept construct is not equivalent to the empty set. All DL constructs can be expressed as tree data structures.

The basic modeling concept of a DL is the axiom, which is a statement relating roles and concepts. Within DLs, a distinction is made between the terminology (TBox) and the assertions (ABox). The TBox describes concept hierarchies, while the ABox contains statements about relations between individuals and concepts, placing individuals within the concept hierarchy.
DL’s do not make the Closed World Assumption, meaning that lack of knowledge does not imply that something is false. DL’s do not make the Unique Name Assumption, meaning that two concepts with different names, are not necessarily two different concepts. An important characteristic of the DL family of languages is that the meaning and syntax of the operations can be consistent across many languages. This facilitates the direct comparison between languages, and often allows the extension of results for one language to another.

### 3.4.4 Correspondence between DL and first-order-logic

Description Logics have been shown to have correspondence to first order logic [88]. Atomic concepts can be represented by unary predicate symbols. Atomic roles, represented by binary predicate symbols, can be used to represent relationships. For example, the term A ∩ B is equivalent to the first-order logic sentence A(x) & B(x) [89]. Grosof et al translated ALL and EL to first order logic as shown below and in Table 8 [90].

- Each atomic concept C, is associated with a unary predicate C(x).
- Each atomic role, P, has an arity, n.
- Each atomic role, P, with n=2, is associated with a binary predicate P(x,y).

<table>
<thead>
<tr>
<th>DL Syntax</th>
<th>First Order Logic [Grosof et al]</th>
</tr>
</thead>
<tbody>
<tr>
<td>a : C</td>
<td>C(a)</td>
</tr>
<tr>
<td>&lt;a,b&gt; : P</td>
<td>P(a,b)</td>
</tr>
<tr>
<td>C ⊆ D</td>
<td>∀x.C(x) → D(x)</td>
</tr>
<tr>
<td>P⁺ ⊆ P</td>
<td>∀x,y,z.P(x,y) P(y,z) → P(x,z)</td>
</tr>
<tr>
<td>C₁ ∩ ... ∩ Cₙ</td>
<td>C₁(z) ∧ ... ∧ Cₙ(x)</td>
</tr>
<tr>
<td>C₁ ∪ ... ∪ Cₙ</td>
<td>C₁(z) ∨ ... ∨ Cₙ(x)</td>
</tr>
<tr>
<td>¬C</td>
<td>¬C(x)</td>
</tr>
<tr>
<td>{a₁,...,aₙ}</td>
<td>x=a₁ ∨ ... ∨ x=aₙ</td>
</tr>
<tr>
<td>∃P.C</td>
<td>∃y.(P(x,y) ∧ C(y))</td>
</tr>
<tr>
<td>∀P.C</td>
<td>∀y.(P(x,y) → C(y))</td>
</tr>
</tbody>
</table>

Table 8. Examples of correspondence between DL and FOL [Grosof].
3.4.5 Limitations of DL’s

There are several important limitations to Description Logics. One consistent limitation is that DL’s are limited to tree structures. DL’s cannot represent arbitrary connectivity relationships, and hence does not support traversal queries on graphs that cannot be represented by tree structures. Another consistent limitation is that only unary and binary relationships can be expressed in DL. Because of this there are clear limits on what can be directly represented. For examples, Entity-Relationship diagrams may not be represented, chemical reaction relationships involving more than two molecules may not be represented, and generally any relationships, where the arity of an equivalent network representation is greater than two, cannot be directly represented. These limitations present a problem for representing health hypotheses, which, as will be discussed in Chapter 4, involve more than binary relationships between concepts. DL’s also do not support representation and reasoning needed to represent and traverse a non-tree, network of domain knowledge or related hypotheses and their supporting information.

Although DL’s have significant limitations, in many other respects, DL’s, as a knowledge representation approach, have advantages when representing and working with health hypotheses. The DL SNOMED-CT, provides excellent access to medical terminology representation and reasoning. The subsumption reasoning approach supported by SNOMED-CT and other DL’s is very appropriate for approximate matching of concepts in an “is-a” hierarchy. For example, matching of a patient profile to clinical trial inclusion criteria, may involve determining whether the patient has a condition that is subsumed by the inclusion criteria, even if the profile descriptions for the clinical trial and the patient are different. The representation of
project information, as well as hypothesis test strategies, can be represented by DL concepts and roles. The need for numeric comparison operators exceeds the capabilities of the EL-expressible SNOMED-CT but can be fulfilled by other more expressible DL’s. Research in DL’s has demonstrated correspondence between DL’s and first order logic, rule-based systems, and graphs, which may enable integration or extension with these other representations [90], [91]. In addition, DL’s have been shown to be amenable to compilation into more computationally tractable languages, such as Ordered Binary Decision Diagrams, that enable efficient query answering in very large knowledge bases. While DL’s as a concept have promise for representing health hypothesis, the problem is that no single DL is sufficient.

3.5 **HRE knowledge representation approach - HDL**

The common representational format of DL’s, the diverse representational needs of the HRE, and the issue of tractability for large scale query answering, suggest a new, innovative approach. Instead of selecting a single DL language, and force-fitting it to all HRE requirements, a problem decomposition approach is taken. The HRE representational approach, which will be referred to as Hybrid Hypergraph Description Logic (HDL), provides a unified framework for connecting concepts in diverse knowledge bases. By decomposing the HRE representation problem into sub-problems, or domains, each with specific representational characteristics, the right level of language can be applied where needed, yet all within a common syntactic and semantic framework. By defining a hypergraph overlay, the capability is added to define arbitrary relationships, which connect concepts in different DL’s. By introducing Boolean formulas as a means to integrate diverse functions, the capability is added to answer queries across different
knowledge representations.

### 3.5.1 HDL Hypergraph overlay and hyperroles

Conceptually, a hypergraph overlay approach adds “seas of connectivity” between a small number of “islands of tractability”. An analogous approach has recently been used in social network analysis, where hyperedges connecting people’s diverse social networks are used to connect diverse networks with different properties, so that connectivity between elements in several different social networks can be analyzed, and new questions answered [92]. This hypergraph connectivity, applied where needed, can provide new query answering capabilities for hypotheses. For instance, it can be useful in evaluating whether a hypothesis concept is linked to a concrete measurable, or which hypotheses use a particular “leads to” construct as rationale for a health research hypothesis.

Recalling, the set-theoretic interpretation of Description Logics, we see that DL’s define concepts as sets of individuals and roles as sets of ordered pairs of individuals, where the individuals are from a specified domain. Compare this to the hypergraph formalism, where a hypergraph is a generalization of a graph.

**Definition. Hypergraph.** A hypergraph, $H$ is a pair $H=(V,E)$, where $V$ is a set of elements, or nodes, and $E$ is a set of non-empty subsets of $V$ called hyperedges or links [93].

This research proposes a DL hypergraph extension, which allows roles that are not limited to sets of ordered pairs of individuals, i.e. the concept of role is extended to sets of bounded sets of individuals. Thus, hyper-roles in HDL are hyperedges that are non-empty subsets of at most n
individuals. Concepts in DL are expressed as trees; hence HDL hyper-roles are hyperedges between trees. HDL hyper-roles may also be typed and include properties such as direction.

**Definition. HDL Hypergraph.** A HDL Hypergraph \( H = (V, E, L, Y) \) is a labeled hypergraph where:
- \( V = \{1, \ldots, n\} \) is a nonempty set of concepts, expressible as trees,
- \( E \subseteq V_1 \times V_2 \times \ldots \times V_n \), where \( E \) is a set of hyperedges, where each hyperedge, \( e \in E \), is a hyperrole
- \( L \) assigns a label to each vertex, \( V \)
- \( Y \) assigns a set of properties, or default types to the hypergraph.
  Types include direction, arity, and DL type.

**Definition. HDL Hyperrole.** A HDL hyperrole, \( e \), is a set \( e = \{l, r, y\} \) where:
- \( r \subseteq V_1 \times V_2 \times \ldots \times V_n \),
- \( l \) assigns a label to each hyperrole, \( e \),
- \( y \) assigns a set of properties, or types to \( e \), that may over-ride the default.
  Types include direction, arity, and DL type.

**Definition. HDL Directed Hyperrole.** A HDL Directed Hyperrole is similar to a directed graph edge in that each hyperrole, \( e \), is defined with \( r = (T, H) \) with the tail, \( T \), being the start node, and the head, \( H \) being the finish node, and \( T \) and \( H \) being subsets of the vertices \( V \).

For example, in a network of chemical reactions represented as [94]:
- (1) \( A \rightarrow B \),
- (2) \( A \& B \rightarrow C \& D \),
- (3) \( E \rightarrow E \)

reactions, such as (1) and (3) may be represented edges in a directed graph, but reaction (2) must be represented as a hyperrole, \( e = \{l, \{\{A, B\}, \{C, D\}\}, y\} \). In chemistry, reactions are also associated with stoichiometric coefficients that indicate the strength of the reaction. Hyperroles may be similarly represented with coefficients that indicate strength of the relationship. In this case, each \( r \) would be defined as \( r = \{T, Ct, H, Ch\} \) where \( Ct \) is the coefficient for the relationship starting from \( T \), and \( Ch \) is the coefficient of the relationship starting from \( H \).

A set of hyperroles, \( E \), for a set of vertices, \( V \), may be expressed as an incidence matrix, where
the rows correspond to the vertices, and columns correspond to the hyperroles. The matrix contents assigned the value of 0 or 1, with 1 corresponding to the participation of a concept in a hyperrole [95]. Adjacency matrices can be constructed that indicate the number of times that a concept participates in a hyperrole. Several hypergraph properties have been studied by leveraging incidence and adjacency matrices. Many operations analogous to graph operations have been defined on hypergraphs. Connectivity operations defined include: determining all the nodes connected to a vertex V, determining whether a vertex V1 is connected to a vertex V2, and determining all the nodes connected to a vertex V1, through V2 [95]. These operations return paths corresponding to a tree rooted at V. Hypergraph operations are usually less computationally tractable than graph operations, but many cases exist where problems are easy to solve.

3.5.2 HDL Boolean expressions and queries

HDL supports Boolean expressions, where the variables are HDL atomic sub-formulas evaluable to \{0,1\} connected by the Boolean operators: and, \(\wedge\), or, \(\vee\), not, \(\neg\), and implies, \(\rightarrow\). Legal atomic sub-formulas are defined as follows:

- Role of type number: role.value \{=, >, <\} numeric constant
- Role of type string.value = string constant
- SNOMED-CT concept: A \(\subseteq\) B (evaluable by a SNOMED-CT classifier)
- Arbitrary DL concepts of the same type, A \(\subseteq\) B.
- Boolean variable: \{0,1\}
- DL variable: A DL term evaluating to the empty set is interpreted as 0, and 1 otherwise.

Queries can be created as Boolean expressions on these sub-formulas. Note that Boolean expressions, converted to CNF, can be converted to hypergraphs or directed hypergraphs [96]. Boolean expressions may also be converted to a tree structure, Ordered Binary Decision Diagram
(OBDD), for efficient evaluation. OBDD’s, discussed in depth in Chapter 6, may be represented as directed hypergraphs of arity=2, which connect concepts through directed edges that are assigned one of two types – regular arc and complement arc. OBDD’s may be represented as trees of concepts connected by DL roles. OBDD’s, which are canonical for a specific ordering of variables, have the property that each node is associated with a function, and every node below is associated with a sub-function of that function [97]. Thus, it is possible to determine whether one function subsumes another within an OBDD.

The purpose of Boolean Formulas in HDL is to define and answer queries that cannot be answered within the context of a single DL representation, such as SNOMED-CT, or through connectivity analysis in hypergraphs. An example of this type of query is the evaluation of inclusion and exclusion criteria for clinical research populations, where criteria include both medical and non-medical parameters, as well as numeric comparisons not representable or evaluable in SNOMED-CT, and where it may be important to know whether one patient population subsumes another.

3.5.3 HDL Knowledge Bases

**HDL Knowledge Bases.** Each model, I, of a knowledge base, K, has three parts – DL axioms, hypergraph structures, and Boolean formulas. DL axioms correspond to tree structures with concepts connected through DL roles. In HDL, a DL role may also be represented as a directed hyperedge, with arity = 2, and assigned a type that corresponds to its DL language. The assignment of types to hyperedges in HDL allows disconnected tree-structures to be constructed from different DL languages. Using this convention, a set of related concepts may be expressed
with the most tractable DL required to represent what is needed. The hypergraph connects tree-structures that may not be connected in a single DL.

In HDL, the TBox and ABox conventions in Description Logics are used, and HBox and BBox conventions are added. The HDL knowledge base, \( K = (T, H, B, A) \) is a 4-tuple where (i) \( T \) is DL TBox, (ii) \( H \) is a hypergraph HBox with \( n \) vertices, (iii) \( B \) is a BBox set of Boolean formulas, and (iv) \( A \) is an ABox that can also contain hypergraph assertions of the form \( A(a_1, \ldots, a_l) \) for \( a_i \in N \). The DL TBox defines the concept hierarchy and tree roles. The ABox defines instantiations of the TBox.

Consider the following HDL axioms and their English translation (in italics). Let \( K = (T, H, B, A) \) be a HDL knowledge base, where \( T \) is the TBox, \( H \) is the HBox, \( B \) is the Boolean Expression Box, and \( A \) is the ABox.

**HDL TBox Terminology Expressions.** The TBox represents hierarchies of concepts. \( T \) contains the TBox axioms.

Tai Chi \( \sqsubseteq \) Exercise  
*Tai Chi is a kind of exercise.*

ParkinsonsDisease \( \equiv \) Disease \( \cap \exists\) involves_anatomy.Nervous_System \( \cap \forall\) has_abnormality Dopamine_Generating_Cell_Death \( \cap \exists\) has_symptom.Impaired_Balance  
*Parkinson's Disease is a disease that involves at least the nervous system, and is always characterized by death of dopamine generating cells, and which has as one of its symptoms, impaired balance.*

ParkinsonsPatient \( \equiv \) Person \( \cap \exists\) has_condition.ParkinsonsDisease  
*Parkinsons Patient denotes the set of individuals that are instances of the concept Person who are connected through the role has_condition to some instance of the concept ParkinsonsDisease.*
TailoredTaiChi ≡ Tai Chi ∩ ∃targeted_function.Balance ∩ ∃targeted_participant.ParkinsonsPatient

Tailored Tai Chi is a Tai Chi that is connected through the role targeted_function to some instances of the concept Balance and through the role targeted_participant to some instances of the concept ParkinsonsPatient

G ≡ ParkinsonsPatient

A ≡ Person ∩ ∃exercise_participation.TailoredTaiChi

A denotes the set of people who participate in at least one exercise, Tailored Tai Chi

B ≡ Person ∩ ∃exerciseParticipation.TaiChi ∩ ∀Balance.Improved

B denotes the set of people who participate in Tai Chi who have improved balanced

C ≡ Person ∩ ∃exerciseParticipation.TaiChi ∩ ∀Balance.Improved

C denotes the set of people who participate in Tai Chi who have reduced falls

**HDL HBox Hypergraph Expressions.** Now that we have defined the DL tree concepts, we want to link these together to state a new relationship: Targeted Tai Chi leads to improved balance and reduced falls in Parkinsons Patients. Note that the term “in Parkinsons patients” implicitly refers to the entire preceding sentence. To be unambiguous, we restate this as: Targeted Tai Chi in Parkinsons Patients leads to improved balance and reduced falls in Parkinsons Patients.

This cannot be stated in the TBox, since the leads-to relationship may involve more than two concepts, and consequently is represented as a hyperrole, which is not defined for a TBox. In K=(T, H, B, A) H contains the HDL axioms. We use the hyper-role concept to denote the “A leads to B and C” relationship: e1 = (leadsTo, {{A ∩ G}, {B ∩ G, C ∩ G}}, DH), where DH designates an edge type, directed hypergraph. While it is possible to designate a new concept to denote A ∩ G, we purposely leave G factored out, so as to explicitly state that this is the
population context, or set of individuals, in which the “leads to” relationship applies. \( e_1 \) may also be stated as: \( e_1 = \langle \text{leadsTo}, \{G \cap \{\{A\}, \{B, C\}\}\}, \text{DH} \rangle \). Expressions such as \( e_1 \) will later be used to represent components of hypotheses, and expressions such as \( G \) will be used to represent populations for which hypotheses are valid.

**HDL BBox. HDL Boolean Expressions.**

The BBox may be used to describe a Boolean expression representing a concept that cannot be described by a single DL. For instance, a simple population definition might be represented as:

\[
\begin{align*}
\text{Population} & \equiv \\
& \text{isPerson} \land \\
& \text{has\_disorder.ParkinsonsDisease} \land \\
& \text{severityHoehnYahrScale.(>,4)} \land \\
& \text{hasAge(>=,40)} \land \text{hasAge(<=85)}. \\
\end{align*}
\]

**HDL ABox Assigning Individuals.** In addition, in \( K=(T,H,B,A) \) ABox axioms assign individuals to concepts and roles:

\[
\begin{align*}
\text{mary:ParkinsonsPatient} & \quad (1) \\
\text{nicholas:Person} & \quad (2) \\
\text{exerciseParticipation(nicholas,TaiChi)} & \quad (3) \\
\text{hasCondition(nicholas,ParkinsonsDisease)} & \quad (4) \\
\text{hasCondition(mary,ParkinsonsDisease)} & \quad (5) \\
\text{hasAge(mary,49)} & \quad (6) \\
\end{align*}
\]

Using the hypergraph representation, we may also represent (3) as:

\( e_2 = \langle \text{exerciseParticipation}, \{\{\text{nicholas}\}, \{\text{TaiChi}\}, \text{DL}\rangle \), where SCT designates an edge type, “DL”.

**HDL Queries.** Within HDL knowledge bases, the allowable operations and properties of each DL language may be applied to axioms in that specific DL language. DL’s support the reasoning
tasks: subsumption, instance checking, relation checking, concept consistency and knowledge base consistency. Hypergraph traversal operations are defined that enable answering of queries regarding connectivity/reachability, common subgraphs, and path length. Boolean formulas may be used to define queries spanning multiple DL’s.

3.6 Conclusions

The NIH clinical trials database is a comprehensive repository of clinical trial projects. However, it does not adequately support the representation of health hypotheses or the need for earlier engagement of health consumers in health research conceptualization and development. This research addresses the following shortcomings of the NIH clinicaltrials.gov database: 1) the current representational framework does not meet the requirements for enabling consistent representation and ease of creation of hypotheses, and 2) the current framework misses opportunities to engage health consumers in suggesting new ideas and important measures, and 3) the current framework does not make productive use of its diverse and comprehensive medical knowledge. The architecture for a Health Research Exchange (HRE) was developed to address these shortcomings. The vision is that by interacting with the HRE, both health consumers and health researchers can propose health research ideas, critique health research ideas in formation, and suggest health measures that are meaningful to them.

This chapter makes three significant contributions: 1) it defines the stakeholders and requirements for a collaborative health research exchange, 2) it defines an architecture for the HRE, and 3) it defines a representation approach, Hybrid Hypergraph Description Logic, HDL,
that expands the definition of Description Logics, to include hyperroles, i.e. to include hyperedges that link more than two concepts, and Boolean formulas. This will enable representation of complex concepts, such as health hypotheses, and will enable connection and traversal between diverse knowledge bases, where each knowledge base may be represented by a description logic, graph, or hypergraph with defined properties. The advantage of this approach is that it can facilitate the use of the best Description Logic, or other formalism, for a knowledge base subset while maintaining connectivity across the entire knowledge base. HDL will be leveraged to create a health hypothesis framework that facilitates representation and reasoning about health hypotheses incorporating realistic medical terminology.
CHAPTER 4    Defining Health Hypotheses

In this chapter, we propose a model for health hypotheses that draws from Karl Popper’s requirements for scientific hypotheses, and from observations of hypothesis formation and analysis in clinical research settings. It also draws from statistical education models for teaching hypothesis formation, as well as ongoing work at UCLA and other organizations that equate hypotheses with semantic webs of information. This approach is based upon the premise that a structured, re-usable representation of health research ideas will enable more people to understand and participate in health research. The proposed representation for health research ideas is a decomposable model of scientific hypotheses, where the hypothesis is defined not only by its conceptual content, but by its connectivity, and specialized functions. The basic concept is that scientific hypotheses have specialized parts and descriptors that act to define the hypothesis, and that these parts work together: 1) to describe the basic ideas in the hypothesis, 2) to connect the hypothesis to the body of available knowledge, and 3) to specify how the hypothesis will be evaluated, or tested. The decomposition of the hypothesis into standard parts and descriptors enables a new paradigm of hypothesis construction, which facilitates the reuse of existing hypotheses, and hypothesis components called Domains. The proposed approach of providing a structure and reuse methodology for hypothesis development is expected to expedite development of hypothesis predictions that are amenable to being tested and evaluated. Connecting the hypothesis to other bodies of knowledge is expected make hypotheses more useful and understandable. Managing the hypotheses as connected, but unique, entities with descriptors enables answering of many new types of queries.
This chapter develops and illustrates an information model, the Molecular Hypothesis Model (MoHM), which enables the communication and sharing of Health Hypotheses and supporting information, among the many stakeholders in health research. Built upon the Hypergraph Description Logic (HDL) described in Chapter 3, the MoHM meets the proposed requirements for a scientific hypothesis: Predictive Utility, Plausibility, Testability, Falsifiability, and State. Inspired by biomolecular models such as proteins, the MoHM is based upon the principles of modularity, connectivity, and domain specificity. Through application of these principles, the MoHM meets the technical challenge of the development of an information abstraction for health research ideas, expressed as hypotheses, which can interface with heterogeneous data and inference methods in medical domains at multiple levels of abstraction. In addition it provides the foundation to address the requirement of incremental specification of hypotheses, such that they can be productively shared from inception to full specification.

Section 4.1 describes foundational work in hypothesis models. Section 4.2 develops the Molecular Hypothesis Model concepts. Section 4.3 presents the MoHM representation, and its Domains. Section 4.4 summarizes the MoHM.

### 4.1 Hypothesis models

#### 4.1.1 What is a Hypothesis?

There are many different definitions and models in the literature for the concept of a hypothesis, each definition depending upon the context in which it is used. Some definitions are **structural**, describing the form of a hypothesis. In logic, hypotheses may be viewed as formal logical propositions, evaluable to true or false [98]. In statistical analysis, hypotheses are statements
about relationships between variables or data [99]. In semantic web research, a hypothesis takes on the form of a web of connected arguments, represented by a graph structure. Prior work in hypothesis models at UCLA, develops the idea of a scientific hypotheses as a web of related arguments and information [100] drawing upon prior work by Bush [101], Whorf, [102] Engelbart [103], Nelson, [104] Kuhn [105], and Quine [106] and others who contributed to the vision of the internet and its interconnected web of knowledge. In this paradigm, a web of related arguments and information are represented as graph structures that can provide “a “skeleton” or “scaffolding” that can be annotated with additional associations to the scientific literature”, and evolve into more rigorous descriptions and models.

Other hypothesis definitions are functional, describing the purpose and the role of a hypothesis in some process. For example, in scientific research, a hypothesis can be defined as a “proposed explanation of observable phenomena” [107] where the explanation has some basis in reality [108]. In scientific research, hypotheses participate in the scientific method, a process by which phenomena are observed, hypotheses are formed that explain the phenomena, and tests are executed which are designed to test the hypothesis. Another functional definition, advocated by Karp [109] in his model of hypothesis formation in molecular biology, proposes that the development of hypotheses is the elimination of differences between prediction and observations, and thus can be treated as a design problem.

Hypotheses may also function as a general tool for thinking. The term hypothesis is derived from a Greek term hypotithenai meaning “to put under” or “to suppose” [110]. Hypotheses have been used at least since the time of the ancient Greeks [111] to explore the ramifications of a
supposed state being true. Hypotheses can be very powerful. They may affect the behavior or thinking of a person, and may color our way of thinking about the world and gathering data. An interesting contrast is the term hypothetical, also from the same root. A hypothetical is a statement that is used as a thinking tool, to explore the ramifications of the statement being true. Unlike a hypothesis, a hypothetical is not necessarily linked to reality, nor is it testable [112]. Like hypotheses, hypotheticals, when expressed as questions, have been shown to impact behaviors, especially in the political arena [113].

Finally, some hypothesis definitions propose properties that hypotheses must possess in order to be meet the basic requirement of a hypothesis. This is the approach taken by Karl Popper, when he proposed that a hypothesis must make a prediction [114]. He stated that in making a prediction, a risk must be taken on, and the more risk that a prediction has, the more powerful the hypothesis. This property will be referred to in this work as predictive utility. Additional properties proposed by Karl Popper are that hypotheses must be testable and falsifiable. The property of testability means that for a prediction made, there is a possible approach for testing it. The property of falsifiability means that, in addition to the testability property, it must be possible to show that the prediction can be proven wrong.

The falsifiability property supports an important pattern of thinking in scientific research. While it is not possible to prove that a statement is true in all cases, it may be possible to prove that a statement is false in at least one case. For example, if there is a hypothesis that states: “Treatment X reduces pain,” it is not possible to prove that Treatment X reduces pain in all cases. However, it may be possible to prove that, “for a defined group of people, Treatment X
may make a difference in the pain felt, that cannot be explained by chance alone.” Since it is possible to prove that a relationship can be false, but not possible to prove it true, a scientific hypothesis is often presented as a statement, known as the null hypothesis, which proposes that there is no effect. A second statement describing the proposed effect is known as an alternative hypothesis. If the null hypothesis is shown to be false, then the alternative hypothesis is assumed to be true. In actuality, a hypothesis can never be proven true, since not all related predictions can be tested. Instead, hypotheses that have been tested and not proven false, remain active as working hypotheses, assumed to be true for the time being. In the example above, if the null hypothesis is shown to be false by showing that a difference in pain cannot be explained by chance alone, then the alternative hypothesis that “Treatment X reduces pain” is supported. In order to simplify discussion, in most cases, the hypothesis being referred to in this document, is not the null hypothesis, but the alternative hypothesis.

In addition to the predictive utility, testability, and falsifiability properties, others have discussed the requirement for plausibility of a hypothesis. Plausibility, informally, means that the hypothesis is consistent with generally known facts and theorems and what we know of the world. Chrysippus, a great Stoic philosopher credited with creating a system of propositional logic [115], made an important point, relevant to plausibility and the importance of linking to prior causes. According to Chrysippus, every proposition is either true or false, and this must apply to future events as well. In addition,

“If any motion exists without a cause, then not every proposition will be either true or false. For that which has not efficient causes is neither true nor false. But every proposition is either true or false. Therefore, there is no motion without a cause. And if this is so, then all effects owe their existence to prior causes.” [116]
If one interprets the Chrysippus point of view with respect to hypotheses, his observation would imply that any hypothesis that predicts an effect, and that is falsifiable, would be dependent on the existence of prior relevant causes. The stated dependency on prior relevant causes is one aspect of what we will refer to as the plausibility property. Plausibility as a property of hypotheses has been studied experimentally in diagnostic problem solving systems [117]-[118].

The many different ways of defining hypotheses offer interesting perspectives, but make the hypothesis difficult to define precisely. However each perspective adds information that aids in understanding the hypothesis, as a concept.

4.1.2 The Hypothesis formation process

Not only is it difficult to define precisely what a hypothesis is, but the act of developing a hypothesis is a very complex, and difficult task. The book *Medawar’s Advice to a Young Scientist* takes the point of view that the hypothesis is a creative act:

"The truth is not in nature waiting to declare itself, and we cannot know a priori which observations are relevant and which are not: every discovery, every enlargement of the understanding begins as an imaginative preconception of what the truth might be. The imaginative preconception – a 'hypothesis' – by a process as easy or as difficult to understand as any other creative act of mind; it is a brain-wave, an inspired guess, the product of a blaze of insight. It comes, anyway, from within and cannot be arrived at by the exercise of any known calculus of discovery. A hypothesis is a sort of draft law about what the world – or some particularly interesting aspect of it – may be like; or in a wider sense it may be a mechanical invention, a solid or embodied hypothesis of which performance is the test." [119]

This is one perspective, and should be carefully considered when developing a computational model of hypotheses. However, it is possible that some hypotheses may address incremental improvements in understanding, and may be created by modifying, expanding, or constricting, previous hypotheses. It is also possible that some “blazes of insight” may be replicated by some
application of transformational rules, heuristics, or patterns of prior hypotheses formation. Gettys and Fisher [120] propose a model of hypothesis generation that triggers when currently held hypotheses are rendered less probable by new data. They developed a plausibility estimation process that suggested that new hypotheses are generated when their plausibility is high enough to make them competitors for the currently active hypotheses.

In the area of education, there is a variety of guidance in writing hypotheses. These expose some of the problems inherent in developing a good hypothesis, and can be valuable in developing methods for structuring hypothesis writing. The National Health Museum facility for Health Education, Access Excellence, established in 1993, has published an on-line student lesson on writing hypotheses [121]. In this lesson, it proposes that informal hypotheses may be written as sentences such as “Salt in soil may affect plant growth”, or “Ultraviolet light may cause skin cancer.” The word “may” is important, since it indicates a tentative relationship. However, the student lesson proposes that a more formally written hypothesis is needed and that should be a sentence that is in the form of an IF-THEN statement. An example hypothesis is “IF skin cancer is related to ultraviolet light, then people with more exposure to ultraviolet light will have a higher frequency of skin cancer”. According to this method, a more formal expression of the hypothesis, the IF portion of the sentence must express some type of relationship, and the THEN part of the statement must have a prediction that can be tested.

Another common theme is that a hypothesis must go through a process that includes a series of activities, ending with the testing of the hypotheses before it becomes a generally accepted theory. This process is often referred to as the Scientific Method. The Scientific Method has
two major sub-processes, hypothesize and test. These processes are highly dependent upon each other. During the **hypothesize process** the following actions often occur: 1) observe phenomena, 2) link observations to existing knowledge to assess how the phenomena may be explained, 3) create a new hypothesis, possibly a creative “aha!” moment, and 4) develop an approach to test the hypothesis. The hypothesize process represents a moderate amount of time and work, is often iterative, and may be difficult to communicate during the process, but the quality of the thinking has a big impact on the following phase, the test process. During the **test process**, the following activities occur: 1) Prepare the testing environment, 2) perform the test and gather appropriate data, 3) analyze the data, and 4) assess the hypothesis. Relative to the hypothesize process, the test process generally entails a much larger amount and duration of work. It is also more amenable to communication, since it is more concrete. Unfortunately, if the planning that takes place during hypothesis formation is poor, much of the work of the test process can go to waste.

One limiting characteristic of the hypothesis is that it is difficult to communicate goals and tradeoffs during hypothesis formation. A hypothesis is often a “black box” until the test strategy has been laid out. An example of a hypothesis development process is shown in Figure 4.
With this black box approach, where the hypothesis is not fully understandable until there is a test strategy, it is usually not possible to know, within a research community, the answers to questions such as: Who is working on what hypothesis? Who will need what data? What research is under consideration? What concepts are being developed or tested, in what populations?

Sometimes, in the case of clinical research, these questions may not be answerable until the test strategy for a research hypothesis is complete, in the form of an approved and distributed protocol. Another drawback of the black box approach to hypothesis development is that, without early patient participation in the hypothesis formation process, clinical research may not address relationships between parameters important to patients, i.e., predictable effects or measure that really make a difference in their well-being or ability to adhere to a proposed therapeutic regimen.
Since there are no centralized sources for determining which hypotheses are being developed and considered within and across research communities, parallel in-depth planning for health research in a patient population, addressing the same questions, may go on for months, without researchers knowing about the overlap. Furthermore, potential subjects, who might be interested in participating in a clinical study, often won’t learn about the health research until a study is ready to enroll subjects. Sometimes the enrollment period is already over before patients learn about a clinical trial of interest. Sometimes qualifying for clinical research requires meeting criteria such as abstinence from specific medications, procedures, or behaviors for a period of time prior to entering a research study. Without advance knowledge of research planning in progress, patients will be unable to prepare themselves to increase their chances for participation in relevant clinical research. The end result is that both collaboration and clinical research participation suffer, because the hypothesis formation thought processes are nearly invisible to the people who might be interested in participating – other researchers and potential research subjects.

The process of hypothesis formation in clinical research, in practice, often relies on the reuse of existing hypotheses or parts of hypotheses. For instance, the creation of hypotheses pertaining to health research may follow commonly accepted use and reuse patterns, such as: 1) Plausibility of a hypothesis may rely upon a clinically meaningful relationship that is currently used and accepted by the medical and regulatory communities. 2) The test of the hypothesis may include outcome measures that are commonly utilized in a field, or for a specific health condition. 3) The hypothesis evaluation strategy may rely on commonly accepted methods of statistical
analysis for each field, or medical condition, and 4) Hypotheses that address related topics often refer to the same or similar prior work. Because of the potential to reuse significant prior work during hypothesis development, it is not uncommon for clinical research protocol writers to use templates, or even copy-paste-modify development methodologies during clinical research protocol creation.

Furthermore, hypotheses that may seem to be attributable to “aha!” moments can often be traced to common scientific patterns of thinking. For instance, if a treatment cannot be shown to be effective in a very wide population, then an attempt may be made to narrow the patient population to those most likely to benefit from the treatment. This “narrowing the population” pattern has been applied to genetic markers that may influence an outcome [122]. So while it may seem to be a “blaze of insight” when a scientist exclaims, “I hypothesize that the KRAS mutation has an effect on colon cancer progression in patients treated with drug A”, at the foundation of that “aha!” moment is a common model of reasoning that has been applied in many cases, and is a currently popular source of many health hypotheses. The KRAS mutation hypothesis takes a “narrowing the population” approach so that the effect on the population under study is more concentrated, and thus, more detectable. It also proposes a “genetic marker” that is a predictor of an outcome – another popular hypothesis pattern. Another type of common pattern of hypothesis generation is the Interventional Hypothesis pattern. In this pattern, there is a hypothesized pathway of reactions that leads to a health condition, and a hypothesized intervention, one that interrupts or enhances a part of that pathway. The interventional hypothesis pattern features a statement that the intervention has a positive or negative impact on the health condition, within a defined population. Similarly, this pattern can be used to generate
hypotheses about what elements participate in a pathway. In another type of hypothesis pattern, the Parameterized Difference pattern, there is a hypothesized difference in outcomes that is a result of a difference in conditions for groups within a population.

The point here is that, while hypothesis formation is currently a “black box”, there are patterns of information and patterns of thinking that are commonly used and re-used in practice. If we can identify common patterns of hypothesis formation, it may be possible to apply and reuse those patterns, as people often do, but do so within an organized, computational framework. This work defines a taxonomy of hypotheses, where families of hypotheses share common structure and characteristics, and where some of these common patterns of hypothesis formation and reasoning are leveraged to facilitate the hypothesis formation process.

4.1.3 The Management of Hypotheses

While the hypothesis is an important tool in science, there is little published work on management of large numbers of hypotheses. One relevant application was found in the context of information fusion of multi-vehicle tracking [123] where tracking is highly dependent on inferring the structure, behavior, and intent of a system from observations. In this work, the problem of hypothesis management is characterized as having two main components: 1) the situation hypothesis problem – describing model instances that best characterize the observations, and 2) the data association problem – determining correspondence between models and observed data. Two systems are described, TCT Reasoner and CADRE, that explicitly represent hypotheses as graphs, relating them to underlying concepts and observed data. The most interesting and relevant aspect of the CADRE system is that context hierarchies are used to
enable generalization of hypothesis trees, enabling reasoning at multiple levels of abstraction.

Another exploration of hypothesis management was in the context of situation recognition [124] using Petri Nets. In both of these cases, management methods were applied to sets of hypotheses, predefined for a particular class of applications.

4.1.4 The Elephant in the Research Lab

One of the most interesting aspects of the concept of the hypothesis is that, in spite of having been in use as a widespread and practical conceptual tool for thousands of years of human history, it seems to elude precise definition. In some ways, the story of the hypothesis parallels the Indian story of the blind men and the elephant. In exploring the contours of the elephant with their hands, one blind man, while holding the tail, thinks that the elephant is like a rope. Another, who wraps his arms around a leg, envisions a tree trunk, while a third who runs his hands along the ears, is sure that the elephant is like a curtain. In one version of the story, where there are six different blind men and six different viewpoints, the response from an observant wise man is this:

"All of you are right. The reason every one of you is telling it differently is because each one of you touched the different part of the elephant. So, actually the elephant has all those features, what you all said." [125]

Like the elephant, the multifaceted hypothesis may be difficult to understand because it may have many diverse, yet connected, structural and functional parts. One question that arises then, is it possible that a hypothesis has many or all of the features that hypothesis researchers have touched, explained, and explored? If so, which of these features are most important and in what contexts? How do those features work together to create an integrated whole?
Will we ever truly know what a hypothesis is? Possibly not. But we do know that more productive application of scientific research is sorely needed. With the development and testing of hypotheses on the critical path of scientific research, more efficient formation and management of hypotheses have the potential to make a meaningful difference. Is it possible that the need for computational support for hypothesis formation is the elephant in the corner of the research lab? Perhaps. Rate and quality of hypothesis formation appear to be unacknowledged and unaddressed factors that may be influencing the progression of scientific research.

If we can develop computational models that enable us to represent and reason, in an integrated way, about the multiple dimensions of a hypothesis, if we can quickly formulate and share hypotheses, and if we can use these models to make faster progress in our explanation and exploration of natural phenomena, we take one step closer to understanding the nature of hypotheses. In turn, computational models will enable us to leverage advances in information technologies, such as semantic web, efficient representation and query mechanisms, and collaboration strategies, to share and discuss hypotheses more widely and earlier in the process of hypothesis formation. This has the potential to lead to more productive use of scientific and health research resources.
4.2 The HRE Molecular Hypothesis Model (MoHM)

As argued in Section 4.1, the many different models of hypotheses may represent various dimensions of a single concept. This work combines several of these dimensions, in an integrated data representation, which will henceforth be designated the Molecular Hypothesis Model (MoHM). The MoHM is a system, defined by three principles – connectedness, modularity, and domain specificity. The connectedness principle states that hypotheses can be represented as patterns of connections that do not stand alone, but are connected to other hypotheses, facts and relationships. These connections provide a context in which the hypothesis can be evaluated and interpreted. The modularity principle states that all hypotheses can be decomposed into reusable, modular parts. The domain specificity principle states that the modular parts of hypotheses may have representations specific to the function of the individual parts. These principles work together, through the MoHM, to address the HRE requirements to represent a variety of types of information, to interface with heterogeneous representations in medical domains at multiple levels of abstraction, and to enable incremental, iterative construction of hypotheses. Implemented in the MoHM, these principles are intended to provide a foundation that will enable faster formulation and earlier sharing of health hypotheses.

The name “Molecular Hypothesis Model” is inspired by biochemical models in the following sense:

- **Molecules are modular** – aggregates of atomic elements; constructed from building blocks of molecular patterns, such as \( \text{CH}_2, \text{OH}, \text{C}_n \). Larger biological molecules are often built from repeating units of smaller molecular motifs, such as base pairs, sugars, fatty acids, and amino...
acids.

- **Molecules are connected** – within each molecule by atomic bonds, and across molecules by atomic and molecular interactions that are governed by the laws of physics. There are various types of bonds, or connections, each with their own properties. Traversing the elements in a molecule involve traversing over different types of bonds. An atom or molecule can be involved in more than one connection or interaction at once.

- **Molecular components are domain specific** – with different representations affecting the function to be accomplished. For example, each DNA molecule is a string of base pairs; proteins are assemblies of amino acids; fats comprise a single glycerol molecule linked to three chains of fatty acids. The molecular representation determines what operations can occur and how efficient these operations are.

Many molecular biology processes are ideally adapted for incorporating a variety of structures – interfacing with heterogeneous molecules for a variety of tasks and across multiple levels of biological abstraction – from atom to base pair to gene to molecular component. One exemplar participant in these processes is the protein. Protein structure is encoded in genes and constructed from amino acid building blocks within the cell environment. Proteins comprise stable domains, which often have distinct functions and structures. “In many cases a domain from a large protein may retain its correct three-dimensional structure even when it is separated from the remainder of the polypeptide chain” [126]. Protein domains may be combined and reused across many different proteins [127], “New sequences are adapted from pre-existing sequences rather than being invented”, [128]. Proteins with [similar sequence] and demonstrably similar function are said to be in the same protein family . . . with a strong evolutionary
relationships as evidenced by conserved structure and function [129].

The Molecular Hypothesis Model draws inspiration from the molecular biology world, exemplifying similar organizational and compositional principles. Similar insights from the field of molecular biology have been recognized by others in the field of computer science. In [130], Batory and Buchmann described a framework for “molecular objects, abstract data types, and data models, which treated collections of heterogenous tuples as objects, in order to provide different representations at different levels of abstractions. Their approach extends the concept of atomic aggregation of atomic objects to that of molecular aggregation of a set of objects.

Another description of a “molecular model” concept in information technology is the patent filed in Dec, 2006 by R.A. Stanley and E.A. Gombocz, [USPTO Publication number US 2002/0156756] [131] as “Intelligent molecular object data structure and method for its application. . . The molecular object is a “. . . software product invented to enable unified user presentation, accessing, routing, and functionally integrated processing of potentially (but not necessarily) diverse data.” This patent has been referenced by over 30 other patents filed by leading information technology corporations including IBM, salesforce.com, Microsoft, Yahoo, SAP, Oracle and HP. The theme shared by these molecular models is the need for dealing with heterogeneous data representations and methods. This theme is continued in the HRE Molecular Hypothesis Model.

Like proteins [126], the MoHM features an integrated structure with several functional domains, each with its unique function, properties, and structure. Without the domains, the hypothesis is not fully functional, just as an automobile without wheels or without an engine is not fully fit to
perform its function, or a protein without an important domain is unable to fold into a functioning structure. The MoHM domains work together, as a system, to create the behaviors and artifacts that are understood to be associated with hypotheses. Furthermore these domains can be reused to create new hypotheses or families of hypotheses with similar, but unique, structure and function.

In the following sub-sections, the MoHM will be described from three different perspectives: 1) the properties perspective, describing the hypothesis property requirements and their implications, 2) the functional perspective, describing the MoHM at a conceptual level, and 3) the structural perspective, describing the artifact itself, and its domains.

### 4.2.1 MoHM – Required Properties and their Implications

Drawing from prior work as described in Section 4.1, and building upon the approach of Karl Popper, we begin to define the MoHM by the properties, or attributes that it must possess. The following properties are a requirement of the MoHM: predictive utility, plausibility, testability, falsifiability, and state. Each property is analyzed with respect to its implications in defining a computational model for hypotheses.

**Property 1. Predictive Utility. Definition:** The Predictive Utility Property requires that the hypothesis must state a relationship, and a predicted effect that is a consequence of that relationship. **Implications:** It must be possible to state relationships, S, between concepts. It must be possible to express predicted effects, R. If S is true, then R must also be true, or the hypothesis is falsified. In addition, it must be possible to express the context, or including
populations, and timeframes, in which the predicted effect is expected to be valid.

**Property 2. Plausibility. Definition:** The Plausibility Property requires that the hypothesis is connected, through links, to other bodies of knowledge. The plausibility property can be assessed for the statement, S, the prediction, R, and the relationship between concepts in S and R. The degree to which the hypothesis is connected to pre-existing bodies of knowledge, is a measure of its plausibility. **Implications:** It must be possible to represent reference links such as citations and urls. It must be possible to connect concepts and to represent and traverse chains of concepts. It must be possible to connect and determine connectivity of hypotheses to other complex information structures.

**Property 3. Testability. Definition:** The Testability Property requires that there is a strategy defined for testing the prediction by generation of observable data that maps to a predicted effect. **Implications:** There exists at least one test strategy, associated with the hypothesis, which assigns groups for comparison, and parameters that can be controlled (independent variables), as well as parameters that are allowed to change (dependent variables). Testability implies that some parameters must be linked to measurable observables.

**Property 4. Falsifiability. Definition:** The Falsifiability Property requires that it is possible to assess whether a hypothesis is false. **Implications:** There exists at least one valid method, associated with the hypothesis, which has the function of determining if the hypothesis is false, by comparing the predicted effect to the observable data. This type of method will be referred, henceforth, as an analysis method. A valid analysis method is one that is consistent with the
number and type of groups and parameters, and that conforms to commonly accepted guidelines for statistical analysis. Parameters used in the analysis method must be connected to measurable parameters.

**Property 5. State.** **Definition:** The state of the hypothesis is a function of the degree to which the following properties are fulfilled: plausibility, predictive utility, testability, and falsifiability. **Implications:** Each of the other properties is a function of connectedness. Consequently, the overall maturity of the hypothesis is a function of the connectedness of concepts and relationships of the hypothesis to other concepts and bodies of information.

In summary, a hypothesis must include a plausible statement, S, and must make a prediction, R, that can be tested and falsified. In order for a hypothesis to be plausible, it must link to other known facts, relationships, or working hypotheses. In order for the prediction to be falsifiable, there must be defined a method or strategy, or plan for attempting to prove it wrong. A fully formed and complete hypothesis must meet all of the requirements. However, during the process of hypothesis design, a hypothesis may be at incrementally increasing states of maturity that are a function of its knowledge connectivity.

### 4.2.2 MoHM – Functional Perspective

Cognitive artifacts are defined as "those artificial devices that maintain, display, or operate upon information in order to serve a representational function and that affect human cognitive performance." [132]
Within the MoHM, the hypothesis is defined as a cognitive artifact that has the function of enabling human cognitive performance in the task of predicting effects based upon prior or existing plausible relationships.

This section describes how the hypothesis functions as a cognitive artifact. The following definitions are proposed.

1. A **parameter**, P, can be expressed as an attribute or role of an object.
2. A **population**, \( G_P \), can be expressed as a set of objects having measurable parameters.
3. A **group**, G, is a subset of a population. A **group parameter set**, \( P_G \), is a set of parameters that partitions a population into disjoint groups.
4. A **concept**, C, is a mental model, or cognitive artifact that denotes a physical or mental object or condition. In MoHM, a concept, C, can be expressed as a DL tree construct, \( C \equiv P \cap T \cap G \)
   where T and G provide a context for P:
   - \( P \) : a set of parameters, or attributes, or DL-roles
   - \( T \) : a set of time points or time intervals
   - \( G \) : a group or population of objects
5. An **observable**, O, is a concept that can be observed or measured.
6. A **hypothesis statement**, S, is a relationship between two or more concepts.
7. A **predicted effect**, R, is a relationship between two or more concepts, where each of the concepts in R has a relationship to a concept in S, as well as connectivity to a defined set of observables.
8. A **hypothesis space** is the set of all hypotheses that can be defined for a set of concepts, where \( C \equiv P \cap T \cap G \).

Referring back to our academic hypothesis teaching example, consider the hypothesis: “IF ultraviolet light is related to skin cancer, THEN people with more exposure to ultraviolet light
will have a higher frequency of skin cancer”. In this example S is “ultraviolet light is related to skin cancer, R is “people with more exposure to ultraviolet light will have a higher frequency of skin cancer.” The hypothesis is “IF S THEN R”. The hypothesis involves an n-way relationship between concepts in S and R.

A hypothesis, H, has the function, h: R → S, of mapping one relation, R to another, S, where if R is true, then S is not shown to be false, and if R is false then S must also be false. The hypothesis is falsified if it can be shown that S is false. The null hypothesis, N, from a statistical perspective, predicts that the observed data from the test of the prediction is compatible with data that could be encountered as a result of chance alone. The prediction, R, is a function of the null hypothesis, p: N → R. If the null hypothesis is shown to be true, then R is false, S is false, and the hypothesis is falsified. If the null hypothesis is shown to be false, then R is assumed to be true, S cannot be shown to be false, and the hypothesis, H, is supported and continues to be active.

A Population Health Hypothesis, PoH, is a hypothesis that is evaluated within the context of a specified population, and which attempts to falsify predictions for that population through statistical analysis on the data for that population. This chapter focuses on Population Health Hypotheses.

Hypothesis Types. The characteristics and mechanics of the mapping function, h: R → S, depend upon the type of hypothesis. In the MoHM framework, three major families, or classes, of hypotheses are proposed, corresponding to three different types of predictions. These are the
Difference Hypothesis, the Correlation Hypothesis, and the Emergent Hypothesis. Because of its widespread usage in health research, as well as the existence and usage of a large number of well-documented and relevant test and analysis strategies, this work focuses on Difference Hypotheses. This work may be extended to other types of hypotheses as future work. The three types of hypotheses are described below.

1. A **Difference Hypothesis** predicts a difference in parameter values that can be detected either across time points or across groups under specified conditions. A difference hypothesis often works by “flattening” the hypothesis space, i.e., by holding some of the concept elements constant. For instance, a difference may be proposed in a parameter, P, across two different groups, G₁ and G₂, at a time, T. Or a difference may be proposed in P across several time points, T₁, T₂...Tₙ, for the same group. Within the class of Difference Hypotheses, there are several variations on this hypothesis space “flattening” approach. Two that will be addressed in this dissertation include the Interventional Hypothesis and the Parameterized Difference Hypothesis.

2. A **Correlation Hypothesis** predicts a correlation, or association, between parameters, time points, or groups that makes them “statistically dependent”.

3. An **Emergent Hypotheses** predicts the existence of groups, such as clusters or hierarchies. These groups are usually a function of P. A sub-type of emergent hypothesis is the **Membership Hypothesis**. If a group is defined by a set of parameters, then membership of an object in a group is defined by closeness of fit to the value of those defining parameters. A test of the hypothesis involves determining whether the object meets the criteria for group membership.

The evaluation of a Population Health Hypothesis, of the type Difference Hypothesis, is
accomplished through the use of statistical operators. Statistical operators operate on data representing observed or derived concepts. They work by determining whether proposed features or differences in a dataset may be achieved by chance alone. The type of statistical operator that can be applied depends upon the types of parameters, time points, or groups that are involved in the relationships. The existence or lack of accepted statistical analysis methods constrain the falsifiable predictions that can be made. In other words, if there is no statistical method that can be used to evaluate the test of a prediction, then the Difference Hypothesis cannot be falsified. Often, with some thought and parameter transformation work, a prediction can be modified such that it is amenable to evaluation by available and accepted statistical analysis methods. If a prediction has been tested, and is not shown to be false, then the corresponding hypothesis is supported, and the hypothesized relationship may be used in predicting related effects.

4.2.3 MoHM Structural Perspective

The properties of the hypothesis indicate a need for several types of connections, or links to diverse types of information. The Predictive Utility Property requires connections between the hypothesis statement, S, and the Prediction, R, and their related concepts. The Plausibility Property requires connections to other hypotheses, to references, and to existing knowledge networks. The Testability Property requires connections from the hypothesis concepts to definable groups and to measurable parameters. The Falsifiability Property requires connections from definable groups and measurable parameters to analysis methods. The State Property is a function of the connectivity of the concepts in each domain. These connections may be one-to-one, or many-to-many.
The important role of connectivity in a hypothesis leads to a conceptualization of the hypothesis as a hypergraph structure, with concepts as tree-nodes, consistent with the Hybrid Hypergraph Description Logic (HDL) introduced in Chapter 3. This research proposes that the hypothesis can be represented as a decomposable, composite hypergraph structure with separate, but related, components, or domains, that fulfill its functional and qualitative requirements. These domains must support efficient reasoning for the functions that they support. The following domains are proposed – Project, Prediction, Population, Plausibility, Test Strategy, and Falsification. 1) The **Project Domain** includes the state of the hypothesis as well as descriptive information about the hypothesis, such as hypothesis type, authorship, and key dates. 2) The **Prediction Domain** represents the hypothesis statement, the prediction, and the relation between the prediction and the hypothesis statement. 3) The **Population Domain** contains population descriptors that define the range of the prediction, and which can be used for efficient matching to health subject populations. 4) The **Plausibility Domain** links the hypothesis to other bodies of knowledge, including links to concept definitions, domain concept networks, references, and urls. 5) The **Test Strategy Domain** describes the path from concepts in the prediction to measureable parameters, as well as the groups that are to be compared. 6) The **Falsification Domain** includes analysis strategies that enable the hypothesis to meet the requirement of falsifiability.

Figure 5 depicts the conceptual structure of the hypothesis, and for each domain, its underlying HDL representation, and contents. During the incremental development of the hypothesis, some parts of the hypothesis may not yet be specified. Hypotheses that have identical hypothesis statements but different predictions are represented as different hypotheses, but may be
aggregated as a hypothesis set.

**Anatomy of a Hypothesis**

![Diagram of hypothesis structure]

**Figure 5. Conceptual structure and content of the MoHM Domains**

### 4.2.4 MoHM Working Model representation

For every hypothesis and associated hypothesis structure, there are two forms of representation – the Working Model Representation (WMR) and the Packaged Model Representation (PMR).

The reason for this is that there are two, very different, requirements – 1) the need to rapidly and incrementally construct hypotheses, and to reuse hypothesis information from multiple components and multiple sources, and 2) the need to share hypotheses across diverse platforms where the hypothesis components must travel together to their destinations. The WMR for a hypothesis is a set of statements that describe the hypothesis, where each statement is labeled with a hypothesis ID, and multiple hypotheses may be concurrently in use in the WMR. Each
WMR statement is expressed either as a first-order-logic equivalent of a HDL statement, or as a template that can be transformed to HDL statements. Concept trees in HDL are represented as lists structures, consistent with SNOMED-CT and other DL’s. The next section describes the hypothesis domains in the context of the WMR. For brevity and clarity, the hypothesis ID is left out of each statement.

4.3 Representing Health Hypotheses with Domains

The MoHM domains exemplify the principles of connectivity, modularity, and domain specificity. Each domain represents a modular component of the hypothesis with a specific domain function and representation. Each domain is a composition of HDL axioms that together specify the domain. Each domain participates, through DL roles and HDL hyperroles, in a connected network of concepts. This section details the representations and operations for each domain in the Molecular Hypothesis Model.

4.3.1 The Project Domain – Representing the Hypothesis Process

The Project Domain includes descriptive information about the hypothesis, such as its state, type, ID, authorship, and key process milestones such as creation and modification dates. It may also include DL concept definitions to be used in the other domains. In many ways, the project domain is similar to the clinicaltrials.gov data model, which contains much project-related information about clinical research ideas that are fully formed. However the MoHM Project Domain may include information about the hypothesis at all stages of its formation.

The Project Domain features states representing the maturity of the hypothesis. The state of a
hypothesis is expressed as a tuple, HState(H, S₁, S₂, S₃, S₄), where H is the state value for the entire hypothesis, and each S element the state value for each of the properties – 1) Predictive Utility, 2) Plausibility, 3) Testability, and 4) Falsifiability. Hypothesis state rules, which can be customized to an organization, assign weights, w, to hypothesis properties to provide an overall state value for H, such that H = Σ( w₁S₁+ w₂S₂+ w₃S₃+ w₄S₄) The default weight for each domain is 1. The value for each state is based upon connectivity and completeness measures. There is no value-judgment made as to whether a hypothesis content is good or bad; the hypothesis state is evaluated based upon how each of the domains conforms to expectations about what types of roles should be filled and what types of concepts should be connected in what way. Currently the state values, S, are assigned as shown in Table 9.

<table>
<thead>
<tr>
<th>Property</th>
<th>Label</th>
<th>Point system for Hypothesis State Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothesis State</td>
<td>S₀</td>
<td>S₀ = Σ( w₁S₁+ w₂S₂+ w₃S₃+ w₄S₄)</td>
</tr>
</tbody>
</table>
| Predictive Utility | S₁      | 1 point for a hypothesis statement  
1 point for a prediction  
1 point for a population definition  
Fraction of concepts in the prediction linked to concepts in the hypothesis statement |
|                   | (max=4) | (max=4)                                                                                                  |
| Plausibility      | S₂      | Fraction of concepts in the hypothesis statement linked to references, urls, or domain concept networks  
Fraction of concepts in the prediction linked as above  
Fraction of concepts in the prediction linked to concepts in the hypothesis statement |
|                   | (max=3) | (max=3)                                                                                                  |
| Testability       | S₃      | 1 point for the identification of groups  
1 point for the identification of independent parameters  
1 point for the identification of dependent parameters  
Fraction of concepts in the prediction linked to observable concepts. |
|                   | (max=3) | (max=3)                                                                                                  |
| Falsifiability    | S₄      | Fraction of concepts in the prediction linked to an analysis method |
|                   | (max=1) | (max=1)                                                                                                  |

Table 9. Hypothesis State scoring approach
A history of hypothesis values is not currently defined, but if defined in the future, could enable the use histories of hypothesis formation to drive future learning about hypothesis formation patterns. In the MoHM, the Project Domain is described using HDL TBox and ABox axioms. Table 10 provides examples of roles contained in the Project Domain and operations on that domain.

<table>
<thead>
<tr>
<th>Item</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>DL role</td>
<td>Unique hypothesis id</td>
</tr>
<tr>
<td>HName</td>
<td>DL role</td>
<td>Hypothesis short name or acronym</td>
</tr>
<tr>
<td>DescriptionTxt</td>
<td>DL role</td>
<td>Textual description of Hypothesis</td>
</tr>
<tr>
<td>Hreason</td>
<td>DL role</td>
<td>Reason hypothesis is proposed (ExpectationFailure</td>
</tr>
<tr>
<td>Author</td>
<td>DL role</td>
<td>Author</td>
</tr>
<tr>
<td>Created date</td>
<td>DL role</td>
<td>Date hypothesis was created</td>
</tr>
<tr>
<td>Modified date</td>
<td>DL role</td>
<td>Last modification date</td>
</tr>
<tr>
<td>Version</td>
<td>DL role</td>
<td>Current version</td>
</tr>
<tr>
<td>Hypothesis</td>
<td>DL role</td>
<td>Textual description of the hypothesis</td>
</tr>
<tr>
<td>Htype</td>
<td>DL role</td>
<td>Type of hypothesis (Difference</td>
</tr>
<tr>
<td>PopulationTxt</td>
<td>DL role</td>
<td>Textual description of the population</td>
</tr>
<tr>
<td>Parameters List</td>
<td>Optional</td>
<td>Compiled list of parameters or concepts that participate in the hypothesis (used for matching based upon content)</td>
</tr>
<tr>
<td>Concept</td>
<td>DL Tree Structure</td>
<td>Concept definitions to be across domains</td>
</tr>
<tr>
<td>create_hypothesis</td>
<td>Operation</td>
<td>Create a new hypothesis with its own ID</td>
</tr>
<tr>
<td>define_concept_sct</td>
<td>Operation</td>
<td>Compose a concept in SNOMED-CT</td>
</tr>
<tr>
<td>list_parameters</td>
<td>Operation</td>
<td>Extract the set of parameters associated with the hypothesis</td>
</tr>
<tr>
<td>list_concepts</td>
<td>Operation</td>
<td>Extract the set of concepts associated with the hypothesis</td>
</tr>
</tbody>
</table>

Table 10. Project Domain information content.

4.3.2 The Prediction Domain – General statements to specific predictions

The prediction domain includes the concepts in the hypothesis statement and prediction, as well
as the relationships between them. The prediction domain is the heart of the hypothesis, and performs its mapping function, \( h: R \rightarrow S \), of mapping one relation, \( R \) to another, \( S \). In practice, the initial statement of the hypothesis statement is often very general, often neglecting to mention population and time point constraints on the hypothesis. The prediction domain relates a general hypothesis statement to a more specific prediction that can be tested. In the class of Difference Hypotheses, the hypothesis statement is supported, and is not falsified, if the prediction of a difference is not shown to be false. For example, the hypothesis statement, \( S \), “Denosumab is effective for treating osteoporosis,” is supported, and is not falsified, if the prediction, \( R \), “Subjects treated with Denosumab have greater bone mineral density than subjects treated with standard medical care, in population \( G_P \) at time \( T \)” is not shown to be false. Figure 5 illustrates the function of the Prediction Domain.

---

**Figure 6. Mapping Hypothesis Statements to Predictions**

Note that the \( S \) and \( R \) do not necessarily refer to the same concepts, although they may. The following hypothesis is an example where the sets \( P, G, \) and \( T \) are not the same for \( S \) and \( R \). The
hypothesis statement, S, is: “Pesticide exposure in people is related to ADHD.” The prediction, R, is: “The level of organophosphate metabolites in the urine is higher in children who have been diagnosed with ADHD within the last year” [133]. In this case, the parameters pesticide exposure and level of organophosphate metabolites in the urine, while related, are not the same. Similarly the group, people, is not the same as the group, children. Finally one year is not the same as an unknown timeframe. Note also, the following: The term “level of organophosphate metabolite in the urine” in R acts as a surrogate for “pesticide exposure” in S. The term “children” is subsumed by the term “people”. The relationship “is related to” can be viewed as subsuming the relationship “is higher”. The term “with ADHD” subsumes the term “diagnosed with ADHD in the last year”.

The concepts and relationships in the prediction are usually much more specific, and measurable than those in the hypothesis statement. To illustrate how this often manifests itself in practice, Table 10 comprises some H-statements and their corresponding predictions, that have been manually extracted from clinicaltrials.gov for a variety of health research projects.
<table>
<thead>
<tr>
<th>NCT</th>
<th>Hypothesis Condition</th>
<th>Relationship</th>
<th>Hypothesis Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCT 01426321</td>
<td>Individualized CRT Therapy</td>
<td>increases</td>
<td>positive response to CRT from 65% to 85%</td>
</tr>
<tr>
<td></td>
<td>Image-guided left ventricle lead position vs. standard care</td>
<td>increases</td>
<td>6 month response to CRT from 65% to 85%</td>
</tr>
<tr>
<td>NCT 01426282</td>
<td>Health education given by nurse to outpatients</td>
<td>improves</td>
<td>Outpatient compliance in cardiovascular outcomes</td>
</tr>
<tr>
<td></td>
<td>15 min. education given by nurse</td>
<td>improves</td>
<td>Blood pressure, lipids, drug adherence, hospitalization, exercise, smoking, and/or weight</td>
</tr>
<tr>
<td>NCT 01426178</td>
<td>Neurally Adjusted Ventilator Assist vs. Pressure Support Ventilation</td>
<td>allows reduction in</td>
<td>asynchrony rate defined as a percentage of total respiratory rate</td>
</tr>
<tr>
<td></td>
<td>Neurally adjusted ventilator assist in two 30 min sets of non-invasive ventilation</td>
<td>reduces</td>
<td>asynchrony interpreted by experts from flow, volume, pressure, EADi</td>
</tr>
<tr>
<td>NCT 01425879</td>
<td>Akt inhibitor MK2206 in Advanced Refractory Biliary Cancer</td>
<td>may stop</td>
<td>growth of tumors by blocking enzymes needed for cell growth</td>
</tr>
<tr>
<td></td>
<td>Akt administration every 28 days if no toxicity or progression</td>
<td>reduces</td>
<td>Overall response rate as defined by RECIST 1.1</td>
</tr>
<tr>
<td>NCT 01425580</td>
<td>Liraglutide Treatment vs. Glimepiride in combination with Metformin</td>
<td>improves</td>
<td>heart function</td>
</tr>
<tr>
<td></td>
<td>Liraglutide Treatment (18 mg s.c.) vs. Glimepiride (4 mg p.o) in combination with metformin (500 mg p.o)</td>
<td>improves</td>
<td>Left ventricle longitudinal function and/or functional reserve during rest and/or after exercise using tissue Doppler echocardiography, 24 h blood pr</td>
</tr>
<tr>
<td>NCT 01472991</td>
<td>TC-5619 in adults with inattentive-predominant ADHD</td>
<td>improves</td>
<td>symptoms</td>
</tr>
<tr>
<td></td>
<td>TC-5619 5&amp;25 mg vs. placebo comparator day 1, wk 2,4,6</td>
<td>improves (with safety)</td>
<td>scores on inattentive subscale of CAARS-INV</td>
</tr>
</tbody>
</table>

Table 11. Example H-Statements and Predictions

The H-Statement and the Prediction may be at different levels of abstraction, in different fields, or at different levels of scale. For instance, the H-Statement “Statins reduce heart attacks,” may be tested with the Prediction “People taking statins show reduced levels of cholesterol.” In this
case, the reasoning chain involves a relationship between levels of cholesterol and heart attacks, and assumes that “relative levels of cholesterol” is a plausible surrogate for “relative number of heart attacks.” Another example from clinical trial NCT0470418, the H-Statement is “NICS-15in is efficacious for treating Alzheimer’s disease”, where the prediction is that “administration of NICAS-15 will produce a difference in cognitive outcomes as determined by [sets of survey questions].” [135]

In MoHM, the Prediction Domain makes extensive use of medical and other terminology. The MoHM facilitates the expression of the H-Statement and the Prediction by providing hypothesis templates that provide a higher level of abstraction for ease of understanding and use, as well as re-use. Furthermore, the population and timeframe contexts of the hypothesis have been factored out and represented in a separate domain, the Population Domain. The hypothesis template corresponding to Figure 5 is:

```plaintext
hstatement_text:
prediction_text:
concepts1: definition
concepts2: definition
s_relationship:
conditions: definition
outcomes: definition
p_relationship:
condition:concept1: {surrogate, subsumedBy, unknown}
outcome:concept2: {surrogate, subsumedBy, unknown}
```

The ADHD pesticide hypothesis, developed in the template, would have an underlying expression in HDL. The concept “children” is the population context for which the hypothesis is expected to be valid, consequently it will be expressed in the Population Domain.
Complex concepts such as “level of organophosphate metabolites in the urine”, or “diagnosed with ADHD within the last year” would be expressed as a DL terms, usually a SNOMED-CT term, and assigned an HDL synonym. For example:

```
concept_def(organoPhosphateLevel ,
  “situation(group(associated finding = organo phosphate metabolites)
   (location context = urine)
   (finding context = known present)
   (temporal context = in the past)
   (subject relationship context = subject of record))”, SCT)
```

The natural language expression of the hStatement and the Prediction are included in the MoHM so as to make it possible to check that the encoding of these is consistent with the intended meaning. This is expressed in the following HDL statements:

```
hStatement_txt(id, “Pesticide exposure in people is related to ADHD”)
Prediction_txt(id, “The level of organophosphate metabolites in the urine is higher in children who have been diagnosed with ADHD within the last year”)
hStatement(id, isRelatedTo,[pesticideExposure],[“ADHD”],DH)
prediction(id,IsHigherIn,[organoPhosphateLevel],[ADHDdiagnosed_1yr],DH)
isRelatedCondition(surrogate,pesticideExposure,organoPhosphateLevel,DH)
isRelatedOutcome(subsumedBy,ADHDdiagnosed_1yr”ADHD”,DH)
```

Note that the Prediction Domain does not attempt to interpret the statements made. The primary purpose of the Prediction Domain is to create hyperlinks between concepts, expressing connectivity of concepts. Table 12 lists example information contents of the Prediction Domain.
<table>
<thead>
<tr>
<th>Prediction Item</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>DL concept</td>
<td>Unique hypothesis id</td>
</tr>
<tr>
<td>hStatement_txt</td>
<td>DL role</td>
<td>Textual representation of the H-Statement</td>
</tr>
<tr>
<td>Prediction_txt</td>
<td>DL role</td>
<td>Textual representation of the Prediction</td>
</tr>
<tr>
<td>hStatement</td>
<td>Hrole</td>
<td>Hyper-role definition for H-Statement</td>
</tr>
<tr>
<td>prediction</td>
<td>Hrole</td>
<td>Hyper-role definition for Prediction</td>
</tr>
<tr>
<td>isRelatedCondition</td>
<td>Hrole</td>
<td>Hyper-role definition for relationship between conditions in the prediction and concepts in the H-Statement</td>
</tr>
<tr>
<td>isRelatedOutcome</td>
<td>Hrole</td>
<td>Hyper-role definition for relationship between outcomes in the prediction and concepts in the H-Statement</td>
</tr>
</tbody>
</table>

Table 12. Prediction Domain information content.

### 4.3.3 The Plausibility Domain: Providing a Basis for the Hypothesis

The function of the Plausibility Domain is to represent and link to domain knowledge that connects the hypothesis to other working hypotheses, theories, and facts. There are three major types of plausibility requirements that, in principle, must be fulfilled for a hypothesis: 1) Plausibility of the Hypothesis Statement, the 2) Plausibility of the Prediction, and 3) the plausibility of the relationship between the H-Statement and Prediction relations. 1) and 2) are included in the Plausibility Domain, while 3) is left for future work.
PLAUSIBILITY REQUIREMENTS
• Hypothesis statement
• Prediction
• Mapping between Hypothesis Statement and Prediction

Figure 7. Plausibility requirement for hypotheses.

The Plausibility Domain includes references, urls, and graph structures representing knowledge networks. Example knowledge networks are represented by the “is-a” and “leads-to” relations in the MoHM. As shown in Figure 7, plausibility elements describe links between the hypothesis domains as well as links to external knowledge. Links from concepts in other domains to the plausibility domain are called Basis Links, or Basis. Table 13 illustrates some of the Plausibility Domain elements and operations.
<table>
<thead>
<tr>
<th>Hypothesis Statement Item</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>DL concept</td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td>DL role</td>
<td>Published literature reference</td>
</tr>
<tr>
<td>URL</td>
<td>DL role</td>
<td>URL reference</td>
</tr>
<tr>
<td>leads-to</td>
<td>H-role</td>
<td>A graph structure of binary relationships between concepts. The leads-to edges signify a relationship that is not necessarily causal, but that appear to be related in time to a progression of events that may occur. For example, a leads-to graph might contain the following relationships. “Mitra-valve collapse leads_to backflow. Backflow leads to heart failure. Heart failure leads to exhaustion. Heart failure leads to death.” Traversing this leads-to tree, one may infer that mitral-valve collapse may leads to death.</td>
</tr>
<tr>
<td>is-a</td>
<td>DL-role</td>
<td>graph structure of binary relationships between concepts. The isa element represents a hierarchical tree of “is a” relationships.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Operation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>add_DCN</td>
<td>Add domain concept network, expressed as a list</td>
</tr>
<tr>
<td>add_reference</td>
<td>Add a url or reference</td>
</tr>
<tr>
<td>add_basis</td>
<td>Link an element in the plausibility domain to a concept in another domain</td>
</tr>
<tr>
<td>check_plausibility</td>
<td>Provide a report of links between R and S, and between concepts, urls, references, and named graph structures</td>
</tr>
<tr>
<td>count_plausibility</td>
<td>Count basis links and return a count or a percentage of concepts connected via Basis links</td>
</tr>
</tbody>
</table>

Table 13. Example Plausibility Domain (Basis) Elements

### 4.3.4 The Population Domain: Specifying Populations and Groups

The Population Domain has the function of constraining the prediction to a subset of all of the possible subjects. Using upon HDL BBox axioms, it features a representation that enables efficient matching between individuals and clinical research hypotheses. The practice of specifying populations in clinical research relies on two sets of criteria: Inclusion criteria and Exclusion criteria. Inclusion criteria, \( I = \cap (c_i, \ldots, c_n) \), are those criteria, \( c \), such that all must be
true in order for an individual to be included in a population. Exclusion criteria, \( E = \neg \bigcup (e_1, \ldots, e_n) \), are those criteria, \( e \), such that none must be true in order for an individual to be included in a population. \( E \) can be transformed to reduce the number of disjunctions: \( E = \neg \bigcup (e_1, \ldots, e_n) = \bigcap (\neg e_1, \ldots, \neg e_n) \). Thus a population may be represented as \( P = \bigcap (c_1, \ldots, c_n) \cap (\neg e_1, \ldots, \neg e_n) \). However, the complex nature of clinical research criteria makes elimination of all disjunctions difficult. Population criteria include medical criteria as well as non-medical criteria, such as: Willingness to participate in an activity, engagement in other clinical research projects, or recent travel to other locations. Examples of common criteria seen in inclusion and exclusion criteria are:

- Demographics such as gender, age group, ethnic background
- Health habits such as smoking, weight category, alcohol use, illegal drug use
- Health conditions that are common in health profiles encountered in physicians offices, such as: pre-existing heart conditions, diabetes, cancers, and prior events, such as, surgeries, injuries, and diagnoses.

In the MoHM, the **Population Domain** describes constructs that define the population. A major challenge of representing clinical research population criteria is the strong need for use of medical terminology in conjunction with common use of negation, disjunction, and numeric comparisons within the criteria stated by clinical research designers. SNOMED-CT and many compatible DL extensions, such as EL++, do not support all of the operations required for clinical research inclusion and exclusion criteria. Hence, the evaluation of whether an individual can be included in a clinical research population, cannot be accomplished purely through DL subsumption queries.

The MoHM addresses this challenge through the use of the HDL BBox, which defines allowable
propositional formulas, evaluating to true or false, connected by Boolean Operators. The propositional formulas may contain comparison expressions such as “(hasAge, >, 30)”, as well as DL operations such as “hasCondition.disease ⊆ a DL expression”. An object is a member of a population if the Boolean expression evaluates to true. For example, in the Parkinson’s Disease Tai Chi study [REF NEJM Feb9,2012], the population criteria are:

**Inclusion Criteria:** “clinical diagnosis of Parkinson’s disease, with a disease severity rating of stage 1 to 4 on the Hoehn and Yahr scale (which ranges form 1 to 5, with higher scores representing more severe disease); an age of 40 to 85 years; at least one score of 2 or more for at least one limb for the tremor, rigidity, postural stability, or bradykinesia items in the motor section of the Unified Parkinson’s Disease rating score (UPDRS) III); stable medication use; ability to stand unaided and walk with or without an assistive device; medical clearance for participation; and willingness to be assigned to any of the three interventions.

**Exclusion criteria:** Current participation in any other behavioral or pharmacologic study or instructor-led exercise program, a Mini-Mental State examination score lower than 24 (indicating some degree of cognitive impairment), debilitating conditions or vision impairment that would impede full participation in the study, and unavailability during the study period.”

The Population Criteria are defined in the HDL BBox as follows:

DebilitatingCondition = (vision impairment (defined in Snomed-CT))
Availibility = between(startDate(yy-mm-dd), endDate(yy-mm-dd))

Population =
  Person ∧
  has_disorder.ParkinsonsDisease ∧
  severityHoehnYahrScale.(>4) ∧
  hasAge(>=40) ∩ hasAge(<=85) ∧
  (∃UPDRS_Rigidity(>2) ∪ ∃UPDRS_PosturalStability(>2) ∪ ∃UPDRS_Bradykinesia(>2)) ∧
  medicationUsageStability.true ∧ hasAbility.standUnaided ∧ hasAbility.walk ∧
  hasMedicalClearance.true ∧ hasWillingness.true ∧
  ¬(currentParticipation.behaviorStudy ∧
   ¬(currentParticipation.pharmacologicStudy ∧
    ¬(currentParticipation.instructorLedExerciseProgram ∧
     ¬ (mini-mental state examination(<,24)) ∧
     ¬ DebilitatingCondition ∧ Availability)
The BBox is also used to specify disjoint partitions of the population. Partitions will be used to specify groups in the Test Strategy Domain. Partitions may be determined by selection on parameter values, by randomization (which can be represented by a parameter value assignment), or by some other criteria. A partition within a population is represented as a sub-population, and is represented as the intersection between the population and the population defined by the conjunction of the inclusion/exclusion group parameters.

A population or partition can be named, copied, reused in another hypothesis, and modified/renamed. If variables are ordered and transformed into OBDD’s (see Chapter 6), it can be determined whether one population definition is subsumed by another. Table 14 indicates some operations for the Population Domain.

<table>
<thead>
<tr>
<th>Item</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>name_population</td>
<td>Operation</td>
<td>Name a population definition</td>
</tr>
<tr>
<td>copy_population</td>
<td>Operation</td>
<td>Copy the population of the current hypothesis to memory</td>
</tr>
<tr>
<td>use_population</td>
<td>Operation</td>
<td>Attach a named population to a hypothesis</td>
</tr>
<tr>
<td>population_subsumption</td>
<td>Operation</td>
<td>Determine whether A is subsumed by B.</td>
</tr>
</tbody>
</table>

Table 14. Example population Domain operations

### 4.3.5 The Test Strategy - Operationalizing Parameters and Groups

In health research, a test strategy specifies how many and which groups will be defined, which parameters will be measured, which parameters will be controlled, the type and expected distribution of the parameters, and the relationship of parameters to each other. In the MoHM, creation of a test strategy comprises definition of parameters and groups and specification of a path from prediction concepts to measurable parameters.
Often a hypothesis prediction is not immediately testable, that is, it is not possible to measure and test the parameters in the prediction. An example of this is shown with the hypothesis, “If P53 gene mutation P53 indicates general drug resistance, then the population with P53 gene mutation will be more resistant to treatment A.” In this case, the term “resistant to treatment” may not be directly measurable. However, it is possible to operationalize this concept, so that resistance to treatment is replaced with a more concrete, measurable concept, such as “percent increase in size of tumor”, where there is proposed relation between the percent size increase in tumor, which can be directly measured, and the resistance to treatment, which cannot be directly measured.

For Difference Hypotheses, a common test strategy is to subdivide the population into mutually exclusive partitions, or groups that can be compared. In this example, the groups will be defined by the value of the parameter “has_P53_gene_mutation”. In the Population Domains, partitions can be defined by the group parameters that subdivide the population into disjoint subsets. In the Test Strategy Domain, partitions are selected and designated as groups. A group parameter may be either derived or measured. For instance, group1 may be the subset of the population that is exposed to an intervention, with group 2 being not exposed. The exposure to the intervention may need to be operationalized, so that it is directly connected to one or more measurable parameters. Parameters are typed through a DL role, measure-type, to be either measured or derived. A parameter is considered to be “measured” only if explicitly stated to be so. Parameters are also typed through a DL role, control-type to be either independent or dependent.
Groups are designated as related or unrelated. Groups are related if they are not statistically independent of each other. For instance, two groups, where each group contains one of a set of siblings, may be designated as related groups because they share common characteristics that might bias the results. Another example of related groups is the definition of two groups, where each group refers to the same set of elements, but at different time points. The number of groups in the test strategy is assigned and is designated by the role, groupn.

Specification of a path from the prediction to measurable concepts is accomplished through two types of roles, the surrogate role and the composition role.

The **Surrogate role** relates concepts through a surrogate link, meaning that one concept is intended to be a substitute for another. A surrogate link is a one-way link between two concepts, such that concept A can replace concept B. For example, “number of social interactions per week” might be a surrogate for “social interaction” in surrogate(social_interaction, “number of social interactions per week”).

The **Composition role** describes concepts that are related to other through composition. For example, a derived concept for “quality of life” might be the weighted sum of several concepts, such as “lack of pain”, “daily exercise”, “social interactions”. The composition element, describes the relationship between the concepts and the composition. For example, the composition relationship may stated with multiple DL axioms: composition(quality_of_life, lack_of_pain), composition(quality_of_life, exercise), composition (quality_of_life),
social_interaction”). Note that this also expressible as a hyperrole, so that the input and output concepts are related and the path of associations can be traced.

**Test Strategy Operations.** The operations for the test strategy focus on determining whether there is a path from concepts in the prediction to measurable parameters. In the MoHM, if there is a path between each concept in the prediction and a measured parameter, the number of groups are specified, and the groups are defined by a group parameter set, $P_G$, then a test strategy exists.

<table>
<thead>
<tr>
<th><strong>Operation</strong></th>
<th><strong>Inputs</strong></th>
<th><strong>Description</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>groupn</td>
<td>DL-role</td>
<td>Number of groups</td>
</tr>
<tr>
<td>groupr</td>
<td>DL-role</td>
<td>Relatedness of two groups</td>
</tr>
<tr>
<td>Check_test_connectivity</td>
<td>Operation</td>
<td>Determine which concepts in the prediction for which a path can be found to a concept that is a measurable.</td>
</tr>
<tr>
<td>Check_test_strategy</td>
<td>Operation</td>
<td>Determine whether a test strategy exists</td>
</tr>
</tbody>
</table>

Table 15. Test Domain operations

4.3.6 **The Falsification Domain: Falsifying the Prediction**

A common theme in statistical methods education is the appropriateness of the analysis to the situation. Historically each analysis type was developed as a solution to a specific research challenge. Over time a wide variety of analyses have been developed.

“The research question one asks determines the design of the study, the number of groups, the type of data, etc. These factors, in turn, define the statistical method that is needed. This . . . is a cookbook summarizing the appropriate statistical tests to use when answering the most commonly asked research questions.” [136]

The type of statistical analysis may vary according to the statistician’s preferences as well as the conventions of a field of research. However, many generally accepted guidelines do exist. Two
of these provided the foundation for the development of analysis rules applied in the MoHM: 1) The UCLA Statistics analysis guidelines [137], 2) A text in medical data analysis [138]. These organizational principles were supplemented with insights from personal correspondence [139].

In the MoHM, a prediction becomes falsifiable when there is a valid analysis strategy. An analysis strategy specifies the method that will be used to evaluate the outcome of the test strategy. HDL analysis strategies are represented in the BBox, as Boolean Expressions, where it is possible to represent more than one valid analysis strategy for a hypothesis. The evaluation of this is through the analysis rules in the BBox. The analysis strategy must be consistent with the characteristics of the analysis parameters, and groups that it is applied to. A hypothesis without a valid analysis strategy does not meet the hypothesis criteria for falsifiability.

**Analysis Data Typing.** In order to utilize analysis rules, each parameter that is considered to be included in an analysis must be typed from the perspective of statistical analysis. This is a different kind and purpose of typing than conventionally used in programming languages or databases. For instance, a data element that is represented as an integer in a data coding schema, may have one of two different analysis types. It could be ordinal, meaning it has a numerical value, or it could be nominal, meaning it has a rank value. The analysis data type for a parameter affects the kinds of analyses that can be performed on that data. Analysis data typing is specified as a DL role on a parameter: antype={ordinal, nominal, interval, categorical}. Analysis parameters do not need to be directly measurable, but must have a path to measurable parameters. Parameters to be used in an analysis must have the following roles associated with them: analysis type, control type, measure type. In addition, a distribution type is assigned, the
indicates the assumed distribution of the parameter. The default distribution type is “normal”.

A HDL analysis rule defines the feature requirements for the validity of a specific analysis strategy. Three categories of features are important in analysis rules: the attributes (or roles) of the independent variable, the attributes (or roles) of the dependent variable, and the population and group characteristics. Within those feature categories, the important attributes are: distribution and variance of the variable, the number and relatedness of groups or levels, the type of variable being considered. Table 16 depicts example analysis rules.

<table>
<thead>
<tr>
<th>Example Analysis Rules (dv=dependent variable, idv=independent variable)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>11</td>
</tr>
</tbody>
</table>

Table 16. Example analysis rules.
Falsifiability Domain Operations. MoHM analysis rules can be used for two purposes: 1) from a top-down perspective, for determining whether a proposed analysis is valid for the current operationalized prediction, and 2) from a bottom-up perspective, suggesting which analyses are valid for defined measures. As with the other components of the HRE, an analysis rule set can be published and shared. Table 17 lists some operations for the Falsifiability Domain.

<table>
<thead>
<tr>
<th>Operation</th>
<th>Inputs</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>analysis_options</td>
<td>P,T,G</td>
<td>Return the possible analyses for a given hypothesis space</td>
</tr>
<tr>
<td>analysis_validity</td>
<td>P,T,G</td>
<td>Determine whether there is a valid analysis for the specified inputs.</td>
</tr>
<tr>
<td>analysis_consistency</td>
<td>A,P,T,G</td>
<td>Determine whether an analysis, A, is consistent with the inputs.</td>
</tr>
</tbody>
</table>

Table 17. Falsifiability Domain operations

4.4 Summary and Conclusions

This chapter begins by providing historical and theoretical context for the hypothesis as a concept, motivating the need for a computational model for hypotheses. The technical contribution of this chapter is the development of a novel approach for representing health hypotheses – one that extends the common perception of a hypothesis as a single statement, to a model of a hypothesis as a system, a cognitive artifact with component parts, or domains, that operate together for form an operational whole. This model – based upon the principles of connectivity, modularity, and domain specificity – is designated the Molecular Hypothesis Model (MoHM). The MoHM concept is operationally similar to molecular models, such as proteins, that have specialized domains with specialized structures and functions.
This approach defines the information representation for health hypotheses, as a set of modular domains, each with its own specific structure and function in the hypotheses. These domains, as well as the concepts within them, are connected through a network of DL roles and HDL hyperroles. This connectivity enables the domains of the hypothesis to work together to ensure that the hypothesis meets the requirements of a scientific hypothesis: the properties of predictive utility, plausibility, testability, falsifiability, and state.

These properties are implemented by the six MoHM Domains: 1) the Project Domain, supporting the property of state, that includes other hypothesis-related meta-knowledge such as hypothesis type and important project timelines; 2) the Prediction Domain, supporting the property of predictive utility, where the concepts in the hypothesis H-Statement and Prediction are related to each other through defined relationships; 3) the Population Domain that provides a context for the prediction; 4) the Plausibility Domain, supporting the plausibility property by including references to known facts, publications, and reasoning chains; 5) the Test Strategy Domain, supporting the testability property by defining the groups and observable parameters that are required for testing the hypothesis; and 6) the Falsification Domain, supporting the falsifiability property, by ensuring that a hypothesis has at least one legal analysis strategy that can be used to determine whether a prediction is false.

The next two chapters, Chapter 5 and 6, demonstrate how these domains and properties can be leveraged to provide the important capabilities of incremental, hypothesis construction and reuse, hypothesis management, and collaboration between health researchers and health consumer.
CHAPTER 5  Leveraging Health Hypotheses

The Molecular Hypothesis Model (MoHM), presented in Chapter 4, provides a framework for conceptualizing and representing health hypotheses. This framework was designed for the purpose of engaging the health consumer in health research design and discourse. At the same time, it aims to be consistent with current principles and practice of clinical research. The problem is that the typical health consumer knows very little about clinical research. The challenge is to find an approach that bridges the conceptual gap that exists between health consumers and health research professionals.

Health hypothesis design is, in practice, iterative, often relying on reusing portions of existing hypotheses, as described in Chapter 4. A hypothesis construction methodology that emphasizes reuse and patterns or templates, is both consistent with clinical research practice and potentially accessible to the health consumer. Many people, when provided with examples or templates, building blocks, and simple rules for construction, are able to construct new, complex artifacts, for example, people who are able to create their own websites, with the aid of an underlying infrastructure that manages the implementation details. The Health Research Exchange provides an underlying health hypothesis construction and management infrastructure, made possible by the Molecular Hypothesis Model (MoHM).

The three organizational principles of the MoHM – modularity, domain specificity, and connectedness – make possible a reuse and management methodology for hypotheses: The first feature is modularity. The innovative modular structure provided by the Molecular Hypothesis
Model (MoHM) acts to divide the problem space of hypothesis formation into smaller, easier to solve, and potentially reusable, problem “chunks”. Most importantly, a modular representation makes it possible to construct the parts, or domains, of a hypothesis, from a set of building blocks for each of the parts. The second feature is domain specificity. Each domain is specialized to a representation that makes it possible to answer common queries pertaining to the domain. The representation used in each domain determines the function of the domain, and the type and efficiency of reasoning that is possible using domain elements. For example, in the plausibility domain, which is represented as a network of concepts, reasoning about connectedness is accomplished by hypergraph traversal. In the population domain, represented by Boolean formulas, reasoning about patient eligibility is accomplished by Boolean formula evaluation. Reasoning across multiple hypotheses, within a single domain, is designed, through the use of Description Logic formalisms and Boolean functions, to be efficient and scalable. The third feature is connectedness. The HDL roles and hyperroles of the MoHM act to connect concepts within and between hypotheses, forming a network of knowledge that is overlaid on the domains. This network of knowledge makes it possible to answer queries such as: which concepts in a hypothesis, or one of its domains, participate in a particular role? This network of knowledge may also enable access to underlying pathways of plausibility that diverse hypotheses may share. In addition, patterns of hypothesis construction can be represented as patterns of connections, analogous to network motifs that may be reused.

This chapter builds on these three features – modularity, domain specificity, and connectedness, to describe three important capabilities in which the MoHM can be leveraged in the context of the Health Research Exchange (HRE). The first capability is a development and reuse
methodology for Population Health Hypotheses, leveraging knowledge embedded in existing hypotheses to make hypothesis formation more efficient and accessible. Section 5.1 describes the hypothesis development and reuse methodology, Case-Network Domain Reasoning, CaNDoR, and the use of Hypothesis Design Patterns. The second capability is hypothesis management for the HRE, supporting hypothesis construction, modification, sharing, and queries. Section 5.2 describes the hypothesis management capabilities and operations provided by a Hypothesis Reasoning Engine (HyRE). The third capability is organization of Personal Health Record data through Personal Health Hypotheses, and Health Condition Profile templates. This is an extension of the MoHM to Personal Health Hypotheses that may be stored, not in a hypothesis database, but in active Personal Health Records of individuals. Section 5.3 describes how the MoHM can be used to describe Personal Health Hypotheses, and how these can provide the foundation for an active Personal Health Record. Section 5.4 summarizes the contributions of this chapter.

5.1 Hypothesis Formation with Case-Network Domain Reasoning

The modular structure provided by the Molecular Hypothesis Model (MoHM) makes it possible to define a method for reusing domain building blocks of prior hypotheses to create new, and possibly innovative, hypotheses. This reuse methodology includes templates for common hypothesis design patterns that capture health research practice. These design patterns may be made available, so that people who are not health researchers are provided with representational and computational aids for participating in the health research definition process.

This research proposes that the health hypothesis formation process is an artifact design problem
that can be modeled as a MoHM design and construction process, utilizing an approach similar to case-based reasoning (CBR). In case-based reasoning, when a problem is to be solved, the approach is to find and retrieve cases with similar problem features, and then reuse the solutions to one or more of these problems with appropriate modifications [140]. Similarity is determined by a one or more similarity measures, or a composite similarity measure, which may be the weighted sum of a vector of several problem features, or may be determined by alternative approaches such as Nearest Neighbor [141]. Modification rules are used to make changes to the reference solution features based upon the difference between the retrieved and new problem features. If the number of reference cases is large, case libraries may be partitioned based upon the type of cases covered, so as to reduce the search space [142].

Case-based reasoning is often used when there is no first-principles design model, where there is insufficient understanding of the domain, and where a purely algorithmic or rule-based approach cannot be used. An example of case-based reasoning is the Weld Process Design Advisor (WPDA) [143] which was developed to design sets of weld process parameters for the Space Shuttle Main Engine. Because of the unique geometries and specialized materials, design of weld process parameters, including temperature, timing, duration, and weld material, was, in practice, iteratively achieved through trial and error. Yet some weld engineers were able to design good first-pass solutions by “reusing” existing parameter sets - selecting relevant prior weld cases and modifying them to fit the current situation. The WPDA CBR approach divided the design space into classes of problems characterized by geometry and material types. Additional similarity measures were defined within each class, including weld-type, thickness, material properties, welding apparatus, and heat sink location and type. The WPDA system
retrieved sets of similar cases, based upon the problem characteristics, automatically constructing new solutions (weld parameter sets) through reuse and modification of existing ones. This system used a hybrid inference approach, implemented in an object-oriented Prolog with a rule-based extension, with similarity measures including both qualitative and quantitative parameters. Case modification was accomplished through inference rules as well as calculated differences in heat effects. There are many other examples of successful application of CBR to difficult design problems including [144],[145],[146].

A problem solving-approach, similar to CBR, is proposed, for HRE hypothesis formation. The modularity principle of the MoHM, with the decomposition of the hypothesis into its component parts, presents us with an opportunity to retrieve and reuse not only entire hypothesis cases, but also hypothesis components – the hypothesis domains. Reflecting the domain specificity feature of the MoHM, each domain has a function, and a corresponding representation specific to that domain’s function. This enables domain-specific building blocks that can be reused across different hypotheses. Thus, the retrieval of similar cases will be able to take advantage of similarity in one or several domains, or at a meta-level for the entire hypothesis. This approach of looking for similarity in linked structures corresponding to the hypothesis domains, will be referred to as Case-Network Domain Reasoning (CaNDoR). There are three major differences between CaNDoR and simple case-based reasoning: 1) the case selection and reuse principles may be applied, not only to the main artifact, but to its components, 2) the reuse of cases in CaNDoR includes not only reuse of attributes or parameters, but extends to reuse of links to other information, and 3) CaNDoR does not automatically create new health hypotheses. Instead it presents to the user for selection, candidate hypotheses or component domains that meet
specified similarity criteria. This approach ensures that the hypothesis author remains “in charge” of the creative aspects of health hypothesis formation.

### 5.1.1 Finding and Using Similar Hypothesis Cases

As in CBR, an important aspect of CaNDoR is defining the problem features that will be the basis of hypothesis case retrieval and reuse. Table 18 lists several CaNDoR hypothesis problem features, selected for their ability to influence hypothesis design. The type of hypothesis prediction affects the structural connectedness of the Prediction Domain. Health conditions affect outcome measures, and their composition from other measures. Population descriptions affect parameter definitions. Medical fields influence parameter definitions as well as their surrogate or composition relationships [134]. The strategy for falsification, including the number and type of groups, and the tests performed, impacts the statistical analysis method.

<table>
<thead>
<tr>
<th>Problem Feature</th>
<th>Associated Solution Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothesis type</td>
<td>• Structural model used to construct the prediction</td>
</tr>
<tr>
<td>(Difference, Correlation, Emergent,</td>
<td>• Types of roles or links that can be reused</td>
</tr>
<tr>
<td>Interventional, Parameterized Difference)</td>
<td></td>
</tr>
<tr>
<td>Health Condition</td>
<td>• Measures</td>
</tr>
<tr>
<td>Health Outcome</td>
<td>• References, urls</td>
</tr>
<tr>
<td></td>
<td>• LeadsTo networks</td>
</tr>
<tr>
<td></td>
<td>• Parameter compositions</td>
</tr>
<tr>
<td>Parameter types</td>
<td>• Test strategy</td>
</tr>
<tr>
<td></td>
<td>• Analysis method</td>
</tr>
<tr>
<td>Population descriptions</td>
<td>• Outcome measures</td>
</tr>
<tr>
<td>Demographics</td>
<td>• Test parameters</td>
</tr>
</tbody>
</table>

Table 18. HRE Hypothesis problem and solution features
There are two methods for detecting similar hypotheses. 1) **Concept-based similarity**, and **Prediction-based similarity**. Concept-based similarity matches and retrieves hypothesis cases based upon shared or connected concepts. Prediction-based similarity matches and retrieves cases based upon categories of hypotheses with structural similarity – i.e., similar networks of relationships or roles, and relationship types.

When a hypothesis has been retrieved based upon some similarity criteria, it can be reused as a whole, or in part. For example, if the new hypothesis statement is: “Rituximab maintenance treatment is effective for Chronic Lymphocytic Leukemia”, and one of the hypotheses retrieved based upon concept matches, due to shared mention of Rituximab as a condition, is: “Rituximab is effective for Lymphoma”, the basic structure of the Prediction Domain may be reused, as well as many of the Plausibility Domain elements, such as urls, references, and leadsTo relationships. Since both CLL and Lymphoma are cancers of B-cell lymphocyte proliferation, some of the condition and outcome measures, such as blood counts, lymph node swelling, and infection rates, may be reused with modification. Since Kaplan-Meir survival analysis is a common analysis in cancer research, even some parts of the test strategy and analysis approach may be partially reusable. However, since the old and the new hypotheses differ in the timeframe for the intervention (treatment therapy versus maintenance therapy, respectively), the structure and timeframe of the outcome measures may need to be significantly modified in the new hypothesis. But even this has some potential for reuse and recombination – of the outcome measure structure for a maintenance treatment for an entirely different health condition.

It may seem, at first glance, that reusing hypotheses might be practical for situations where the
new hypothesis is very similar to a prior one, but not for truly ground-breaking, creative hypotheses. But even ground-breaking, creative hypotheses about new effects must be grounded in reality. Consequently there may be prior causes of those effects that may be shared and reused. For instance, suppose that a clinical researcher would like to propose a new hypothesis about drug TC5619 that has shown cognitive enhancing effects in animal studies. The new hypothesis is “TC5619 will have an effect on ADHD” [147]. Retrieving hypothesis cases by concept similarity of outcomes, may yield, due to shared reference to the concept “ADHD” as an outcome, the existing hypotheses: “Maternal smoking contributes to the development of childhood ADHD” [148] and “Cognitive training will have an affect on pediatric ADHD” [149]. There is one immediately obvious reuse opportunity – population and measure definitions for ADHD from either or both of these hypotheses may be reusable. But there is another, less obvious, reuse opportunity. If the researcher includes, in the Plausibility Domain of the new hypothesis, the surrogate relationship, “TC5619 acts as a surrogate for nicotine,” and the Basis of the maternal smoking hypothesis contains leadsTo relationships involving smoking and nicotine levels in the bloodstream, then the underlying nicotine connection becomes relevant. Consequently, there may be references, urls, and leadsTo relationships of interest in the existing Maternal Smoking hypothesis that may be reused in the new TC5619 hypothesis. Practical definitions, such as nicotine blood concentration levels and measurement strategies, may also be found in the maternal smoking hypothesis. It may also be possible to identify other common measures important at various points in the physiological pathways, since the effects of nicotine is an underlying mechanism in common to both hypotheses.

**Concept-Based Similarity.** As illustrated by the example above, concept-based similarity
involves more than matching keywords within a monolithic hypothesis; it leverages the MoHM structure and its Description Logic foundations to find concepts that participate in specific domains or roles in the hypothesis. Specification of concept-based similarity involves specification of the concept, the domain(s), and any desired role-participation as:

\[ \text{similar\_concept}(\text{concept}, \text{domain}, \text{role}) \]. Concept-based similarity leverages the connectedness feature of the MoHM.

For example, the query: \text{similar\_concept}(\text{“ADHD”}, \text{prediction}, \text{has\_outcome}) retrieves hypotheses where the concept “ADHD” occurs in the Prediction Domain, and participates in an has\_outcome role. To specifically search for hypotheses that pertain to nicotine pathways, the query: \text{similar\_concept}(\text{nicotine}, \text{plausibility}, \text{LeadsTo}) finds hypotheses where the concept “nicotine” participates in a set of LeadsTo relationships in the Plausibility Domain. The intersection of these two queries reduces the search space of similar hypotheses, yielding more relevant hypotheses including the hypothesis addressing maternal smoking and ADHD.

Determining similarity of medical terminology is an important aspect of concept-based similarity. There are many ways to describe medical conditions and interventions, so that terms that look very different, are in fact, very similar, for example “heart attack” and “cardiac infarction”. The clinicaltrials.gov database successfully handles medical terminology synonyms through UMLS. SNOMED-CT, selected for the HRE because of its DL foundations, also provides for handling of medical terminology synonyms. We will assume that a SNOMED-CT engine is available to identify medical terminology synonyms that can be applied in searches for similar hypotheses. In addition, through subsumption queries, a SNOMED-CT engine may
search health hypotheses for concepts that are subsumed by the concepts specified in a query.

**Prediction-Based Similarity.** Prediction-Based Similarity retrieves hypothesis cases based upon the type and structure, or connectedness, of the prediction being made. Example hypothesis predictions types are Interventional Difference and Parameterized Difference. An Interventional Difference prediction predicts that there will be an effect in one group as the result of an intervention that will not be seen in a group without the intervention. A Parameterized Difference prediction predicts that there will be an effect in one group and not another, where one group has a set of parameters whose values are significantly different from the other group. The type of prediction affects the underlying relationship and hypergraph structure of the hypothesis. Hypotheses with the same prediction type will have similar connectivity. A hypothesis structure can be reused even if there is no concept similarity between hypotheses. Retrieving a hypothesis by prediction, or structural, similarity is complementary to concept similarity retrieval.

The HRE does not currently search for similar, arbitrary network structures; this is a possible area for future research. Instead, it defines common hypothesis network patterns that can be used as templates for hypothesis formation and reuse. These patterns, designated as Hypothesis Design Patterns, act as scaffolding upon which domain components can be assembled. The idea is to use the templates to make hypothesis prediction formation more efficient and more accessible to non-research professionals.
5.1.2 Hypothesis Design Patterns

It is not the intent of this research to replace the creative aspect of hypothesis formation. Instead, the goal is to assist in the design process of transforming relatively unstructured hypothesis statements into testable, falsifiable hypotheses. Contributing to this goal are Hypothesis Design Patterns that capture and reuse the creativity of existing hypotheses and their authors. These are conceptually similar to knowledge maps [150], where meta-patterns of domain knowledge are repeated and reused, to provide a reusable template for construction of a reasoning pattern. This section presents the Parameterized Difference Hypothesis that is a common pattern in hypothesis formation. The underlying idea of this predictive model is that a difference in parameters, given otherwise nearly identical conditions, will lead to a difference in outcomes. A conceptual diagram for the Parameterized Difference Hypothesis Model (PDHM) is shown in Figure 8.

As can be seen by the figure, the prediction includes two sets of conditions and outcomes. The relationship between each of the condition/outcome pairs is a set of leadsTo relationships, which is represented as a directed hypergraph, with nodes of class condition and outcome, and leadsTo edges. The leadsTo relationship set does not imply causality, but approximate correlation, with the source usually occurring in time before the sink. An example of this is: Mitral valve prolapse “leadsTo” Backflow “leadsTo” Heart Attack [151]. In the PDHM, there are disjoint groups defined by a difference in their conditions, specified by parameter values. For instance, for the mitral valve replacement example, the difference might be a difference in median ages of the groups, or a difference in gender, or existence of other health factors. The use of a template is intended to simplify the representation for hypothesis authors, while retaining the rich networked structure of the hypothesis in HDL.
**Figure 8. Hypothesis Design Pattern: Parameterized Difference Hypothesis**

**Hypothesis Design Pattern Template.** Represented as a template, a PDHM hypothesis design pattern can be represented by a set of statements, having an underlying HDL representation.

Hypothesis Statement: {concept A} {relatedTo} {concept B}
Population description: {BBox}
Group parameters = [p1, p2, p3]
Group1: {BBox}
Group2: {BBox}
Condition: C
Outcome: O
Outcome measure: P
Statistical measures = [avg, med, stdev, etc.]
surrogate(P,O)
leadsTo(C,C2), leadsTo(C2,C3), leadsTo(C3,O)

**HDL Representation.** Represented in HDL, the PHDM specifies the types of roles and concepts involved in the prediction:

A population, G_P, is defined by a set of population parameters, P_G, and BBox Boolean expressions:
The set of parameters that subdivide a population into disjoint groups, $G$, are group parameters, $P_G = \{p_1, p_2, \ldots, p_n\}$. In HDL, each group represents a HDL concept.

$G_1 \equiv G \cap \text{BBox}_1(P_G)$
$G_2 \equiv G \cap \text{BBox}_2(P_G) \cap \neg G_1$
$\ldots$
$G_n \equiv G \cap \neg G_1 \cap \neg G_2 \cap \cdots \cap \neg G_{n-1}$

Every member of the group is connected by a role has_condition to $C$, a DL tree concept in the class Condition.

$C \subseteq \text{Condition}$

$G_1 \cap \forall \text{has_condition}.C$

A group may be connected to concepts by more than one instance of the has_condition role.

Some members of the group are connected by a role has_outcome to a concept $O$, a DL tree concept in the class of Outcomes.

$C \subseteq \text{Outcome}$

$G_1 \exists \text{has_outcome}.O$

Some concept, $C$, is connected through a set of plausibility domain hyperroles to a concept, $O$.

$e_1 = (\text{leadsTo}, \{G \cap \{\{C\}, \{C_2\}\}\}, \text{DH})$.
$e_2 = (\text{leadsTo}, \{G \cap \{\{C_2\}, \{C_3\}\}\}, \text{DH})$.
$e_3 = (\text{leadsTo}, \{G \cap \{\{C_2\}, \{O\}\}\}, \text{DH})$.

An outcome is connected through the role has_measure to an instance of the concept, Parameter.

$G_1 \cap \exists \text{has_outcome}.(O \cap \exists \text{has_measure}.P)$

Defining new concepts, $X_1$ and $X_2$, denoting groups whose members participate in has_condition and has_outcome roles, and where the outcome concept participates in a has_measure role with an instance in the class Parameter.
\[ X_1 \equiv G_1 \cap \forall \text{has_condition}.C \cap \exists \text{has_outcome}.(O \cap \exists \text{has_measure}.P) \]
\[ X_2 \equiv G_2 \cap \forall \text{has_condition}.C \cap \exists \text{has_outcome}.(O \cap \exists \text{has_measure}.P) \]

Statistical measures are defined:

\[ \text{hasCount, hasAverage, hasMedian, hasStdDev, hasVar, hasDistrib} \subseteq \text{hasStatParam} \]

Note that statistical measures are not properties of individuals; statistical measures are properties of a group or set of individuals. Using hyperroles, we can designate statistical relationships, \( Z \), where the statistical measure is a relationship between a group, or set of individuals \( \{G\} \), and an instance of the class Parameter.

\[ \{\text{set of statistical tests}\} \subseteq \text{StatisticalTest} \]
\[ \{\text{Difference, NoDifference}\} \subseteq \text{StatisticalResult} \]

The HDL representation utilizes roles and hyper-roles to form an underlying network of knowledge connections implementing the Hypothesis Design Pattern.

**Tai Chi Example.** The Parkinsons Disease TaiChi study population, expressed as in Section 4.3.4, is designated in this example as \( G \). The groups are defined by the value of their role exerciseParticipation, either “Tailored Tai Chi” or “Flexibility Exercises”. The values of the role exerciseParticipation are the conditions of interest for this population. Balance measures and number of falls are the outcomes of interest for this population. As a simplification of the actual
study, two groups and one outcome measure will be illustrated.

TailoredTaiChi ⊆ BalanceExercise ⊆ Condition
FlexibilityExercise ⊆ Condition
Balance ⊆ Outcome
Falls ⊆ Outcome

\[ G_1 \equiv P \cap \exists exerciseParticipation.TailoredTaiChi \]
\[ G_2 \equiv P \cap \exists exerciseParticipation.FlexibilityExercises \cap \neg G_1 \]

Each group is connected by the role has_condition to Parkinsons in the population definition, and by the role has_outcome to Falls, that have a measure NumberOfFalls.

\[ X_1 \equiv G_1 \cap \forall has\_condition.Parkinsons \cap \exists has\_outcome.(Falls \cap \exists has\_measure.NumberOfFalls) \]
\[ X_2 \equiv G_2 \cap \forall has\_condition.C \cap \exists has\_Parkinsons.(Falls \cap \exists has\_measure.NumberOfFalls) \]

hasCount, hasAverage, hasMedian, hasStdDev, hasVar, hasDistrib ⊆ hasStatParam

\[ Z_1 \equiv \{X_1\} \cap \exists hasAverage.NumberOfFalls \]
(Equivalent to hyperedge, \( e = (\text{hasAverage, \{\{X_1\},\{NumberOfFalls\}\}}, \text{ DH}) \).)

\[ Z_2 \equiv \{X_1\} \cap \exists hasAverage.NumberOfFalls \]

Testing the null hypothesis, we determine if there is no difference.
Result \( \equiv \{Z_1, Z_2\} \cap \exists\text{Test.NoDifference} \)

If result NoDifference is true, then the Result is false, and the prediction is falsified, as well as the hypothesis.

The Parameterized Difference Hypothesis Model may be slightly modified to support several types of clinical research, including Case Control studies and Cohort studies. Case Control Studies identify differences in conditions or parameter values in two groups, one with an outcome, the other without the outcome, where the outcome is often a disease condition. A case-control study requires a separate hypothesis for each condition that is examined. Cohort Studies look for differences in outcome in one group at two different points in time, relating risk factors to outcomes. A more specialized form of the Parameterized Difference Hypothesis Model is the
Interventional Difference Hypothesis Model. In this model, the only intended significant difference between groups is the application of an intervention. This is a common health research model used in Interventional Clinical Trials. Randomization is often used to ensure that there are no effective differences between the groups other than the intervention. Group assignment by randomization is treated as a difference in a randomization parameter, \( P_G = R \).

5.1.3 Constructing Health Hypotheses

Using CaNDoR, the process for constructing health hypotheses is as follows:

1. Create a natural language description of the hypothesis in the Project Domain.
2. Assign values to the project domain roles (attributes) such as author and dates.
3. Create a hypothesis statement, a relationship between concepts, in the Prediction Domain.
4. Create a list of concepts and measures that may be of importance or have relevance.
5. Find similar hypotheses based upon concept similarity. Review these.
6. Determine the Hypothesis Design Pattern to be used.
7. Select a hypothesis that is similar in structure and content (prediction and concept similarity) to the hypothesis being designed.
8. Reuse domain elements from hypotheses with similar concepts that are applicable, attaching them to the scaffolding of the selected Hypothesis Design Pattern.
9. Review the hypothesis maturity state to determine the current maturity of the hypothesis.
10. Check for missing information for the Hypothesis Design Pattern, until the pattern is complete, and the hypothesis adequately represents the research idea.

One of the advantages of using a Hypothesis Design Pattern is that, for a subset of hypothesis types, it transforms the process of developing a health hypothesis, an otherwise creative process,
into a template-based process that provides some assistance for those who may not have years of experience in hypothesis design. It does this through the identification of the type and structure, and connectedness of information needed, and by identifying missing information. The underlying complexity of the HDL knowledge network is not exposed to the hypothesis author, but the network is created, providing a foundation for hypothesis queries and reasoning.

5.2 Managing Health Hypotheses with HyRE

The Hypothesis Reasoning Engine (HyRE) provides essential hypothesis management functions for the Health Research Exchange. The modularity feature of the MoHM ensures that operations within a single hypothesis, across its domains, are straightforward. The domain specificity feature of the MoHM ensures that operations within a single domain, but across large numbers of hypotheses, are efficient and scalable. These functions include:

- Constructors and destructors for hypothesis data structures represented in the MoHM
- Methods for retrieving and reusing similar health hypotheses, and domain components
- Translation of MoHM elements between a working memory representation and a packaged representation that is translatable to XML
- Answering of queries about the set of hypotheses being managed.

These functions of HyRE are implemented in a SWI-Prolog environment, with additional libraries, including the SGML/XML Parser library for translation between prolog structures and XML files [152]. The following additional capabilities, not currently addressed, would be required for a HRE production environment.

- Interfaces with existing DL-reasoning engines, for SNOMED-CT concept construction, or DL queries, such as subsumption.
• A graphical user interface for ease of communication with the HRE.
• A web-based discussion forum for proposing and discussing hypotheses.
• Storage and archive solutions, for managing a large number of evolving hypotheses.

5.2.1 Hypothesis Operations

Some operations included in HyRE have already been introduced in Chapter 4. Table 19 includes some of these as well as additional hypothesis operations.

<table>
<thead>
<tr>
<th>OPERATION</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>create_HStatement</td>
<td>Create new HStatement with ID</td>
</tr>
<tr>
<td>create_prediction</td>
<td>Create prediction for hypothesis ID</td>
</tr>
<tr>
<td>create_surrogate</td>
<td>Create a surrogate relationship for a concept</td>
</tr>
<tr>
<td>create_composition</td>
<td>Create a composition relationship for a set of concepts</td>
</tr>
<tr>
<td>add_LeadsTo</td>
<td>Add LeadsTo relationship to Plausibility Domain</td>
</tr>
<tr>
<td>concept_plausibility_check</td>
<td>For a concept, is it connected to a concept in the plausibility domain?</td>
</tr>
<tr>
<td>hypothesis_plausibility_check</td>
<td>Are all concepts in a hypothesis statement or prediction connected to at least one concept in the Plausibility Domain</td>
</tr>
<tr>
<td>connected</td>
<td>Is concept A connected through role(s) R to concept B?</td>
</tr>
<tr>
<td>add_population_criterion</td>
<td>Add new population criterion</td>
</tr>
<tr>
<td>intervention_model_check</td>
<td>Determine whether a hypothesis conforms to the specification of an interventional hypothesis</td>
</tr>
<tr>
<td>hypothesis_testability_check</td>
<td>Is each concept in the prediction connected to a measurable? (via composition, surrogate, or isa relationships?)</td>
</tr>
<tr>
<td>analysis_options</td>
<td>List analysis options for a set of parameters and groups</td>
</tr>
<tr>
<td>analysis_validity</td>
<td>Determine whether the selected analysis is consistent with the analysis parameter types</td>
</tr>
<tr>
<td>hypothesis_falsifiability_check</td>
<td>Analysis validity = true Hypothesis testability check = true</td>
</tr>
<tr>
<td>similar_concept</td>
<td>Find hypotheses that contains concept C in domain D and participates in the role, R.</td>
</tr>
<tr>
<td>similar_prediction</td>
<td>Find hypotheses that are of the prediction type, T.</td>
</tr>
<tr>
<td>package_hypothesis</td>
<td>Package a set of WMR elements corresponding to one hypothesis, to a list structure, and convert to an XML tree.</td>
</tr>
<tr>
<td>unpackage_hypothesis</td>
<td>Unpackage a hypothesis or hypothesis component XML tree into its corresponding prolog database terms</td>
</tr>
</tbody>
</table>

Table 19. MoHM Operations
5.2.2 Two forms of representation in HyRE

For every hypothesis and associated hypothesis structure, there are two forms of representation – the Working Model Representation (WMR) and the Packaged Model Representation (PMR). These satisfy two, very different, requirements – 1) the need to rapidly and incrementally construct hypotheses, and to reuse hypothesis information from different domains and multiple hypotheses, and 2) the need to share hypotheses across diverse platforms where the hypothesis components must travel together to their destinations.

The WMR, designed to facilitate the construction and reuse of hypotheses, is represented as a set of HDL relations. When a hypothesis is created, it is assigned a numeric ID, which uniquely identifies it. As the hypothesis components are added, every component of the hypothesis is associated with this ID. Components of many existing hypotheses may be in the WMR for reference and reuse. In this form, incremental development and reuse of hypotheses entails adding and editing individual relations and associating them with the appropriate hypothesis ID. A hypothesis that has been falsified does not disappear. Instead it may be removed from the WMR, stored for future use, and either modified or referenced to prevent redundant efforts.

<table>
<thead>
<tr>
<th>WMR</th>
<th>PMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>project(ID, P).</td>
<td>hypothesis(ID, ‘myhypothesis’,</td>
</tr>
<tr>
<td>hStatement(ID,S…).</td>
<td>project(P),</td>
</tr>
<tr>
<td>prediction(ID, R…).</td>
<td>hstatement(S),</td>
</tr>
<tr>
<td>population(ID, P…).</td>
<td>prediction(R),</td>
</tr>
<tr>
<td>teststrat(ID,T…).</td>
<td>population([]),</td>
</tr>
<tr>
<td>anrule(ID, A…).</td>
<td>anrule([]),</td>
</tr>
<tr>
<td>plausibility(ID, B…).</td>
<td>plausibility([]))</td>
</tr>
</tbody>
</table>

Table 20. WMR and PMR concept comparison
The PMR, designed to enable sharing and reuse across diverse platforms, is represented as a tree structure implemented as a list, which has an equivalent representation in XML. HyRE provides methods for translating between the WMR representation of the hypothesis and the PMR representation. Constructors and de-constructors convert the hypotheses between the two representations. Corresponding PMR and WMR conceptual representations are depicted in Table 20.

5.2.3 Health Hypothesis Exchange Language (H2XL)

The HRE will need to interact with current health research and health data standards. For example, the NIH clinicaltrials.gov website utilizes a multi-level XML structure for sharing information about clinical research. In addition, personal health records, such as those offered by Google (until January 1, 2012) and by Microsoft, rely on the XML-based CCR and/or CCD as the primary exchange medium for health data. Consequently, a HRE environment that incorporates XML datasets and structures is essential. Thus, an important operation will be transforming between the WMR, a modular form of HDL that enables rapid conceptual modeling, the PMR, a packaged tree structure, and a form expressible in XML.

The Health Research Exchange includes an XML-based Health Hypothesis Exchange Language (H2XL), for exchanging health hypotheses. Designed to support the exchange of all information related to health hypotheses, H2XL is a XML representation for a HDL hypothesis structure. The mapping is facilitated by a H2XL Data Type Definition, similar to a XML DTD, which specifies the mapping of the tree structure to HDL relationships. Note that H2XL does not
provide a generic mapping between DL knowledge bases and XML; that is a topic of ongoing research not addressed here.

The HDL PMR hypothesis structures, implemented in SWI-Prolog, are consistent with the structure utilized by the SWI-Prolog’s SGML/XML parser. This structure is a Prolog tree of element(Name, Attributes, Content). Each Prolog structure, after translation, becomes a H2XL (XML) file. In order to avoid ambiguity in interpreting the HDL structure, reference to a H2XL DTD should be used in both in HDL-to-H2XL and H2XL-to-HDL conversions. Use of the existing SGML/XML parser and definition of a DTD enables continued development and evolution of the HRE data structures and operations, with only a change to the DTD needed, in order to maintain compatibility with H2XL.

Using H2XL, it is possible to share health hypotheses, Hypothesis Design Patterns, population specifications, and plausibility domain elements. Preparing a hypothesis for publication requires conversion to H2XL. Subscribing to the content of a hypothesis or its domains is implemented by the import of the H2XL structure into the HDL PMR and then parsing the list structures into individual WMR relations tagged with hypothesis ID. Each of the component Hypothesis Domains may be packaged either within a hypothesis, or can be separately encoded through H2XL. For example, a set of LeadsTo relationships may be included in a Hypothesis Structure, or packaged separately as a Plausibility Domain Structure. H2XL supports the exchange of the current working state of a health hypotheses, whether it be in the early formative stages without a fully-fully formed Prediction or Plausibility Domain, or later when the hypothesis is fully formed and the Test Strategy Domain and the Falsification Domain and analysis strategy are complete.
5.2.4 Answering questions about Hypotheses

In addition to providing basic hypothesis construction, transformation, and similarity matching operations, HyRE may be used to answer questions relevant to managing a large number of health hypotheses. This is especially useful in the context of large research databases, such as Phase III outcomes research, where a common data resource is shared between large numbers of diverse stakeholders, and ownership of ideas and data is an important factor. Data mining can be used to identify and develop new, interesting hypotheses by identifying patterns in data. It is an important tool in clinical research, often used as a catalyst to form new research hypotheses for clinical exploration. However, today, there is no automated platform for managing the set of mining hypotheses. This can be problematic when multiple data mining scenarios are being investigated concurrently, or when any data set has a large number of variables or models.

HyRE may be used to answer questions concerning the overlap or cover of existing hypotheses, as well as the connection between research hypotheses and research goals. Answering an expanded range of questions requires that information relevant to the project management questions must be included in the hypothesis Project Domain, including: research goals, organizational and functional affiliations, versioning of concept definitions, interest and belief scores, as well milestone dates. In addition, meta-information about concepts represented in the data mining effort, such as information type and analysis type should be included. With this additional information, HyRE may support many of the queries essential to managing a large data mining effort. Some of these are listed below:

- Which data elements are covered by a set of hypotheses?
- What is the overlap of hypotheses across organizations?
• What hypotheses refer a plausibility or biological pathway?
• Who is working on what hypotheses?
• What hypotheses (and who) are affected by a change in a concept definition?
• What types of analysis methods were utilized on a set of research projects?
• How does a new hypothesis overlap ongoing research?

5.2.5 Summary

The Hypothesis Reasoning Engine (HyRE) provides hypothesis management capabilities, enabling their construction, transformation, matching, sharing, and query answering. HyRE leverages the Molecular Hypothesis Model modular structure, domain specific DL foundations, and hyperedge connectivity, providing reusable templates and knowledge structures for Population Health Hypotheses.

5.3 Active PHRs with Personal Health Hypotheses

Previous chapters have focused on Population Health Hypotheses, hypotheses concerning relationships in populations of individuals. This section proposes a conceptual plan for organizing health data around Personal Health Hypotheses. It begins by motivating the need for a PHR organizational principle. It then describes Personal Health Hypotheses and their role in transforming Personal Health Records to active participants in the health process. Finally it extends the notion of Hypothesis Design Patterns to health hypotheses patterns for specific health conditions, or Health Condition Profile (HCP) templates.

Medical technology is rapidly advancing, and with it the diversity, complexity, and volume of information. Many different sources and types of medical records are utilized, transferred, and
stored during medical care, including handwritten notes, text, databases, graphics, images, and other visual representations of information. Consider, for example, a patient that has a high white blood cell count, and that undergoes a diagnostic analysis and subsequent treatment for Chronic Lymphocytic Leukemia (CLL). The tests for this may be performed by several external laboratories with the results combined into one comprehensive lab report with the following information: 1) text describing the patient, the patient history, and clinical observations; 2) a table of red and white blood cell counts with appropriate units; 3) graphical representations of the actual ranges versus the normal ranges; 4) graphical plots showing the distribution of cells at each level of maturity, 5) an embedded photograph of stained blood and bone marrow cells, with abnormal cells indicated by a superimposed arrow; 6) pictures of a set of human chromosomes; and 7) a textual assessment of any chromosomal mutations or biomarkers that may be relevant to the diagnosis. This lab report may then be combined with annotated photos of the patient showing any visible manifestations of CLL and a typewritten physician’s assessment of the diagnosis and recommended treatment strategy. A report will be prepared each time the patient undergoes an assessment of status. If the physician searches external literature to determine the best course of treatment, additional published papers or industry standard guidelines may be referenced in the medical notes. When the patient is hospitalized for any reason, the medical file will also accumulate potentially hundreds of pages of doctor’s orders, doctor’s status assessments, and manually-annotated event and medication histories. If the patient has other medical conditions, the records for those conditions may be collected and stored at the point of care, which may at be a completely different institution. If the patient does not request that medical records are transferred between institutions, the patient’s physician may not have the necessary information to perform an accurate assessment of health or to prepare a treatment plan.
Over a patient’s lifetime, he or she is likely to accumulate a series of patient records, most probably stored at a multitude of clinics and institutions. Due to the complexity of information and the health industry and information systems fragmentations, the organization and management of personal health information is a daunting endeavor and is one of the most challenging unsolved information management problems today.

However, if one looks closely at the information that is included in a medical record, there is one thing that most of the content has in common – the tests and evaluations that comprise the medical record largely result from the need to create and test hypotheses that seek to explain some symptoms or behaviors of an individual. Other than routine health procedures such as vaccinations, and regular checkups, people typically don’t visit a health care practitioner unless they have a known problem or suspect that something is amiss or might potentially become a problem. An ethical physician doesn’t order non-routine procedures unless there is, under consideration, a hypothesis that seeks to explain a set of symptoms or conditions – a hypothesis that should be tested.

Consider a new paradigm, the active PHR, which organizes health information around a set of current and prior personal health hypotheses. In this new paradigm, several assumptions are made. 1) The person is at the center, and all other entities are viewed as peripheral and transient service providers; 2) Just as a person’s body integrates, through biology, all of the events and medical interventions it experiences, the person’s medical information is integrated through the set of health hypotheses that are associated with the individual over time; 3) Health information about a person can, with the person’s permission, be shared with other external entities, within
the context of the health hypotheses that drove the collection of the health information; 4) The personal health record becomes more than a passive recording of events, as current hypotheses under consideration may trigger active intervention by the PHR on behalf of the individual – to search for published information relevant to the hypothesis, to identify missing information needed to test a personal health hypothesis, to find relevant clinical research, and to facilitate engagement of the individual in meaningful discussions with health researchers.

The Health Consumer and PHR as active participants in the HRE

![Health Consumer and PHR diagram]

Figure 9. The active Personal Health Record (aPHR).

5.3.1 The Personal Health Hypothesis (PrH)

The Personal Health Hypothesis (PrH) is a specialized type of health hypothesis that addresses the personal state of health for an individual. It is similar in structure to the Population Health
Hypothesis (PoH) described in Chapter 4, and Chapter 5.1-5.2. As with Population Health Hypotheses, the MoHM features of modularity and connectivity are important for Personal Health Hypotheses. However, the property of domain specificity becomes especially important for PrH, as the population domain of the PoH, becomes a major avenue for enabling connectivity, or engagement between population health hypotheses and the health profiles of health consumers. Like the PoH, the PrH includes a Hypothesis Statement, at least one Prediction that can be tested, a Plausibility Domain with links to relevant supporting information, and at least one strategy for testing and falsifying the hypothesis. Unlike the PoH, which is evaluated within the context of a specified population, and which features a plan to falsify a prediction for that population through statistical analysis, the PrH has a population of one, and so cannot be falsified through statistical analysis. Instead, the PrH may have many predictions and tests for one health condition for a single individual. The PrH postulates a health condition that explains symptoms that an individual has and makes falsifiable predictions that are likely to be true if the health condition is present. Alternatively, the PrH can be viewed as a hypothesis about population membership, i.e., a hypothesis that states that a person is a member of a group that shares a specific set of conditions. Using this viewpoint, the population domain specifies the set of conditions that defines the set of individuals that are members of this group. The tests, data, and analyses that are a result of the PrH are associated with it, and linked to it, thus organizing health data around Personal Health Hypotheses. A person’s PHR may have many different health hypotheses, some of which overlap in that they may compete to explain the same sets of symptoms. As long as the hypothesis has not been falsified and the health condition ruled out, it may remain active as a PrH.
Here is a simple, ad hoc example. If Joe’s toe is very painful after stubbing it and it is swollen, then Joe might hypothesize a PrH H-Statement, “My toe may be broken.” For that H-statement, the Prediction might be, “X-rays will show a break in my toe bone.” The Plausibility Domain for that PrH would include a statement, “Broken bone LeadsTo pain and swelling.” The test strategy is a visit to the doctor and a request for an x-ray. The analysis strategy, in this case, is to rely upon the doctor to make the assessment. Absence of a fracture or break in the x-rays would falsify the prediction, and thus the hypothesis. The doctor’s notes and x-ray results, when completed, will reside in Joe’s aPHR, linked to the hypothesis described as “My toe may be broken.” When the doctor designates that the toe bone is healed and is no longer broken, it becomes an inactive hypotheses, but remains in the historical record of the PHR, linked to the x-ray data. The x-ray always carries with it the context of the hypothesis in which it was requested and evaluated. For the more complex CLL example in this chapter’s introduction, many different predictions and tests would be made to assess and respond to the CLL hypothesis.

5.3.2 Health Condition Profiles (HCP)

Recasting the Personal Health Hypothesis as a hypothesis about membership in a population leads to the following approach – defining a Health Condition Profile (HCP) template that comprises the set of predictions and results that are usually associated with determining whether an individual is a member of that population.

This HCP template is a specialization of the Hypothesis Design Pattern concept. However, the HCP template goes further than defining the relationship structure; it specifies and organizes the set of predictions, tests and measures associated with commonly encountered health hypotheses.
Each HCP template, ideally, would be created and published by healthcare experts. Each template includes the relevant information necessary to describe the health condition, as well as the recommended tests and analysis strategies that should be used to determine whether a person is a member of the specified health condition population. A PrH need not use a HCP template. However, creation of a PrH can be much more efficient, fast, and accurate if a HCP template is available.

With the HCP template, the PHR becomes empowered with information about what is expected and what is missing for a health condition profile. The missing information may trigger actions by the PHR to remind the individual about what is missing or to seek more information about the condition on behalf of the health consumer. The actual data elements, such as images, diagnostic test results, and physician summaries, may reside in specialized data repositories with uniform representation for efficiency and performance, but the linkage to the HCP acts to connect each data item to the context of specific hypotheses.

Early research in integration architectures that support and inform this concept was performed at Stanford University in the mid 1990’s. A component-based integration method was developed where components were defined as conditions or disease specific bodies of knowledge. Basic time-stamped data from Electronic Medical Records were evaluated to determine the relevant higher-level concepts, such as disease states. Domain ontologies were used to guide further knowledge acquisition of relevant information. A library of problem solving methods was developed including the planning of structured, protocol-directed therapy. Treatment protocols were selected and used to develop patient-specific treatment plans. Patient response to
treatment was evaluated against these protocols [153].

The aPHR may participate in searching for relevant clinical research for a HCP. A HCP can be used by an aPHR to create a “message” that can be shared, containing only specific parameters and values relevant to that health condition. In addition, the HCP makes it possible to include default values in that message, when actual values are unknown or deemed not sharable. A message created from a HCP is a structured method for expressing a health condition. As described in Chapter 6, a HCP can be used as an information-rich expression of interest in a health research endeavor, when engaging in productive discussions with health researchers or with others with a similar condition. A HCP message might be initiated by a health consumer, or, with a health consumer’s permission, by an aPHR.

5.4 Contributions

The Molecular Hypothesis Model (MoHM), presented in Chapter 4, provides a framework for conceptualizing and representing health hypotheses, which has the distinguishing features – modularity, domain specificity, and connectedness. This framework was designed to engage the health consumer in health research design and discourse, while remaining consistent with the principles and practice of clinical research. Building upon the MoHM, Chapter 5 develops the inference and organizational approaches needed to make the MoHM useful.

This chapter illustrates how the MoHM features are leveraged to create a reuse and management methodology for health hypotheses. The MoHM methodology aims to narrow the conceptual gap between health consumer and health researcher, making a shared understanding possible.
The MoHM methodology makes possible three important HRE capabilities: 1) an approach similar to case-based design for constructing and reusing hypotheses, 2) a hypothesis management infrastructure, and 3) a conceptual plan for using health hypotheses to organize the diverse information in personal health records. The contribution of this chapter is to illustrate how the MoHM and its characteristic features, makes these capabilities possible.

The first capability is a methodology for development of Population Health Hypotheses that leverages the vast knowledge contained in existing hypotheses, reusing this knowledge to make hypothesis formation more efficient and accessible. The methodology, Case Network Domain Reasoning (CaNDoR), is enabled by the modularity, domain specificity, and connectedness of the MoHM. Decomposing hypotheses into modular domains enables the reuse of hypotheses in a variation on case-based design. Each domain has a function, and a corresponding representation specific to that domain’s function, enabling domain-specific building blocks that can be reused. Building blocks, design patterns, and inference methods are made available, so that people who are not health researchers are provided with representational and computational aids for participating in the health research definition process. Connectedness enables two types of similarity matching: concept similarity and prediction-based similarity, a kind of structural similarity. Hypothesis design patterns provide templates, represented as patterns of connectivity, for study designs. Two design patterns are the parameterized difference hypothesis model and the interventional difference hypothesis model. An important feature of CaNDoR is the ability to incorporate realistic medical terminology and subsumption reasoning through SNOMED-CT. Through the MoHM, the average health consumer may have access to many examples of reusable health hypotheses, as well as reusable links to underlying biological pathways.
The second capability is the management of large numbers of hypotheses. HyRE, the Hypothesis Reasoning Engine, supports hypothesis management functions including construction, retrieval, re-use, translation, sharing, and querying. It supports the packaging and sharing of health hypotheses through XML tree structures. The MoHM feature of connectedness permits the answering of important questions about hypotheses such as overlap, or cover, as well as sharing of underlying plausibility chains. Domain specificity ensures that operations within a single domain, but across multiple hypotheses are efficient and scalable.

The third capability is the organization of Personal Health Records around Personal Health Hypotheses (PrH). PrH’s are intended to be stored, not in a hypothesis database, but in the individual personal health records of individuals, enabling the conceptualization of an active Personal Health Record. PrH’s organize and provide context for personal health data in PHRs. A Health Condition Profile (HCP) extends the concept of templates to the Personal Health Hypothesis, with HCP messages acting as a data sharing mechanism. Modularity and connectedness are important for both population and personal health hypotheses. However, the property of domain specificity becomes especially important for personal health hypotheses, as the population domain becomes a major avenue for enabling connectivity, or engagement, between population health hypotheses and the HCP’s of health consumers.

The MoHM is the central model that provides a foundation for the transformation of health hypothesis development to a form more accessible to the average health care consumer. The MoHM is designed to enable health consumers to state health hypothesis ideas in a way that will
interest, engage, and even educate, clinical research professionals. Leveraging the MoHM, the amorphous process of hypothesis formation – which currently relies heavily on “inspiration” or a “blaze of insight”, as well as years of clinical research training – becomes more defined. Management of large bodies of hypotheses becomes, through modularity, domain specificity, and connectivity, more feasible. Integrating health hypotheses into Personal Health Records, as an organizational principle, further reinforces and integrates the concept of health hypotheses with the health consumer. Through the MoHM, the health consumer will have the opportunity to participate in a shared understanding of health research ideas and projects. This shared understanding is essential to fulfilling an important requirement for the HRE – collaborative development of health hypotheses – that will be addressed in Chapter 6.
CHAPTER 6  Health research collaboration in the HRE

One currently recognized problem in health research is insufficient opportunity for health consumers and patients to provide input to clinical research design. This is especially true with respect to patient-centered outcomes, i.e., clinical outcomes important to patients [154], and patient-centric participation, i.e., patient-initiated participation [155]. An important unaddressed issue is the collaboration process. Individual research projects may sponsor focus groups, but there is no standard health research collaboration process that individuals can initiate or be involved in. This chapter proposes a model for the collaboration process in health research, describing a framework for connecting and engaging health researchers with health consumers and with other health researchers. This model addresses the following questions: How can health researchers collaborate with health consumers and other health researchers to obtain their ideas relevant to a specific health hypothesis? How can health consumers initiate discussions with health researchers in mutually relevant areas? How can matching research hypotheses to potential patients be accomplished on a large scale?

This chapter introduces a collaboration model, the Hypothesize/Engage Protocol (HYPEG), which facilitates the interaction of health researchers and potential research participants while research is still in the formative stages. It does this by centering discussion around the Molecular Hypothesis Model (MoHM) and leveraging its capabilities for hypothesis construction, management, and messaging. The HYPEG protocol is innovative in that it focuses on different perspectives, or domains, of the hypothesis, as the type of interaction changes. HYPEG enables HRE participants: 1) to hypothesize research ideas and find relevant hypotheses (utilizing
CaNDoR and HyRE), 2) to match health condition profiles to population inclusion and exclusion criteria, and 3) to engage in in-depth discussion of a specific research hypothesis. While HYPEG is targeted at people-to-people interactions, it recognizes the Personal Health Record as an important element in the collaboration process. The active Personal Health Record may participate in HRE by organizing its information around personal health hypotheses, expressed as Health Condition Profiles, and by automatically constructing some of the messages essential to HYPEG. The HYPEG protocol also addresses issues of scale with its use of a knowledge compilation approach, featuring Ordered Binary Decision Diagrams (OBDD).

### 6.1 Models for Collaboration

Collaboration can be difficult to define. People have differing views of what collaboration really is, ranging from general communication to developing a deep agreement on issues. Some argue that collaboration requires a team to “gather together in the pursuit of shared goals” [156], while others believe that both structured and unstructured approaches to collaboration have value. Collaboration requires a shared understanding of the whole [157]. Collaboration differs from cooperation, in that in cooperative environments, work is shared by breaking it down into smaller pieces, and a shared understanding of the whole is not necessarily essential. Some have suggested more formal definitions for collaboration. One definition of collaboration is “a coordinated, synchronous activity that is the result of a continued attempt to construct and maintain a shared conception of a problem” [158]. Rochelle argued that “the crux of collaboration is the problem of convergence”, and that the process of convergence is characterized by the development of a deep-featured situation . . . through the constructive use of interactive cycles of conversational turn-taking, constrained by the application of progressively
higher standards of evidence for convergence”. Since there are many different definitions and models for collaboration, what becomes important is a shared understanding of the collaboration model itself – making clear what is expected or not expected from participant [159].

6.1.1 Collaborative Design

Collaboration has been studied in depth in the area of collaborative design, where artifacts are designed by teams of experts in diverse areas, with the focus on ensuring that all perspectives of the problem are considered concurrently. The premise is that collaborative design produces a better design that meets a wide variety of requirements and that can be manufactured within defined constraints of cost, time and quality. A large body of literature is available describing the challenges, and proposing a variety of models and systems for collaborative design [160].

While at first glance, collaborative design and collaborative research are very different tasks, clinical research shares many of the same problem characteristics as collaborative design – a very large space of possibilities, a hierarchy of domain concepts, the definition of connections between concepts, the need for interdisciplinary collaboration and structured process, and the need for sharing ideas, and scoring, tracking and evaluation functions. These are described in [160]. Collaborative design has been successfully applied in many different fields.

6.1.2 Collaboration Environments

The definition of collaborative environments has evolved considerably with the adoption of new technologies, such as email, instant messaging, on-line videoconferencing, electronic white boards, and new information-sharing forums such as Facebook and Twitter. As such, the field of collaborative environments is evolving rapidly. The important features of a collaboration
environment vary depending upon the collaboration goal - whether for education, social networking, or for accomplishing work in a distributed environment. Where accomplishing work in a distributed environment is a goal, important collaborative features include team decision-making to determine a course of action, developing shared understanding, and intelligence analysis, or data mining of group-owned data [161], [162].

There are several examples of collaborative environments in health research. Alzforum, the Alzheimer’s Disease Research Forum (www.alzforum.org) [163] is a research community-supported site that manages information about the research community, literature, and scientific resources for Alzheimer’s disease, with the goal of promoting communication, research efficiency, and collaborative multidisciplinary exploration. Its SCF (Social Collaboration Framework) [164] is a “Web 2.0” platform for collaborative biomedical research, supporting online discourse and data sharing. Alzforum is aimed at community building and knowledge management, and attempts to maintain a constant sense of momentum and community focus. It acts as a library of the current state of the art in Alzheimers disease. It’s ALZSWAN application serves as an annotated library of all known Alzheimer’s hypotheses.

SWAN (Semantic Web Applications in Neuromedicine) is a project supporting Alzforum with RDF-based knowledge bases, including a textual knowledge base of current hypotheses about Alzheimer’s Diseases. Within SWAN, there are diverse hypotheses about the nature and etiology of the disease. The basic element in SWAN is the “claim,” where each claim may be annotated with evidence that supports, counters, or is neutral, with respect to the claim [164]. Evidence entries also may be annotated claims. A result is an annotated network of hypotheses.
Collaboration can be an important aspect of hypothesis formation for two functions important to health research – data analysis and data mining. Cross-Industry Standard Platform for Data Mining (CRISP-DM) [166] is a collaborative data mining process and model developed by a consortium of data mining suppliers and users, http://www.crisp-dm.org [166], where well-defined roles and processes, as well as a centralized model of evaluation, were used to improve communication. The power of collaboration in data mining has been demonstrated by the Netflix Prize experience over the past few years. This data mining contest (http://www.netflixprize.com), initiated in 2007 [167], offered a $1M prize for a very clear goal – improving on their existing methods by 10%. The Netflix community developed a social structure that incentivized contribution. The leaderboard – their overall scoreboard – became a central resource for evaluating competing hypotheses, helping to identify strategies, providing information on analysis turnaround time, and prioritizing efforts of the community [168].

Collaboration environments have in common the notion of agreed-upon collaboration patterns. These patterns may be very structured, or semi-structured, but must provide some guidelines and formats for how information is to be shared, so that a shared understanding of the task at hand is achieved. While there are many different types of communication patterns, the HRE takes the approach of leveraging the Publish/Subscribe Communication Pattern, customizing it to the needs of health research collaboration, and to the requirement of large-scale matching of health research ideas to potential research participants.
6.1.3 Publish/Subscribe Communication Pattern

Publish/subscribe architectures feature a messaging paradigm where information producers publish their information, and information consumers express their interest in, or subscribe to, a subset of relevant messages. A key advantage of the publish/subscribe paradigm is the ability to separate, both in space and time, producers and consumers of information. This ability is advantageous in remote collaboration efforts where participants may be in different time zones, or who may discover a need for information at different points in time. Publish/subscribe systems are generally implemented as middleware, located between applications and network protocols, providing a core set of functionality that makes application building more concise and efficient [169].

Several variations of publish/subscribe architectures have been introduced. In subject-based publish/subscribe [170] messages are tagged with subjects, which consumers subscribe to. In content-based publish/subscribe, subscribers subscribe to queries, which are executed over the entire set of messages [171]. Two key design choices in a publish/subscribe system are: 1) the methods for filtering and matching of subscriptions to messages, and 2) the strategy for distribution of the messages. Many challenges in the use of the publish/subscribe framework become apparent as the volume and frequency of messages increases. In content-based publish/subscribe, performance of the subscribed queries can be a major challenge, and have been addressed in [172]. Publish/subscribe approaches to facilitate collaboration have been explored, including use of a quickly-configurable, blackboard-like model [173], and an object-based event model [174] where users receive only objects of certain types, depending upon their role. In addition, new communication platforms, such as Twitter, are being explored for
collaboration and crowdsourcing on mobile platforms [175].

6.2 HRE Hypothesize/Engage for Collaboration

The HRE’s Hypothesize/Engage (HYPEG) collaboration framework provides a communication protocol that is based upon the Publish/Subscribe paradigm. It differs in that the type of information matched, and methods for matching changes with the level of engagement. The collaboration model is depicted in Figure 10. This framework features a three-phase approach, with successive cycles of interaction and engagement by health consumers and health researchers. Different domains of the MoHM are engaged at each phase. The goal of the HYPEG collaboration framework is to facilitate a shared conceptualization of health research problems, and to engage health researchers and health consumers in working together to achieve high-quality representations of health research ideas.

6.2.1 Phase One. Hypothesize

In Phase One, Hypothesize, interactions are similar to those of a hybrid topic-based publish-subscribe system. Health Researchers or health consumers transform their research ideas into MoHM population health hypotheses in a process called Hypothesize, then publish these health hypotheses to selected HRE topics. Published hypotheses may be in any state of formation, from initial H-Statement to fully formed MoHM. The hypothesis formation and posting aspect of the collaboration model is not limited to health research professionals; health consumers may also publish health hypotheses. Those who publish health hypotheses register with the system, identifying a level of expertise as health researchers: university, corporation, organization, or health consumer.
A health researcher, consumer, or active Personal Health Record may search for existing health hypotheses, and ask to be notified of any updates to selected hypotheses. Search criteria may include predefined topics, as well as concepts listed in the MoHM Project Domain and the Prediction Domain. In addition to searching for health hypotheses, health consumers may search for a Health Condition Profile (HCP) by topic. The HCP, when included in the active Personal Health Record (aPHR), acts as a hypothesis that the health consumer is a member of the selected health condition population.

6.2.2 Phase Two. Profile

In Phase Two, Profile, HYPEG interacts with the MoHM Population Domain information and representation to match health hypotheses to health condition profiles. In this phase, the system behavior is similar to a content-based publish-subscribe protocol, where health hypotheses are represented as queries and health profiles are represented as HCP messages. Health consumers post HCP messages indicating a specific set of health parameters and values, relevant to that person’s hypothesized health condition. A health consumer with an aPHR may direct it to regularly post HCP messages. A health consumer that does not have an aPHR may post HCP messages, supplying only that information that is known or deemed shareable. A person may also post a profile unrelated to his or her own health to learn about health research for a friend or relative.

After a Health Condition Profile is anonymously posted to the HRE, it can be matched through a query. Researchers who have queried for health profiles that are matches to a MoHM Population
Domain are notified when a match occurs. Query results may be in the form of enumerated solutions where each match is a published health profile, or in the form of a count of matches found. A researcher will not have access to the identity associated with a health profile, but will have access to de-identified HCP data. Through the HRE, the researcher may invite a researcher or health consumer to engage in a health research topic. The invitee may choose to remain anonymous, using a screen name instead, for identification. Health Consumers are also notified when a match occurs, and informed of the health hypotheses they have been matched to. A health consumer may request to be “engaged” in the research discussion for those health hypotheses.

### 6.2.3 Phase Three. Engage

While the researcher and health consumer will not have access directly to each other, they can post a request for engagement through the HRE. When both health researcher and health consumer agree to a state of “engagement”, a two-way exchange of information is made possible. In Phase Three, the Hypothesize/Engage collaboration model takes on the characteristics of a forum specialized to a Population Health Hypothesis, augmented with a list-based publish-subscribe protocol. Other researchers or health consumers may submit information to the forum, and also may post their contributions to the forum – hypothesis components such as references, urls, small domain concept networks, and strategies for hypothesis testing and analysis. While the owner of a hypothesis determines which information is included, contributors who are engaged will be able to see updates to the developing hypothesis. The engagement process is monitored, tracked, and mediated, in order to ensure that the HRE is being used for its intended purpose.
6.2.4 Collaboration Events and Queries

There are several types of collaboration events: Event types include: 1) creation or revision of a hypothesis in the HRE, 2) posting of a Health Condition Profile (HCP) message, 3) matching of a Population Health Hypothesis to a HCP message, 4) Submission or acceptance of a request for engagement, and 5) posting of MoHM elements to a specific hypothesis forum when Engaged. Queries are expressed as HDL BBox Boolean expressions. The primary query types include: 1) topic-based searches for Population Health Hypotheses and Health Condition Profiles at Phase One, and 2) content-based queries, matching HCP’s to Health Hypothesis Population Domains at
Phase Two. Issues of scale for content-based queries will be addressed further in this chapter in Section 6.3. When events take place that match existing queries, notification occurs. Messages are distributed via email or social media communication platforms, or are stored in the HRE registered user’s account. Table 21 demonstrates a simplified population domain and a corresponding matching HCP message.

<table>
<thead>
<tr>
<th>Hypothesis Population Domain</th>
<th>Health Condition Profile Message</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age&gt;21 ∧ condition=&quot;ADHD&quot; ∧</td>
<td>(Condition=&quot;ADHD&quot;) = true</td>
</tr>
<tr>
<td>last_hospitalization&gt;12 (months) ∧</td>
<td>(Age &gt;21) = true</td>
</tr>
<tr>
<td>currently_not_taking=&quot;drug x&quot;</td>
<td>(condition=&quot;ADHD&quot;) = false</td>
</tr>
<tr>
<td>Age&lt;65 ∧ condition=&quot;CLL&quot; ∧</td>
<td>(Age&lt;65) = true; condition=&quot;CLL&quot;;</td>
</tr>
<tr>
<td>wbc&gt;30,000 ∧ not(prior_chemotherapy)</td>
<td>(wbc&gt;30,000)=true;</td>
</tr>
<tr>
<td></td>
<td>prior_chemotherapy=false</td>
</tr>
</tbody>
</table>

Table 21. Hypothesis query

6.2.5 Example Hypothesize/Engage scenario

The following hypothetical case describes an interaction between Carl, a health consumer, the HRE, through HYPEG, and a health researcher: Carl notices that he is having difficulty breathing after exercise. He consults with his physician, who, after doing several tests, diagnoses him with severe mitral valve prolapse. The completed test results are linked to Carl’s aPHR in the context of a severe mitral valve prolapse hypothesis. The doctor indicates that Carl’s heart valve may need to be replaced, but explains that there may be other options. Carl decides learn more about the condition, so that he can make an informed decision. In addition to searching online and reading material about mitral valve prolapse, he searches the NIH clinical trials database to learn if there is ongoing clinical research evaluating other less invasive alternatives.
From the search he learns that there are a total of 89 studies that are in progress or have been recently completed, each addressing a different aspect of mitral valve prolapse treatment. Only the trials that are currently enrolling include contact information. While studying the descriptions of trials, he learns that there are extensive and technical inclusion and exclusion criteria that describe the eligibility for each of the research projects. He realizes that he would have to thoroughly understand these lists to understand if the trials are relevant to him. At that point Carl feels overwhelmed. Not only does he have a serious health problem, he has to face the decision of what to do, when he feels he doesn’t have enough information. Other than his time-constrained doctor, he doesn’t have anywhere else to turn to talk about his condition.

Carl consults with the Health Research Exchange (HRE) website. Through the HRE, he posts an interest in research in mitral valve prolapse in order to be notified any time there is health research news for this condition. He also posts a separate interest in the intervention, heart valve replacement, because he is interested in learning about complications and side effects of the surgery. In addition to receiving information, he would like to be able to discuss his condition with others who are interested in mitral valve prolapse. He learns that if he posts a Health Condition Profile (HCP) message, with his health profile information relevant to mitral valve prolapse, health researchers will have a means to connect with him. He may also be able to match his health condition with health research in the planning stage.

He decides that he would like to post a HCP message. From within the HRE he selects the condition, mitral valve prolapse, and adds information that he does know about his condition. Carl submits a request for a Health Condition Profile (HCP) template. The HRE returns the best
HCP template match for the information he provided. The HCP includes criteria and default values for mitral valve prolapse. Carl can edit the profile before posting a HCP message to the HRE for consideration by researchers. Carl may also ask his aPHR to provide his recorded health data for all or some HCP fields. Since each health profile is template-based and has the necessary information for that health condition, Carl has to do the minimal work needed to publish his HCP message. The profile is already in the form needed for efficient matching in the HRE. Carl reviews and posts his HCP message to the HRE.

In this case, Carl’s HCP message matches a particular health hypothesis in the formulation stage that has a population that includes the condition mitral valve prolapse, and that will test an alternative to heart valve replacement. The protocol is still being developed, so it would not be submitted to clinicaltrials.gov for several months. The health researcher, who is the author of the hypothesis, would like to gain some feedback from potential participants regarding the formulation of the “quality of life” measures that will be used as one of the criteria for efficacy of the alternative treatment. The researcher, seeing that Carl’s HCP message is very relevant to his research, decides to invite Carl to engage in the health hypothesis through the HRE. After “engaging” in this forum through a screen name that does not reveal his true identity, Carl provides input to the researcher on his views of what should be measured as “quality of life” as an outcome. He also points out to the researcher that one of the planned repeated measures for the clinical study would require, for many potential participants, several two-hour round-trip drives to a specialized radiology center at a university, which would be extremely inconvenient and would be a deterrent to people who might be interested in enrolling. The researcher takes this into account and, realizing that the radiology measure, while useful, was not essential to the
study goals, removes that measure from the study plan. The future enrollment of the study has been positively affected by input from Carl, a health consumer. Carl, through these interactive cycles in the forum with the health researcher, is contributing to a convergence in the definition of a research question, research measures, and a Population Health Hypothesis.

During the interaction, Carl learns that a medication, X, that he is taking for pain, is counter-indicated in the evolving research protocol that he is interested in. He requests an alternative from his doctor that is acceptable, several months in advance of the research protocol being approved and ready for enrollment. By doing so, Carl leaves open the possibility for participation in the research investigating the experimental alternative to mitral valve replacement. Otherwise, he would have unknowingly and unnecessarily caused himself to be ineligible for the clinical trial, by matching with one of the exclusion criteria – administration of medication X, within the last six months. Carl, through his interaction with the HRE, has increased the potential population and enrollment rate for a clinical research study.

The researcher also engages other researchers and health consumers in the health hypothesis topic, asking for feedback. They may post references, population criteria recommendations, recommended measures, and composite endpoint definitions, as a collaborative contribution to the evolution of the research hypotheses. If these are posted in MoHM format, they can be directly incorporated as building blocks into the evolving health hypothesis. Carl and other participants are able to see the shared contributions and discussion. Through the HRE, and by engaging in a health hypothesis topic specific to his condition, Carl has the opportunity to learn more about the issues and tradeoffs for his condition, and to participate in a shared conception of
the problem of mitral valve prolapse. By engaging in several health hypothesis topics, Carl begins to realize that there are no easy answers for his condition, but he feels more connected, more informed, and more able to deal with the decision process for his own health condition. Carl, as an informed health consumer, has become part of a community of connected and engaged health researchers.

6.3 **Design for Scale – From MoHM to OBDD’s**

One issue that must be addressed early in the design of the HRE is system performance. Opening up health research discourse to both researchers and consumers will result in significant issues of scale. There are already over 100,000 research projects described in the clinical trials database, each one with multiple inclusion and exclusion criteria. With increased participation expected in the generation of health research ideas, scaling up is an issue that should be included in the architecture design, even if it is not included in an initial implementation. An individual must be able to efficiently find research interests specific to their health and that of their families. Conversely, as health researchers seek to identify the number of potential subjects during research design, and then identify and communicate with potential participants during the study design, recruitment and enrollment phase, the large number of possible participants can be in the hundreds for rare diseases, to millions for more widely applicable research such as vaccine clinical trials.

6.3.1 **Knowledge Compilation for Performance**

There are two major approaches to addressing performance problems: 1) development of domain-specialized algorithms optimized for efficiency, and 2) compilation of a knowledge base
to a target language that can answer the target questions in polynomial time. Instead of developing specialized add-on mechanisms for performance enhancement that may change as the HRE evolves, a more efficient and general approach is proposed – to design a representation for the HRE that enables the use of knowledge compilation algorithms and rapid processing of large numbers of queries, where large-scale matching is required. In the HRE, it must be possible to do this compilation efficiently, as the knowledge base is continually changing. This is particularly important for the matching of population criteria to potential research subjects.

The Molecular Hypothesis Model, through the Population Domain provides the foundation for an efficient, innovative method for matching health hypotheses to health consumers. The BBox representation for the Population Domain is ideally suited, not only for expressing population criteria, but also for expressing population queries that can be efficiently executed.

Many other domain problems involve tens of thousands of interacting factors and multiple outcomes. These include traditional problems such as digital logic design, testing and verification, as well as prediction of chemical structure-activity relationships, and more recent applications such as discovery and verification of relationships between gene expression and biological activity. Sequences of Boolean functions are used to model many of these complex domains, the most prevalent being digital logic design and verification. A variety of methods have been developed for representing and operating on Boolean functions. One method, Ordered Binary Decision Diagrams (OBDDs) is used extensively in the VLSI design and verification field due to its compact, canonical representation and computational efficiency. The HRE leverages OBDD’s and their efficient matching capabilities for matching health research
population criteria (expressed in the HDL BBox) to Health Condition Profile messages.

6.3.2 Ordered Binary Decision Diagrams (OBDD’s)

Binary decision diagrams are represented as directed acyclic graphs (DAGs) where each internal node corresponds to the variables defining the functions and terminal vertices take on the values 0 or 1. BDDs have the decision property, where each internal node, also known as a decision node, represents an or-node of the form \((X \land \alpha) \lor (\neg X \land \beta)\), where \(X\) is an input variable, and \(\alpha\) and \(\beta\) are other decision nodes. In the syntax of BDDs, each decision node \((X \land \alpha) \lor (\neg X \land \beta)\), is more concisely denoted as \(\alpha \leftarrow_{\text{high}} X \rightarrow_{\text{low}} \beta\). The value of the BDD function is achieved through traversal of the graph, starting at the root and traveling to each node based upon the value assigned to each variable. BDDs that have the restriction of one occurrence of each variable along a path (test-only-once property) are called FBDDs. The additional restriction of a total ordering \(<\) on the variables results in an OBDD<. The language known as OBDD is the union of all OBDD< for a function. One advantage of OBDDs is the ability to achieve a reduced, canonical representation, so that for each order there is only one form of the function. Reduction to the canonical form is achieved through removal of duplicate terminals, removal of duplicate non-terminals and removal of duplicate tests. The resulting graph is sometimes referred to as a Reduced Ordered Binary Decision Diagram (ROBDD) [176].

6.3.3 OBDD Queries and Transformations

OBDDs are extensively utilized due to their ability to answer many classes of queries in polynomial time. These classes of queries include consistency, validity, clausal entailment, sentential entailment, implicant, equivalence, model counting and model enumeration. The
classes of queries that can be answered by a language in polynomial time influences what applications the language can be used for. BDDs with neither test-once nor ordering properties cannot be guaranteed to answer any of the above queries in polynomial time [177].

Another important feature of OBDDs is the class of transformations that can be applied in polynomial time. For instance, the transformations of conditioning, singleton forgetting and negation can be applied to OBDDs in polynomial time. The polytime transformations that can be achieved impact the suitability of the OBDD for certain types of applications including planning, diagnosis, belief-revision and building compilers for target languages. The ability of OBDDs to answer queries and perform transformations in polynomial time make them an excellent medium for large problems where computational tractability is key. For this reason, much work has been done on OBDDs to extend their application and to prove or disprove relevant properties [178],[179].

### 6.3.4 OBDD Variable Ordering

A key feature of OBDDs is the consistent ordering of variables in the decision tree. The ordering of OBDD input variables has a significant effect on OBDD size to the extent that a “bad” ordering may make a problem computationally infeasible. The selection of the optimal ordering for an OBDD has been shown to be an NP-Complete problem. OBDDs for multiple output functions can be combined into single OBDDs when part of the functionality is shared between the outputs. For some applications, such as circuits, this is essential. This is also essential for efficient matching of large numbers of queries, such as in publish-subscribe architectures, where there may be a large amount of overlap in the matching criteria. When outputs share much of the
same logic, the best ordering for one output may be similar to the best ordering for other outputs. However, when outputs share little logic, it may be the case that an ordering for one output may be in conflict with the ordering for another output. Although determination of an optimal ordering for an OBDD has been proven to be an NP-Complete problem, there are many practical problems for which an effective ordering can be determined, by utilizing information about the domain [180], [181].

There are many different methods of variable ordering. These are generally classified as either static variable ordering (SVO) or dynamic variable ordering (DVO). Static Variable Ordering approaches rely on structural or topological information as a basis for inferring functional dependency and impact. Several examples of SVO methods include: Depth first search (DFS) [182], Dynamic weight assignment [183], Depth first search with interleaving [184], and a sampling approach [185]. Domain-specific ordering methods, based upon circuit partitioning and placement [186], have shown good results by representing circuits as hypergraphs and minimizing the net length and max cut for each circuit. Dynamic variable ordering (DVO) is a class of approaches that begins with an initial order, which is analyzed and modified at internal points such that the size of the OBDD is minimized. There is no single SVO or DVO method that works for all types of circuits. Machine learning has also been used to address the ordering problem. [187]-[188] One key consideration with machine learning methods is the time that compilation takes.

### 6.3.5 OBDD’s in publish-subscribe systems

Research at Microsoft and CMU [189] defined a mapping of filtering in publish-subscribe
system to OBDD’s in order to address issues of scale. The general idea of this approach is to represent many subscription queries by a multiple-output OBDD, where each output, $o_i$, corresponds to a query, $Q_i$, and each node corresponds to atomic sub-formulas of the queries. The atomic sub-formulas are defined by a subscription language that comprises typed attributes, $v$, with type $E\{\text{INT, DBL, STR}\}$, atoms, $c$, and operators $<, >, =, \text{and substring(s)}$, that can be applied to pairs of atoms. Queries are defined by Boolean expressions on these atomic sub-formulas. Messages are defined by assigning values to a sequence of attributes. Atomic formula are mapped to propositional variables, $x_i$, labeled with the natural numbers. The adjacency matrix representing the OBDD has the form: $\text{node}[i] = (\text{low}[i], \text{high}[i], \text{label}[i], \text{value}[i])$, consistent with the OBDD labeling described earlier.

An example message is represented as:

$$V = \langle \text{company, product, price} \rangle \text{ over the typed event schema } \langle \text{STR, STR, DBL} \rangle \text{ as }$$

\begin{itemize}
    \item \text{<company> IBM </company>}
    \item \text{<product> PC AT </product>}
    \item \text{<price> 1000 </price>}
\end{itemize}

An example query is:

(\text{company}="\text{IBM}\) or \text{company}="\text{Dell}\) or \text{company}="\text{Siemens}\)\) AND \text{"PC\} (substring of) product AND price<100.

Matching is achieved by running a bottom-up evaluation procedure, EvalBDD, on the multiple-output OBDD. EvalBDD assigns the message value for the variable to the variable, and then evaluates the OBDD node. Incomplete messages (those that do not contain all attributes) may be handled by assigning default values. An output that evaluates to 1 indicates that the corresponding subscription matches the message. In this approach, using OBDD’s, the time for evaluation of messages is linear in the number of OBDD nodes. In addition, the size of the
OBDD’s was deemed acceptable; consequently, ordering of variables for this domain was not considered to be an issue and was not explored.

**6.3.6 The unknown problem with OBDD’s**

One of the challenges of using OBDD’s for matching health profiles with health hypotheses is the problem of unknown values. This is important because a person hoping to search for and find a relevant health hypothesis may not know or have access to health parameters values needed for matching to population inclusion and exclusion criteria. Furthermore, a health researcher looking for people who meet specific criteria, may because of missing information in health profiles overlook many potential subjects. The problem is that, since a node in an OBDD must yield either 1 or 0, true or false, representing one or more unknown values may lead to multiple OBDD’s, additional variables, or multiple path traversals. When there are a large fraction of unknowns, a large number of alternate traversals of an OBDD, may result, transforming an efficient traversal and solution into a search problem in a large search space with many solutions.

Some researchers have equated unknown values to don’t care values [190], created new representations, such as Ordered Ternary Decision Diagrams. Other researchers have used attributed edges and additional variables to represent don’t care values [191]. The problem with creating entirely new representations of OBDD’s to handle unknowns is that, by diverging from the mainstream of OBDD infrastructure, one leaves behind a large number of tools and methods for efficiently creating OBDD-based capabilities.
6.3.7 OBDD’s in the HRE

The HRE, through HYPEG, applies and extends a content-based Publish/Subscribe approach in a novel way. In HYPEG, hypothesis Population Domains (represented in the HDL BBox) take on the role of subscriptions and are abstracted as queries, while HCP’s are the source of messages that are a vector of assigned values. When hypotheses are posted to the HRE through the Hypothesize process, the population criteria in the BBox are precompiled into a multi-output OBDD, where each output represents a hypothesis. When there is a match of a message and a query (a HCP message to a population hypothesis) there are two possible distribution operations: a) The hypotheses (queries) that match are delivered to the publisher of the message, the health consumer, or b) The health profile (message) is delivered to the health hypothesis (subscriber). Distribution operations may only occur if the owner has agreed to the distribution.

HRE defines types \{\text{INT, DBL, SCT}\} where type SCT, refers to a SNOMED-CT term. The use of SNOMED terms instead of strings as a data type is expected to be superior for medical terminology, since a string-matching approach may not find matches between terms that are equivalent but not expressed or composed in the same way. HYPEG recognizes and leverages the fact that SNOMED-CT represents medical terms as compositions and that subsumption is a tractable operation supported by SNOMED-CT.

The HRE does not entirely solve the problem of unknown values, but it does reduce the impact through the following three strategies: 1) leveraging of SNOMED-CT’s terminological expressions for expressing unknown, not-found, and incomplete findings, 2) use of HCP templates enabling the assignment of a default value for every health concept and measure in a
health profile, and 3) grouping of health hypotheses by health condition families, thus increasing the number of OBDD’s but decreasing the number of variables in each one.

6.4 Conclusion

The Hypothesize/Engage (HYPEG) protocol enables a multi-phase exchange of information, which becomes progressively more interactive and constructive, as the depth and volume of information exchanged increases. Central to HYPEG, is the Molecular Hypothesis Model representation for hypotheses, used as a basis for discussion and collaboration. In addition, HYPEG leverages the MoHM capabilities of hypothesis construction and reuse (CaNDoR), management and queries (HyRE), and messaging (HCP). The three HYPEG phases – Hypothesize, Profile, and Engage – provide a process for health consumers and health researchers to interact and collaborate in health research. The HYPEG collaboration protocol is a variation on Publish/Subscribe, where the type of interaction depends upon the HYPEG phase.

Through this process, and through the Molecular Hypothesis Model, the HRE will enable people, who are not medical experts, to participate in a shared understanding of the health research formulation process, and to contribute to convergence in defining population health hypotheses. By precompiling research population criteria into OBDD’s and matching them against HCP messages, very large scale matching of health hypotheses to potential participants can occur. This large scale matching enables researches to gain a better and earlier view of research population characteristics, and gives health consumers more opportunities to find and engage in relevant clinical research. The HRE approach to collaboration is consistent with an incremental,
and collaborative development process for clinical research hypotheses, where hypotheses are built through a series of additions and iterative refinements.
CHAPTER 7  Summary and Contributions

This dissertation introduces a new paradigm in health research – the Health Research Exchange (HRE) – an informatics platform that enables the health consumer to play a significant and essential role in contributing to clinical research ideas, populations, patient-relevant outcome measures, and enrollment success. It does so by defining a model for health research hypotheses – based upon the organizational principles of modularity, connectedness, and domain specificity. This model is innovative in that it enables reuse and recombination of existing hypothesis elements, and the assembly of new hypotheses from elemental building blocks. This model, named the Molecular Hypothesis Model (MoHM) for its organizational similarity to biomolecules, provides a model for representation and communication of health research hypotheses, which are the foundation of every health research study. By providing a representation and communication platform, new capabilities are enabled, including hypothesis management and collaboration during the hypothesis formation or design process.

The MoHM is significant in that it will enable health consumers to suggest health research ideas for consideration by health researchers, and to keep track of their own personal health hypotheses through their personal health records. It will enable researchers to stay in touch with what other researchers are planning or are engaged in, and to exchange concrete contributions of references, biological pathways, research population definitions, and hypothesis testing methods. It will enable health providers, such as physicians, to follow the state-of-the-art of ongoing research in their area of interest, and to understand underlying links to other research.
This chapter summarizes the challenges and gaps that this research addresses, and the research contributions of this work. It ends by describing possible future work that extends these ideas and capabilities.

### 7.1 Challenges

Advancement in medical therapies requires clinical research studies to test a therapy’s efficacy and safety in humans, yet this process can be slow, inefficient and error-prone. Two key problems resulting in long delays are 1) the translation of research ideas into testable hypotheses, and 2) the enrollment of subjects in clinical research that test these hypotheses. A crucial need for computational support is indicated by several challenges – the complex, multi-level nature of the medical field, the large number of possible health hypotheses, and the large size, variety, and complexity of clinical research information.

Strong social and economic forces are now changing the way healthcare works. The organization, duration, and relevance of clinical research is evolving to emphasize increased efficiencies, productivity, and inclusion of health-consumer-centered outcomes. The global trend toward Evidence-Based Medicine is one reflection of this shifting momentum. Another reflection in the United States is the new Patient Centered Outcomes Research Institute (PCORI), which this year published and initiated funding for a high-profile agenda on national priorities in Patient-Centered Outcomes Research.

In the period of change, a key technical challenge is the specification and development of an information model and architecture that will enable people to collaboratively participate in the
health research formulation process. People, who may not be medical experts, but who in their daily lives note health patterns and associations that could be valuable to health researchers, must be able to propose research ideas, share them with others, and find research ideas that are relevant to their interests. The information model must support a representation for health ideas that is useful to a health researcher, who must work in the context of stating and evaluating scientific hypotheses. The representation must accommodate the practical need and practice of incremental development, such that research ideas can be productively shared from early inception to full specification. The information model must be able to interface with heterogeneous representations and inference methods in medical domains at multiple levels of abstraction. The architecture must support a scalable process for collaboration through information sharing and engagement, and should be able to leverage and integrate Personal Health Records in the collaboration process. Because of the size and complexity of clinical research endeavors, the collaboration process must be efficient. It must be able to support matching of millions of potential research participants, to hundreds of thousands of research ideas, via combinations of hundreds of thousands of concepts.

The National Institute of Health (NIH) clinical trials database partially addresses some of these challenges, as a comprehensive repository of clinical trial projects. However, it does not adequately support the representation of health hypotheses or the need for earlier engagement of health consumers in health research conceptualization and development. This research addresses the following shortcomings of the NIH clinicaltrials.gov database: 1) the current representational framework does not meet the requirements for enabling consistent representation and ease of creation of hypotheses, and 2) the current framework misses
opportunities to engage health consumers in suggesting new ideas and important measures, and
3) the current framework does not make productive use of its diverse and comprehensive medical
knowledge. The architecture for a Health Research Exchange (HRE) was developed to address
these shortcomings. The vision is that by interacting with the HRE, both health consumers and
health researchers can propose health research ideas, critique health research ideas in formation,
and suggest health measures that are meaningful to them.

7.2 Contributions

This research defines an architecture for health research collaboration – the Health Research
Exchange (HRE) – that supports and encourages engagement of health consumers in the
definition of health research. The architecture includes an efficient, modular representation, a
health hypothesis management capability, and a conceptual approach for collaboration. The
HRE is built on the premise that a system-like, structured representation of health research
endeavors will enable more people to understand and participate in health research, and will
facilitate the development of relevant, testable health research hypotheses. The basic concept is
that research ideas, expressed as scientific hypotheses, have standard parts that act to define the
hypothesis; these parts act to connect the hypothesis to the body of available knowledge.
Template and domain knowledge resources are provided to speed the development of
hypotheses, and to enable improved communication among health researchers. Connecting the
hypotheses to other bodies of knowledge makes them more understandable, and makes it
possible to identify underlying pathways of reasoning that hypotheses may share. Managing the
hypotheses as connected, but unique, entities makes it possible to answer many new types of
queries. Defining a structured collaboration protocol enables a shared understanding of how health researchers and health consumers can productively exchange ideas and information through a neutral intermediary, the Health Research Exchange.

Central to the HRE is a novel approach for representing health hypotheses – one that extends the common perception of a hypothesis as a single statement, to a model of a hypothesis as a system. In this model, the hypothesis is a cognitive artifact with component parts, or domains, that operate together to form an operational whole. This model – the Molecular Hypothesis Model (MoHM) – is based upon the principles of connectivity, modularity, and domain specificity. The MoHM meets the requirements of a scientific hypothesis – predictive utility, plausibility, testability, falsifiability, and state – by specifying the representation of information needed to support these requirements. The MoHM, inspired by protein domain structure and function, represents and integrates a variety of information types, with major functional components organized by domain. Like proteins, its domains, when separated from the parent structure, may retain their useful structure and function, and can be reused to construct other hypotheses.

The underlying representational formalism for the MoHM is the Hybrid Hypergraph Description Logic, HDL. The purpose of HDL is to work within a efficient Description Logic framework to represent complex concepts, such as hypothesized relationships, biological pathways, and research population inclusion/exclusion criteria that are essential to a computational model for health hypotheses. This extension to the Description Logic formalism includes a BBox of Boolean formulas and a HBox of hyperedges, i.e. hyperroles that link more than two concepts. These are complementary to the ABox and TBox elements characteristic of DL’s. The
SNOMED-CT DL integrates with HDL; complex medical concepts in SNOMED-CT are represented as tree-nodes, connected by HDL hyperroles. The advantage of the HDL approach is that, while each functional domain is connected and functions as part of the hypothesis system, its representation is specialized to the function that it must perform. That specialization is intended to provide advantages in computational efficiency and representational fit. Hypergraph traversal of concepts across a single hypothesis is not computationally demanding. However answering queries, such health consumer relevance, across multiple hypotheses, or patient-eligibility, across multiple research participants, must be computationally efficient, due to the large numbers involved. The utilization of efficient DL formalisms, as well as a transformation to OBDD’s for patient-eligibility queries, addresses the practical issues of performance by building this into the HRE architecture design.

In the MoHM, there are six domains – the Project Domain, the Prediction Domain, the Population Domain, the Plausibility Domain, the Testability Domain, and the Falsifiability Domain. Each domain in the hypothesis is characterized by its function, and its representation. All medical terminology is expressed in SNOMED-CT.

- The function of the **Project Domain** is to provide a summary of the hypothesis. Its goals are similar to the clinical trials database in that it is project-focused. However, by including the property of state, it provides insight into the degree of maturity of a research hypothesis. The Project Domain includes other project-related meta-knowledge such as hypothesis type and important project timelines.
- The function of the **Prediction Domain** is to support the property of Predictive Utility,
where the concepts in the hypothesis H-Statement and Prediction are related to each other through a network of relationships. The Prediction Domain is represented in a hypergraph of relationships between domain concepts.

- The function of the **Population Domain** is to represent inclusion and exclusion criteria for the populations for which the hypothesis is applicable. The Population Domain provides a context for the hypothesis prediction. It is represented as a BBox in HDL. Its domain representation is specialized to match with Health Condition Profile (HCP) messages, which express parameter values for Health Condition Profiles. This enables efficient matching of population criteria to individual health profiles.

- The function of the **Plausibility Domain** is to ground the hypothesis, connecting it to existing facts, references, relationships, and other active hypotheses, in order to support the Plausibility Property. The Plausibility Domain is a network of DL tree-concepts connected by DL roles and HDL hyperroles.

- The function of the **Test Strategy Domain**, is to support the testability property by subdividing the population into groups and defining parameters that are required for testing the hypothesis. The Test Strategy Domain enables the operationalization of concepts in the prediction, linking abstractions to measurable observable parameters. The test strategy domain is represented via standard ABox and TBox axioms.

- The function of the **Falsifiability Domain** is to support the Falsifiability Property, by ensuring that a hypothesis has at least one legal analysis strategy that can be used to determine whether a prediction is false. Legality is determined by analysis rules that assess consistency between analysis methods and the proposed groups and analysis parameters.
The MoHM makes possible three important HRE capabilities: 1) an approach similar to case-based design for constructing and reusing hypotheses, 2) a hypothesis management infrastructure, including a conceptual plan for using health hypotheses to organize the diverse information in personal health records, and 3) a collaboration platform. This dissertation illustrates how the MoHM and its characteristic features, makes these capabilities possible.

The first capability is a methodology for development of Population Health Hypotheses that leverages the vast knowledge contained in existing hypotheses, reusing this knowledge to make hypothesis formation more efficient and accessible. The methodology, Case Network Domain Reasoning (CaNDoR), is enabled by the modularity, domain specificity, and connectedness of the MoHM. Decomposing hypotheses into modular domains enables the reuse of hypotheses in a variation on case-based design, with domain-specific building blocks that can be reused. Building blocks, design patterns, and inference methods are defined, so that people who may not be health researchers are provided with representational and computational aids for participating in the health research definition process. Connectedness enables two types of similarity matching: concept similarity and prediction-based similarity, a kind of structural similarity. Hypothesis Design Patterns provide templates, represented as patterns of connectivity, for study designs. Two design patterns are the parameterized difference hypothesis model and the interventional difference hypothesis model. An important feature of CaNDoR is the ability to incorporate realistic medical terminology and subsumption reasoning through SNOMED-CT. Through the MoHM, the average health consumer may have access to many examples of reusable health hypotheses, as well as reusable links to underlying biological pathways.
The second capability is the management of large numbers of hypotheses. HyRE, the Hypothesis Reasoning Engine, supports hypothesis management functions including construction, retrieval, re-use, translation, sharing, and querying. It supports the packaging and sharing of health hypotheses through XML tree structures. The MoHM feature of connectedness permits the answering of important questions about hypotheses such as overlap, or cover, as well as sharing of underlying plausibility chains. Domain specificity ensures that operations within a single domain, but across multiple hypotheses, are efficient and scalable.

The third capability is a collaboration methodology. The Hypothesize/Engage (HYPEG) protocol enables a multi-phase exchange of information, which becomes progressively more interactive and constructive, as the depth and volume of information exchanged increases. HYPEG leverages the MoHM capabilities of hypothesis construction and reuse (CaNDoR), management and queries (HyRE), and messaging (HCP). The three HYPEG phases – Hypothesize, Profile, and Engage – provide a process for health consumers and health researchers to interact and collaborate in health research. The HYPEG collaboration protocol is a variation on Publish/Subscribe, where the type of interaction depends upon the HYPEG phase.

### 7.3 Looking Back

This research began with two questions: 1) What is a hypothesis? and 2) How can a better understanding of hypotheses lead to improvements in health research? These questions have been addressed by proposing a model for hypotheses and then evaluating that model in the context of health research. While there are many different models for hypotheses, and it is
possible that we may never truly know what a hypothesis is, this dissertation proposes:

“The hypothesis is a system.”

In addition, this dissertation makes a testable prediction,

“IF the hypothesis is a system, THEN it can be represented as an artifact fulfilling a purpose and having the characteristics of a system including structure, interconnectivity, behavior, and decomposability.”

Referring back to the classification of hypotheses defined in this work, this hypothesis is a Membership Hypothesis, a subset of the Emergent Hypothesis class of hypotheses. This research proposes the MoHM as an exploratory test to determine if hypotheses can be constructed as interconnected, functional systems of decomposable parts. The population used for the test was health hypotheses, with the test set being a representative sub-set of health hypotheses contained in, but not explicitly represented as health hypotheses in, the clinicaltrials.gov database. Addressing the second question: How can a better understanding of hypotheses lead to improvements in health research? If the hypothesis is a system, and if hypothesis construction is a kind of system design problem, than hypothesis system design through reuse, recombination, and collaboration will likely lead to a better product. This is an idea that should be tested in the context of real health research, in an environment similar to that proposed for the Health Research Exchange.

7.4 Going Forward

As in most research, one hypothesis leads to another, and there is more work to be done. The ultimate hypothesis for this line of research is that the MoHM approach, in the context of the Health Research Exchange, can lead to a paradigm change in health research collaboration.
Referring back to Medawar’s Advice to a Young Scientist:

*A hypothesis is a sort of draft law about what the world – or some particularly interesting aspect of it – may be like; or in a wider sense it may be a mechanical invention, a solid or embodied hypothesis of which performance is the test.*

In the wider sense, the Health Research Exchange, based upon the MoHM, must become a mechanical invention, a solid or embodied hypothesis of which performance is the test. In other words it must undergo the reality of implementation, testing by real users, and incorporation of large-scale information.

In order for the HRE to become reality, several additional requirements should be met, each one through a process that hypothesizes a solution and tests it through performance.

- The most pressing need is for a web-based interface for health hypothesis development, that is flexible and easy to use, and that enables the end user to select from health strategies and reusable components. The flexibility and usability of this interface will be important to the success of the HRE.
- The HRE will require additional templates and encoding of common hypothesis patterns to open the gate to faster development of research hypotheses in a field.
- The addition of project management information, organizations and timelines, would enable the HRE to be expanded, similar to clinicaltrials.gov, to work as a health research project management system, while retaining its capability as a health hypothesis management system.
- With additional capabilities to visually navigate across health hypotheses, the HRE may form
the foundation for a health research navigator – that enables one to explore how various health hypotheses are connected through concepts, population, test strategies, and former work.

- Additional work is needed to extend the definition of hypothesis similarity measures, beyond the structured methods in the HRE, based on MoHM relationships. Unstructured methods, based upon text and image matching algorithms, such as those leveraged by internet search engines, may provide additional capabilities for detecting similar health hypotheses, both inside and outside the HRE.

- By defining the path from H-statement to prediction to measureable parameter, the MoHM may enable a way forward for aiding privacy maintenance, by being able to identify exactly the data fields and records that are needed to support a specific scientific hypothesis or health research project, thus potentially providing some controls and traceability on the dispersal of private health information.

- Other future work might include incorporation of other medical terminologies, in addition to SNOMED-CT.

### 7.5 Wrap-up

The MoHM is the central information model that provides a foundation for the transformation of health hypothesis development to a form more accessible to the average health care consumer – with the proposed mechanism for accessibility being templates, patterns, reuse, and structured communication – informatics approaches often used to make complex tasks simpler. The expected outcome is that accessibility will enable health consumers to share their observations about the world with health researchers, thus improving the quality and relevance of health
research. This is consistent with what we are seeing in other fields, where social media and the internet have made widespread participation in a variety of tasks newly possible.

The MoHM is designed to enable health consumers to state health hypothesis ideas in an incremental, modular approach that will interest, engage, and even educate, clinical research professionals. Leveraging the MoHM, the amorphous process of hypothesis formation – which currently relies heavily on “inspiration” or a “blaze of insight”, as well as years of clinical research training – becomes more defined. Management of large bodies of hypotheses becomes, through modularity, domain specificity, and connectivity, more feasible. Integrating health hypotheses into Personal Health Records, as an organizational principle, further reinforces and integrates the concept of health hypotheses with the health consumer. Through the MoHM, the health consumer will have the opportunity to participate in a shared understanding of health research ideas and projects. This shared understanding is essential to fulfilling an important requirement for the Health Research Exchange – collaborative development of health hypotheses. The HRE supports and extends Patient-Centered Outcomes Research, such as that defined by PCORI, in that it provides an informatics platform for patients and healthcare providers to suggest health research and health research outcomes important to them.
CHAPTER 8 References


<http://googleblog.blogspot.com/2008/02/google-health-first-look.html>


196


[70] Overfelt, M. “A mother’s race for a cure, One entrepreneur plunged into biotech to save her daughter’s life.” *Fortune Small Business* 100 2007, CNNMoney.


[72] Elashoff, M., AV Matveyenko, B Gier, R. Elashoff, PC Butler. “Pancreatitis, Pancreatic,
and Thyroid Cancer With Glucagon-Like Peptide-1-Based Therapies.” *Gastroenterology* 141, 150-156, 2011.


[139] Karlamangla, A. Data Analysis Strategies course notes. UCLA School of Medicine, 2006.


[149] “NCT01133418, Efficacy of a Cognitive Training Intervention for Children With


[190] Jennings, G., Symbolic incompletely specified functions for correct evaluation in the presence of indeterminate input values, 28th Hawaii Int’l Conf. on System Science (Vol.