UNIVERSITY OF CALIFORNIA

Los Angeles

A Methodology to Apply Evidence from Scientific Literature to Guide Individually-tailored Evidence-based Medicine

A dissertation submitted in partial satisfaction of
the requirements for the degree Doctor of Philosophy
in Biomedical Engineering

by

Juan Wu

2016

ABSTRACT OF THE DISSERTATION

A Methodology to Apply Evidence from Scientific Literature to Guide Individually-tailored Evidence-based Medicine

by

Juan Wu

Doctor of Philosophy in Biomedical Engineering
University of California, Los Angeles, 2016
Professor Alex Anh-Tuan Bui, Chair

Knowledge about the biology, etiology, staging, and treatment of a disease can be found in a growing and disparate set of sources, including observational clinical data, scientific literature, and clinical guidelines. However, effectively utilizing these sources to support clinical decision making remains a challenge. One of the challenges stems from the need to integrate knowledge from multiple sources that have heterogeneous representation, while locating and appraising evidence relevant to an individual patient can also be an issue. The objective of this dissertation is the formulation of an intermediate representation that logically consolidates and standardizes knowledge fragments across these sources, along with the definition of operators on this representation that generate, in a principled manner, the needed elements to facilitate answering clinical queries to support evidence-based medicine (EBM). The contributions of this work are: (1) a standardized representation, called Phenomenon-Centric Data Model Plus (PCDM+), which adopts the probabilistic entity-relationship model and captures and structures information about a disease drawn from scientific literature and patient records, emphasizing population-level observations and evidence;

and (2) a set of operators that retrieve and infer information about individual patients from the PCDM+ to inform clinical queries. This work is demonstrated and evaluated in the domain of intracranial aneurysm.

The dissertation of Juan Wu is approved.

Denise R. Aberle

Ricky Kiyotaka Taira

Gary R. Duckwiler

Alex Anh-Tuan Bui, Committee Chair

University of California, Los Angeles

2016

This is dedicated to my savior, Jesus Christ.

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

First, I would like to thank my Lord Jesus Christ, who died for me, cleansed my sins, and brought me back to God’s family. Becoming God’s daughter was the most significant source of support during my Ph.D. training. I thank God for his everlasting love and mercy. For the past four years, every morning, I would start my work by praying in my office, so that my mind was renewed, then I could transform any negative emotion to a positive attitude and start the day with a grateful heart. Without God’s love and grace, I would have not been able to finish this work.

I would like to thank my committee Drs. Alex Bui, William Hsu, Ricky Taira, Gary Duckwiler, and Denise Aberle for their support and guidance. Each of them inspired me by their dedication to research and teaching. I would also like to acknowledge Drs. Fernando Vinuela, David Geffner, Lisa Bolden, Jennifer Long, Dizon Manuel, and Ji Zhang for insightful discussions and domain expertise. The individuals who participated in the evaluation tasks, and who reviewed and provided feedback on this manuscript, also deserve my gratitude.

I would like to thank the Medical Imaging Informatics (MII) group, including the previous and current faculty, staff, and students.

I wish to thank my advisor, Dr. Alex Bui, for his patience, generosity, and financial support during my graduate studies. He gave me constructive suggestions on my research direction, presentations, and writing. I appreciate his talents in teaching, scientific writing, creating polished slides, and baking. I am grateful to him for paying the bills for the nice lunches and dinners during the MII mingle time. He made my experience at UCLA easier than it would have been without his support and guidance.

I would also like to take this opportunity to thank my mentor Dr. William Hsu, who walked along with me for the past seven years. He taught me how to tell an elevator story and how to present a journal article. He accompanied me to the conference where I gave my scientific presentation to a large audience for the first time in my life. He helped me to revise my first conference paper eleven times. He gave up the "bruin 5k run/walk" in order to attend my baptism. He wrote me long emails that inspired me to think deeper about my research. He also actively organized student activities such as Friday Breakfast. Just like my advisor, he is the person who reminded me of God's grace.

I am also indebted to my "professor dad" Dr. Ricky Taira, who instructed me to have a scientific attitude towards my research. His question "what is the nugget" always drove me crazy, but made me stay focused on the innovations I would add to the field. He is a great research idea generator and he came to my office at random times to share brilliant ideas with me. He also encouraged me to do physical exercises, which helped my brain to stay active.

I would also like to thank Dr. William Speier, who was my peer and taught me Java programming and helped me debug my java codes many times during my first year. He set an example for me as a good student researcher. I still remember the time when, realizing that I faced many frustrations from research, he took me for a walk near the office and told me that my happiness and health were more important than the research results.

I am also highly appreciative of Dr. Maurine Tong, especially her sisterhood friendship and her willingness to walk with me on the path to where we are today. We joined the program at the same time, and graduate at the same time. She showed me what resilience is. Facing as many challenges as I did, she never gave up, and continued writing good papers. Her prayers for me were so earnest and I will never forget the time she prayed for me before my oral defense.

I would also like to acknowledge Drs. Jean Garcia-Gathright, Kyle Singleton, and Mary McNamara for their help and company as graduate students. I am very glad to have shared this journey with them.

I would also like to thank Lew Andrada for the time he listened to me nag about the big and small things and provided me with free counseling. His friendship means a lot to me.

I wish to say “thank you” to Isabel, Larry, Shahin, and Anne-Marie for the administrative support during my graduate study.

I would like to express my sincere gratitude to my American family: Mana, Judy, Greg, Ling, Monica, Raphael, Scot, and Jerry. They are not related to me by blood, but they showed me their love and support during my years at UCLA, which made me understand what grace is.

I would also like to thank my church family at CBCWLA. I am particularly indebted to my pastor for his teaching in God’s words. I am grateful to my brothers and sisters for their prayers and care, especially during the hard times.

I would like to thank my family in China, especially my parents and my younger brother, for their support and understanding. I am particularly grateful to my mom Wenlan, for she encouraged me to come abroad to study.

This work is funded in part by the National Institutes of Health/National Institute of Biomedical Imaging and Bioengineering (NIBIB) R01 EB000362.

VITA

2007	B.S., Biomedical Engineering Central South University Changsha, Hunan, China
2007-2009	Graduate course and research, Biomedical Engineering Central South University Changsha, Hunan, China
2009-2015	Graduate Student Researcher Medical Imaging Informatics Group University of California, Los Angeles
2012-Present	Ph.D. Candidate, Biomedical Engineering University of California, Los Angeles

PUBLICATIONS AND PRESENTATIONS

***Wu JA**, Hsu W, Taira RK, Bui AAT. PCDM+: integrating evidence from literature for clinical decision making. NSF I/UCRC for semantic computing Annual Planning Workshop, Irvine, CA; June 2014. Poster.

Wu JA, Hsu W, Bui AAT. An Approach for Incorporating Context in Building Probabilistic Predictive Models. 2nd IEEE Int Conf on Healthcare Informatics, Imaging and Systems Biology (HISB), 2012.

Wu JA, *Hsu W, Bui AAT. Extracting relevant information from clinical records: Towards modeling the evolution of intracranial aneurysms. Proc AMIA Fall Symp, 2012. p.2005. Poster.

*Loo R, **Wu JA**, Hsu W. Harnessing observational clinical data to improve patient care: Studying intracranial aneurysms. UCLA Engineering Tech Forum, 2012. Poster.

*Hsu W, Tong M, **Wu J**, Lin M, Bui AAT, Taira RK. Tools for modeling medical imaging and molecular biology correlates using published literature. RSNA Annual Meeting, Chicago, IL; Nov 2011. p.314. Education exhibit.

*Tong M, **Wu J**, Chae S, Chern A, Speier W, Hsu W, Taira RK. A tool to formalize information from clinical trials for disease modeling. RSNA Annual Meeting, Chicago, IL; Nov 2011. p.316. Education exhibit.

*Tong M, ***Wu J**, Chae S, Chern A, Speier W, Hsu W, Bui AAT, Taira RK. Computer-assisted systematic review and interactive visualization of published literature. RSNA Annual Meeting, Chicago, IL; Nov 2010. p.556. Education exhibit.

*Arnold C, Speier W, Hsu W, Garcia-Gathright J, **Wu J**, Tong M, Taira R. UCLA summary for i2b2/VA 2010 NLP shared-task challenge. Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA; Nov 2010. Poster.

CHAPTER 1. Introduction

Physicians need to answer certain questions about a disease for clinical tasks. What are the causes of the disease? How can it be prevented? How does it progress? Which treatments are effective in controlling such progression? To answer these questions, knowledge about a disease may be obtained from multiple sources, including observational data from hospitals, which record the treatment courses of patients; scientific literature, which reports findings from clinical trials or other studies; and clinical guidelines, which summarize available evidence to guide clinical decision making in specific conditions. However, physicians need a significant amount of expertise and time to integrate these pieces of knowledge in order to answer clinical questions, especially when attempting to individualize the information to a specific patient's situation.

Physicians can be assisted in answering these questions by constructing comprehensive disease models, such as a Bayesian belief network (BBN), to encode what is known about the etiology and progression of the disease and to generate predictions about diagnosis and outcome. In spite of current work in areas such as meta-analysis [Chen et al., 2013; Zhong et al., 2013; Buffart et al., 2013] and disease modeling [Mant et al., 2005; Harrison and Kennedy, 2005; D'Amelio et al., 2010; Newton et al., 2012], no systematic way exists to effectively organize available information across multiple sources and translate pertinent data into a comprehensive disease model that is consistent with the knowledge provided.

Currently, the process of constructing comprehensive disease models is hindered by two issues. First, existing data models focus on capturing facts such as findings, diagnosis, and treatment but fail to maintain the context in which these facts were obtained (e.g., the purpose of data collection, the measurement of each finding, the certainty of the findings). Such context is necessary for interpretation of the facts, as loss of context leads to the inappropriate secondary use of the data (e.g., without knowing the imaging modality, the sentence, "the mass is 5.1 × 2.3 cm," provides little information about the implications of these dimensions). The heterogeneous representation of such facts and contexts across multiple sources presents further challenges. For example, knowledge yielded by clinical observational data and controlled trials differs in the purpose of data collection, data storage, targeted users, knowledge delivery, and representation. Such variation calls for a data model that standardizes and captures partial knowledge yet minimizes the bias for constructing a comprehensive disease model. Second, information collected from existing sources is not in a form that is conducive to constructing a disease model. For example, the parameters of a BBN are conditional probabilities of a child node given its parent nodes (e.g., P(C | A, B)), and the data source may only provide partial knowledge (e.g., P(C | A) or P(A, B), and qualitative semantic relationships); the model requires a more specific biomarker measurement (e.g., nicotine intake, hormone level), while the sources only contain broader patient-level information (e.g., smoking history, gender). Thus, systematic transforms are needed to integrate and translate original knowledge fragments to parameters needed for model construction.

1.1 Research and Contributions

To solve the aforementioned problems, the aim of this dissertation research was to:

[Aim 1] Create an intermediate representation that logically consolidates and standardizes knowledge fragments and the associated context across the sources; and

[Aim 2] Develop operators that translate evidence into knowledge elements to inform clinical decision making relating to a specific patient.

The evaluation was performed in the domain of intracranial aneurysm (ICAs).

1.1.1 PCDM+

To address the first problem noted above, I designed a conceptual representation, called *Phenomenon-Centric Data Model Plus* (PCDM+), to standardize and organize observations and findings pertinent to a disease to facilitate evidence-based medicine. Compared with the original PCDM [Bui and Taira, 2010], PCDM+ emphasizes evidence collected at the population-level (e.g., distribution of patient characteristics, survival curves, subgroup analyses). PCDM+ was designed using a probabilistic entity-relationship (PER) model as its graphical language, which contains five types of classes:

1. Entity. An entity holds key concepts, which are used to tell a story of a disease on a patient or a popu-

lation. The PCDM+ core entities are source, phenomenon, theory, evidence, finding, intervention, behavior, research study, population, and statistical analysis.

2. Relationship. A relationship structures the hierarchy and interactions among entities. Examples of PCDM+ core relationships include *is-a, part-of,* and *has-a* that represent inheritance, aggregation, and composition relations, respectively. Examples of relationships representing interactions are *measure, support, analyze,* and *affect*.
3. Attribute. An attribute is a property that describes entities and relationships, emphasizing the context of findings such as observation type, unit, accuracy; the certainty of measurement; and the certainty of existence. Observation is a class designated to holding such context. Every entity has one or several attributes.
4. Arc. An arc delineates relations among attributes that are reported is scientific studies. Relation type and conditions where the relation was discovered (e.g., data source, study purpose, location of the study, sample size) are stored in an arc class. Hypothesis is the class used to encode such arcs.
5. Local distribution. A local distribution represents the probability distribution of attributes (e.g., descriptive statistics of the attribute values over a population) and strength of arcs (e.g., statistical method and significance) from each data source as a means of capturing population-level evidence. Local distributions are captured in probability, distribution, and statistical analysis classes.

The current version of PCDM+ consists of three components: PCDM-Clinic, which encodes clinical observational data; PCDM-Literature, which captures the evidence from scientific paper; and an Inference-Layer, consisting of a set of operators that link the PCDM-Literature to PCDM-Clinic to facilitate evidence-based medicine.

1.1.2 Operators

To overcome the second problem noted earlier, transforming unstructured knowledge obtained from data sources into formatted knowledge that evidence-based medicine requires, I created operators to translate the knowledge encoded within the PCDM+ into a form that can be used to facilitate answering clinical queries. In this work, I designed a set of operators to help enable two applications: (1) facilitate decision making in clinical practice by retrieving and aggregating knowledge fragments from published studies for a given patient; and (2) to construct BBNs that can help in prognostic tasks. BBNs were chosen as a predictive modeling framework due to their ability to handle the uncertainty inherent in clinical evidence and their capacity to generate predictions with incomplete knowledge.

Motivated by the first application, I designed three operators to retrieve and synthesize evidence from PCDM+ and instantiate:

1. Patient-population matching. When clinicians appraise evidence from published clinical trial reports or observational studies to a specific patient, they first assess if this patient meets the eligibility criteria of this study. This operator links a specific patient case to the populations that this patient is eligible for, allowing all findings pertinent to these populations to be applied to the individual. Based on the standard classes shared by PCDM-Clinic and PCDM-Literature, my initial implementation of the operator employs rule-based algorithms and Boolean logic to assess a patient's eligibility by defining a weighted scoring function.
2. Relation extraction. To facilitate the examination of variables for an outcome of interest (e.g., risk factors for rupture, growth, survival) and to elucidate the relations among the variables, this operator is designed to retrieve hypotheses stored in PCDM+ that have been tested in published studies and are supported by results yielded. Each hypothesis has one or more statistical analyses associated with it. Details about each statistical analysis including the pertinent observations, statistical methods, statistical results, significance level, analyzed time, population, and interpretation are also linked and retrieved to help physicians assess the evidence strength. This operator is used to assist in topology specification in BBN construction.
3. Probability retrieval. This operator helps the physician to estimate the associated (conditional) probabilities and statistics for a particular patient (e.g., rupture risk and treatment risk) by indexing all of the probabilities in PCDM+ that are related and inform the patient's case. This operator creates a probability pool to facilitate parameter estimation for building BBNs.

Construction of BBNs requires three elements: variables, topology, and parameters. Building upon the operators defined in the first application, I designed operators to transform PCDM+ into BBNs:

1. Variable selection and discretization. This operator translates PCDM+ entities and attributes into random variables in the BBN by following designed inclusion rules. A discretization strategy for a variable is chosen among the encoded discretization methods enumerated within PCDM+ based on a minimized entropy method.
2. Topology specification. In specifying the model's structure, an operator is provided to utilize available evidence from PCDM+ relationships and arcs to draw the most consistent and efficient structure of the BBN. Dependency among variables is drawn and examined within PCDM+ before mapping to the BBN. This operator is related to the relation extraction operator described earlier.
3. Parameter estimation. To estimate the conditional probability tables, partial statistics captured as local distribution classes in PCDM+ are used. The partial evidence provides constraints for generating an estimated distribution for each variable in the BBN. A Bayesian approach is used to update parameters when new evidence is input into PCDM+. This operator parallels the operator for retrieving related probabilities given above.

1.2 Organization of the Dissertation

The remainder of this thesis is organized as follows. Chapter 2 provides a literature review on topics related to this work, while the PCDM+ design, instantiation, and its implementation are described in Chapter 3, emphasizing PCDM-Literature and PCDM-Clinic. The developed operators are described in Chapter 4 and their use demonstrated through a patient case to answer clinical queries. Finally, Chapter 5 concludes this dissertation and provides recommendations for future work.

CHAPTER 2. Background

A literature review was conducted, focusing on works related to the aims of this dissertation, including data sources, evidence-based medicine (EBM), case-based retrieval systems, relational data modeling, and Bayesian belief network (BBN). The aims of the review of pertinent literature presented the subsequent sections are noted below:

- To explore the challenges in integrating multiple sources to facilitate evidence-based medicine, I firstly examined several data sources that are available for clinical research and clinical decision making, including observational data, scientific literature, curated database, existing models, and clinical guidelines. The results of this part of the review are given in Section 2.1.
- A literature review on EBM was conducted to specify the steps required to achieve EBM. In addition, because my work focused on two data sources, observational data from medical records and research findings from the scientific literature, I reviewed the existing work on utilizing published literature to aid decision making in clinical practice. This is described in Section 2.2.
- Case-based retrieval (CBR) systems in the medical domain are described in Section 2.3. CBR has been widely used in clinical settings to retrieve similar patient cases to facilitate decision making on new patients. The sources reviewed in this section elucidate the potential use of an underlying representation with sufficient context (i.e., PCDM+) to improve similar patient matching.
- Evidence from published literature contains many relations and probabilities. In order to choose a language to implement PCDM+, I explored the existing languages for implementing relational and probabilistic models. This analysis is described in Section 2.4.
- One aim of PCDM+ is to facilitate BBN building. Therefore, current methods of variable selection, structure learning, and parameter estimation for constructing BBNs are reviewed in Section 2.5.

2.1 Data Sources

Multiple sources are available to construct a disease model, including observational data from daily practice, published literature, existing structured datasets, other disease models, and clinical guidelines. The types of information provided by each data source and the challenges of utilizing them to construct a disease model are briefly described below.

2.1.1 Observational Data

Observational clinical data are generated as a result of clinical consultations, lab tests, image scans, biopsies, and other procedures during patient visits or hospitalizations. These data are informative as they capture the entire treatment course and provide detailed information on disease etiology, progression, and treatment for individual patients. Nevertheless, challenges exist in utilizing these data for creating predictive models as: (1) variations in demographic and clinical characteristics, procedures, treatment courses, and visit frequencies are common; (2) information is not standardized and is often reported separately in different documents without sufficient context; and (3) missing data can occur due to loss during follow-up (e.g., patient missing or cancelling an exam), limitation of the local hospital (e.g., lack of a modality, procedure or treatment), and/or the patient's medical condition (e.g., a patient suffering from memory loss may not able to provide medical and social history).

2.1.2 Scientific Literature

Published papers on randomized controlled trials provide details about the process undertaken to validate a (causal) hypothesis on a study population. Such publications provide a significant body of evidence supporting the efficacy of treatments on different study populations, identifying potential adverse events, and characterizing risk factors that may lead to poor survival outcomes. Despite the richness of information, several difficulties arise when trying to translate published findings into probabilistic models as: (1) while the study design, patient eligibility, data collection process, and statistical analysis are reported following certain criteria (e.g., CONSORT [Schulz et al., 2010]), this information remains highly unstructured and unstandardized; (2) numerical evidence is often not properly considered, and (3) conflicting information is reported by different authors.

2.1.3 Curated Databases

Researchers have created an abundance of structured datasets by curating unstructured data into research databases, collecting data as a part of randomized clinical trials and constructing large repositories of associated clinical observations. These repositories are significant in that they provide a large set of cases for a disease, allowing models to be trained on an extensive amount of data. However, depending on how the data is collected and aggregated, the consistency in the format in which information is reported varies. For example, gene expression data may not be normalized across sites and machines, resulting in values that are not comparable without adequate pre-processing [Herrero et al., 2003].

2.1.4 Existing Models

Given the increase in the volume and scope of available data, the use of machine learning techniques has expanded in order to model, relate, and classify available data (e.g., rule learning, decision trees). In addition, qualitative models such as ontologies that capture the variables, relationships, and attributes, can be a source of information in designing a model. For example, the @neurist ontology [Boeker et al., 2007], developed to accommodate data collection across multiple sites in Europe on intracranial aneurysm patients, provides a means to standardize the representation of variables and their states in a model, facilitating the integration of data from additional sites.

2.1.5 Clinical Guidelines

Clinical guidelines (CGs), which are usually produced by (inter)national medical associations or governmental bodies, summarize the highest quality of evidence on prevention, diagnosis, prognosis, therapy, risk/benefit, and cost-effectiveness. While CGs provide a high level of evidence on a disease, their impact on clinical practice is arguably limited. Possible reasons include [Broughton and Rathbone, 2001]: (1) CGs cannot assist clinicians in tailoring care to patients' individual needs, particularly in presence of comorbidities and/or conditions characterized by great variation; (2) CGs are time-consuming and hard to follow because most of them are too long to read and provide unclear recommendations; and (3) conflicting guidelines provided by different professional bodies can confuse and frustrate practitioners.

2.1.6 Challenges and Hypothesis

As the number of data sources is large, there are important considerations that need to be addressed prior to integrating the data into a model. First, existing datasets have typically been collected for a primary purpose (e.g., to test a specific hypothesis or answer clinical question), and could be biased when utilized to answer secondary questions. For instance, given a database primarily developed to determine the safety and effectiveness of endovascular coiling, the patient population will not be sufficient to answer other questions associated with aneurysms when treatment is a risk factor. Therefore, the context of original data collection (e.g., purpose of the original study, patient eligibility) is needed to identify potential sources of bias. A second consideration is the issue of missing data. Missing data remains a common problem in creating clinical datasets and may be due to a variety of causes (e.g., information intentionally unreported, measurement error, variable added after data collection started). While missing data may be addressed through imputation or sampling techniques, the appropriate technique should take into consideration whether or not the data is missing at random. A potential solution to address the issues of bias and missing data can be obtained by leveraging information that is captured across multiple sources, supplementing each other. Combination of these data sources can improve the prediction accuracy of a comprehensive disease model. The challenge in adopting this approach is that each knowledge source has bespoke characteristics and unique knowledge representation. Consider, for instance, data collected observationally versus data from controlled trials. Observational data from daily clinical practice are patient-oriented and individual level data, while scientific literature is usually study-specific and hypothesis-driven, with findings reported at population level. Clinical data are captured across disparate medical documents, images, pathologic samples, lab tests, genetic tests and other databases; while clinical trial results, together with its population characteristics, study design and statistical methods, are presented in the published literature. Clinical data are collected for patient healthcare delivery and are written in a language that physicians, nurses, and practitioners can understand, whereas clinical trials are usually conducted to test a hypothesis through statistical analyses.

2.2 Evidence-based Medicine

Evidence-based medicine (EBM) is a widely used term in medical practice. It is defined as, “a systematic approach to clinical problem solving which allows the **integration** of the **best available research evidence** with **clinical expertise** and **patient values**” [Sackett et al., 2000].

2.2.1 Steps to Practice EBM

The Five-Step Model of Evidence-Based Medicine is widely used to educate medical students [Akobeng, 2005], and is summarized below. These steps inform the design and usage of PCDM+ in subsequent chapters.

Step 1. Formulation of answerable clinical questions. In this step, physicians need to convert their information needs into answerable clinical questions. To help clinicians to formulate answerable questions, PICO (**P**atient/**P**roblem, **I**ntervention, **C**omparison, **O**utcome) framework was proposed [Huang et al., 2006], comprising of four elements that should be addressed, namely the patient or problem in question; the intervention, test, or exposure of interest; the comparison interventions; and the outcome(s) of interest. For example, when a physician encounters a new patient with an unruptured intracranial aneurysm (ICA), he/she needs to answer questions such as, “*How likely will this aneurysm grow and rupture*?” and, “*Which is a better treatment for this patient, surgical clipping or endovascular coiling?*”

Step 2. Finding the evidence. Physicians need to efficiently determine the best evidence with which to answer the clinical questions. Today, with the increasing number of research papers that are accessible through curated resources such as PubMed, physicians can obtain newly reported evidence that can be applied to patients under their care.

Step 3. Appraising the evidence. Critically appraising collected evidence for its validity and usefulness is a time-consuming and often difficult task. The process comprises of evaluating the quality of the evidence provided by each paper given its validity (i.e., are the results valid), importance (i.e., what are the results), and applicability to the patient of interest (i.e., are the results useful).

Step 4. Applying the evidence. Application of the results of this appraisal in medical practice must incorporate a patient's preferences/values and clinical circumstances (e.g., availability of the treatment, clinical expertise for certain surgery).

Step 5. Evaluating performance. Lastly, the performance of EBM in terms of measurable outcomes needs to be evaluated to determine the utility of the approach.

2.2.2 Applying EBM in the Clinical Setting

Physicians often turn to publications to gather evidence in order to make treatment suggestions. However, the time required to search, read, and summarize evidence from multiple papers can be difficult to find in a busy schedule. To address this issue, physicians often rely on resources such as clinical guidelines; expert-curated summaries like UpToDate [Jaeschke, 2000]; and meta-analyses such as Cochrane Reviews [Levin, 2001], which summarize a body of literature into brief, salient points. Nevertheless, when making a decision for a given patient case, the physician is left with the task of recalling the evidence pertinent to each recommendation.

Significant efforts are being dedicated to integrating evidence to answer clinical questions that arise in the course of care. For example, the HL7 Context-Aware Knowledge Retrieval (InfoButton) standard aids clinicians and patients in answering clinical questions by providing personalized links to online resources based on information found in the electronic health record (EHR) [Del Fiol et al., 2012]. The linkage between the EHR and external resources is managed by an InfoButton Manager, which maintains an explicit knowledge base of terms with mappings to external resources via a uniform resource locator (URL).

Furthermore, with the increasingly sophisticated natural language processing and machine learning techniques, it becomes possible to automatically structure the free-text and extract the key findings to apply to individual patients. Systems that assist clinicians with answering clinical questions based on evidence reported in literature have been widely explored. For instance, CDAPubMed aids clinicians in retrieving publications related to their patient by automatically identifying MeSH terms with the medical record and searching PubMed using these terms [Perez-Rey et al., 2012]. AskHERMES analyzes complex questions and outputs summaries from indexed resources such as MEDLINE abstracts and PubMed articles [Cao et al., 2011].

While the aforementioned systems perform information retrieval at the document or sentence level, a standardized representation will enable concept-level integration and permit users to identify papers that are related to a given concept (e.g., aneurysm rupture) and retrieve associated context across all relevant papers. The Translational Medicine Ontology (TMO) is one example of such a representation. It is a high level, patient-centric semantic representation that is used to integrate EHR data with Linked Open Drug Data (LODD) to answer questions relating to clinical trial recruitment and personalized medicine [Luciano et al., 2011]. While both approaches support expressive queries that leverage the underlying knowledge base, PCDM+ is distinguished from TMO by two important characteristics: (1) PCDM+ organizes data pertaining to a "phenomenon" (i.e., a symptom or a diagnosed medical problem) and maintains the context surrounding each medical finding; and (2) PCDM+ is more broadly conceived to answer a wide range of diagnostic and treatment selection questions, while TMO is targeted towards matching specific drug recommendations to individual patients. In addition, in this work, in contrast to the Info-Button approach, PCDM+ incorporates external knowledge as a part of its representation. In a recent paper, Garcia-Gathright et al. [2016] introduced "contextualized semantic maps," a graphical design that incorporates study population information as context for a particular publication. While this paper is a result of research effort in summarizing and linking published literature to individual patients using context information, the context the authors used is limited to patient demographics, risk factors, treatment history, and tumor features. PCDM+ provides a wider range of contextual fragments, emphasizing the probabilities and statistics.

Wider application of extracting evidence from literature and applying it to individual patients still faces several challenges, including: (1) a standard mechanism to assess which publications contain good evidence is presently lacking; (2) important contextual details about a published study are often lost during the process of summarization, which may lead to incorrect interpretation of the conclusion provided by the authors; and (3) including published literature only will introduce bias, as articles published in peer-reviewed journals tend to focus on positive results and new findings. Nevertheless, these problems may be overcome by linking high-quality evidence from systematic research to clinical practice, thereby aiding in the effort to make informed medical decisions.

2.3 Case-based Retrieval in Medical Decision Making

Medical decision making largely relies on clinical guidelines for well-studied diseases. Clinical guidelines are statements that suggest procedures for the diagnosis, management, or prevention of specific diseases or conditions, which have been approved by expert panels. However, for diseases where inter-patient variability is extremely high (e.g., non-small cell lung cancer, NSCLC), diagnostic and therapeutic decisions always need to be properly tailored to the individual patient's situation. Examples of non-compliance with guidelines are often reported, despite their proved efficacy in improving patient care. Reasons for this can be an improper or weak guideline definition, e.g., due to the presence of biases, changes in evidence or, more frequently, obsolescence of data and/or procedures [Montani, 2009]. While the guidelines are not sufficient for medical decision making, physicians need to identify and solve health problems for patients with limited observations from the patient, expertise resulting from years of training, as well as experience from previously treated patients. A case-based reasoning (CBR) system can identify similar cases among a large number of previous cases and facilitate physicians in making clinical decisions.

CBR is an approach that capitalizes on past experience to solve current problems. Human beings are robust problem-solvers despite limited and uncertain knowledge, and their performance improves with experience. These same qualities are desirable in CBR systems. CBR has proved to be especially applicable to decision support in medicine; even when guidelines or models are available for certain diseases, historical cases can provide key background and evidence for proper interpretation.

Therefore, reasoning from examples is natural for clinicians and case histories have long been essential in the training of healthcare professionals. Owing to the increasing amount of medical data yielded by clinical practice and scientific experiments, use CBR systems to enable automatic learning from previous cases can facilitate physicians in making medical decisions.

2.3.1 Case-based Retrieval Systems in Medicine

For decades, researchers have been incorporating case experience in clinical reasoning models to facilitate diagnosis and treatment planning [Kolodner and Kolodner, 1987]. A number of sophisticated medical CBR systems are presently in use, e.g., SHRINK [Kolodner, 1983], CASEY [Koton, 1988], FLORENCE [Bradburn, 1994], CASE-PARTNER [Bichindaritz et al., 1998], CASEREC [Balaa et al., 2003], and geneCBR [Jaulent et al., 1997]. The purposes of these systems vary, and include diagnosis, classification, tutoring, planning, and knowledge acquisition/management. SHRINK was one of the first CBR systems applied in health sciences in the 1980s, where it was used for psychiatric diagnosis and treatment [Kolodner, 1983]. This system uses a structure called DIAGNOSTIC MOPs to represent knowledge pertaining to a particular disorder, including signs and symptoms, treatments, and relations to other disorders in the same category to aid differential diagnosis. This enables SHRINK to learn from success and failures to update its case memory and improve performance. CASEY, another early medical application, integrates CBR techniques with an expert system, called the Heart Failure model, to manage patients with cardiac disease [Koton, 1988]. The Heart Failure model provides causal explanations for the findings obtained (i.e., observable features and their values) and identified states (e.g., presence of a disease, therapy, or qualitative assessments of physiological parameters). CASEY uses the Heart Failure model to evaluate the significance of the difference between the new case and a retrieved case. If differences are insignificant, the solution of the retrieved case is adapted. Otherwise, CASEY uses the Heart Failure model to generate a new solution.

In the 1990s, more advanced expert systems, such as FLORENCE [Bradburn, 1994], ALEXIA [Bergmann et al., 2005], and ROENTGEN [Nilsson et al., 2004] were developed. FLORENCE, for example, models the reasoning of an expert in advising on diagnosis, prognosis, and prescription within a nursing domain, using both rule-based and case-based reasoning.

Most recently, Montani et al. presented a CBR system capable of retrieving similar patients from a database in order to provide a suggestion on the revision of diabetic patient's therapy scheme. In their system, a case is defined as a set of features observed during a visit, with an associated prototypical class, which is the situation that occurs during diabetic patient monitoring. First, an input case is classified to a predefined prototypical class, allowing similar cases in that class to be retrieved using nearest neighbor techniques. This work combines situation assessment with CBR, and incorporates prior knowledge with naïve Bayes classification. However, the taxonomic knowledge of the classes is still provided by an expert, and is hence not part of the system logic. Schmidt and Gierl [2003] combined temporal abstraction with CBR and applied it on the prognosis of kidney function, as well as the temporal spread of infectious disease such as influenza or bronchitis. However, as their work is based on complete and well-structured information, obtaining the trend descriptions of kidney function, for example, is relatively straightforward. Context and situation awareness are two concepts commonly employed in user modeling [Zimmermann, 2003], and have recently been adapted to CBR within ambient intelligent systems in tourist domain [Kofod-petersen, 2006] and flow assurance control domain [Nwiabu, 2011].

Historically, early medical CBR systems were typically used in pilot testing or clinical trials. However, most CBR systems that have been developed in the past decade are aimed at clinical evaluation and daily clinical use. A survey conducted by Begum et al. [2011] reveals a clear trend of multipurpose and multimodal CBR systems in healthcare in recent years. Additionally, the data types used to build the case base are varied, ranging from text, image and signal, to microarray data. Owing to these advances, CBR applications in bioinformatics have become promising. Consequently, a need to incorporate other data science techniques (e.g., data mining, information extraction) in CBR systems has emerged to structure information and to deal with the high dimensionality.

2.3.2 Case-based Retrieval Components

Despite differences in purpose and/or data type, all CBR systems comprise of four components: (1) a case memory with previous cases stored in suitable format; (2) a case retrieval mechanism to match similar cases; (3) a case adaptation process to adjust the solutions of similar cases to that of the new case; and (4) a functionality to add the new case to the case base to enable dynamic learning. In this literature review, the focus in on extant studies addressing the first two components of current CBR systems, as these are related to the PCDM+ and operator design developed in the present study.

In the medical domain, a case usually refers to a given medical procedure applied to a patient, such as a visit, a treatment, or the execution of a full set of clinical guidelines [Montani, 2011]. Current medical CBR systems represent a case in different ways, which can be classified into four main categories: feature vector, frame-based, object-oriented, and textual representations [Bergmann et al., 2005]. The feature vector representation is the most popular, as any observation collected for the diagnosis or treatment planning can be represented as a feature, e.g., age, blood pressure, tumor grade, etc. Some features are aggregated from several signs. For example, the Glasgow Coma Scale (GCS) comprises eye, verbal, and motor responses. Hence, a case C with a collection of features (f1, f2, . . . fn) and their values (v1, v2, . . . vn) can be represented as a feature vector C = <f1 = v1, f2 = v2, . . . fn = vn>. More sophisticated systems use hierarchical representations or generalized cases, where cases are clustered into various prototype classes at an abstract level in the case base. Thus, a new case will be first classified into a class, from which similarity cases will be selected. However, all these representations merely focus on the objective measurements, and their contexts are not stored in those representations, e.g., when the measurement was collected, what was the patient situation by then, etc.

The second component of a CBR system, the retrieval mechanism, is highly related to its case representation. The goal of similarity judgment is to determine which cases are most usefully similar, and with high adaptability, given the desired results of the CBR process. Even though there are different approaches to measuring similarity, such as fuzzy logic and distance metrics, the objective is to classify cases according to some features that allow the use of these cases in similar situations [Jurisica, 1993].

2.3.3 Adding Context to Case-based Retrieval Systems

The context surrounding medical findings is very important in the medical decision-making process. Ignoring context may lead to undesirable events. For example, if it is known that *Drug A* is effective for NSCLC patients, but harmful to those who have liver cancer, then if “having liver cancer” is not included in the case representation, for a new patient who has both NSCLC and liver cancer, potentially detrimental decision may be made (i.e., administering *Drug A* to the patient). This example also highlights the importance of representing another type of context in a CBR system: new evidence from basic science or clinical trials (e.g., the toxicity of *Drug A* to liver cancer patients may be discovered later). Integrating the breadth of a patient's overall health situation, as well as entering pieces of evidence from most recent works for the disease into the CBR systems, can improve the treatment planning. Additionally, context plays an important role in the adaptation process in CBR systems. Once similar cases are selected, the next step is to adapt them to a new case. If the context is not stored in the case base, it is difficult to accurately adapt solutions to new cases. One obvious example is treatment availability. If all the previous cases were collected when *Treatment A* was not available in a particular hospital, the treatment selection will be biased, as the system would not suggest the use of *Treatment A*. Context thus provides important knowledge to a CBR system, particularly when considering EBM, by incorporating the patient's situation, the full range of observations, and the evidence yielded by available literature (e.g., clinical trial publications, new knowledge about the etiology or progression of the disease from basic science experiments) that properly inform similarity metrics and ultimately decision making.

2.4 Relational Data Modeling

To choose appropriate formalism for PCDM+, several existing models that can represent concepts and relations with probabilities were reviewed.

2.4.1 Entity-relationship (ER) Model

One of the longstanding representations for data modeling is the *Entity-relationship (ER) model* [Chen, 1976], which is commonly used to describe databases containing relational data. The building blocks of ER models are entities, relations, and attributes. An entity corresponds to a concept or an object; a relationship specifies an interaction among entities; and an attribute corresponds to a variable that describes the properties of an entity or a relationship. Figure 2.1 depicts an example of an ER model [Heckerman et al., 2007]. An ER model represents the structure of a database graphically by defining the entity classes (square), relationship classes (diamond), attribute classes (oval) and their interactions, providing a conceptual view of the database. A relationship has one of the following cardinality ratios: 1:1 (e.g., each lecturer has a unique office), 1:M (e.g., a lecturer may tutor many students, but each student has just one tutor); M:M (e.g., each student takes several modules, and each module is taken by several students). The cardinality of a relationship is presented at the end of the link as an arc.

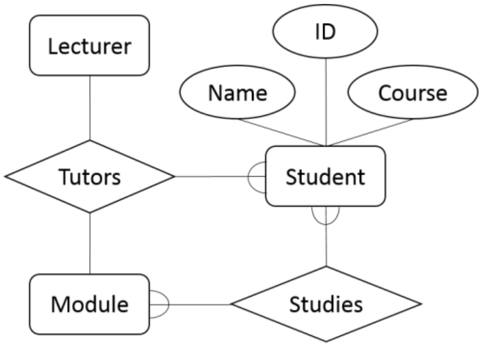


Figure 2.1 An example of the ER model [Heckerman et al., 2007].

Although the ER model introduces some semantic meaning via cardinality, it is limited to representing concepts and relations with rich semantic meanings (e.g., an *is-a* relationship cannot be represented using a classic ER framework) and it also fails to capture the uncertainty of an existence of an entity, a relationship, or an attribute.

2.4.2 Probabilistic Relation Model (PRM)

Friedman et al. [1999] proposed *probabilistic relation models (PRMs)* to model the uncertainty of some attribute values, and to specify the probabilities of relations. As an example, let X, Y, and Z represent object classes (i.e., entity classes); A, B, and C represent attribute classes; and a dot represent a possession relationship between an object and its attributes. Thus, X.A refers to the attribute class A of entity class X. A PRM consists of two components: the qualitative dependency structure and the parameters associated with it. The dependency structure is defined by associating with each attribute X.A a set of parents Pa(X.A). The parents are attributes that can influence X.A directly. They will be instantiated with different values for different objects. X.A's parent can be another attribute in X (e.g., X.B), or another attribute of related objects (e.g., if X and Y are related, then, X.A and one of Y's attributes Y.C may have a dependent relationship). Given a set of parents Pa(X.A) for X.A, a local probability model for X.A can be defined. As Friedman et al. [1999] pointed out, a PRM is more expressive than standard models, such as Bayesian networks, and it also extends the algorithms for learning Bayesian network to learn the structure and parameter for PRMs.

Although the PRM model can be a good candidate for implementing PCDM+, this paradigm would complicate PCDM+ as a database schema. The conditional relationships and temporal data cannot be well represented using PRM.

2.4.3 Probabilistic Entity-relationship Model (PER)

Heckerman et al. [2007] introduced a graphical language for relational data called the probabilistic entity-relationship (PER) model. A PER model comprises of five class types: entity, relationship, attribute, arc, and local distribution. The definitions of entity, relationship and attribute classes are similar to those in ER models. Arc classes are used to represent the relationships among attributes, and local distribution classes store the canonical distributions of attributes. Compared to ER models and PRMs, PER models are more expressive in defining different representations of relationships, such as a "restricted" relationship, self-relationships, and probabilistic relationships. Constraints of a relationship and partial relationship existence can also be represented properly in PER models.

However, in the course of the literature search, very few examples of PER models were found. Nonetheless, I determined that the PER graphical language is suitable for implementing PCDM+. With arc classes, the relationships extracted from literature can be recorded; and with local distribution classes, the distribution of attributes can be stored. In this work, I augment the local distribution class to record the probabilities of arcs as well (i.e., relationships among attributes).

2.4.4 Phenomenon-centric Data Model (PCDM)

The phenomenon-centric data model (PCDM) is a representation that organizes observational data in medical records, modeled after the investigative process in clinical practice [Bui and Taira, 2010]. The motivating design principle of PCDM is to view the practice of medicine as a scientific experiment. When a patient visits the hospital with a medical problem (i.e., phenomenon), for instance, “severe headache,” the clinical process involves investigating the possible reasons (i.e., theory) for such a phenomenon, making a diagnosis, and selecting an optimal treatment. Theories are supported by evidence, which is derived from EHRs or external resources. A phenomenon is thus presented as medical findings (e.g., an aneurysm), and a finding has properties at different levels (e.g., aneurysm size, wall shear stress).

The design of PCDM is aligned with Rudolph Virchow’s (1821-1902) understanding of illness. The illness is not an entity, but a process. Feinstenian supported this view. Uffe Juul Jensen [2007], commenting on Feinstainian’s opinion, observed that, “disease should be understood as evolving entities, entities with ‘a natural history’.” PCDM models illness as an evolving phenomenon medical practitioners observe until it is diagnosed as a disease. Despite its core entities, PCDM still lacks an explicit representation to comprehensively integrate evidence from external resources. For example, PCDM defines an entity named the external source and a link between external source and evidence, yet it lacks constructs to fully represent the evidence yielded by external sources. This work extends PCDM to encode evidence sourced from the literature to facilitate clinical decision making. In this work, I also introduce probabilities to support belief propagation processes in PCDM+.

2.5 Bayesian Belief Networks (BBNs)

Bayesian belief networks (BBNs) are a type of probabilistic graphical model and have been widely used in the medical domain to facilitate prediction tasks [Hoot and Aronsky, 2005; Stojadinovic et al., 2009]. In a BBN, the variable that we make inference on is called a “target variable,” and the other variables contributing to the distribution of this target variable are called “evidence variables.” As a directed acyclic graph, a BBN comprises nodes, edges, and conditional probability tables (CPT). A node represents a random variable, an edge from node A to node B that represents the dependency between these two variables, and the conditional probabilities between A and B are stored in a CPT. A is called as a parent node of B, and B is a child node of A. BBNs hold the assumption that the configuration of a node is only dependent on its parent(s). Two important concepts, the Markov blanket and d-separation, are introduced as they are frequently used to examine the conditional dependence and independence in BBNs. In this work, I adopt these two concepts in PCDM+ and will explain their use in the method section.

1. **Markov Blanket.** A Markov blanket for a target variable T (see Figure 2.2), denoted as MB(T), is the set of nodes that includes T's parents, its children, and its children's other parent (i.e., T's spouse nodes). MB(T) contains all the variables that shield node T from the rest of the network, allowing it to maintain the only knowledge needed to predict the behavior of T [Pearl, 1988].

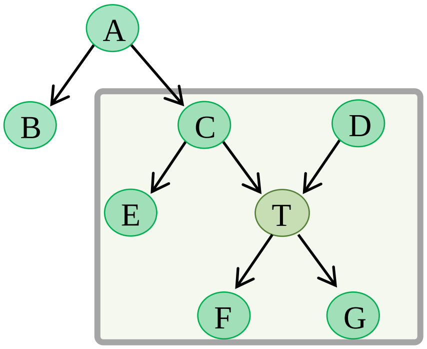


Figure 2.2 The Markov blanket of variable T (inside of the rectangle)

2. **D-separation**. D-separation is an important concept in a graphic model and examines conditional independency ("D" stands for "dependency"). When T is d-separated from A, it implies no information flows from A to T; thus, they are independent. D-separation occurs in four possible situations, as shown in Figure 2.3: (a) indirect causal effect; (b) indirect evidential effect; (c) common cause; and (d) common effect [Koller and Friedman, 2009]. In the first three cases, T is said to be d-separated from A if B is observed; in the last case, T is said to be d-separated from A if B is not observed and

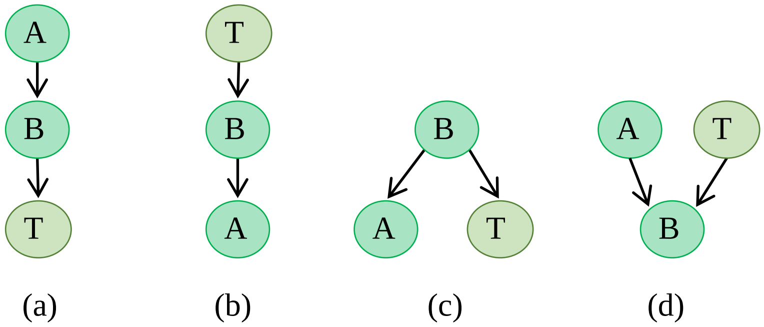


Figure 2.3 (a-c) T is d-separated from A if B is observed; (d) T is d-separated from A if B is not observed and none of B's child nodes are observed.

none of B's child nodes are observed.

My work focuses on using BBNs to model the knowledge of a disease for three reasons. First, a belief network's ability to encompass the uncertainty inherent in clinical evidence makes it superior to other models in the medical domain. A BBN incorporates prior knowledge with data to efficiently propagate the evidence through the network in order to elucidate the causal relations and make inferences in future cases. Such a unique characteristic results in its wide use for prediction tasks on diagnosis and prognosis [Zhao and Weng, 2011; Hoot and Aronsky, 2005]. Second, a BBN decreases the computational complexity of encoding a joint distribution over a high-dimensional space and enables informed decision making with incomplete knowledge. The etiology and progression of a disease are highly complex phenomena and a large number of risk factors are involved. The heterogeneity among patients increases the problem complexity. As one type of a probabilistic graphical model, the BBN holds an assumption that distribution of a child node depends solely on its parent node(s). Conditional independence exists among nodes without links, and the joint distribution can be computed using a much smaller number of conditional probabilities. Third, several extended versions of the BBN are available to address issues that can arise in medical data and make the extension of current work feasible. For example, a dynamic belief network (DBN) provides a way to encode the information of time-variant variables (e.g., aneurysm dome size). The hierarchical Bayesian network, another extension of the BBN, can relax or embed reasonable assumptions of the network to model the heterogeneity of data sources or features at different levels (e.g., genetic, tissue, or system).

BBN construction involves three main steps: variable selection and discretization, structure learning, and parameter estimation. Current methods for implementing these steps are summarized in the following sections.

2.5.1 Variable Selection and Discretization

Variable selection. Variable selection, also referred as "feature selection" in the data mining field, is the first step of building a BBN. A good variable selection algorithm reduces the dimensionality of feature space, especially when dealing with a large number of features (e.g., genomic data). Adding a new variable to the network increases the cost of its construction, as the data quantity and number of parameters increase. Variable selection is the process of finding a subset of predictors with the strongest predictive power at a minimized cost. A thorough review of feature selection in bioinformatics can be found in extant literature [Saeys et al., 2007].

Stepwise regression and genetic algorithms are two examples of current techniques that have been developed to select a subset of variables automatically from data. Stepwise regression has been widely used in high-dimensional models [Wasserman and Roeder, 2009; Ing and Lai, 2011]. Regression can be forward or backward. Forward selection involves starting with no variables in the model, testing the addition of each variable using a chosen model comparison criterion, adding the variable (if any) that improves the model the most, and repeating this process until adding further variables does not improve the model performance. In contrast, backward elimination starts with a full set of variables, eliminates a variable each time that least contributes to the model, and repeats until the model can no longer be improved. Genetic algorithms, on the other hand, transform the variable selection into a natural selection process. Initiated with a random set of solutions (i.e., variables/topology), each solution may evolve with “crossover” or “mutation” operators to produce the next generation. Solutions are selected according to some fitness function. Stopping criteria are designed to terminate the process. A genetic algorithm modified for feature selection [Leardi et al., 1997] has been widely used for variable selection for a wide range of data types, such as on microarray [Xuan et al., 2011], mass spectrometry [Li, 2008], imaging [Handels, 1997] and other mixed data sources. Applications of genetic algorithms for BBN construction can be found in pertinent literature [Gevaert et al., 2007; Larranaga et al., 2007; Correa et al., 2011].

A variety of Bayesian variable selection methods based on Gibbs sampling have been proposed, including the stochastic search variable selection (SSVS) [George and McCulloch, 1993], the unconditional priors (UP) approach [Kuo and Mallick, 1998], and the Gibbs variable selection (GVS) [Dallaportas et al., 2000]. These Bayesian approaches estimate the marginal posterior probability that a variable should be included in the model. A review of Bayesian variable selection was given by O'Hara and Sillanpää

[2009]. In contrast to regression methods, Bayesian approaches can facilitate the integration of ancillary information regarding variables under study through prior probability distributions. For example, Hill et al. [2011] proposed an approach to incorporate pathway- and network-based information to aid prior elicitation for Bayesian variable selection studies in order to identify a subset of molecular features that may jointly influence cancer drug response.

Several researchers [Koller, 1996; Cooper, 1997] have suggested that the Markov blanket of a target variable is a key concept for solving the variable selection problem. Aliferis et al. [2003] introduced a Markov blanket algorithm, called HITON, for optimal variable selection. While this algorithm can find MB(T) automatically, its soundness is limited by the sample size and other strong assumptions.

In summary, current algorithms permit automatic variable selection, but the choice of the comparison or termination criteria can be subjective. In addition, they may overfit the data and fail to represent the domain in the real world. In my work, I augment domain knowledge from scientific literature with the clinical data to facilitate variable section.

Discretization. Variables in BBNs can be categorical (e.g., tumor stage, gender) or continuous (e.g., age, tumor size). Currently, two approaches can be adopted to deal with continuous variables in BBNs: assign the variables to specific families of parametric distributions (e.g., Gaussian distribution), or discretize them and learn a network structure over the discretized domain. There is a tradeoff between these two methods. In hybrid BBNs, where continuous variables and categorical variables exist simultaneously, great effort has been dedicated to estimating the distribution of continuous variables and addressing the inference within such mixed models [John and Langley, 1995; Cobb et al., 2007]. With discretization, information loss is introduced and inference is thus approximated. In this work, I choose to discretize variables to simplify the inference tasks.

A discretization algorithm repeats the process of selecting a cut point and evaluating the accuracy until the optimal cut point within the range of data value is obtained. The speed and accuracy of the method are usually used as criteria when comparing different algorithms. A method can be supervised or unsupervised depending on whether the class information is exploited when discretizing the data. For example, Fayyad and Irani [1993] proposed an entropy-based supervised method to minimize the joint entropy of the continuous variable and the classification variable. Discretization can be local or global. Local discretization methods deal with each variable independently, while global discretization methods determine the discretization of a set of variables simultaneously. Chmielewski and Grzymala-busse [1996] presented a method to transform any local discretization method (e.g., equal interval width method, equal frequency per interval method) into a global one. Chou et al. [2008] proposed a global approach based on minimized entropy in rough sets classification. A review of discretization methods was provided by Liu [2002].

Discretization in BBNs can be integrated as a part of the structure learning process. Friedman and Goldszmidt [1996] introduced a method based on the Minimal Description Length (MDL) principle for choosing a threshold for the discretization while learning the network structure. This method starts with two partitions of a continuous variable and then iterates the partitioning until there is no further improvement in the MDL score. Given a BBN structure, this method discretizes each continuous variable in the Markov blanket of the target variable. Clarke and Barton [2000] proposed an algorithm for partitioning continuous variables before and during BBN construction using Bayesian or MDL metrics.

In my work, PCDM+ maintains a collection of possible discretization strategies of variables from each data source. These discretization strategies are derived from scientific experiments that allow researchers to draw conclusions with significant relationships. Therefore, I assume that they provide a deeper clinical meaning beyond just fitting the data well. I subsequently employ some of the existing algorithms to determine the optimal partitioning way.

2.5.2 Structure Learning

Two current approaches to specifying the structure of the network are based on (1) asking experts to specify the topology, or (2) learning the structure automatically from data. Most structure learning algorithms can be classified into one of three groups: score-based, constraint-based, or hybrid structure learning. Score-based algorithms search for a structure that best matches the data by introducing a score function such as a minimum description length-based scoring function [Larn and Bacchus, 1994], and BDe Metric, which is a score equivalent to a Dirichlet posterior density [Heckerman et al., 1996]. Constraint-based algorithms use constraints (i.e., conditional independence statements) to form the network, determined by statistical tests (e.g., Pearson's chi-squared test). Ji et al. [2005] presented a hybrid algorithm that integrates an independence test with a scoring metric. Stajduhar and Dalbelo-Basić [2010] applied a search-and-score hill-climbing algorithm and a constraint-based algorithm in order to learn a BBN from censored survival data and compared the performance of different BBNs. Tang and Srihari [2012] proposed two algorithms to assess dependency between variables using the chi-squared test of independence between pairs of variables and the log-likelihood evaluation criterion for the network.

In sum, current methods suffer from two important limitations: expert-defined topology may not account for hidden variables, and constraint- or score-based structure learning algorithms that learn the topology from data directly may result in an unreasonable relationship that will lead to inaccurate inference and explanation. The literature search performed as a part of the present study revealed paucity of studies that incorporate relations extracted from literature to facilitate structure learning. In my work, by integrating evidence from clinical observational data and scientific literature, I explore how the topology provided by PCDM+ can yield more granular relationships that cannot be learned from clinical data alone. I put forth the idea that the topology learned from PCDM+ can have greater clinical significance when compared with a data-driven topology, and potentially more granularity relative to that yielded by an expert-defined topology.

2.5.3 Parameter Estimation

Some authors considered learning the structure and parameters at the same time (e.g., Bayesian model averaging [Hoeting et al., 1999]). However, I mainly focus on how to estimate parameters with a known topology. As new evidence of some nodes becomes available, CPTs are updated to reflect that the confidence of belief on the distribution of those nodes has been updated upon reviewing the evidence. Parameter estimation is an active learning process and has been discussed previously [Tong and Koller, 1997; Bauer et al., 1997].

With complete data. Different techniques have been developed to estimate parameters from data directly. Among them, Maximum Likelihood Estimation (MLE) and Bayesian estimation are two classic methods that can be adopted to learn parameters with complete data [Heckerman, 2008]. When using MLE, it is assumed that a parameter theta is unknown but fixed. An ML estimate of theta is the value that maximizes the likelihood of data (i.e., P(data | theta)). When using the Bayesian approach, we treat theta as a random variable, assume a prior probability, and use data to compute its posterior probability. Thus, the difference between these two techniques is that MLE is a point estimator while Bayesian estimation provides a posterior distribution of the parameter. A drawback of MLE is that, when available samples are small, the estimation is inaccurate; in contrast, different priors may need to be tested in Bayesian estimation. In my work, PCDM+ integrates multiple data sources, resulting in large sample size, and the probabilities from literature can be used to formalize a proper prior if Bayesian estimation is used.

With missing data. Dealing with missing data is an integral step of parameter estimation. Missing data in clinical observations or studies can occur for various reasons, leading to different "missingness" patterns [Guideline on Missing Data in Confirmatory Clinical Trials, 2009]:

- Missing Completely at Random (MCAR) occurs if the probability of an observation being absent does not depend on observed or unobserved measurements. A typical example is a patient moving to another city for non-health reasons. This patient could be considered a random and representative sample drawn from the total study population.
- Missing at Random (MAR) describes a situation when the probability of a missing observation depends on observed measurements only. For example, when patient attrition occurs due to lack of efficacy, it would be appropriate to impute poor efficacy outcomes subsequently for this patient.
- Missing Not at Random (MNAR) occurs if the probability of an observation being absent depends on unobserved measurements. For example, after a series of visits with good outcomes, a patient drops out due to lack of efficacy. In this scenario, the value of the unobserved responses depends on information not available for the analysis, and thus, future observations cannot be predicted by the model without bias.

Eekhout et al. [2012] provided a recent review on how missing data are reported and handled in extant research. Although not all missing data in longitudinal studies are missing at random, the assumption of MAR holds in most cases. Several parameter fitting algorithms, such as multiple imputation [Patrician, 2002], expectation maximization (EM) [Shiaikh et al., 2010], Gibbs sampling [Chen et al., 2012] and Markov chain Monte Carlo (MCMC) [Mao and Li, 2005], have been used to impute missing data from longitudinal studies. Most of them assume that MAR holds.

A number of techniques have been developed to address missing data and estimate parameters for BBNs directly from observational data; however, little work has been done to utilize the partial statistics reported in scientific literature to compute or update the probabilities in BBNs. Zhao and Weng [2011] developed a weighted Bayesian network by combining electronic health records and PubMed knowledge, yet their work is limited to using only the occurrence of certain concept pairs in PubMed journal abstracts as a prior conditional probability. Nikovski [2000] discussed how to combine partial statistics and domain-dependent constraints to construct a BBN for medical diagnosis. In my work, I utilize a Bayesian approach to combine the clinical observational data and partial statistics from scientific literature to impute missing data in order to facilitate parameter estimation.

CHAPTER 3. Phenomenon-Centric Data Model Plus (PCDM+)

This chapter describes how I achieved Aim 1 of this dissertation:

[Aim 1] *Create an intermediate representation that logically consolidates and standardizes knowledge fragments and the associated context across the sources.*

The Phenomenon-Centric Data Model Plus (PCDM+) is a conceptual representation aiming to logically consolidate fragmented evidence from published literature and medical records. Using PCDM+, I link the evidence to medical records to facilitate the clinical decision-making process for individual patients. PCDM+ is an extension of the original PCDM framework [Bui and Taira, 2010]. PCDM was built to organize observational data in medical records pertinent to a phenomenon of interest (e.g., symptoms) to illustrate the investigative process in clinical practice (e.g. from symptoms to diagnosis). PCDM+ builds upon this foundation to integrate multiple data sources (e.g., clinical observational data and scientific literature) to further facilitate evidence-based medicine. PCDM+ adopts the core entities, attributes, and relationships from PCDM, but emphasizes the population-level evidence reported from clinical trials with new constructs, described herein.

3.1 PCDM+ Development

PCDM+ was developed using previously reported methodologies that were employed to create other biomedical ontologies [Luciano et al., 2011; Smith et al., 2005]. A requirements analysis was performed to enumerate the types of information to be incorporated into PCDM+. Additional entities were added to PCDM based on the results of the requirements analysis and the manual annotation of the medical records of 15 patients with ICA and 15 published papers. A systematic approach was formulated to instantiate the PCDM+ with the clinical scenario outlined above.

3.1.1 Understanding User Requirements

The objective of PCDM+ is to support physicians in the following tasks [Sackett et al., 2000]: pose a clinical question; acquire relevant literature related to answering the question; appraise each study in the collected literature; and apply relevant evidence to the individual patient in relation to the original question.

To understand the user requirements, five board-certified clinicians from a range of subspecialty disciplines were informally interviewed to identify differences in information seeking behavior when reading published papers. Four of the five clinicians indicated that they read published literature regularly to keep abreast of new treatments and techniques. Additionally, they searched for evidence from the literature when having low confidence in their own decision, especially when presented with rare diseases or complex situations. These practicing physicians looked not only at study conclusions but also assessed patient characteristics and study design to ascertain whether the reported findings were applicable to their patients. While all of the clinicians had a general desire to understand associated statistical analyses in published studies, they found interpreting this information difficult due to the lack of expertise. Three requirements were identified and used to guide the design of PCDM+ entities that: (1) capture key findings from medical records and clinical literature; (2) represent the context surrounding the key findings, including statistics; and (3) semantically relate entities to support a wide range of queries.

3.1.2 PCDM+ Design

To create a standardized representation that structures and consolidates knowledge and associated context across medical records and scientific literature, the PCDM+ design aimed to achieve three sub-goals:

1. Determining the type of information needed from each source that can be used to support evidence-based medicine;
2. Establishing data modeling constructs involving entities, relationships, and their attributes to standardize this information from each source; and
3. Defining classes that link medical records to the relevant findings reported in the literature using the data modeling constructs.

As such, core entities in PCDM+ were identified based on a combination of top-down and bottom-up approaches. In the top-down approach, commonly referenced guidelines were reviewed, including: the PICO (Population, Intervention, Comparison, and Outcome) framework [Huang et al., 2006], which is widely used as an organizational strategy for posing clinical questions to improve retrieval results;

Schardt et al. [n.d.], which summarizes common types of clinical questions and identifies the best study type to answer each (Table 3.1); and Guyatt et al. [1994], which suggests a set of questions that readers should ask when assessing and applying evidence from the literature (Table 3.2). In addition, existing reporting guidelines were examined to enumerate key characteristics of randomized controlled trials (e.g., CONSORT [Schulz et al., 2010]), observational studies (e.g., STROBE [Vandenbroucke et al., 2007]), case reports (e.g., CARE [Gagnier et al., 2013]), and systematic reviews (e.g., PRISMA [Moher et al., 2009]).

Table 3.1 Categories of clinical questions and the types of research studies used to answer these questions, derived from [Schardt et al., n/a].

Common type of questions	Type of study
Diagnosis. Select and interpret diagnostic tests.	Prospective, blind comparison to a gold standard or cross-sectional
Therapy. Select treatments that minimize harm and cost, while maximizing positive changes.	Randomized controlled trial; cohort study
Prognosis. Estimate the patient's likely clinical course over time and anticipate potential complications or factors that influence response to treatment.	Cohort study; case control; case series
Etiology. Understand the origin of a disease or condition.	Cohort study; case control; case series

Table 3.2 Key questions to pose when evaluating extant literature on therapies (reproduced from Guyatt et al. [1994]).

Are the results of the study valid?
Was the assignment of patients to treatments randomized?
Were all patients who entered the trial properly accounted for and attributed at its conclusion?
Was follow-up complete?
Were patients analyzed in the groups to which they were randomized?
Were patients, health workers, and study personnel "blind" to treatment?
Were the groups similar at the start of the trial?
Aside from the experimental intervention, were the groups treated equally?
What were the results?
How large was the treatment effect?
How precise was the estimate of the treatment effect?
Will the results help me in caring for my patients?
Can the results be applied to my patient care?
Were all clinically important outcomes considered?
Are the likely treatment benefits worth the potential harms and costs?

To assist with the interpretation of reported statistics, guidelines for reporting statistics (e.g., SAMPL [Lang et al., 2013]) were reviewed and key concepts were noted for incorporation as properties in

PCDM+. For instance, PCDM+ aims to assist in identifying statistical information reported in the literature (e.g., sample sizes, p-values, confidence intervals), organizing this information in a manner that assists a “lay” audience with querying and validating statistical results (e.g., assessing whether a given statistical test is appropriate). From the existing reporting guidelines, a list of entities and attributes that are commonly required for each study type was created, and entities that are unique to certain study type were also recognized. Complementing the literature review, expert opinions were consulted and a list of variables that are considered as important clinical features was provided by an expert in the domain of intracranial aneurysm.

In the bottom-up approach, medical records of 50 aneurysm patients and 50 scientific publications on intracranial aneurysm, its natural history and treatment comparison were reviewed. The aim was to verify the list of entities and attributes obtained from the top-down approach, to understand their representation formats in each source, and to identify variables that are required, but were not identified in the top-down approach.

Figure 3.1 provides a high-level view of the base PCDM+ schema that resulted from the aforementioned process. PCDM+ comprises of three components: (1) PCDM-Clinic, (2) Inference Layer, and (3) PCDM-Literature. PCDM-Clinic and PCDM-Literature are used to encode evidence sourced from medical records and published literature, respectively. They share a set of classes (e.g., intervention) and have their own unique classes (e.g., probability). Inference Layer is designed to enable the linkage between evidence yielded by the patient’s records and the statistics from published literature to support EBM. This is achieved with a set of operators, described in Chapter 4. Table 3.3 enumerates entity names, descriptions, examples, and the source from which the entity was derived.

Table 3.3 PCDM+ core entities (listed in alphabetical order).

Entity	Description	Example	Source
Assessment	A statistical or clinical judgment with what have been observed	This patient has an enlarged aneurysm	PCDM
Behavior	The way a phenomenon evolves over time	Aneurysm growth; aneurysm rupture	PCDM
Evidence	Information used to support a theory	A brain aneurysm is observed in CTA images	PCDM
Finding	An observed manifestation of a phenomenon	Brain aneurysm	PCDM
Hypothesis	Hypothesized relations between variables that can be used to explain the disease etiology and progression	Older females have a higher risk of developing aneurysms	PCDM, STROBE, CONSORT, PRISMA
Intervention	An act having a preventive, diagnostic, therapeutic, or rehabilitative effect on a phenomenon	CTA imaging; surgical clipping	PCDM, PICO, CONSORT
Observation	Information collected by observing or measuring a property or a behavior	4.5 mm AP x 2.1 mm TR x 3.6 mm CC	PCDM
Patient	A person with a history of medical problems	Jane Smith	PCDM, PICO, CARE
Phenomenon	A medical problem of interest	Headache; subarachnoid hemorrhage	PCDM, PICO
Population	A group of patients that are selected for a specific purpose	Patients from UCLA Medical Center that satisfy criteria for ICA coiling.	PICO, STROBE, CONSORT
Probability	Reported percentage or conditional probability	p(size < 7 mm) = 30%; p(grow=yes | size < 7 mm) = 0.10	SAMPL, STROBE
Property	A feature of a medical finding	Aneurysm size	PCDM, STROBE
Research Study	The process of acquiring and analyzing data, which yields evidence that supports or refutes a theory	A study to examine the factors related to aneurysm rupture	PICO, EBM tutorial
Statistical Analysis	A component of data analytics to assess a hypothesis, including input, method, result, and assessment		SAMPL
Study Variable	A feature, either observed or theoretical, that is examined in a research study	Patient age, aneurysm size, hypertension	STROBE, CONSORT, PRISMA
Source	A resource from which evidence is derived	EHR, literature	PCDM
State	A snapshot of findings, observations, and properties for a specific time point/encounter	Aneurysm size, shape, and location as measured during a single imaging study	PCDM
Stream	Temporal ordering of entities over time	A patient's finding history	PCDM
Theory	A possible explanation of a phenomenon	A growing aneurysm that puts pressure on surrounding areas, causing a headache	PCDM

In PCDM+, a contextual fragment is defined as being any supporting information that characterizes the quality and process by which evidence for a finding is collected. Motivating examples for capturing this contextual information include: (1) deciding if the findings yielded by a study (e.g., initial aneurysm size has a significant correlation with aneurysm growth) can be applied to a specific patient (i.e., whether the patient satisfies the study's eligibility criteria); (2) appreciating the context related to a risk factor (e.g., aneurysm size), such as measurement units (e.g., millimeters), level of certainty (e.g., definite, appears to be, less likely, unlikely, does not exist), and data type (e.g., continuous, categorical); (3) clearly delineating differences in assumptions, interpretation, and measurement error related to how information is acquired (e.g., magnetic resonance imaging versus conventional angiography); and (4) assessing the strength of information for a given relation based on sample size, statistical significance level (e.g., α = 0.01 vs. 0.05) and evaluation metric (e.g., p-value). Collectively, these "fragments" are semantically related together (e.g., interprets, described by, has observation) in PCDM+, allowing users to retrieve the context around a given observation. For instance, given a behavior, PCDM+ may return all relevant properties and associated study hypotheses.

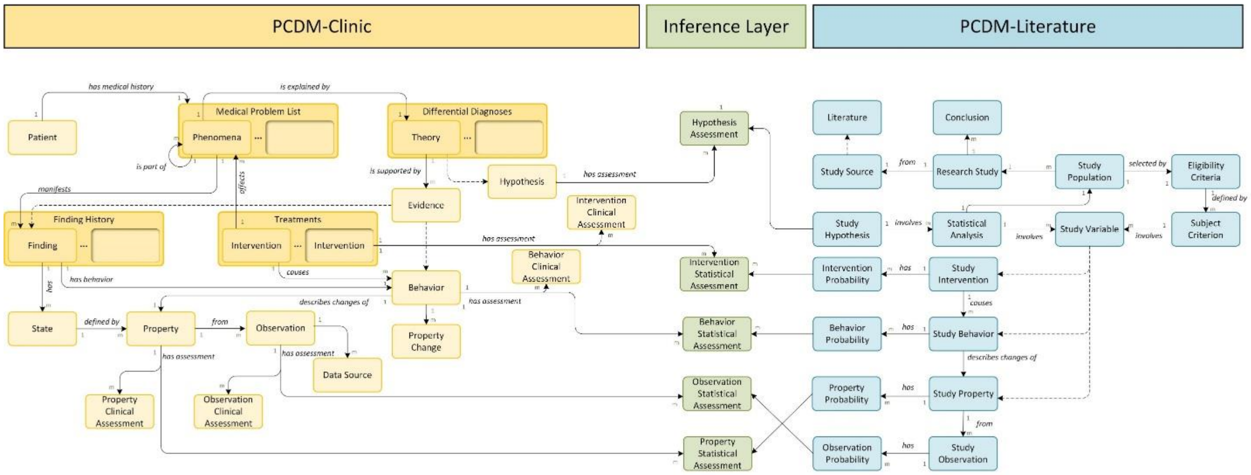


Figure 3.1 PCDM+ core entities. For clarity, only the major entity constructs, relationships, and attributes are illustrated. Dashed lines indicate is-a class inheritance.

Figure 3.2 shows a portion of the PCDM+ that has been instantiated for the ICA domain. Rectangles with thickened borders represent entities that have been instantiated with information from a patient case and are derived as subclasses of base PCDM+ entities (given in italics). In this illustration, the PCDM+ has been instantiated with information about the patient's smoking status and observed change in aneurysm size. The observation captures the individual value of a property from the medical record. For example, a property (e.g., aneurysm size) may have an observation (e.g., 6.3 mm AP × 6.6 mm TR × 5.3 mm CC) from a patient's radiology report in the EHR. The probability (e.g., *P(aneurysm size < 7 mm) = 75%*) from the literature based on a reported study of 165 patients is captured in the probability entity. Findings reported in literature (e.g., *"smoking leads to aneurysm growth"*) are evidence that can be utilized for clinical decision making, which are encoded in *study hypothesis* class. Each *study hypothesis* may have multiple *statistical analyses* reported from different studies, whereby each *statistical analysis* is subsequently linked to the contextual fragments such as *input*, *population*, *significance assessment*, and *statistical result*.

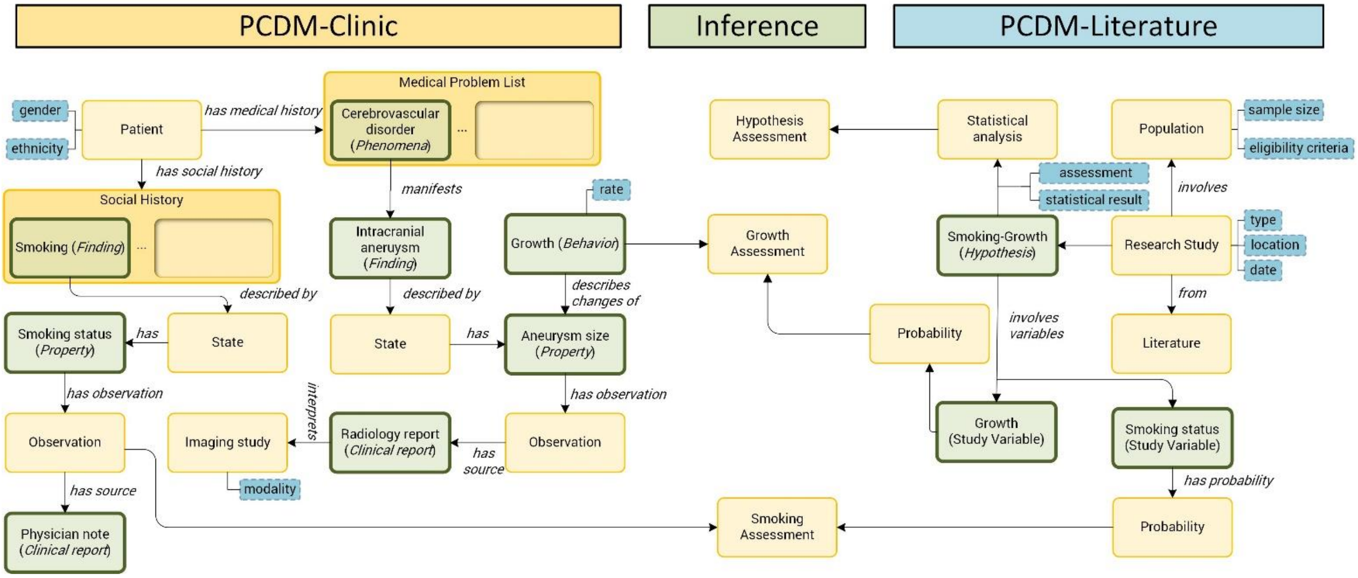


Figure 3.2 A demonstration of PCDM+ in aneurysm domain with a patient case and one paper. Some attributes and entities are omitted in order to accentuate the main aspects of the paradigm.

3.2 Probabilistic Entity-relationship (PER) Model

One way to implement PCDM+ is to design it as a probabilistic entity-relationship (PER) model [Heckman, 2004]. As mentioned Chapter 2, there are several advantages of PER models relative to entity-relationship (ER) and probabilistic relational models (PRMs). The PER model structure makes it suitable for meeting the expressiveness requirement of PCDM+. There are five class types in a PER model: entity, relationship, attribute, arc and local distribution. While the entity, relationship, and attribute classes are basic constructs and correspond to those in ER models, the arc class is used to represent the relationship between attributes, and the local distribution class serves to capture the distribution of an attribute. The details of each class in PCDM+ are described in the following sections.

3.2.1 Entity, Relationship, and Attribute

In entity-relationship (ER) models, an entity refers to a concept or an object; a relationship is defined among entities; and attributes are properties of entities or relationships. The core entities, relationships, attributes, together with the stream concept in PCDM, have been adopted and redefined in PCDM+. These classes not only help standardize the representation of key findings; they are also designed to maintain the context of the observed phenomenon.

Core entity classes

- Source: A person, publication or other record or document that provides information.
- Phenomenon: A problem of interest under investigation in clinical practice or in clinical research. *Symptom* and *Disease***,** for example, are two subclasses of *Phenomenon*. *Phenomenon* is hierarchical, (i.e., several lower-level phenomena can be grouped together to form a higher-level phenomenon and "part-of" relationship is maintained between entities at different levels). Phenomenon can evolve over time as a disease progresses or interventions take effect, e.g., symptoms can evolve into a medical problem. This change of phenomenon is captured in a temporal stream.
- Theory: An explanation of the phenomenon. *Hypothesis* is a subclass of theory. When new evidence is introduced into the model, the belief about each theory will be updated. Thus, the theory entity is

also placed into a stream to maintain the dynamic updating of its certainty.

- Evidence: Scientific experiments or observation data that support a theory. Evidence bridges a data source to the theory it supports.
- Finding: A physical manifestation of a phenomenon at a certain biological level. A *Finding* is usually revealed by exam(s) and its attributes are measured based on results of the exam(s). A *Finding* is put into a "finding history" stream to track the change of its attributes along time.
- Intervention: Any event having preventive, diagnostic, therapeutic or rehabilitative aims. *Exam* and *Treatment* are two subclasses of *Intervention*. For instance, *Exam* is the superclass for *Blood Test*, *Image Scan*, *Biopsy* and other tests. Likewise, *Treatment* is the superclass for *Medication*, *Radiotherapy*, *Surgery*, *Chemotherapy*, and other therapies. An intervention entity is input into a stream to form an "*intervention history*" that has an "*exam history*" and "*a treatment history*" as sub-streams.
- Behavior: Any significant change of a finding that results in a clinical presentation. *Aneurysm Rupture* and *Aneurysm Growth*, for example, are subclasses of *Behavior*. Note that behavior defined in PCDM+ is different from a human behavior, such as intervention, and is rather a behavior of a disease.

PCDM+ adds new entities to capture population-level evidence from scientific literature:

- Research Study: The process of acquiring and analyzing patient data. A study reveals evidence to update (either supporting or failing to support) a theory. It is linked to source information and is mapped to at least one population, from which the study was conducted. *Observational Study* and *Experiment* are two subclasses of study. Similar to a phenomenon entity, *Study* is a hierarchical entity. One study can comprise several (smaller) studies, for example, subgroup analysis in a clinical trial.
- Population: A specific collection of patients under study. Thus, *Patient* is the only element class in a population class. *Population* is also a hierarchical entity (e.g., the experimental group in a trial is a sub-population of the recruited population); a subgroup can be a sub-population of the experimental

group. A list of constraints on *Patient's* and/or *Finding's* attributes serves as inclusion criteria to a certain population.

- Statistical Analysis: A collection of methods used to process large amounts of data in a study. Instances of *Statistical Method* are used to make up a *Statistical Analysis*.

Core relationship classes

A relationship class in PCDM+ is used to describe the relationship among entities. Such a relationship can be a physical, spatial, functional, temporal, or conceptual relation. I emphasize three core relationship classes to represent relations among abstract entity classes as they play important roles in the specification of graphical models later (e.g., the topology of a belief network). These relations are not innovative but are rather adopted from existing descriptive logic:

- *is-a*: An inheritance relationship. This relationship exists among a class and its subclasses. If class X is a specialization of a generalization class Y, I define X is a subclass of Y, where Y is a super-class of X. Examples are "hypothesis is a theory" and "treatment is an intervention." Within an "X is-a Y" relationship, attributes of the superclass Y are inherited by the subclass X. For example, attributes of Intervention such as "intervener," "duration," "location," and "device" are inherited by Treatment. Similarly, Surgery, as a subclass of Treatment, inherits these attributes and other attributes Treatment has. This leads to a hierarchy of superclass/subclass relationships.
- *part-of*: An aggregation relationship. The relationship exists when multiple classes are aggregated to form a new entity class. Examples are "aneurysm location *is part of* morphology" and "aneurysm size *is part of* morphology." A special form of the "*part-of*" relation occurs when a collection of the same entity class generates a new class (e.g., "A patient is *part of* a population"). This is the case because composition is a special form of aggregation. I use the "*has-a*" relation to represent composition and use "*part-of*" to represent aggregation but not a composition in PCDM+.

- *has-a*: A composition relationship. This relationship exists when a class X comprises class Y (i.e., the class Y is the only element of class X). Examples are “population *has a* patient” and “statistical analysis *has a* statistical method.” It is neither a “*is-a*” relationship nor the inverse of “*is-a”* (e.g., “population *has a* patient” does not mean “population is a patient” or “patient is a population”). Certain attributes will be transferred from Y to X within a relation “X *has a* Y,” but the collection X has new attributes that its element Y does not have (e.g., *Patient* does not have the attribute “*sample size*” of *Population*). Broadly speaking, it can be an inverse of “part-of” relation. However, in PCDM+, while one entity class can be in multiple “*part-of*” relationships (e.g., *Morphology*), it can be in no more than one “has-a” relationship (e.g., *Population*).

The abovementioned abstract relationship classes may be placed under different superclasses to indicate different relationships (e.g., *part-of* can be *physically-part-of* or *conceptually-part-of*). Additional semantic relationships that are common in the clinical domain among entity classes are also defined. Examples are: *measure,* e.g., “exams measure findings”; *support*, e.g., “evidence support theory”; *analyze,* e.g., “statistical methods analyze findings”; *affect,* e.g., “intervention affects finding.” Some of these relations may also have subclasses. For example, *affect* has its subclasses *treat, prevent* and *interact_with;* and *analyze* has its subclass *assess*.

Core attribute classes

While the entity or relationship classes provide containers to hold key findings and their relations, such findings and relations are characterized by properties, which are stored in attribute classes. Every entity or relationship class has a unique set of attributes. PCDM+ defines attribute classes to not only represent a given feature, but also to document the context of the feature measurement. For example, in the statement, “*this aneurysm measures 3.85 mm CC × 3.53 mm TR in size*,” from a patient's medical documents, the feature is mapped to the entity instance *aneurysm dome size*; and its context includes: (1) this is a *quantitative* measurement; (2) the unit is *mm*; (3) the precision of the measurement is +/- *0.02*; and (4) the diameter in anterior-posterior (AP) dimension is *missing*, but the transverse (TR) diameter is *3.85*, and the craniocaudal (CC) diameter is *3.53*. If given more context about this statement, we may find the modality used is computed tomography (*CT*), and the exam is a *CT angiogram*. All this information is captured in PCDM+ for an accurate description of the observation. As another example, from a clinical trial paper, the statement, "*mean size of the aneurysms was 5.7 mm*," does not convey its complete significance or meaning unless we read the entire paper to find which dimension it refers to, what modality was used to measure it, and how many patients were included in the study. Thus, to capture complete knowledge of intracranial aneurysm morphology, besides the dome size, observational data of aneurysm shape, location, neck size, and neck orientation are also needed to instantiate the aneurysm morphology entity. PCDM+ defines these entities, a "part-of" relation among these entities and morphology, along with the corresponding attributes.

In summary, the core entity, relationship, and attribute classes in PCDM+ serve as standardized containers for findings and observations in order to maintain precise and comprehensive knowledge. When new knowledge becomes available (e.g., a new source/population/patient/finding), PCDM+ instantiates these classes and updates the theories that help to explain the phenomenon.

3.2.2 Arc and Local Distribution

While entity, relationship and attribute classes may be sufficient to represent individual-level observations, arc and local distribution classes in PCDM+ are used to maintain population-level evidence and the associated context that is frequently reported in scientific literature. An arc records a relation between attributes, and a local distribution quantifies an attribute's distribution. I extended the local distribution to also quantify an arc regarding its statistics.

Arc. Two subtypes of arc exist in PCDM+ to capture population-level observations and evidence, respectively:

- Observational arc: An observational arc A→B in PCDM+ indicates conditional probabilities among attributes A and B have been calculated/obtained from some observational data, but no analysis is conducted using this data. For instance, in the UCLA aneurysm research database, it can be established that the conditional probability of (rupture=yes|age<50) for a sample of 1,000 patients is 0.12.

Thus, an observational arc does not provide direct evidence of a relation, but rather captures the partial statistics that can be reused for parameter estimation in a probabilistic or statistical model.

- Experimental arc: An experimental arc A→B indicates that a relation between A and B has been discovered in one or multiple studies. For example, in concluding the paper, Juvela [2000] stated, "*active smoking status at the time of diagnosis was a significant risk factor for aneurysm rupture*." This assertion can be mapped to PCDM+ as an arc between "smoking status" and "aneurysm rupture." The relation type can be associated, dependent, or causal; and can also be negative (e.g., "the study shows that aneurysm size is not an independent indicator of rupture").

When an arc is instantiated by a study, it is also related to a "context slot" that stores the conditions from which the arc was derived (e.g., the study purpose, the study time and location, and the studied sample size). A context slot is an instantiation of a subset of entity-relation-attributes in PCDM+. For example, a context slot can be instantiated with information retrieved from statistical analysis, observation, population and source entities.

Local distribution. In PCDM+, a local distribution class is augmented to capture the partial statistics from different data sources, including attribute distributions, conditional probabilities among attributes, and statistics of a discovered relation. It quantifies an attribute or an arc with descriptive or inferential statistics. Correspondingly, two subtypes of local distribution class exist in PCDM+:

1. Local attribute distribution (LtD): A LtD class is used to capture the descriptive statistics of an attribute or an entity (e.g., age, gender, aneurysm location, aneurysm size) from observational data and available literature. The aim of capturing these partial statistics is to estimate the distribution parameters and reuse the data for integration. The most frequently reported descriptive statistics include range/minimum/maximum values and percentages of each state in the sampled population. These descriptors can be typically found in the textual body of published articles (e.g., "*Of the 6,697 aneurysms studied, 91% were discovered incidentally. Most aneurysms were in the middle cerebral arteries [36%]*"), as well as the tables of patient characteristics (Table 3.4). With a large number of patient

data (e.g., from the electronic health record), it is also possible to summarize and calculate descriptive statistics and include them into LtD classes.

Table 3.4 Example of statistics reported in clinical trial literature [Juvela et al., 2000].

Characteristic	Ruptured Aneurysm	Unruptured Aneurysm	All Patients
no. of patients	33(23%)	109(77%)	142
woman	22(29%)	54(71%)	76(54%)
age (yrs)			
median	36.8	43.6	41.9
range	22.6-57.6	14.6-60.7	14.6-60.7
aneurysmdiameter (mm)			
mean(SD)	5.6(4.9)	4.9(3.2)	5.1(3.7)
median(range)	4(2-25)	4(2-26)	4(2-26)
2-6	23(20%)	93(80%)	116(82%)
7-9	6(37%)	10(63%)	16(11%)
10-15	2(33%)	4(67%)	6(4%)
16-20	1(50%)	1(50%)	2(1%)
21-26	1(50%)	1(50%)	2(1%)

2. <u>Local arc distribution (LrD)</u>: A LrD class stores inferential statistics that quantify the certainty of a relation. For instance, Juvela et al. [2000] reported, *"RR = 1.46, 95% CI = 1.04-2.06, p = 0.033," to* quantify the relation stated in, "*active smoking status at the time of diagnosis was a significant risk factor for aneurysm rupture.*" These statistics can be stored in an LrD associated with the arc *smoking status* → *aneurysm rupture*. The context of these statistics—such as the statistical method (e.g., log-rank test), significance level (5%), and reported way for determining a p-value (e.g., two-tailed test)—is important in interpreting the statistics. This methodological information is also associated with the statistics in LrD classes. While most of the clinical trial literature reports on frequentist statistics such as a p-value, a few researchers employ Bayesian statistics (e.g., Bayesian factors) to report a relation. Therefore, LrD is designed in a way to capture probabilities from both perspectives.

In summary, arc and local distribution classes capture population-level evidence qualitatively and quantitatively and enable PCDM+ to integrate knowledge from scientific literature to update the theories and evidence.

Example 3.1 Mapping from data to PCDM+. Key findings and their context derived from clinical documents for aneurysm patients are mapped into PCDM+ to instantiate the attribute and local distribution classes, while the evidence and the associated context obtained from selected published literature on aneurysm rupture are used to instantiate the arc and local distribution classes. Below is an example of data from multiple clinical documents, including physician notes and radiology reports, which record the treatment course of a patient with an unruptured aneurysm:

> *"This patient is a 45-year-old man with an incidentally detected wide-necked 7 mm anterior communicating artery aneurysm based on the anatomy appreciated from the CTA. Social History…He has been smoking one-half packs per day since his teenage years. He does not drink alcohol. Family History: no family history of intracranial aneurysms or subarachnoid hemorrhage…We performed stent-assisted coil-embolization on November 6, 2011. The coil-embolization is successful. A small amount of residual aneurysm opacification remains at the right base of the aneurysm. The patient will return for a followup MRI and contrast enhanced MRA in 3 months, and a catheter angiogram in 6 months. "*

Through a PubMed search of "aneurysm rupture risk factors," a list of journal articles was acquired, with two supplying the following evidence:

1. "Cigarette smoking, size of the unruptured intracranial aneurysm, and age, inversely, are important factors determining risk for subsequent aneurysm rupture. Active smoking status as a time-dependent covariate was an even more significant risk factor for aneurysm rupture (adjusted RR 3.04, 95% CI 1.21–7.66, p = 0.02)" [Juvela et al., 2000].

2. “The risk of rupture increased with increasing size of the aneurysm. As compared with aneurysms in the middle cerebral arteries, those in the posterior and anterior communicating arteries were more likely to rupture” [Morita et al., 2012]

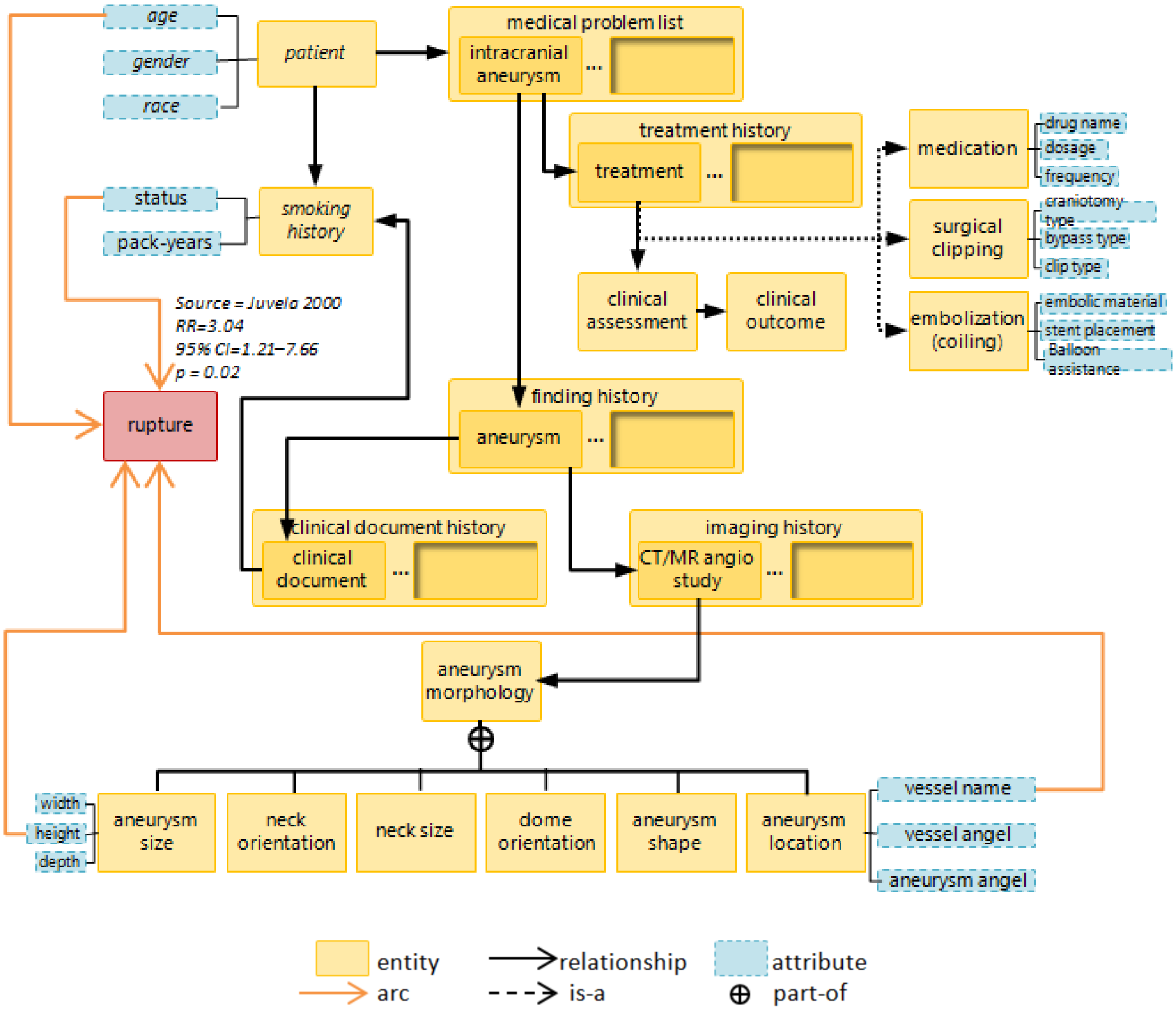


Figure 3.3 A demonstration of PCDM+ [some attributes are eliminated]

As shown in Figure 3.3, key concepts from these clinical documents can be captured in PCDM+ entities and attributes; the discovered relations among smoking, age, aneurysm size, aneurysm location and rupture from literature are recorded in PCDM+ arcs.

3.2.3 Limitations of a PER Model and an Alternative Solution

While a PER model satisfies most of the requirements of PCDM+ design, it has its limitations. First, in a PER model, an arc only defines a relation between two attributes. But in PCDM+, where the aim is to capture any relation reported in a published paper, it is also necessary to express such a relation between two entities, or between an attribute and an entity. Moreover, arcs in the PER model fail to capture multivariate analyses, involving relations among multiple variables. Second, though the local distribution class, a PER model is also able to capture the distribution of an attribute, whereas PCDM+ aims to capture the statistics associated with a reported relation as well. The standard PER model does not allow a local distribution to be associated with an arc. Third, there are no existing tools to practically implement a PER model.

To mitigate these limitations and implement PCDM+ with existing tools, the current version of PCDM+ uses an extended ER model with entities and attributes that fulfill the functions of the arc and local distribution classes in a PER model: the arc class was replaced by the hypothesis class, which is able to represent relations among entities, between an entity and attribute or among attributes; and multivariate analyses are fully captured. The local attribute distribution (LtD) class is replaced by probability and distribution classes in PCDM+; the local arc distribution (LrD) class is replaced by the attribute "*statistical result"* of the statistical analysis entity.

3.3 Clinical Scenarios in Intracranial Aneurysm (ICA)

PCDM+ provides a generalizable framework for any disease. In this dissertation, intracranial aneurysm was selected as the disease domain to demonstrate the applications of PCDM+. In the following sections, I describe the formulation of PCDM+ based on the results yielded by a requirements analysis and describe four core entities (Population, Study, Probability, and Statistical Analysis) that are introduced in PCDM+ to capture population-level evidence sourced from literature. Clinical scenarios drawn from the domain of ICAs are presented as driving examples, based on which use cases and applications are described, demonstrating different aspects of PCDM+ use in answering pertinent clinical questions.

Approximately 3.2% of the population has an unruptured brain aneurysm [Juvela, 2011]. Most individuals with a brain aneurysm may never notice its presence, but if an aneurysm ruptures, subarachnoid hemorrhaging may result with an associated mortality rate of up to 50% [Suarez et al., 2006]. Unruptured aneurysms are detected incidentally when the patient undergoes an imaging study of the brain, typically for other clinical indications. Identification of risk factors for rupture is critical in managing such patients, as it remains unclear whether intervention is required (i.e., in many cases, the aneurysm may neither grow nor rupture). Patients with unruptured aneurysms typically have three treatment options: (1) observation, which consists of routine follow-ups to assess whether the aneurysm grows; (2) surgical clipping, which is an invasive procedure that involves performing a craniotomy and affixing a clip around the aneurysm neck; or (3) endovascular coiling, which is a minimally invasive procedure in which detachable coils are inserted into the aneurysm using a micro-catheter. In deciding on the appropriate treatment, a clinician is tasked with understanding the patient's medical history, along with weighing the risks and benefits associated with each treatment option.

3.3.1 Patient Cases

I selected two patient cases to demonstrate the design and application of PCDM+ to link evidence obtained from published literature in order to facilitate answering clinical decision-making questions. These are real patient cases from UCLA Medical Center and their records shown here were documented in an interventional neuroradiology consultation notes.

Patient Case 1: "*A 70-year-old white woman with incidental finding of right posterior communicating artery aneurysm came to our institution for consultation. She has a medical history of head and neck cancer. She was a smoker from age 20 to 25. She denies any alcohol consumption or recreational drug use. Her blood pressure is 106/65. She has a family history of stroke and hypertension. The aneurysm parameters are 6.3 mm AP × 6.6 mm TR × 5.3 mm CC according to the CT angiogram.*"

Patient Case 2: "*This is a 52 y.o. male who was referred for evaluation of a cerebral aneurysm . . . . A CTA of his brain demonstrated a 3×6 mm aneurysm of the A1 segment of the right anterior communicating artery. He has a strong family history of aneurysm and SAH. His father and brother have been diagnosed with cerebral aneurysms. He is a nonsmoker and denies alcohol use.*"

Treatment suggestions by the interventional neuroradiologist for each case are also recorded: “I have discussed these findings with the patient and recommended a diagnostic cerebral angiogram and treatment with coil embolization. As there may be a wide neck, we will pretreat with ASA and Plavix.”

3.3.2 Clinical Decision Making

According to the clinical narrative, variables such as history of cancer, smoker status, family history, and aneurysm characteristics have relevance in calculating a patient’s risk of rupture. While many of these variables have been widely documented in extant research studies, clinicians lack tools to easily incorporate this information into rupture risk assessment. While risk calculators have been published [Killeen and Kockro, 2013] risk is modeled using data from specific trials and patient populations, which result in a risk calculation that may not be accurate, or even applicable to an individual. A more objective approach towards risk assessment is to tailor the retrieval of scientific evidence to answer the following questions based on the unique characteristics of an individual:

- Which studies are relevant to the patient so that evidence from those studies can be applied to the patient?
- What evidence is available to explain the etiology of the aneurysm?
- What are the reported risk factors that are prognostic for aneurysm growth and rupture?
- Which risk factors are important to make predictions on aneurysm growth and rupture for this particular patient?
- Which is a better treatment for this patient, surgical clipping or endovascular coiling, based on the supporting evidence?

Hence, the goal of PCDM+ is to encode and match findings reported in relevant literature to a patient’s medical record, thereby assisting physicians in answering these questions.

3.4 PCDM+ Instantiation

The following sections describe how the PCDM+ model is instantiated with data from the EHR and published literature. Generally, the process involves the following steps: (1) selecting papers related to brain aneurysms from PubMed; (2) adding results and associated context from each paper; and (3) incorporating details about a patient that are drawn from his/her medical record. Each step is described in detail below.

3.4.1 Paper Selection

Patient Case 1. A PubMed search was conducted in April 2014, returning over 20,000 papers on intracranial aneurysms. The initial search results were further filtered based on the availability of full-text content, whether the study was performed on humans, and the presence of the search terms “unruptured” and “growth,” resulting in 71 papers. All 71 papers were manually reviewed to obtain a broad understanding of potential properties, behaviors, relations, and observations that are relevant to aneurysm growth and rupture. This resulted in a subset of 22 papers selected as representative studies that would be used to instantiate the PCDM+. For each selected paper, an annotator worked to create instances of appropriate core entities, encoding information from the paper in the model. This information was then used to answer the clinical questions listed in Section 3.1 as a part of the evaluation.

Patient Case 2. For Patient Case 2, I used another set of 20 papers. The reason for selecting a set of papers that are different from those for Patient Case 1 is that the patient characteristics of these two patients are different. These papers were carefully pre-selected by using the following criteria: (1) seminal papers on natural history and treatment risk assessment were selected and the significance of 12 papers was approved by a physician at UCLA Interventional Radiology; and (2) PubMed search conducted in January 2015, with keywords “brain aneurysm, wide neck, coiling” with filters “free full text,” “published in recent 5 years” and “human” yielded 31 papers. To obtain evidence of strong relevance, only 8 out of these 31 papers were selected according to the sample size (>100). Finally, 20 papers were selected, 12 of which are seminal papers with approval from a domain expert, and the remaining 8 are recent papers published in PubMed.

3.4.2 Paper Mapping

The mapping of a paper’s contents aims to identify all contextual fragments in terms of PCDM+ entities (e.g., Population, Study) and relations. The process of manually annotating each paper and mapping this information to PCDM+ is summarized in the following five steps (Figure 3.4).

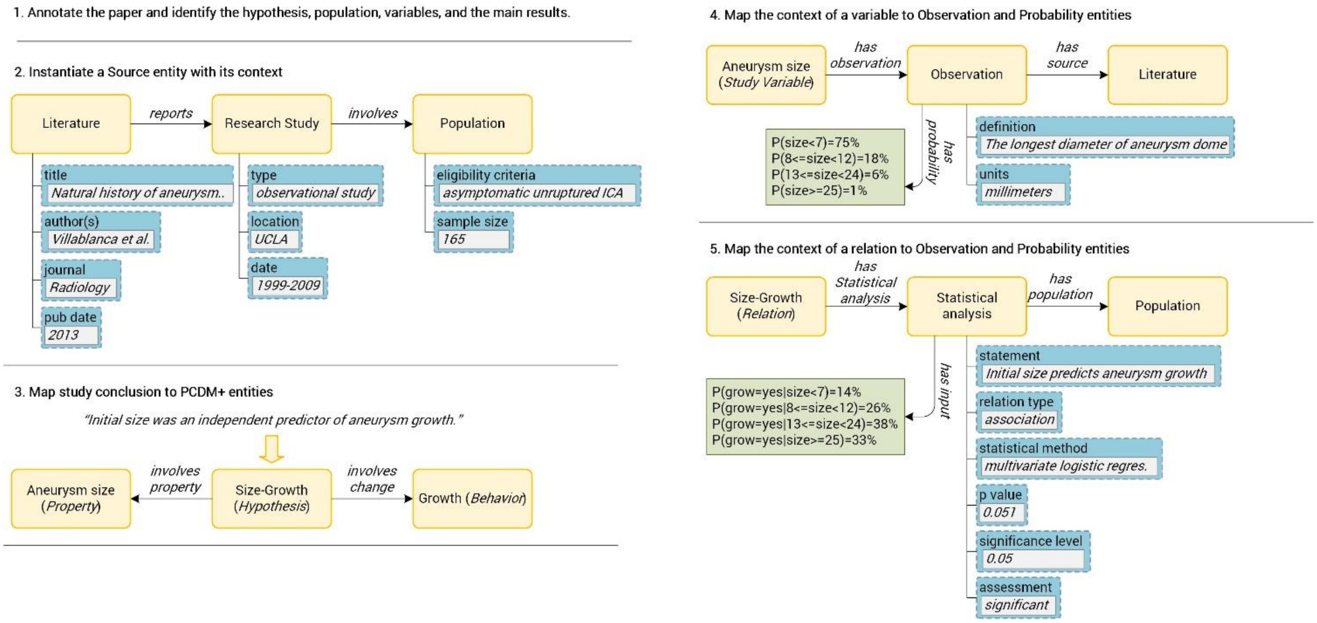


Figure 3.4 Mapping a paper into PCDM+ to identify all contextual fragments. The paper used in this example is “Natural history of asymptomatic unruptured cerebral aneurysms evaluated at CT angiography: Growth and rupture incidence and correlation with epidemiologic risk factors” [Villablanca et al., 2013].

The paper used in this example is titled “Natural history of asymptomatic unruptured cerebral aneurysms evaluated at CT angiography: Growth and rupture incidence and correlation with epidemiologic risk factors” [Villablanca et al., 2013].

- For a given study, annotate the paper and identify the hypothesis, population, variables, and the main results.
- Create a Source instance and add the context including paper metadata, study design, eligibility criteria, sample size, and other context.
- Find the conclusion of the study and map any relation statement to a PCDM+ Hypothesis instance (e.g., the hypothesized statement “smoking is associated with growth” is represented as “smoking-

growth"); and map the variables to PCDM+ entity instances (e.g., "aneurysm size" as a Property instance and "growth" as a Behavior instance).

- For each entity instance (e.g., "aneurysm size", "growth"), add an Observation instance to maintain context such as definition, unit, data type, discretized states, technique/modality, and descriptive statistics; and add a Probability instance to capture the marginal probabilities.
- For each Hypothesis instance, add a Statistical Analysis instance to capture context fragments including statistical method, p-value, assessment, and significance level; link to the corresponding Probability instance that records associated conditional probabilities or related measures as the input of statistical analysis.

3.4.3 Patient Case Mapping

Incorporating information from the patient record involves mapping patient demographics, vital signs, and encounter information to the PCDM+ and extracting structured information (e.g., location, shape, measurements) from clinical narratives. The Division of Interventional Neurosurgery within the Department of Radiological Sciences completes a case report form (CRF) to collect structured information on each individual seen at our academic medical center. This form provides a list of data elements that are considered important information in the management of aneurysm patients, including demographics, medical history, social history, clinical presentation, imagining follow-up, hemodynamics, treatment, and outcome (see Table 3.5). At institutions that do not utilize CRFs as part of their workflow, information will need to be extracted from clinical documents, a longstanding challenge [Cambria and White, 2014] that is beyond the scope of this work; results of an automated pipeline were previously reported for aneurysm-related information [Wu et al., 2012].

Table 3.5 Aneurysm Case Report Form data element (the categorical states for each data element and more details about hospital course are omitted for space limitation).

Section	Data Element
Demographics	Medical record number, date of birth, gender, ethnicity, country of birth, weight, height, blood pressure
Medical history	Hypertension, aneurysm, arteriovenous malformation, diabetes, heart disease, head injury, inherited disease, current medications
Family history	Hypertension, aneurysm, arteriovenous malformation, diabetes, heart , disease, head injury,

	inherited disease
Social history	Smoking history, alcohol use, recreational drug use
Clinical presentation	Present illnesses, Fisher CT score, Glasgow coma score, rupture status, date collected
Treatment	Date of treatment, anatomic result, immediate clinical outcome, complications, assistance, treatment type
Hospital Course	Date admitted, date discharged, days in ICU, days in hospital, vasospasm, seizure, shunt
Image analysis	Modality, date of scan, number of aneurysms, side, aneurysm shape, aneurysm location, sac_AP, sac_TR, sac_CC, neck_AP, neck_TR, neck_CC, dome/neck ratio, vessel angle, inclination angle
Clinical follow-up	Date of follow-up, Glasgow coma score, modified Rankin score, Karnofsky performance score (KPS), clinical outcome
Hemodynamics	Flow pattern, flow stability, flow division, flow impact, flow impingement, flow jet concentration, flow jet section

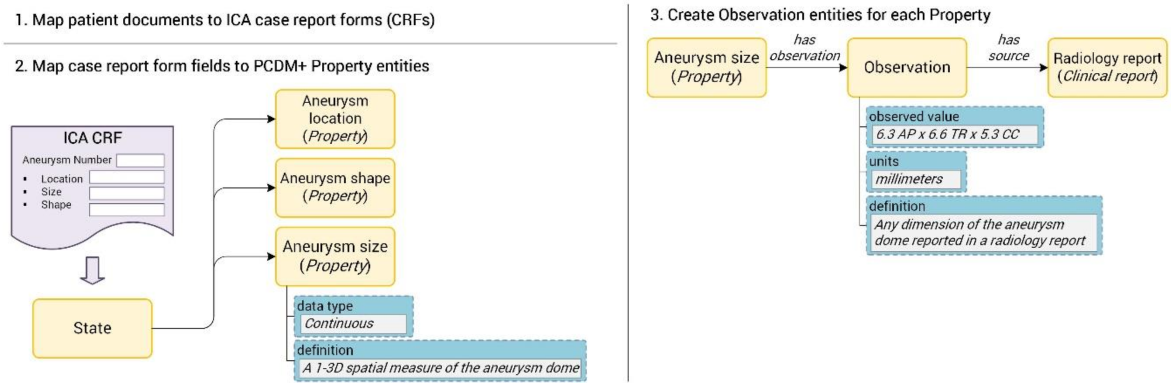

1. Map patient documents to ICA case report forms (CRFs)
2. Map case report form fields to PCDM+ Property entities
ICA CRF
Aneurysm Number
Location
Size
Shape
State
Aneurysm location (Property)
Aneurysm shape (Property)
Aneurysm size (Property)
data type
Continuous
definition
A 1-3D spatial measure of the aneurysm dome
3. Create Observation entities for each Property
Aneurysm size (Property)
has observation
Observation
has source
Radiology report (Clinical report)
observed value
6.3 AP x 6.6 TR x 5.3 CC
units
millimeters
definition
Any dimension of the aneurysm dome reported in a radiology report

Figure 3.5 Mapping a patient case into PCDM+. The Patient Case 1 described in Section 3.1 is illustrated.

The following steps were used to map the patient case into PCDM+ and are also depicted in Figure 3.5:

- Map the patient's documents to the CRF.
- Instantiate a Patient entity by adding attributes such as ID, gender, name, birth date, and ethnicity.
- Create instances in PCDM+ corresponding to the data elements in the CRF, e.g., "intracranial aneurysm" as a Finding instance and "aneurysm size" as a Property instance.
- For each entity instance (e.g., aneurysm size), create an Observation instance to capture its observed value and the context, including definition, data type, unit, and modality/technique.

3.5 PCDM+ Implementation

This section describes the tools and programming languages used to implement the PCDM+. A description of the types of clinical queries that can be answered using PCDM+ is also given.

3.5.1 PCDM+ in OWL Format

PCDM+ was initially implemented using Protégé 4.3 [Del Fiol et al., 2012]. Currently, the model contains 7,289 axioms; 5,989 logical axioms; 76 entities; 32 relationships; and 59 attributes with the 22 papers selected to perform evidence-based medicine on Patient Case 1. The experiments and evaluations conducted on Patient Case 1 were based on the Protégé version of PCDM+.

Figure 3.6 is a graphical depiction of the PCDM+ instantiated with information from Patient Case 1 and a paper on the relationship of "*initial aneurysm size is a risk factor for aneurysm growth*," as described in previous sections. According to the medical record, the patient had an aneurysm that measured 6.3 AP × 6.6 TR × 5.3 CC mm. The authors reported that 75% of unruptured aneurysms were less than 7 mm, and there is a 14% chance that the aneurysm will grow.

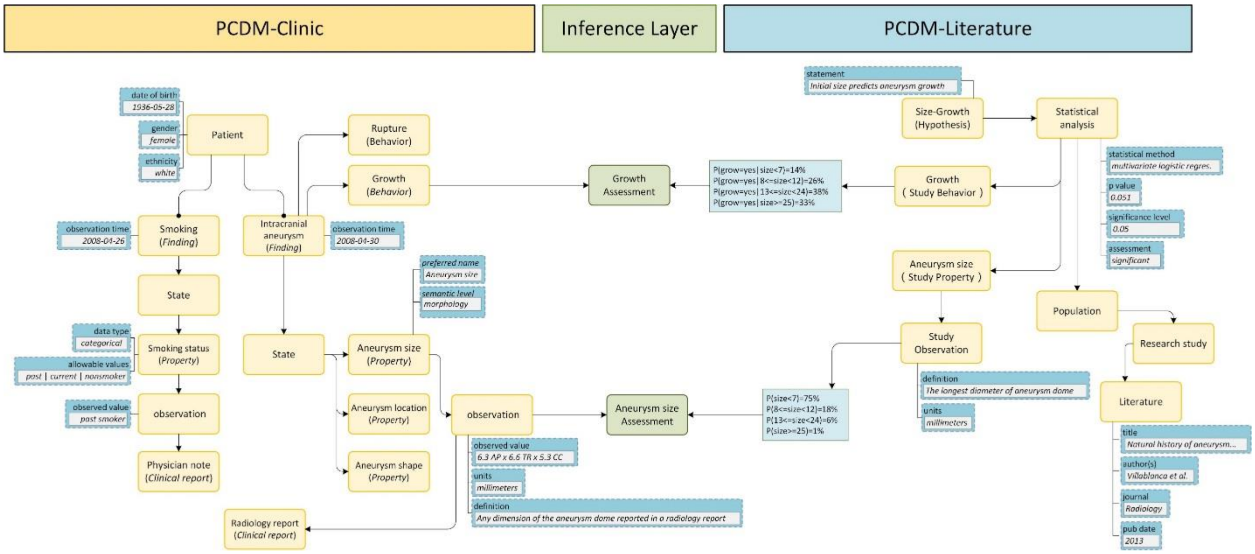
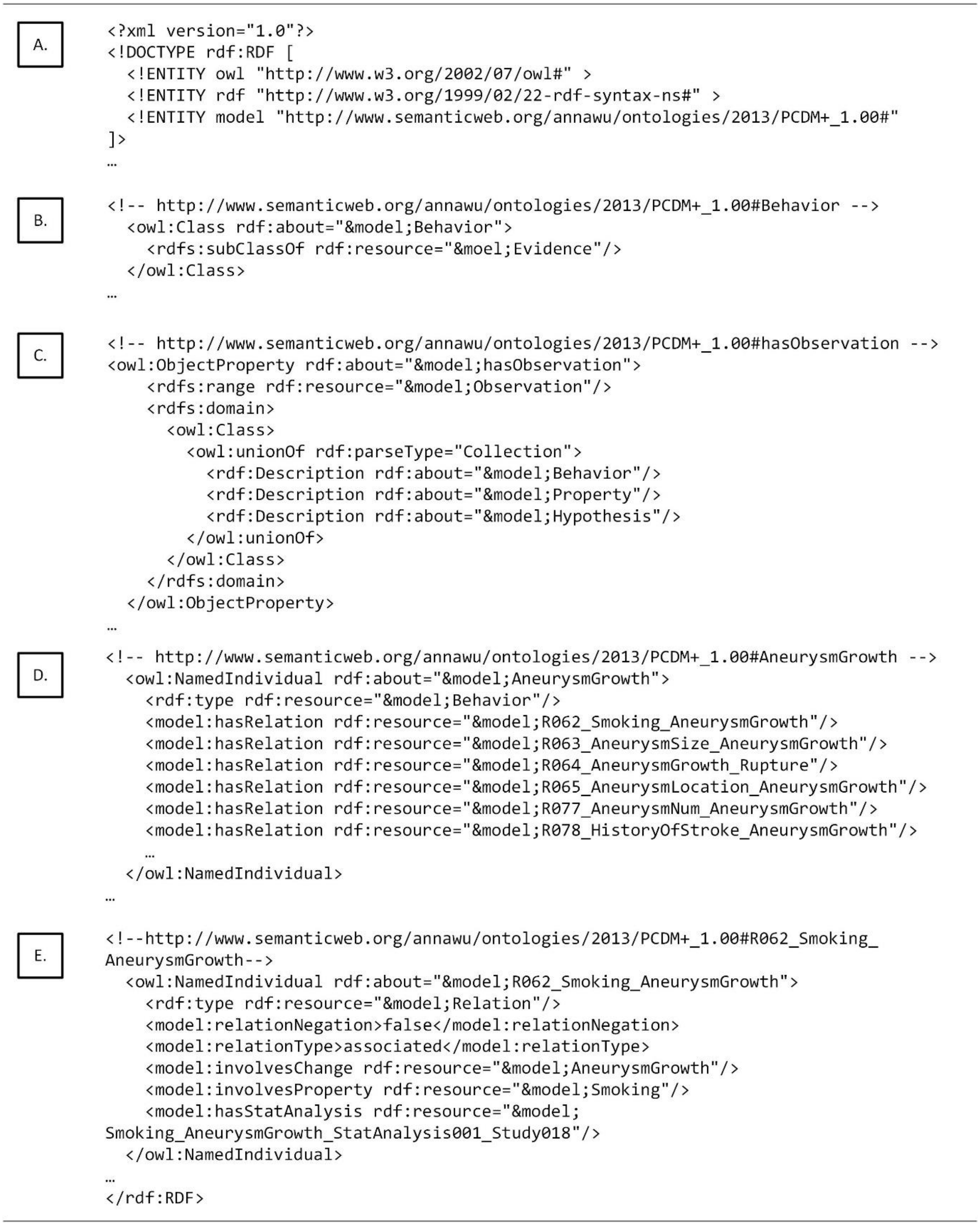


Figure 3.6 Integrate the patient case and a paper in PCDM+. The link between hypothesis "size-growth" and "Aneurysm size (as a Study variable)" is omitted. Some other instances are omitted as well due to the space limitation.

As shown in Figure 3.6, the aneurysm measurements are captured as a property called *Observed Value* related to an Observation, while the chance of aneurysm growth (i.e., 14%) is represented as a conditional probability statement P(grow = yes | size < 7 mm) as an instance of a Probability. As evidence from multiple papers and patient records is added to the PCDM+, information from each data source is maintained as an independent instance. As a result, PCDM+ is capable of returning all evidence related to a given entity. Figure 3.7 provides a portion of this encoding.



A.

```
<?xml version="1.0"?>
<!DOCTYPE rdf:RDF [
  <!ENTITY owl "http://www.w3.org/2002/07/owl#" >
  <!ENTITY rdf "http://www.w3.org/1999/02/22-rdf-syntax-ns#" >
  <!ENTITY model "http://www.semanticweb.org/annawu/ontologies/2013/PCDM+_1.00#"
]>
…
```

B.

```
<!-- http://www.semanticweb.org/annawu/ontologies/2013/PCDM+_1.00#Behavior -->
  <owl:Class rdf:about="&model;Behavior">
    <rdfs:subClassOf rdf:resource="&moel;Evidence"/>
  </owl:Class>
…
```

C.

```
<!-- http://www.semanticweb.org/annawu/ontologies/2013/PCDM+_1.00#hasObservation -->
<owl:ObjectProperty rdf:about="&model;hasObservation">
    <rdfs:range rdf:resource="&model;Observation"/>
    <rdfs:domain>
      <owl:Class>
        <owl:unionOf rdf:parseType="Collection">
          <rdf:Description rdf:about="&model;Behavior"/>
          <rdf:Description rdf:about="&model;Property"/>
          <rdf:Description rdf:about="&model;Hypothesis"/>
        </owl:unionOf>
      </owl:Class>
    </rdfs:domain>
  </owl:ObjectProperty>
…
```

D.

```
<!-- http://www.semanticweb.org/annawu/ontologies/2013/PCDM+_1.00#AneurysmGrowth -->
  <owl:NamedIndividual rdf:about="&model;AneurysmGrowth">
    <rdf:type rdf:resource="&model;Behavior"/>
    <model:hasRelation rdf:resource="&model;R062_Smoking_AneurysmGrowth"/>
    <model:hasRelation rdf:resource="&model;R063_AneurysmSize_AneurysmGrowth"/>
    <model:hasRelation rdf:resource="&model;R064_AneurysmGrowth_Rupture"/>
    <model:hasRelation rdf:resource="&model;R065_AneurysmLocation_AneurysmGrowth"/>
    <model:hasRelation rdf:resource="&model;R077_AneurysmNum_AneurysmGrowth"/>
    <model:hasRelation rdf:resource="&model;R078_HistoryOfStroke_AneurysmGrowth"/>
    …
  </owl:NamedIndividual>
…
```

E.

```
<!--http://www.semanticweb.org/annawu/ontologies/2013/PCDM+_1.00#R062_Smoking_
AneurysmGrowth-->
  <owl:NamedIndividual rdf:about="&model;R062_Smoking_AneurysmGrowth">
    <rdf:type rdf:resource="&model;Relation"/>
    <model:relationNegation>false</model:relationNegation>
    <model:relationType>associated</model:relationType>
    <model:involvesChange rdf:resource="&model;AneurysmGrowth"/>
    <model:involvesProperty rdf:resource="&model;Smoking"/>
    <model:hasStatAnalysis rdf:resource="&model;
Smoking_AneurysmGrowth_StatAnalysis001_Study018"/>
  </owl:NamedIndividual>
…
</rdf:RDF>
```

Figure 3.7 PCDM+ in RDF/OWL format. (A) definition of abbreviations in the entity namespace (e.g., “model” is short for the URI of PCDM+); (B) definition of entities (e.g., Behavior is defined as a subclass of Evidence); (C) definition of relationships among entities (e.g., *hasObservation* is defined as a relationship between *Behavior* or *Property* and *Observation*); (D) instantiation of entities, linked with hypotheses (e.g., *Aneurysm Growth* is added as an instance of *Behavior*, with relations to *Smoking*, *Aneurysm Size*, *Rupture*, *Aneurysm Location*, *Aneurysm Number*, and *History of Stroke*); and (E) instantiation of hypothesis with statistical analysis conducted (e.g., *Smoking_AneurysmGrowth* is an instance of *Hypothesis* with an statistical analysis from Study 018).

3.5.2 PCDM+ in XML Format

During the first experiment with Patient Case 1, I experienced issues when using the Protégé user interface to define probabilities. Therefore, an XML format of PCDM+ was generated. MySQL workbench (version 6.3) (Oracle Corporation, 2016) was used to create a database with 10 tables to store different fragmented evidence. The experiments conducted on Patient Case 2 were based on the XML version of PCDM+. After mapping 20 papers into XML format, 60 unique study variables, 78 unique hypotheses, 182 statistical analyses, and over 1,000 probabilities were recorded.

3.6 PCDM+ Evaluation

This section describes the tasks and the results of PCDM+ evaluation. Two experiments were performed to assess: (1) the utility of PCDM+ in selecting papers of interest; (2) the completeness and correctness of the PCDM+ design. The performance among annotators was also compared when they manually extracted relations and concepts from papers. The experiments were conducted using Patient Case 1 described in Sections 3.3.1 and the 22 papers (the paper selection process was described in Section 3.4.1).

3.6.1 Experiment Design

- Experiment 1. In Experiment 1, an expert with domain knowledge of ICAs was asked to review 22 papers, design a set of specific questions (Table 3.6), and identify papers that can provide answers to these questions. PCDM+ was then queried to assess its ability to identify the relevant studies, with precision/recall metrics computed accordingly. The objective was to retrieve papers that can help the physician gain more information to assist in EBM for a specified patient. The criteria used to filter papers in this experiment include: (1) selecting papers pertaining to studies that examined relationships among certain variables (e.g., Questions 2 & 3) and further to select those in which authors report significant relationships only (e.g., Question 1); (2) to select papers reporting on studies that had a sample population to which the patient is similar (e.g., Question 4); and (3) to select papers that reported outcome analysis for certain treatment (e.g., Question 5).
- Experiment 2. In Experiment 2, eight biomedical informatics students were recruited to annotate 11

papers retrieved from PCDM+ in Section 4.3.1 (Table 4.7). These annotators were given a tutorial on intracranial aneurysms; therefore, they had a basic understanding of this disease domain. Each paper was randomly assigned to three different annotators to independently generate a complete list of the relations, concepts and observations; as such, each annotator reviewed four full-text papers (one was given five papers). The annotators were asked to complete a table capturing: relation statements reported in the papers (including those with and without statistical tests); the concepts involved in each relation; and the statistical significance (yes/no/NA) of each relation. I subsequently analyzed the answers provided by all the annotators to generate a full list of concepts, relations, and observations for each concept and relation. This list was then compared against the information captured in the PCDM+ instantiation of the papers.

3.6.2 Evaluation Results

Table 3.6 shows the questions related to Experiment 1. In the assessment task, for the first criterion (i.e., Questions 1-3), precision and recall were 73.7%（14/19）and 93.3% (14/15), respectively. For the second criterion (Question 4), precision was 83.3%（10/12）and recall was 100% (12/12). Finally, for the third criterion of Experiment 1 (Question 5), both the precision and recall were 100%. Overall, PCDM+ achieved a precision of 96.8% (30/31) and a recall of 81.1% (30/37), respectively, in Experiment 1. This suggests that physicians can query PCDM+ to select papers that satisfy their criteria.

Notably, there was one false positive and seven false negatives in the results. Error analysis was conducted to elucidate the reasons for these false results, as noted below:

- **False positives.** One false positive was returned, involving a paper that evaluates the impact of hypertension and nicotine on the size of ruptured intracranial aneurysms. This paper (Paper ID: 020) reveals that aneurysm size is a dependent parameter on hypertension and cigarette smoking in ruptured aneurysms. The encoded relation in PCDM+ is *Hypertension_Smoking_Aneurysm-Size_Rupture*. While this paper was returned by the model to answer Question 3, the expert did not consider that the aim of this study was evaluating initial aneurysm size as a risk for rupture. Its authors did not solely examine whether initial aneurysm size is predictive of rupture, but rather compared the size of ruptured aneurysms from patients with and without hypertension and a history of smoking. In hindsight, the relation encoded in the model should instead be *Hypertension_Smoking_Aneurysm-Size* with ruptured aneurysms as context.

- **False negatives.** There were seven false negatives that PCDM+ failed to retrieve. For Question 1, two papers were not retrieved by PCDM+, as it recorded “low wall shear stress” as a property different from “wall shear stress.” For Question 3, three relevant papers assessing the influence of initial aneurysm size on rupture were not identified as their authors did not report results in the result or discussion/conclusion section of the paper (hence, PCDM+ did not capture these relations in the mapping process). Lastly, two relevant studies were not retrieved in answering Question 4, as the authors did not directly report the mean age of the studied population; instead, they reported the mean age of subgroups. PCDM+ captured the subgroup information, but did not combine those subgroups to form information pertaining to the entire studied population.

Table 3.6 Evaluation Task 1 questions.

1. **Which studies report significant relationships between wall shear stress and aneurysm rupture?**
2. **Which studies evaluate the relationship between aneurysm growth and rupture?**
3. **Which studies assess if initial aneurysm size is a risk factor for aneurysm growth or rupture?**
4. **Which studies involve population with average age >50?**
5. **Which papers report the outcome of endovascular coiling?**

Based on this analysis, PCDM+ is capable of retrieving the majority of papers that satisfy a query. Experiment 1 also supports the premise that the contextual information in PCDM+ is useful for filtering papers. In Experiment 2, the completeness and correctness of the PCDM+ design was evaluated. A list of concepts, relations, and their observations were generated by summarizing the answers provided by the eight annotators. In the 11 papers, 72 unique relations were found, with 216 observations. Among all relation observations, 85 were reported to be statistically significant, 114 were found not to be statistically significant, and 18 were not supported by statistical tests. In addition, 66 unique concepts were used in these relation statements, with 118 observations.

Analysis conducted as a part of Experiment 2 revealed that all annotators and PCDM+ were able to capture the majority of the relations reported in the papers. PCDM+ captured 61 relations (84.7%) and 197 observations of these relations (91.2%) (Table 3.8). The relations that PCDM+ failed to capture can be found in Table 3.7. Markedly, differences among annotators were due to two reasons: (1) in some cases, relation statements were mentioned in a paper (e.g., in its discussion section) but were not directly concluded by the authors (e.g., references to a secondary study), and thus it was not clear if such relations should be captured (or not); and (2) the representation of relations differed between annotators (e.g., in one study, a multivariate analysis was performed to evaluate the influence of six morphological factors on aneurysm rupture—some annotators captured this as six different relations, while others reported it as a single relation).

Table 3.7 Relations and the number of observations retrieved by annotators versus PCDM+ from 11 eligible papers (in the PCDM+ column, when PCDM+ captured all the observations, the cell is empty; when PCDM+ captured some observations, the cell is filled with the number of observations captured by PCDM+; when PCDM+ missed the entire relation, the cell is filled with a "NO")

Unique Relation	number of observations	PCDM+
Age_Growth	1	
Age_Rupture	10	8
Age_SmokingHistory	1	NO
Age_SurgicalOutcome	1	
AlcoholUse_AneurysmFormation	2	
AlcoholUse_Growth	1	
AneurysmalSymptoms_SurgicalOutcome	1	NO
AneurysmClinicalPresentation_Rupture	1	NO
AneurysmLocation_EndovascularOutcome	1	
AneurysmLocation_Growth	4	
AneurysmLocation_Rupture	7	
AneurysmLocation_SurgicalOutcome	1	
AneurysmMultiplicity_Growth	2	
AneurysmMultiplicity_Rupture	3	
AneurysmShape_Growth	1	NO
AneurysmSize_Age	2	
AneurysmSize_EndovascularOutcome	1	

Unique Relation	number of observations	PCDM+
AneurysmSize_Growth	3	
AneurysmSize_Rupture	28	25
AneurysmSize_SurgicalOutcome	1	
ArtrialFibrillation_AneurysmFormation	2	
AspectRatio_Rupture	9	
BloodPressure_Rupture	3	
BodyMassIndex_AneurysmFormation	2	
CoronaryArteryDisease_Growth	2	
Diabetes_AneurysmFormation	2	
Diabetes_Growth	2	
EllipticityIndex_Rupture	3	
EnergyLoss_Rupture	1	
FamilyHisotryOfAneurysm_Growth	3	
FamilyHistoryOfMyocardialInfarction_AneurysmFormation	2	
FamilyHistoryOfSAH_Rupture	1	
FamilyHistoryOfStroke_AneurysmFormation	2	
FlowSpeed_Rupture	1	NO
Gender_Growth	3	
Gender_Rupture	1	
Growth_Rupture	6	
HeartDisease_AneurysmFormation	2	
HeightWidthRatio_Rupture	5	
HistoryOfSAH_Rupture	2	
HistoryofIschaemicCerebrovascularDisease	1	
HistoryOfSAH_Growth	3	
HistoryOfStroke_Growth	2	
HistoryOfTIA_Growth	2	
Hypercholesterolemia_AneurysmFormation	2	
Hypertension_AneurysmFormation	2	
Hypertension_Growth	3	
Hypertension_Rupture	2	
Hypertension_SmokingHistory_AneurysmFormation	1	
InflowAngle_FlowPenetration	2	NO
InflowAngle_FlowSpeed	2	NO
InflowAngle_Rupture	5	
InflowAngle_WSS	2	
Migraine_AneurysmFormation	2	
MuralCalcification_Growth	1	NO
NonsphericityIndex_Rupture	8	
NumberOfAneurysm_Growth	2	

Unique Relation	number of observations	PCDM+
NumberOfVortices_Rupture	3	
OSI_Rupture	3	
PhysicalExercise_AneurysmFormation	2	
ResidentTime_Rupture	2	NO
SizeRatio_Rupture	8	
SmokingHistory_AlcoholUse	1	
SmokingHistory_AneurysmFormation	2	
SmokingHistory_Growth	4	
SmokingHistory_Rupture	3	
SystolicBP_Rupture	1	
Thrombus_Growth	1	NO
Treatment_AneurysmalSymptoms	1	NO
Treatment_SAH	1	
UndulationIndex_Rupture	2	
WSS_Rupture	13	

Table 3.8 Relations and their observations retrieved from 11 eligible papers (only statistical significance, definition or state as a context, and paper ID are presented due to space limitation).

Relations	Significance	Definition/ state used in the relation	Paper ID
Age_Growth	N		11
Age_Rupture	N	age at the time of diagnosis	12
Age_Rupture	N		14
Age_Rupture	N	age = (31-40 yrs) vs. (<30 yrs)	22
Age_Rupture	N	age = (41-50 yrs) vs. (<30 yrs)	22
Age_Rupture	N	age (>51) vs. (<30 yrs)	22
Age_Rupture	Y	age as a continuous variable	22
Age_Rupture	N	age = (31-40 yrs) vs. (<30 yrs)	22
Age_Rupture	Y	age = (41-50 yrs) vs. (<30 yrs)	22
Age_Rupture	N	age (>51) vs. (<30 yrs)	22
Age_Rupture	Y	age as a continuous variable	22
Age_SmokingHistory	Y		22
Age_SurgicalOutcome	Y	age>=50 vs. <50 yrs, outcome = poor vs. good	14
AlcoholUse_AneurysmFormation	N	alcohol >=18 U/week	16
AlcoholUse_AneurysmFormation	N	alcohol >=18 U/week	16
AlcoholUse_Growth	Y		11
AneurysmalSymptoms_SurgicalOutcome	Y	aneurysmal symptoms other than rupture	14
AneurysmClinicalPresentation_Rupture	N	rupture = SAH, aneurysm clinical presentation: symptomatic	22

Relations	Significance	Definition/ state used in the relation	Paper ID
		aneurysm, incidental aneurysm, and prior SAH	
AneurysmLocation_ EndovascularOutcome	Y	location = posterior circulation vs. anterior circulation	14
AneurysmLocation_Growth	N		10
AneurysmLocation_Growth	Y	posterior circulation vs. non-posterior circulation	10
AneurysmLocation_Growth	N		11
AneurysmLocation_Growth	N		18
AneurysmLocation_Rupture	N	location= ICA vs. =ACA	13
AneurysmLocation_Rupture	N	location= ICA vs. =MCA	13
AneurysmLocation_Rupture	N	location= ICA vs. =VABA	13
AneurysmLocation_Rupture	Y	location=anterior circulation vs. =posterior circulation	13
AneurysmLocation_Rupture	Y	location = tips of basilar artery vs. internal carotid artery,rupture = haemorrhage	14
AneurysmLocation_Rupture	Y	location = cavernous artery vs. internal carotid artery,rupture = haemorrhage	14
AneurysmLocation_Rupture	Y	location = posterior communicating artery vs. internal carotid artery,rupture = haemorrhage	14
AneurysmLocation_SurgicalOutcome	Y	location = posterior circulation vs. anterior circulation	14
AneurysmMultiplicity_Growth	N	alpha = 0.05	10
AneurysmMultiplicity_Growth	N		18
AneurysmMultiplicity_Rupture	NA		12
AneurysmMultiplicity_Rupture	N		13
AneurysmMultiplicity_Rupture	N	rupture = SAH	22
AneurysmShape_Growth	Y		18
AneurysmSize_Age	Y		22
AneurysmSize_Age	N		22
AneurysmSize_EndovascularOutcome	Y	diameter >12 mm	14
AneurysmSize_Growth	N	size>10 mm vs. size<=10 mm	11
AneurysmSize_Growth	N	average initial size	12
AneurysmSize_Growth	Y	initial size	18
AneurysmSize_Rupture	N		4
AneurysmSize_Rupture	N		4
AneurysmSize_Rupture	N		4
AneurysmSize_Rupture	Y	average of maximum diameter	5
AneurysmSize_Rupture	Y	height	5
AneurysmSize_Rupture	N	average of maximum diameter	5

Relations	Significance	Definition/ state used in the relation	Paper ID
AneurysmSize_Rupture	N	height	5
AneurysmSize_Rupture	Y	average of maximum diameter	5
AneurysmSize_Rupture	Y	height	5
AneurysmSize_Rupture	N	average of maximum diameter	5
AneurysmSize_Rupture	N	height	5
AneurysmSize_Rupture	N	average of maximum diameter	5
AneurysmSize_Rupture	N	height	5
AneurysmSize_Rupture	Y	size at the end of follow-up	12
AneurysmSize_Rupture	N	size=(5-9.9 mm) vs. (<5 mm)	13
AneurysmSize_Rupture	Y	size=(10-24.9 mm) vs. (<5 mm)	13
AneurysmSize_Rupture	Y	size=(>25 mm) vs. (<5 mm)	13
AneurysmSize_Rupture	NA		14
AneurysmSize_Rupture	NA		14
AneurysmSize_Rupture	Y	maximum diameter = 7-12 mm, rupture = hemorrhage	14
AneurysmSize_Rupture	Y	maximum diameter >12 mm, rupture = hemorrhage	14
AneurysmSize_Rupture	N	size = (7-9 mm) vs. = (2-6 mm)	22
AneurysmSize_Rupture	N	size = (10-26 mm) vs. = (2-6 mm)	22
AneurysmSize_Rupture	Y	size as a continuous variable	22
AneurysmSize_Rupture	N	size = (7-9 mm) vs. = (2-6 mm)	22
AneurysmSize_Rupture	Y	size = (10-26 mm) vs. = (2-6 mm)	22
AneurysmSize_Rupture	Y	size as a continuous variable	22
AneurysmSize_Rupture	Y	size >7 mm vs. <7 mm	22
AneurysmSize_SurgicalOutcome	Y	diameter >12 mm	14
ArtrialFibrillation_AneurysmFormation	N		16
ArtrialFibrillation_AneurysmFormation	N		16
AspectRatio_Rupture	N		4
AspectRatio_Rupture	N		4
AspectRatio_Rupture	N		4
AspectRatio_Rupture	N		5
AspectRatio_Rupture	N		5
AspectRatio_Rupture	Y		5
AspectRatio_Rupture	N		5
AspectRatio_Rupture	N		5
AspectRatio_Rupture	NA	AR>1.6 --> rupture	6
BloodPressure_Rupture	N	Here it refers the BP at the beginning of the follow-up. mean arterial blood pressure = diastolic BP+(systolic BP-	22

Relations	Significance	Definition/ state used in the relation	Paper ID
		diastolic BP)/3.	
BloodPressure_Rupture	Y	here it refers to the BP at the end of the follow-up	22
BloodPressure_Rupture	N	mean arterial pressure after adjusted for age	22
BodyMassIndex_AneurysmFormation	N	Body mass index >=30	16
BodyMassIndex_AneurysmFormation	N	Body mass index >=30	16
CoronaryArteryDisease_Growth	N		10
CoronaryArteryDisease_Growth	N		10
Diabetes_AneurysmFormation	N		16
Diabetes_AneurysmFormation	N		16
Diabetes_Growth	N		10
Diabetes_Growth	N		10
EllipticityIndex_Rupture	Y		4
EllipticityIndex_Rupture	N		4
EllipticityIndex_Rupture	N		4
EnergyLoss_Rupture	Y		6
FamilyHisotryOfAneurysm_Growth	N	family history of ICA	10
FamilyHisotryOfAneurysm_Growth	N		11
FamilyHisotryOfAneurysm_Growth	N	family history of ICA	10
FamilyHistoryOfMyocardialInfarction_AneurysmFormation	Y		16
FamilyHistoryOfMyocardialInfarction_AneurysmFormation	N		16
FamilyHistoryOfSAH_Rupture	NA		12
FamilyHistoryOfStroke_AneurysmFormation	Y	Stroke here includes ischemic and hemorrhagic stroke but excludes subarachnoid hemorrhage	16
FamilyHistoryOfStroke_AneurysmFormation	Y	Stroke here includes ischemic and hemorrhagic stroke but excludes subarachnoid hemorrhage	16
FlowSpeed_Rupture	NA		6
Gender_Growth	N		10
Gender_Growth	N		10
Gender_Growth	N		11
Gender_Rupture	N	rupture = SAH	22
Growth_Rupture	N		11
Growth_Rupture	Y	absolute diameter growth	12
Growth_Rupture	Y	growth percentage for aneurysms with and without rupture	12
Growth_Rupture	Y		18
Growth_Rupture	Y		18

Relations	Significance	Definition/ state used in the relation	Paper ID
Growth_Rupture	N	annual growth rate	12
HeartDisease_AneurysmFormation	N		16
HeartDisease_AneurysmFormation	N		16
HeightWidthRatio_Rupture	N		5
HeightWidthRatio_Rupture	N		5
HeightWidthRatio_Rupture	Y		5
HeightWidthRatio_Rupture	N		5
HeightWidthRatio_Rupture	Y		5
HisotryOfSAH_Rupture	Y		13
HisotryOfSAH_Rupture	N		13
HistoryofIschaemicCerebrovascularDisease	Y		14
HistoryOfSAH_Growth	N	prior aneurysmal SAH	11
HistoryOfSAH_Rupture	NA		12
HistoryOfSAH_Rupture	Y	rupture rate	14
HistoryOfStroke_Growth	Y		10
HistoryOfStroke_Growth	N		10
HistoryOfTIA_Growth	N		10
HistoryOfTIA_Growth	Y		10
Hypercholesterolemia_AneurysmFormation	N		16
Hypercholesterolmia_ _AneurysmFormation	Y		16
Hypertension_AneurysmFormation	Y		16
Hypertension_AneurysmFormation	Y		16
Hypertension_Growth	N		10
Hypertension_Growth	N		10
Hypertension_Growth	N		11
Hypertension_Rupture	NA		12
Hypertension_Rupture	N	at the beginning of the follow-up. Hypertension is defined as a systolic pressure repeatedly greater than 160 mm Hg, diastolic pressure greater than 95 mmHg, or as the use of antihypertension medication.	22
Hypertension_ _SmokingHistory_AneurysmFormation	Y		16
InflowAngle_FlowPenetration	NA		5
InflowAngle_FlowPenetration	NA		5
InflowAngle_FlowSpeed	NA		5
InflowAngle_FlowSpeed	NA		5
InflowAngle_Rupture	Y		5
InflowAngle_Rupture	N		5

Relations	Significance	Definition/ state used in the relation	Paper ID
InflowAngle_Rupture	Y		5
InflowAngle_Rupture	N		5
InflowAngle_Rupture	Y		5
InflowAngle_WSS	NA		5
InflowAngle_WSS	NA		5
Migraine_AneurysmFormation	N		16
Migraine_AneurysmFormation	N		16
MuralCalcification_Growth	N		18
NonsphericityIndex_Rupture	Y		4
NonsphericityIndex_Rupture	N		4
NonsphericityIndex_Rupture	N		4
NonsphericityIndex_Rupture	N		5
NonsphericityIndex_Rupture	N		5
NonsphericityIndex_Rupture	Y		5
NonsphericityIndex_Rupture	Y		5
NonsphericityIndex_Rupture	N		5
NumberOfAneurysm_Growth	N		10
NumberOfAneurysm_Growth	Y		10
NumberOfVortices_Rupture	Y		4
NumberOfVortices_Rupture	N		4
NumberOfVortices_Rupture	N		4
OSI_Rupture	Y	average OSI	4
OSI_Rupture	Y	average OSI	4
OSI_Rupture	Y	average OSI	4
PhysicalExercise_AneurysmFormation	Y	exercise >= 3 times /week	16
PhysicalExercise_AneurysmFormation	Y	exercise >= 3 times /week	16
ResidentTime_Rupture	Y	relative resident time	4
ResidentTime_Rupture	NA		6
SizeRatio_Rupture	Y		4
SizeRatio_Rupture	Y		4
SizeRatio_Rupture	Y		4
SizeRatio_Rupture	N		5
SizeRatio_Rupture	N		5
SizeRatio_Rupture	Y		5
SizeRatio_Rupture	N		5
SizeRatio_Rupture	Y		5
SmokingHistory_AlcoholUse	Y		22
SmokingHistory_AneurysmFormation	Y		16
SmokingHistory_AneurysmFormation	Y		16

Relations	Significance	Definition/ state used in the relation	Paper ID
SmokingHistory_Growth	N	current or previous cigarette smoking	10
SmokingHistory_Growth	N	current or previous cigarette smoking	10
SmokingHistory_Growth	N		11
SmokingHistory_Growth	Y		18
SmokingHistory_Rupture	NA		12
SmokingHistory_Rupture	Y	at the time of diagnosis	22
SmokingHistory_Rupture	Y	as a time-dependent covariate	22
SystolicBP_Rupture	N	after adjusted for age	22
Thrombus_Growth	N	intraluminal thrombus	18
Treatment_AneurysmalSymptoms	NA	Treatment = operation (surgery and coiling) vs. observation	14
Treatment_SAH	NA	Treatment = operation (surgery and coiling) vs. observation	14
UndulationIndex_Rupture	Y		4
UndulationIndex_Rupture	N		4
UndulationIndex_Rupture	N		4
WSS_Rupture	Y	average WSS	4
WSS_Rupture	Y	maximum intra-aneurysmal WSS	4
WSS_Rupture	Y	low WSS area	4
WSS_Rupture	N	WSS gradient	4
WSS_Rupture	Y	average WSS	4
WSS_Rupture	N	WSS gradient	4
WSS_Rupture	N	maximum intra-aneurysmal WSS	4
WSS_Rupture	N	low WSS area	4
WSS_Rupture	Y	average WSS	4
WSS_Rupture	N	WSS gradient	4
WSS_Rupture	N	maximum intra-aneurysmal WSS	4
WSS_Rupture	N	low WSS area	4
WSS_Rupture	N	time-averaged WSS	6

Concepts representation and extraction. Of the 66 concepts and 118 observations, PCDM+ successfully captured 58 concepts (87.9%) and 104 observations of concepts (88.1%). Table 3.9 shows the 58 unique concepts encoded in PCDM+. The concepts that were annotated by the annotators but missed by PCDM+ include length of time, death, aneurysmal symptoms other than rupture, flow penetration, mural calcification, intraluminal thrombus, inflow inlet, and resident time. The results yielded by Experiment 2 suggest that PCDM+ is capable of representing the predominance of concepts and observations that annotators also noted. Moreover, PCDM+ captured more than the average number of observations that the annotators were able to identify in many of the papers (9 of 11 papers).

Consistency among annotators. Experiment 2 also tested the consistency among annotators when they structured evidence from published literature into relations, concepts, and statistics. Table 3.10 shows the concepts that appear in each paper and the performance of each annotation and PCDM+. Based on this information, Table 3.11 was created to show the numbers of concepts annotated by annotators as well as the number of concepts encoded in PCDM+. The mean and variance of this measure are also provided. Notably, the variance among annotators changed across papers: marked variances exist among 4 of the 11 papers, with negligible variances were noted in the remaining papers. The percentage of agreement also varies according to the papers. One possible reason is that each paper had a different level of complexity relating to reporting structure, statistical analyses performed, and a number of hypotheses examined. The results indicated that inconsistency among annotators existed and varied across papers. PCDM+ captured the same (or greater) average number of observations as that achieved by users for most papers.

Table 3.9 List of 66 unique concepts in alphabetical order retrieved from 11 eligible papers by PCDM+.

age of patient	family history of stroke	medical history of diabetes
alcohol use	flow velocity/inflow speed	medical history of heart disease
aneurysm diameter	gender of the patient	medical history of hypercholesterolemia
aneurysm height	growing aneurysms	medical history of migraine
aneurysm height width ratio	growth	mural calcification
aneurysm location	growth rate	nonsphericity index
aneurysm multiplicity	hemorrhage	number of aneurysm
aneurysm shape	history of aneurysmal SAH	number of vertices
aneurysm size	history of hypertension	OSI
aneurysm type	history of ischaemic cerebrovascular disease	penetration of flow
aneurysm volume	history of SAH	regular physical exercise
aneurysmal symptoms other than rupture	history of stroke	relative resident time
aspect ratio	history of transient ischemic attack	risk of SAH

average-wss	image follow-ups	risk of UIA
body mass index	inflow angle	rupture
death	inflow inlet	rupture rate/ cumulative rate of bleeding
ellipticity index	intraluminal thrombus	size ratio
endovascular outcome	length of time	smoking history/status
energy loss	low wss area	surgical outcome
family history of ICA	maximum-wss	systolic BP
family history of MI	mean arterial pressure	undulation index
family history of SAH	medical history of AF	wss gradient

Table 3.10 Concepts that appear in each paper and the performance of each annotation and PCDM+ (indicated with an "X")

Paper ID	Reported Concepts	Annotation 1	Annotation 2	Annotation 3	PCDM+
4	aneurysm size		X	X	
	size ratio	X	X	X	X
	undulation index	X	X	X	X
	ellipticity index	X	X	X	X
	nonsphericity index	X	X	X	X
	average wall shear stress	X	X	X	X
	maximum wall shear stress	X	X	X	X
	wall shear stress gradient			X	X
	low WSS area	X	X	X	X
	average OSI	X	X	X	X
	number of vertices	X	X	X	X
	relative resident time	X	X	X	X
	rupture	X	X	X	X
5	Dmax	X	X	X	X
	height	X	X	X	X
	height width ratio	X	X	X	X
	size ratio	X	X	X	X
	inflow angle	X	X	X	X
	nonsphericity index	X	X	X	X
	sidewall aneurysm	X	X	X	X
	bifurcation aneurysm	X	X	X	X
	aspect ratio		X	X	X
	penetration of flow		X		
	flow velocity		X		
	wall shear stress		X		
	rupture	X	X	X	X
6	aspect ratio		X		

Paper ID	Reported Concepts	Annotation 1	Annotation 2	Annotation 3	PCDM+
	energy loss	X	X	X	X
	wall shear stress	X	X	X	X
	inflow speed	X	X		
	inflow inlet		X		
	resident time	X	X		X
	rupture	X	X		X
10	growth rate	X	X	X	X
	number of aneurysm	X	X	X	X
	a history of stroke	X	X	X	X
	growth	X	X	X	X
	aneurysm location	X	X	X	X
	a history of transient ischemic attack	X	X	X	X
	gender of the patient		X		X
	initial aneurysm size		X		X
	length of time			X	
11	excessive alcohol consumption	X	X	X	X
	age of the patient	X		X	X
	growth	X	X	X	X
	gender of the patient	X		X	X
	smoking status	X		X	X
	history of hypertension	X		X	X
	prior aneurysmal SAH	X		X	X
	family history intracranial aneurysms	X		X	X
	aneurysm size	X		X	X
12	aneurysm location		X		X
	length of time		X		
	aneurysm size		X	X	X
	rupture	X	X	X	X
	growth	X	X	X	X
	age of the patient		X	X	X
	annual growth rate	X	X	X	X
	history of hypertension			X	X
	smoking history			X	X
	family history of SAH			X	X
	history of previous SAH			X	X
	aneurysm multicity			X	X
13	a history of SAH	X	X	X	X
	aneurysm location	X	X	X	X
	aneurysm size	X	X	X	X
	rupture	X	X	X	X

Paper ID	Reported Concepts	Annotation 1	Annotation 2	Annotation 3	PCDM+
14	age of patient	X	X	X	X
	surgical outcome	X	X	X	X
	aneurysm location	X	X	X	X
	aneurysm size	X	X	X	X
	unruptured aneurysm	X	X	X	X
	previous ischaemic cerebrovascular disease		X		X
	aneurysmal symptoms other than rupture		X	X	
	endovascular outcome			X	X
	rupture rate			X	X
16	smoking history	X	X	X	X
	history of hypertension	X	X		X
	family history of stroke	X	X	X	X
	medical history of hypercholesterolemia	X	X	X	X
	regular physical exercise	X	X	X	X
	alcohol use	X			X
	body mass index	X			X
	medical history of diabetes	X			X
	medical history of AF	X			X
	medical history of heart disease	X			X
	medical history of migraine	X			X
	family history of MI	X			X
	risk of UIA	X	X	X	X
	risk of SAH		X		X
18	rupture	X	X	X	X
	growing aneurysms	X		X	X
	aneurysm size	X	X	X	X
	aneurysm volume	X		X	X
	growth	X		X	X
	(tobacco) smoking history	X	X	X	X
	image follow-ups	X			
	aneurysm location		X		X
	aneurysm multiplicity		X		X
	intraluminal thrombus		X		
	mural calcification		X		
	aneurysm shape		X		X
22	aneurysm diameter (size)	X	X	X	X
	age of patient	X	X	X	X
	smoking status	X	X	X	X

Paper ID	Reported Concepts	Annotation 1	Annotation 2	Annotation 3	PCDM+
	rupture rate				X
	rupture	X	X	X	X
	risk of SAH	X			X
	gender of the patient	X			X
	multiplicity of aneurysm	X			X
	cumulative rate of bleeding (rupture)	X		X	X
	systolic BP	X			x
	mean arterial pressure	X			x
	hemorrhage		X		X
	death		X		
	aneurysm location			X	X
	alcohol consumption(use)			X	X
	a history of hypertension			X	X

Table 3.11 Observations captured by each annotator and PCDM+.

Paper	PCDM+ observations of concepts	# observations of concepts by annotators		
		Mean	Variance	Percent agreement
Paper 1	**12**	12	0.67	84.62%
Paper 2	**10**	12	2.89	69.23%
Paper 3	**4**	5	4.22	28.57%
Paper 4	**8**	7	0.67	66.67%
Paper 5	**9**	7	10.89	22.22%
Paper 6	**11**	6	6.22	25%
Paper 7	**4**	4	0	100%
Paper 8	**8**	7	2.67	55.56%
Paper 9	**14**	8	11.56	35.71%
Paper 10	**9**	7	0.67	25%
Paper 11	**15**	8	2.67	25%

CHAPTER 4. Operators

This chapter describes the work performed to meet the Aim 2 of this dissertation:

[Aim 2] *Develop operators that translate evidence into knowledge elements to inform clinical decision making relating to a specific patient.*

After PCDM+ is instantiated with the knowledge fragments of a disease (e.g., intracranial aneurysms) or a phenomenon (e.g., rupture) from multiple sources, the next step is to use the evidence it encodes to inform clinical decision making relating to individual patients. As an intermediate representation aimed at integrating knowledge about the disease from multiple sources, PCDM+ has potential in two applications: (1) to enhance clinical research on understanding the disease by guiding disease modeling, and (2) to improve clinical practice by facilitating individually-tailored medical decision making. The knowledge elements needed to realize these two applications are captured in the PCDM-Inference Layer, and several operators are designed to achieve the knowledge integration and translation. Based on the PCDM+, I developed three operators to guide disease modeling (i.e., for the construction of Bayesian belief networks, BBNs). I also created operators (overlapping with the BBN construction operators) to retrieve and link fragments of evidence found in literature relevant to an individual patient. This work is demonstrated by answering clinical queries that arose during the medical decision-making process in ICA treatment.

4.1 Operators for Evidence-based Medicine

When physicians practice evidence-based medicine (EBM) by integrating evidence from observational studies and randomized clinical trials (RCTs), they need to search among a large volume of literature, extract the useful evidence, and assess if it is relevant to the targeted patients. The utility of PCDM+ stems from its ability to integrate information from patient records and published literature to answer clinical questions. The development of PCDM+ focused on three types of queries:

1. Patient-paper matching. Given the patient characteristics, which papers are relevant for a clinician to consider as part of the decision-making process?

2. Disease etiology and behavior. What is the patient's risk of forming an aneurysm? How likely will this patient's aneurysm grow?
3. Treatment planning. Which treatment is better for this patient, endovascular coiling (EC) or surgical clipping (SC)? What are the prognostic variables that may indicate whether a patient will have improved or worsening neurologic outcomes?

To pose a query to the model, the question can be deconstructed into a set of variables and relationships. For example, the question, "How likely is it that this patient's aneurysm will grow?" may be broken into two separate questions pertaining to information sourced from literature and the patient record: (1) what are the reported risk factors for aneurysm growth (from the literature); and (2) what risk factors does the given patient have? For example, based on the structure of PCDM+, the entity instance "*growth*" is associated with the relation instance "*size-growth*" that has been referenced in several papers. By retrieving all the hypothesis tests associated with the "*size-growth*" relationship, information from papers in which researchers discuss potential risk factors that contribute to aneurysm growth is retrieved. By leveraging the semantic relationships, all observations can be linked to their sources (e.g., a specific paper and population) and population characteristics (e.g., characteristics of the individuals that were studied). Therefore, to utilize the evidence that has been structured in PCDM+ to answer these three types of clinical queries, several operators were designed to aid in patient-paper matching, hypothesis examination, and risk estimation. These operators are called: *patient-population matching, relation extraction,* and *probability retrieval*.

4.1.1 Patient-population Matching

In the first step in evaluating evidence for an individual patient, the papers that contain pieces of knowledge that are applicable to the given patient are selected.

> **Query 1:** *"Given the patient characteristics, which papers are relevant for a clinician to consider as part of the decision-making process?"*

While a few papers may pertain to case studies, most report research findings obtained by studying a sample population. In PCDM+, every piece of evidence is linked to the study population from a published study it originates from. Given that a single paper may reference more than one group of participants that took part in the studies it reports on, I call this operator *patient-population matching* instead of *patient-paper matching* operator.

One method to achieve patient-population matching is to use the eligibility criteria for this population as selection criteria. The eligibility criteria of a study are encoded in PCDM+ as an attribute of a population. By following these steps, one can use the *patient-population matching* operator:

1. The method getEligibilityCriteria (Population pop) is used to query PCDM-Literature to retrieve the eligibility criteria of the population pop (see Table 4.1);
2. Encode each subject criterion as a "feature name-logic operator-feature value" form (see Table 4.2).
3. Use the methods getFeatureName (Eligibility Criterion ec), getLogicOperator(Eligibility Criterion ec), and getFeatureValue (Eligibility Criterion ec), return the feature name, logic operator, and feature value related to eligibility criterion ec.
4. Use the method getObservedValue(Study Variable sv) to retrieve the observed value of study variable sv from PCDM-Clinic (see Table 4.3);
5. Assess if the observed value from the patient satisfies the subject criterion reported in the literature. This process is accomplished by (1) linking the subject variables in PCDM-Literature to the corresponding feature in PCDM-Clinic; (2) finding the values for the feature in both PCDM-Clinic and PCDM-Literature; (3) using a rule-based algorithm to assess if the value from the patient case meets the defined criterion; (4) returning a value indicating that the criterion is not satisfied (e.g., 0); satisfied (e.g., 1); or is missing (e.g., N/A) for each subject criterion (see Table 4.4);
6. If there are missing values:
    a. When one of the criteria is unsatisfied, return 0;
    b. When all criteria are satisfied except for the missing ones, this can be mitigated by weighting the matching score with the ratio of the number of criteria that were satisfied to the total number of

the criteria. A threshold is chosen for a matching score function to determine whether to include that population or not. For example, in the case shown here, a matching score with one missing value out of five criteria is 4/5 = 0.8.

Table 4.1 Eligibility criteria are recorded in the population table.

pop_id	sample_size	eligibility_criteria	subpopulation
pop001	142 patients, 181 aneurysms	(1)Patients with unruptured intracranial aneurysm were included. (2)All patients with a conservatively treated ruptured aneurysm were excluded. (3)Patients with mycotic or fusiform atherosclerotic aneurysms were excluded. (4)Patients with uncommon intracavernous carotid artery aneurysms were excluded. (5)Patients with symptomatic aneurysm were included in the study only if SAH was excluded by a lumbar puncture within a few days after the onset of symptoms.	pop001.subpop001, pop001.subpop002, pop001.subpop003.

Table 4.2 Eligibility criteria are represented in feature name-mathematical symbol-feature value form in a Boolean model.

```
Aneurysm.rupture = no AND
Aneurysm.treatment = no AND
Aneurysm.shape != fusiform AND
Aneurysm.property != mycotic or atherosclerotic AND
((Aneurysm.clinicalPresentation !=symptomatic) OR
(Aneurysm.clinicalPresentation = symptomatic&SAH = no))
```

Table 4.3 Patient case 1

Patient	Lily Bruin
Age	70 years old
Gender	female
Clinical presentation	incidental
Aneurysm rupture	no
Aneurysm location	right PCoA*
Aneurysm size	6.3×6.6×5.3 mm
Aneurysm shape	saccular
Modality	CTA**
Family history of stroke	yes
Family history of hypertension	yes
Medical history	head and neck cancer
Smoking status	former-smoker
Alcohol use	none
Blood pressure	106/65
Treatment	No

(*PCoA: Posterior Communicating Artery; **CTA: Computed Tomography Angiography)

Table 4.4 Matching result for each eligibility criterion.

Eligibility Criteria	Matching Score (1/0)
Aneurysm rupture = no	1
Aneurysm treatment = no	1
Aneurysm shape != fusiform	1
Aneurysm property != mycotic or atherosclerotic	N/A
(Clinical presentation != symptomatic)) OR ((Clinical presentation = symptomatic AND SAH = no))	1

4.1.2 Relation Extraction

In clinical practice, physicians' aim is to identify features that are risk factors for a certain outcome (e.g., rupture), so that given the observed characteristics of a patient, the physician can estimate the risk. In the domain of ICA, two types of risk are considered: rupture risk and treatment risk (i.e., the risk of mortality and morality).

Query 2: *"What are the risk factors for aneurysm rupture? Is the aneurysm this patient has at the high risk of rupture?"*

If the rupture risk is low, taking other factors into consideration as well (e.g., age), a routine observation rather than clipping or coiling of the aneurysm (i.e., watchful waiting) may be chosen as an optimal treatment. If the risk of rupture is high, then treatment risk needs (e.g., the risk of clinical complications) to be estimated. For example, if the risk associated with performing endovascular coiling is higher than that related to surgical clipping treatment for a particular patient, clipping should be selected as the optimal treatment. To answer queries relating to risk estimation, two operators were designed in this work: *relation extraction* and *probability retrieval*. This section illustrates the former via an operator that embeds information retrieval functions based on the PCDM+ design.

The relation extraction operator contains two main methods:

(1) A method called getRelation(Study Variable sv) that returns all relations with a study variable sv recorded in PCDM+ from the Hypothesis table;

(2) A method called getStatAnalyses(Hypothesis h) that returns all statistical analyses associated with hypothesis *h* from PCDM+. Each statistical analysis is retrieved from PCDM+ with the context, in-

cluding the statistical assessment (i.e., significant or not), the statistical result (e.g., p = 0.002), statistical method, analysis time, population, and source information.

For example, if a physician would like to explore the risk factors of aneurysm rupture, the method getRelation(aneurysmRupture) is used to retrieve all the hypotheses that state relations between aneurysm rupture and risk factors (see Table 4.5). Among these relations, if a physician is interested in knowing details about a specific relation (e.g., *aneurysm size* and *aneurysm rupture*), the getStatAnalyses ("aneurysm size is a risk factor of *aneurysm rupture*") method is used to retrieve multiple statistical analyses that have been performed to assess the significance of this relationship in various studies (see Table 4.6).

Table 4.5 Relations of "aneurysm rupture" retrieved from PCDM+

Study variables	Hypothesis statement
gender, aneurysm rupture	"patient gender is a risk factor of aneurysm rupture"
age, aneurysm rupture	"patient age is a risk factor of aneurysm rupture"
aneurysm size, aneurysm rupture	"aneurysm size is a risk factor of aneurysm rupture"
aneurysm location, aneurysm rupture	"aneurysm location is a risk factor of aneurysm rupture"
daughter sac, aneurysm rupture	"daughter sac is a risk factor of aneurysm rupture"
hypertension, aneurysm rupture	"hypertension is a risk factor of aneurysm rupture"

Table 4.6 Statistical analyses of "aneurysm size is a risk factor of aneurysm rupture" retrieved from PCDM+

Statistical method	Statistical result	Statistical assessment	Analysis time	Population	Study
Cox Model with a stepwise procedure	p=0.036	significant	at the end of the follow-up	pop001	study001
NA	p=0.03; p<0.001	significant	at the end of the follow up	pop002.subpop001.group001	study002
…	…	…	…	…	…
multivariate analysis with the proportional hazards methods	p=0.01; p<0.0001	significant	at the end of the follow-up	pop004.subpop001	study004

4.1.3 Probability Retrieval

While the *relation extraction* operator is able to return the reported relations and the associated statistical analyses, the *probability retrieval* operator is designed to retrieve associated probabilities and distributions. The goal of this operator is to provide the currently available evidence to facilitate answering queries such as:

Query 3: *"How likely will this patient's aneurysm grow and rupture?"*

In PCDM+, marginal probabilities (e.g., P(age), P(aneurysm size)) and distributions of certain patient characteristics are captured. These probabilities and distributions are helpful in assessing whether a given patient's information is within the distribution a study characterizes. In addition, PCDM+ maintains conditional probabilities (e.g., P(rupture = yes| aneurysm size = 2-6mm), P(rupture = yes| age <50 year old)) from multiple papers. While these probabilities are not full joint conditional probabilities that can be used to answer the query directly, they provide important insights into how much a given risk factor can influence risk estimation. The *probability retrieval* operator is thus designed to retrieve the marginal probabilities, distributions, and conditional probabilities from PCDM+ that are related to a particular case. To achieve this, multiple methods were created:

1. Method getObservedValue(Study Variable sv, Patient p) returns the observed value of a study variable sv from patient p. Study Variable is a class in PCDM-Literature and it is a superclass of Behavior, Property, and Intervention entities.
2. Method getDistribution(Study Variable sv, Population pop) returns the marginal probability and/or distribution of study variable sv from population pop.
3. Method getProbability(Study Variable sv1, Study Variable sv2) returns the conditional probabilities between these two variables, sv1 and sv2 (e.g., *aneurysm size* and *aneurysm rupture*).
4. Method getProbability (Study Variable sv1, Study Variable sv2, Value v) returns the Prob(sv1|sv2=v).

As an example, when a patient has an aneurysm 6 mm in size, the returned marginal probability of aneurysm size is obtained, as shown in Figure 4.1, while Figure 4.2 shows the returned conditional probability of rupture given aneurysm size.

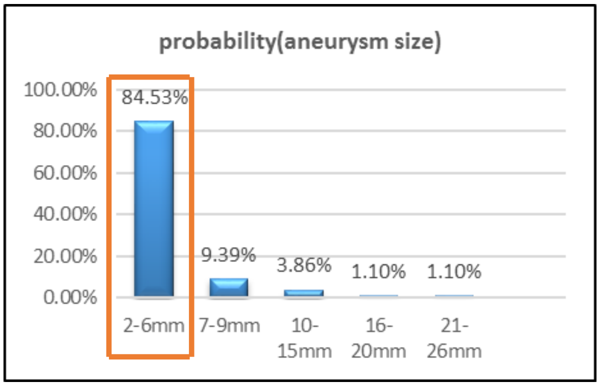


Figure 4.1 Retrieved distribution of aneurysm size

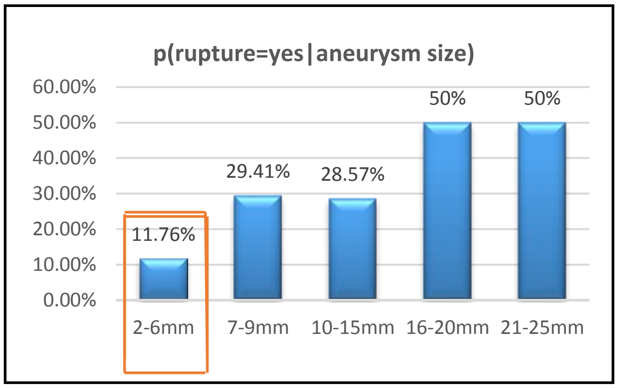


Figure 4.2 Retrieved probability of aneurysm rupture conditioned on aneurysm size

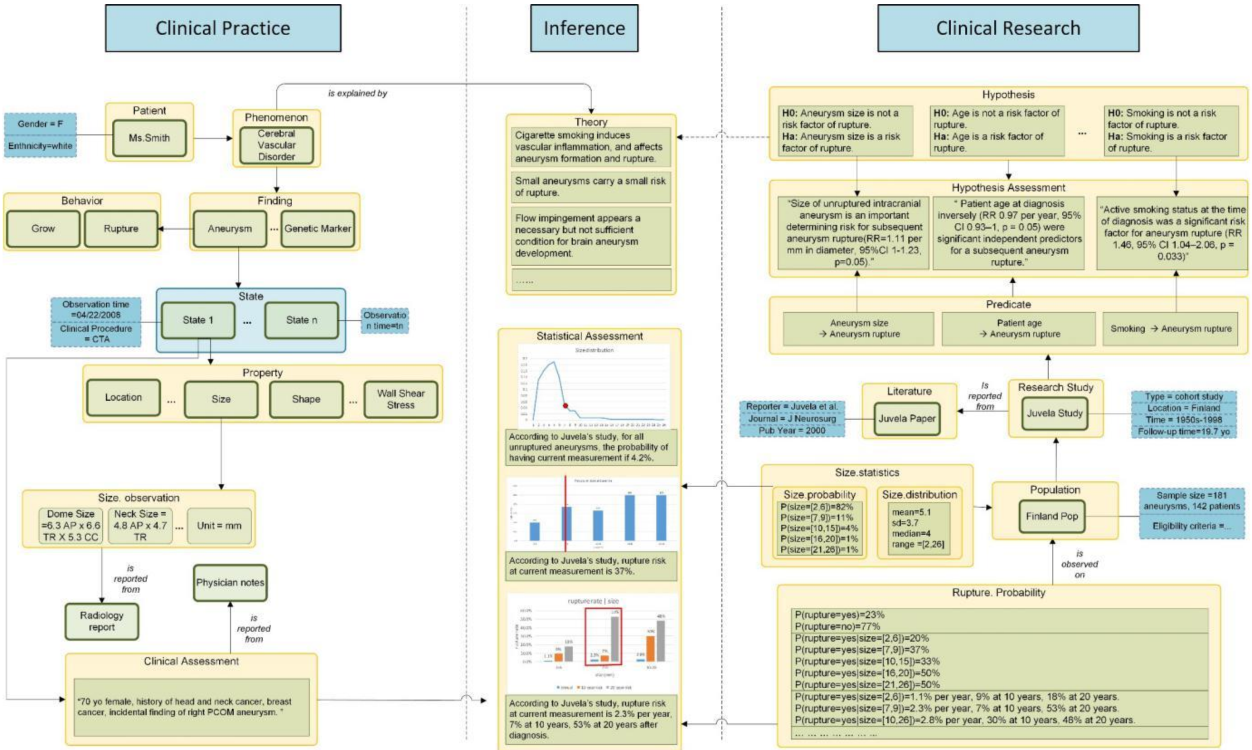

Clinical Practice
Inference
Clinical Research
is explained by
Gender = F
Enthnicity=white
Patient
Ms.Smith
Phenomenon
Cerebral Vascular Disorder
Behavior
Grow
Rupture
Finding
Aneurysm
...
Genetic Marker
State
Observation time =04/22/2008
Clinical Procedure = CTA
State 1
...
State n
Observation time=tn
Property
Location
...
Size
Shape
...
Wall Shear Stress
Size. observation
Dome Size =6.3 AP x 6.6 TR X 5.3 CC
Neck Size = 4.8 AP x 4.7 TR
...
Unit = mm
is reported from
Radiology report
Physician notes
is reported from
Clinical Assessment
"70 yo female, history of head and neck cancer, breast cancer, incidental finding of right PCOM aneurysm. "
Theory
Cigarette smoking induces vascular inflammation, and affects aneurysm formation and rupture.
Small aneurysms carry a small risk of rupture.
Flow impingement appears a necessary but not sufficient condition for brain aneurysm development.
......
Statistical Assessment
According to Juvela's study, for all unruptured aneurysms, the probability of having current measurement if 4.2%.
According to Juvela's study, rupture risk at current measurement is 37%.
rupturerate | size
According to Juvela's study, rupture risk at current measurement is 2.3% per year, 7% at 10 years, 53% at 20 years after diagnosis.
Hypothesis
H0: Aneurysm size is not a risk factor of rupture.
Ha: Aneurysm size is a risk factor of rupture.
H0: Age is not a risk factor of rupture.
Ha: Age is a risk factor of rupture.
...
H0: Smoking is not a risk factor of rupture.
Ha: Smoking is a risk factor of rupture.
Hypothesis Assessment
"Size of unruptured intracranial aneurysm is an important determining risk for subsequent aneurysm rupture(RR=1.11 per mm in diameter, 95%CI 1-1.23, p=0.05)."
" Patient age at diagnosis inversely (RR 0.97 per year, 95% CI 0.93–1, p = 0.05) were significant independent predictors for a subsequent aneurysm rupture."
"Active smoking status at the time of diagnosis was a significant risk factor for aneurysm rupture (RR 1.46, 95% CI 1.04–2.06, p = 0.033)"
Predicate
Aneurysm size → Aneurysm rupture
Patient age → Aneurysm rupture
Smoking → Aneurysm rupture
Reporter = Juvela et al.
Journal = J Neurosurg
Pub Year = 2000
Literature
Juvela Paper
Is reported from
Research Study
Juvela Study
Type = cohort study
Location = Finland
Time = 1950s-1998
Follow-up time=19.7 yo
Size.statistics
Size.probability
P(size=[2,6])=82%
P(size=[7,9])=11%
P(size=[10,15])=4%
P(size=[16,20])=1%
P(size=[21,26])=1%
Size.distribution
mean=5.1
sd=3.7
median=4
range =[2,26]
Population
Finland Pop
Sample size =181 aneurysms, 142 patients
Eligibility criteria =...
is observed on
Rupture. Probability
P(rupture=yes)=23%
P(rupture=no)=77%
P(rupture=yes|size=[2,6])=20%
P(rupture=yes|size=[7,9])=37%
P(rupture=yes|size=[10,15])=33%
P(rupture=yes|size=[16,20])=50%
P(rupture=yes|size=[21,26])=50%
P(rupture=yes|size=[2,6])=1.1% per year, 9% at 10 years, 18% at 20 years.
P(rupture=yes|size=[7,9])=2.3% per year, 7% at 10 years, 53% at 20 years.
P(rupture=yes|size=[10,26])=2.8% per year, 30% at 10 years, 48% at 20 years.
... ...

Figure 4.3 Evidence-based Medicine based on PCDM+

Using these three operators, the evidence encoded in PCDM+ can be translated into elements needed to answer physicians' queries relating to specific patients. Figure 4.3 depicts the evidence-based medicine process after patient cases and published literature are mapped into PCDM+. Importantly, the Inference Layer is where the operators are employed. A systematic demonstration of those operators is given in Section 4.3.

4.2 Operators to Build BBNs

BBNs are commonly used for prediction tasks. In the domain of intracranial aneurysms, when a patient is found with an unruptured aneurysm, it is crucial to make a prediction on the rupture risk. The example below outlines how a BBN is created and instantiated to make predictions on aneurysm rupture based on PCDM+. Three sets of operators were designed to translate knowledge from PCDM+ into BBN elements, including the variables, topology, and conditional probabilities needed to instantiate a graphical model.

4.2.1 Variable Selection and Discretization

When constructing a BBN, it is first necessary to select variables and define their possible states. PCDM+ entities and attributes provide a pool for variable selection. I consider two issues when defining inclusion criteria in the variable selection step: (1) whether a variable has a relation with the target variable; and (2) whether the data for the variable and the relation is available and sufficient to perform parameter estimation (i.e., the conditional probabilities). The operator for variable selection starts with the search for the Markov blanket (MB) of a target variable T in PCDM+ and gradually expands the search space to other variables determined by MB(T). Next, each variable is examined on its "missingness" in the data sources. Here, "missingness" refers to the proportion of data absent for a certain variable within a population. When a variable exceeds a predefined level of missingness (e.g., variables that have few or no recorded observations), semantically related variables can be identified using PCDM+ (e.g., super- or subclasses can be considered). The steps for variable selection are given below:

1. Define the target variable T and search for a corresponding attribute A or entity X in PCDM+. Set A as a decision node.
    1.1. The user defines a target variable T whose state will be predicted based on observations given for the other evidence variables.
    1.2. The user queries PCDM+ to search for entity or attribute which corresponds to T.

1.3a If T does not exist in the PCDM+, the user can create a new entity or attribute of an existing entity. By assigning the relationship with other existing entities, some of the attributes or arcs can be transferred from existing entities to the new one;

1.3b If T exists in PCDM+ as an entity X, select one of its attributes, A, to form a target variable in the BBN. Set A as a decision node and the process continues at step (2); or

1.3c If T exists in PCDM+ as an attribute A, set A serves as a decision node. Find the entity it belongs to, e.g., X, and go to step (2).

2. Find the Markov blanket of A and X in PCDM+ and extend it through *is-a*, *has-a,* and *part-of* relationships to get a superset.
   - 2.1. Place any entity that has a relationship with entity X into an entity set Y; assign any attribute that shares an arc with A to the attribute set B. Find all the elements of Y and B.
   - 2.2. Examine the dependency strength intuitively. If the relation is weak (e.g., with insignificant p-value, or the number of studies revealing this relation is less than a predefined threshold), a substitute entity or attribute in PCDM+ is recommended (e.g., a super- and/or subclass) to continue to the following step.
   - 2.3. Collapse the entities at different levels to allow different views of the desired Markov blanket.
3. Examine the missingness of each element in the extended Markov blanket.
   - 3.1. For each attribute B_i belonging to **B** (i=0, 1, 2, ... n), examine its missingness by retrieving its Local attribute Distribution (LtD), which encodes the distribution of the observations of this attribute in a sampled population.
   - 3.2. Calculate an integrated LtD using designed function with all the recorded LtD of B_i from multiple sources. Enable the user to select data sources to perform this step.
   - 3.3. If the missing data can be ignored (e.g., below some predefined threshold, such as 90%), B_i becomes a BBN node; otherwise, consider omitting B_i or finding a substitute entity or attribute to form a BBN node, e.g., another attribute of the same entity, or if it is an aggregated entity, use one of its components.

Functions designed to facilitate these steps include:

- Methods for collecting and returning the entities and attributes that have relationships or arcs with a specific attribute. Constraints are allowed to filter the returned results (e.g., at least two studies indicated the relation).
- Methods used to calculate the missingness percentage of a selected attribute on each population and/or data source.

An example of variable selection. Figure 4.4 demonstrates how PCDM+ entities and attributes (left) are mapped to BBN nodes (right). First, we find the *rupture* in PCDM+ and set it as a decision node; we then find that *smoking history*, *size*, *location*, and *age* have associated arcs. We examine the missingness of these attributes and entities in the observational clinical data and find that "*pack-years*" is seldom reported in smoking history, but "*smoking status"* is reported for 92% of the patients. Therefore, "*smoking status*" is mapped into the BBN as a node. Next, we examine the context of the *size* entity among journal articles and notice that 96% of articles refer to "*the longest diameter of the aneurysm*" when defining "*size*," thus a node called "*maximal diameter*" is defined in the BBN. By applying the function Max(height, width, depth), the data pertaining to "*size*" in PCDM+ is translated into a "*maximal diameter*" in the BBN.

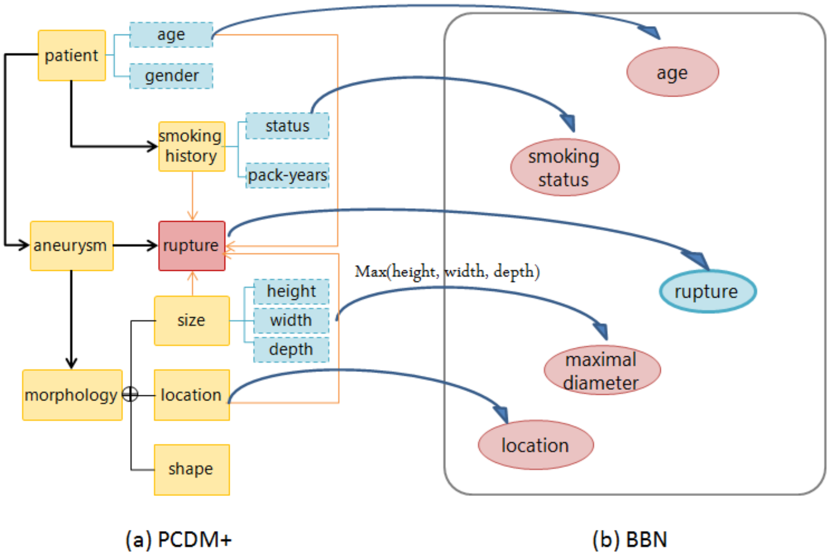


Figure 4.4 Mapping PCDM+ entities and attributes to BBN nodes

In considering discretization, a finite number of states for each BBN node needs to be defined. The goal of discretization is to simplify parameterization (i.e., keep the number of discrete states as small as possible) while minimizing information loss (i.e., approximate the original distribution as much as possible). Methods designed to facilitate the discretization process are:

- Merging and splitting operators that can be applied to merge/split states within the studied population to check the distribution of observational data for categorical variables.
- A function for selecting the optimal discretization among available strategies in PCDM+. Allow users to select the decision criteria (i.e., based on mutual information with decision node only or with all related nodes).

In this work, the following two variable types are considered:

1. Categorical variables. For categorical variables (e.g., "*smoking status*" and "*aneurysm location*"), PCDM+ has already defined standardized categories for each. The problem is that there may be considerable number of missing data instances in certain categories (e.g., aneurysms are seldom found in the basilar artery but commonly develop in anterior/posterior communicating arteries). In this case, the merging operator is used to merge the sparse states properly to simplify parameterization.
2. Continuous variables. For continuous variables, PCDM+ records discretized states over different studies and maintains a list of discretization strategies with the operator to select an optimal strategy. Local and global discretization methods are encoded in this operator for the user to select. The local discretization method is based on minimizing entropy; the global discretization method discretizes each variable to maximize its mutual information with respect to all directly related variables.

An example of discretization. We define the states of the variables selected in the last step: *age, smoking status, maximal diameter, location,* and *rupture*. In PCDM+, *rupture* is a decision node and the probability of a rupture is a function of time P(rupture)=f(t). The possible states of *rupture* at a given time are either "yes" or "no." Different possible discretization strategies of *age* are stored in PCDM+ from the indexed articles. For instance, Juvela [2000] discretized age into <30, 31-40, 41-50, and 51-60 groups while UCAS Japan [2012] used only two categories (<70, >=70). The unit of *age* is years old. We discretize the age in UCLA aneurysm dataset and compare the entropy of age after using these two different partitioning strategies. Classification (<70,>=70) is selected as a better strategy as it leads to a smaller entropy in the age variable, where entropy is calculated using the following equation:

$$Entropy(age) = -\sum_{i=1}^{m} p_i \, log_2 p_i,$$

where p_i refers to the percentage of people within the population whose age belongs to the discretized state i when age is discretized into several states (e.g., <30, 30-40, 41-50, 51-60). The defined states of each variable in the BBN are shown in Figure 4.5.

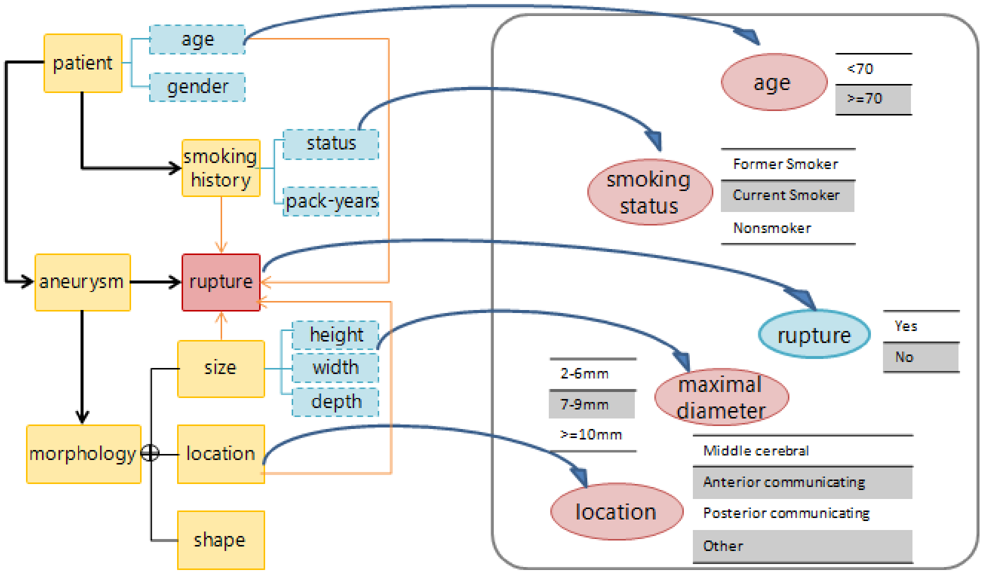


Figure 4.5 Getting discrete states for each variable from PCDM+.

4.2.2 Topology Specification

After selecting the variables and defining the possible states, the next step is to specify the topology (i.e., to decide the (in)dependencies among variables). With PCDM+, topology specification becomes a process of selecting and mapping the relationships and arcs from PCDM+ to edges in BBN. However, a mismatch exists between PCDM+ and BBNs: edges in a BBN indicate dependency, while relationships and arcs in PCDM+ entail a semantic meaning. Thus, to make this transition successful, I focus on: (1) specifying which relationships and arcs represent dependency relations; (2) assessing the existence of the dependency and independency; and (3) determining the level of detail the network should have.

Relation examination. In PCDM+, two constructs represent a relation: relationships and arcs. The former represents relations among entities, while the latter captures relations among attributes. Most relationships are predefined and a new relationship may be added when new knowledge is entered into PCDM+ (e.g., a new relationship may be adopted by mapping an external ontology to PCDM+). Arcs are instantiated when evidence from scientific literature is mapped into PCDM+. In topology specification, I examine each relationship and arc to determine if it is a dependency relation. For example, relationships among entities such as *effect, cause,* and *treat* indicate dependency, while *measure, has source,* and *analyze* do not. Arcs among attributes are named with the reported relation type and not all of them are dependent relations.

Relation assessment. I assess the dependency between any attributes of interest (e.g., the decision node T and other attributes and entities from PCDM+).

- Dependency. Methods that can be adopted to assess dependency include: reassessing the significance of a trial result using Bayesian statistics (e.g., Bayes factor [Dienes, 2011; Goodman, 1999]); and integrating multiple arcs into one to assess the strength of evidence and determine if a dependency exists. The examination includes those attributes that have conflicting evidence reported. The steps performed when integrating multiple relations among two attributes in PCDM+, A and B, to determine the relation dependency between them, are outlined as follows:
    1. Translate frequentist statistics reported in each article that reports a relation between A and B, such as a p-value and confidence interval (CI), to Bayes factor (BF) by adopting the method proposed by Goodman [1999]. Store the BF in the Local arc Distribution (LrD). If conflicting arcs exist, go to step (2); otherwise, continue at step (3).
    2. Use a Bayesian approach to integrate the Bayes factors from multiple journals into a single BF. The method is adopted from meta-analysis. While meta-analysis usually takes a hazard ratio (HR) as an input parameter, here, BF serves as an input.
    3. Set a threshold or a range for BF, and check if the integrated BF is above the threshold or within bounds to determine the dependency.

4. If the dependency is confirmed, draw the corresponding edge in BBN; if not, do not draw the edge.

- Independency. While conditional dependencies are recorded in PCDM+ from different sources as arcs and conflicting arcs between two attributes can be integrated using the abovementioned method, the independency can be inferred from the structure in PCDM+. I use the concept of d-separation from BBNs to examine the independencies among attributes in PCDM+ (examples of d-separation are given in Chapter 2). PCDM+ records the restrictions of certain conditional probabilities. Thus, in topology specification, the restrictions are taken into consideration. An example is given in Figure 4.6, where B is a deterministic function of C and D (i.e., if C and D are observed, the state of B is determined). In this case, observing C and D can infer the independency of A and T. This is translated into a constraint to specify the topology.

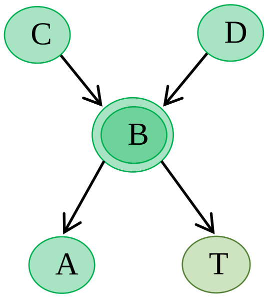


Figure 4.6 An example of independency between A and T with added restriction among the attributes.

Topology adjustment. After finding the (in)dependent relations with the decision nodes, the topology can be adjusted by considering: (1) missing data and (2) desired details. As missing data was discussed in Chapter 4.2.1, only the latter issue is discussed here.

The following methods designed to facilitate the topology specification process:

- A function to retrieve LrD for a given attribute pairs (A, B).
- A function to translate P value to BF for each LrD instance and store it.
- A function to integrate multiple BFs to an integrated BF.

- A function to pass a defined threshold or range of BF to make the final judgment of dependency between A and B.

PCDM+ is designed using an object-oriented approach. Therefore, PCDM+ is able to facilitate construction of an object-oriented BBN. In an object-oriented BBN, nodes belonging to the same class or at the same level in the hierarchy can be aggregated and collapsed to simplify the computation. This can be achieved through *is-a* and *part-of* relationships in PCDM+, as more literature-based evidence is included into PCDM+.

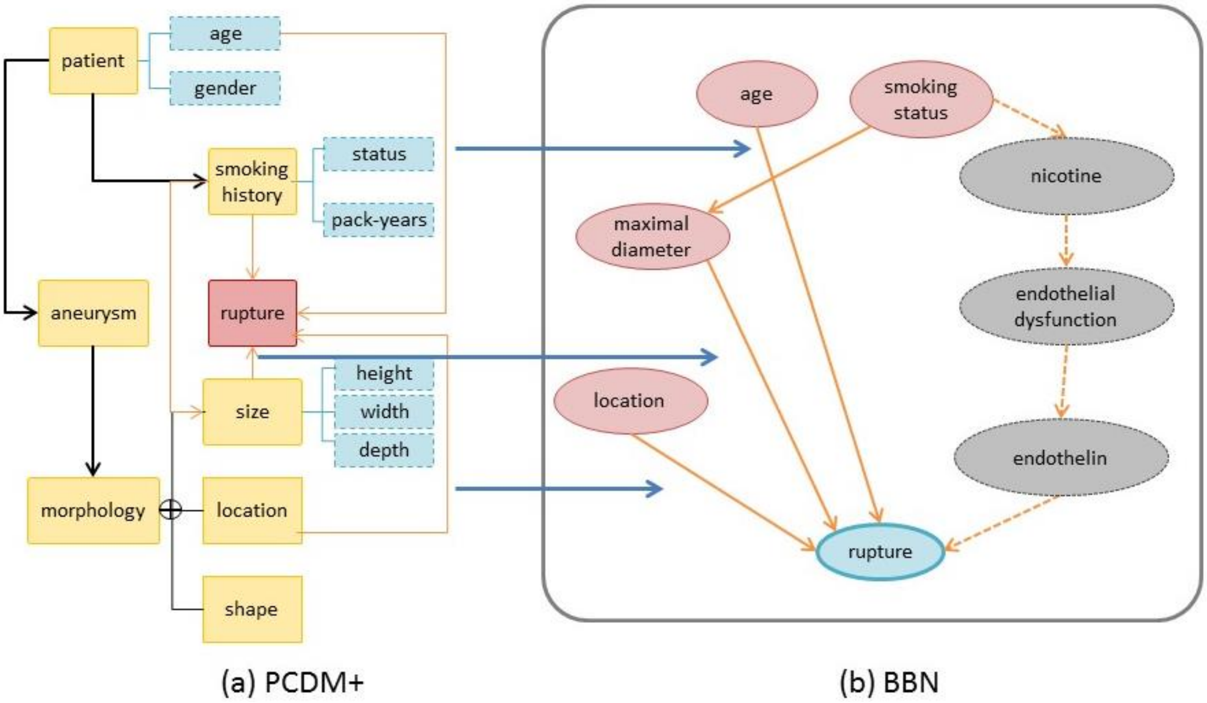


Figure 4.7 Mapping PCDM+ arcs to BBN edges

An example of topology specification. Continuing with the running use case, we first examine the dependencies between rupture and other entities and attributes in PCDM+. The dependency between rupture and risk factors age, smoking history, and size is confirmed. The independency among age, smoking history and size is reassessed and a correlation between smoking history and aneurysm size is found in one journal article; hence, an edge from *smoking status* to *maximal diameter* is drawn. As *smoking status* and *rupture* is d-separated if the *maximal diameter* is observed, we change the edge from *smoking status* to *rupture* into a dashed line to indicate conditional independence. Moreover, with more evidence from literature entering into PCDM+, the dependency between *smoking status* and *rupture* is further explored on a molecular level (as is shown by the dash lines). Figure 4.7 provides an illustration of this translation.

4.2.3 Parameter Estimation

PCDM+ not only stores individual data that instantiates each attribute, it also contains population-level statistics from LrD and LtD classes, including marginal probabilities, such as P(age<55), and conditional probabilities, like P(age<55 | rupture=yes), from different studies. The challenges with parameter estimation from PCDM+ are threefold:

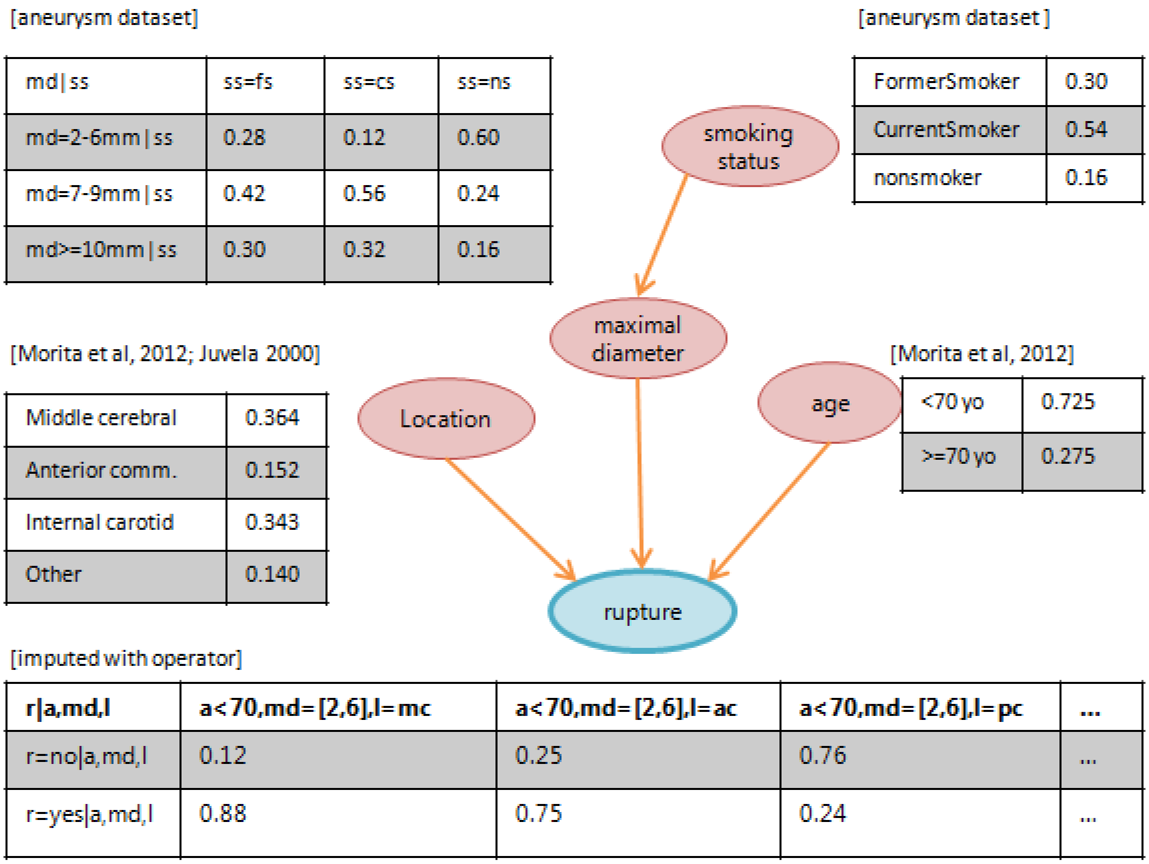
1. At the individual patient level, any missing data needs to be imputed to compute an accurate conditional probability.
2. At the population level, the statistics from sources cannot be directly used to populate the conditional probability tables and the desired joint conditional probabilities need to be computed based on partial statistics (e.g., to compute P(A|B,C) given P(A|B) and P(A|C)).
3. After obtaining the conditional probabilities from observational clinical data and scientific literature, it is necessary to determine how they should be combined and how the system can actively learn the parameter when new data sources are mapped to PCDM+.

The methods developed to address these three issues are described below.

Missing data. As missing data is a major problem when dealing with a large amount of clinical data, I focus on creating an operator to impute missing data with the information encoded in PCDM+ to facilitate parameter estimation. PCDM+ has two features that can be employed to impute missing data: (1) it captures the context of findings at the individual level, which can be used to examine missingness; and (2) it stores domain-dependent constraints and conditional relationships that can be used to compute missing data. While the current algorithms developed to deal with missing data neither ignore the context of findings nor take into consideration domain-dependent constraints, the developed operator can improve current approaches (e.g., parameter estimation based on maximum entropy).

After imputing the missing values, a Bayesian approach is used to combine the probabilities from literature and the frequencies calculated from individual-level data. The parameters can be updated when new evidence is entered into PCDM+.

Example parameter estimation. The partial statistics sourced from pertinent literature are stored in the PDCM+ LrD or LtD classes. We apply the operator with multiple imputation algorithms to address missing values and calculate the required conditional or marginal probabilities from the observational data. After using the probabilities from the literature as prior knowledge, we update the parameters by incorporating the probabilities from observational data to obtain posterior probabilities. Each parameter is updated independently using the described Bayesian approach. Figure 4.8 shows the estimated parameters and their sources. The topology is obtained from the last step described in Section 4.2.2.



[aneurysm dataset]

md| ss	ss=fs	ss=cs	ss=ns
md=2-6mm|ss	0.28	0.12	0.60
md=7-9mm|ss	0.42	0.56	0.24
md>=10mm|ss	0.30	0.32	0.16

[aneurysm dataset]

FormerSmoker	0.30
CurrentSmoker	0.54
nonsmoker	0.16

[Morita et al, 2012; Juvela 2000]

Middle cerebral	0.364
Anterior comm.	0.152
Internal carotid	0.343
Other	0.140

[Morita et al, 2012]

<70 yo	0.725
>=70 yo	0.275

[imputed with operator]

r|a,md,l	**a<70,md=[2,6],l=mc**	**a<70,md=[2,6],l=ac**	**a<70,md=[2,6],l=pc**	**...**
r=no|a,md,l	0.12	0.25	0.76	...
r=yes|a,md,l	0.88	0.75	0.24	...

Figure 4.8 Estimating BBN parameters from PCDM+ Local Distribution classes

4.3 Use Cases

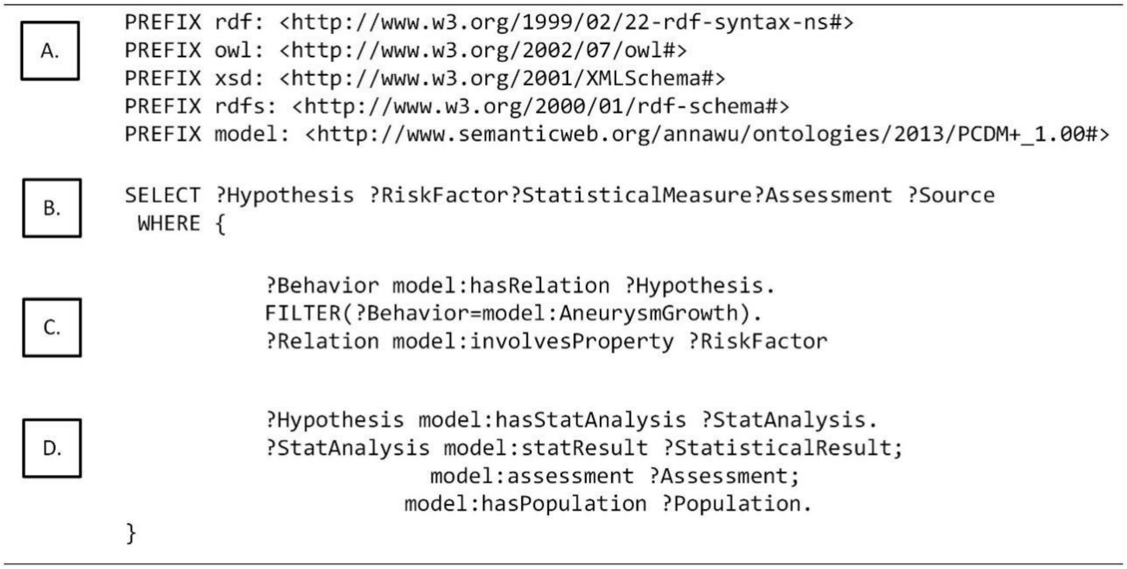
To provide an illustration and pilot validation of the operators designed as a part of this study, I performed queries a PCDM+ that was instantiated with Patient Case 1 and 2, and a set of 22 papers for Patient Case 1 (see Chapter 3 for details on paper selection and paper mapping process). For convenience, I restate the Patient Case 1 here:

Patient Case 1. "*A 70-year-old white woman with an incidental finding of right posterior communicating artery aneurysm came to our institution for consultation. She has a medical history of head and neck cancer. She was a smoker from age 20 to 25. She denies any alcohol consumption or recreational drug use. Her blood pressure is 106/65. She has a family history of stroke and hypertension. The aneurysm parameters are 6.3 mm AP × 6.6 mm TR × 5.3 mm CC according to the CT angiogram.*"

Clinical queries relating to this patient case that I aimed to facilitate answering are:

1. Patient-population matching. Given the patient characteristics, which populations are the ones this patient is eligible for so that the evidence acquired from these populations can be appraised to this patient?
2. Disease etiology and behavior. What are the risk factors for aneurysm formation, growth, and rupture? How likely will this patient's aneurysm rupture?
3. Treatment planning. Which treatment is better for this patient, endovascular coiling (EC) or surgical clipping (SC)? What are prognostic variables that may indicate whether a patient will have improved or worsening neurologic outcomes?

For Patient Case 1, PCDM+ was expressed in RDF/OWL (Resource Description Framework/Web Ontology Language). Queries were formulated using SPARQL and were executed against the model using the Jena library, a Java-based implementation of an OWL-reasoner. Figure 4.9 depicts an example SPARQL query. The query starts with the definition of the prefix used in the SPARQL code (Fig 4.9A), upon which a "SELECT…WHERE…" clause is used to retrieve the associated data elements (Figure 4.9B). This sample code is used to retrieve the risk factors of aneurysm growth (Figure 4.9C) and all the observations of those relations (Figure 4.9D) across the eligible papers.



```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX model: <http://www.semanticweb.org/annawu/ontologies/2013/PCDM+_1.00#>

SELECT ?Hypothesis ?RiskFactor?StatisticalMeasure?Assessment ?Source
 WHERE {

           ?Behavior model:hasRelation ?Hypothesis.
           FILTER(?Behavior=model:AneurysmGrowth).
           ?Relation model:involvesProperty ?RiskFactor


           ?Hypothesis model:hasStatAnalysis ?StatAnalysis.
           ?StatAnalysis model:statResult ?StatisticalResult;
                        model:assessment ?Assessment;
                      model:hasPopulation ?Population.
}
```

Figure 4.9 SPARQL query sample code to retrieve reported relations involves aneurysm growth. A. The Prefix defines the abbreviation used in the SPARQL code; B. a "SELECT…WHERE…" clause is used to retrieve the associated data elements. C. Risk factors of aneurysm growth are retrieved; D. All the observations of those relations are retrieved across the eligible papers.

4.3.1 Patient-Population Matching

Using the *patient-population matching* operator, 11 of the 22 papers were returned as relevant to the patient's specific case, including papers on aneurysm etiology, growth, rupture, and treatment (Table 4.7).

Table 4.7 Relevant papers retrieved from PCDM+

Paper ID	First Author	Title
004	Xiang, J.	Hemodynamic-morphologic discriminants for intracranial aneurysm rupture.
005	Baharoglu, M.I.	Identification of a dichotomy in morphological predictors of rupture status between sidewall- and bifurcation-type intracranial aneurysms.
006	Qian, Y.	Risk analysis of unruptured aneurysms using computational fluid dynamics technology: preliminary results.
010	Chien, A.	Enlargement of small, asymptomatic, unruptured intracranial aneurysms in patients with no history of subarachnoid hemorrhage: the different factors related to the growth of single and multiple aneurysms.
011	So, T.Y.	Risk of growth in unruptured intracranial aneurysms: a retrospective analysis
012	Chmayssani, M.	Relationship of growth to aneurysm rupture in asymptomatic aneurysms ≤ 7 mm: a systematic analysis of the literature.
013	Ishibashi, T.	Unruptured intracranial aneurysms: Incidence of rupture and risk factors.
014	Wiebers,D.O.	Unruptured intracranial aneurysms: natural history, clinical outcome, and risks of surgical and endovascular treatment.
016	Vlak, M.H.M.	Independent risk factors for intracranial aneurysms and their joint effect: a case-control study.
018	Villablanca, J.P.	Natural history of asymptomatic unruptured cerebral aneurysms evaluated at CT angiography: growth and rupture incidence and correlation with epidemiologic risk factors.
022	Juvela, S.	Natural history of unruptured intracranial aneurysms: probability of and risk factors for aneurysm rupture.

The other papers were excluded for several reasons, including (1) the patient did not satisfy the eligibility criteria of the study (e.g., a study aimed to study middle artery aneurysms, but the patient has a posterior communicating artery aneurysm); (2) the patient had missing observations for features mentioned in the study eligibility criteria (e.g., “no vasospasm”); (3) the treatment comparison studies that have “receiving a treatment” as an eligibility criterion were excluded as the patient had not yet received treatment; and (4) reviews that do not provide statistics were excluded.

4.3.2 Aneurysm Etiology, Growth, and Rupture

Using SPARQL queries to execute the *relation extraction* operator, all the relations involving aneurysm formation, aneurysm growth, and aneurysm rupture were retrieved, respectively. For each retrieved relation, statistical analyses associated with each relation, including the statistical results, statistical assessment, source, and population, were also retrieved. Table 4.8, 4.9, and 4.10 show the risk factors retrieved from PCDM+ for aneurysm formation, growth, and rupture, respectively. “NA” in the significance column indicates that no statistical measure was stated in the paper. The context column provides information on the definition of variable, states of categorized variables, and type of statistical analysis used.

Table 4.8 Some risk factors for aneurysm formation and associated context

Risk factor examined	Statistical result	Statistical assessment
Smoking	OR=3.0 (95%CI =2.0-4.5)	Significant
Regular Exercise	OR=0.6 (95%CI =0.3-0.9)	Negatively significant
Hypertension	OR=2.9 (95%CI =1.9-4.6)	Significant
Diabetes	OR=0.9 (95%CI =0.4-2.1)	Not significant
Family History of Stroke	OR=1.6 (95%CI =1.0-2.5)	Significant
Gender	Not reported	Significant
Hypercholesterolema	OR=0.5 (95%CI =0.3,0.9)	Negatively significant

Table 4.9 Risk factors for aneurysm growth examined in the literature

Risk factor examined	Significance	Paper ID	Context
Age	No	011	
Alcohol Use	Yes	011	
Aneurysm Location	No	010	Location = posterior circulation vs. non-posterior circulation; growth = single aneurysm growth (yes/no)
	Yes	010	Location = posterior circulation vs. non-posterior circulation; growth = multiple aneurysm growth (yes/no)
	No	011	

	No	018	
Aneurysm Multiplicity	No	010	
	No	018	
Aneurysm Shape	Yes	018	
Aneurysm Size	No	011	size>10 mm vs. size<=10 mm
	No	012	average initial size
	Yes	018	initial size
Coronary Artery Disease	No	010	
Diabetes	No	010	
Family history of Aneurysm	No	010	
	No	011	
Gender	No	006	
	No	010	
History of SAH	No	011	
History of Stroke	Yes	010	single aneurysm growth
	No	010	multiple aneurysm growth
History of TIA	No	010	single aneurysm growth
	Yes	010	multiple aneurysm growth
Hypertension	No	010	
	No	011	
Mural Calcification	No	018	
Number of Aneurysms	No	010	single aneurysm growth
	Yes	010	multiple aneurysm growth
Smoking History	No	010	current or previous cigarette smoking
	No	010	current or previous cigarette smoking
	No	011	
	Yes	018	
	NA	012	
	Yes	022	smoking status at the time of diagnosis
	Yes	022	smoking as a time-dependent covariate
Thrombus	No	018	

Table 4.10 A detailed list of risk factors for aneurysm rupture reported from literature.

Risk factor examined	Significance	Paper ID	Context
Age	No	012	age at the time of diagnosis
	No	014	
	No	022	age = (31-40 yrs) vs. (<30 yrs)
	No	022	age = (41-50 yrs) vs. (<30 yrs)
	No	022	age (>51) vs. (<30 yrs)

	Yes	022	age as a continuous variable
	No	022	age = (31-40 yrs) vs. (<30 yrs)
	Yes	022	age = (41-50 yrs) vs. (<30 yrs)
	No	022	age (>51) vs. (<30 yrs)
	Yes	022	age as a continuous variable
Aneurysm Clinical Presentation	No	022	rupture = SAH; Aneurysm clinical presentation=symptomatic aneurysm vs. incidental aneurysm vs. prior SAH.
Aspect Ratio	No	004	univariate analysis
	No	004	multivariate analysis with other morphological factors
	No	004	multivariate analysis with other morphological and hemodynamic factors
	No	005	among sidewall aneurysms only; univariate analysis
	No	005	among bifurcation aneurysms only; univariate analysis
	Yes	005	all aneurysms. univariate analysis
	No	005	among sidewall aneurysms only; multivariate analysis
	No	005	among bifurcation aneurysms only; multivariate analysis
	NA	006	AR>1.6 vs. <1.6
Blood Pressure	No	022	BP at the beginning of the follow-up was used. Mean arterial blood pressure = diastolic BP + (systolic BP − diastolic BP)/3
	Yes	022	here it refers to the BP at the end of the follow-up
	No	022	mean arterial pressure after adjusted for age
Ellipticity Index	Yes	004	univariate analysis
	No	004	multivariate analysis with other morphological factors
	No	004	multivariate analysis with other morphological and hemodynamic factors
Energy Loss	Yes	006	
Flow Speed	NA	006	
Gender	No	022	rupture = SAH
	No	011	
	Yes	012	absolute diameter growth
Growth	Yes	012	growth percentage for aneurysms with and without rupture
	Yes	018	all aneurysms in this study
	Yes	018	among saccular aneurysm only
	No	012	annual growth rate
Height Width Ratio	No	005	among sidewall aneurysms only; univariate analysis
	No	005	among bifurcation aneurysms only; univariate analysis
	Yes	005	all aneurysms. univariate analysis
	Yes	005	among sidewall aneurysms only; multivariate analysis
	No	005	among bifurcation aneurysms only; multivariate analysis
History Of SAH	Yes	013	among all aneurysms in the study

	No	013	Only among the small aneurysms < 5 mm
	NA	012	
	Yes	014	rupture rate
Hypertension	NA	012	
	No	022	Here it refers to the status at the beginning of the follow-up. Hypertension is defined as a systolic pressure repeatedly greater than 160 mm Hg, diastolic pressure greater than 95 mmHg, or as the use of antihypertension medication.
Inflow Angle	Yes	005	among sidewall aneurysms only; univariate analysis.
	No	005	among bifurcation aneurysms only; univariate analysis
	Yes	005	all aneurysms; univariate analysis
	Yes	005	among sidewall aneurysms only; multivariate analysis
	No	005	among bifurcation aneurysms only; multivariate analysis
Nonsphericity Index	Yes	004	univariate analysis
	No	004	multivariate analysis with other morphological factors
	No	004	multivariate analysis with other morphological and hemodynamic factors
	No	005	among sidewall aneurysms only; univariate analysis
	No	005	among bifurcation aneurysms only; univariate analysis
	Yes	005	all aneurysms; univariate analysis
	No	005	among sidewall aneurysms only; multivariate analysis
	Yes	005	among bifurcation aneurysms only; multivariate analysis
Number Of Vortices	Yes	004	univariate analysis
	No	004	multivariate analysis with other morphological factors
	No	004	multivariate analysis with other morphological and hemodynamic factors
Oscillatory Shear Index	Yes	004	average OSI, univariate analysis
	Yes	004	average OSI, multivariate analysis with other morphological factors
	Yes	004	average OSI, multivariate analysis with other morphological and hemodynamic factors
Size Ratio	Yes	004	univariate analysis
	Yes	004	multivariate analysis with other morphological factors
	Yes	004	multivariate analysis with other morphological and hemodynamic factors
	No	005	among sidewall aneurysms only; univariate analysis
	No	005	among bifurcation aneurysms only; univariate analysis.
	Yes	005	all aneurysms; univariate analysis
	Yes	005	among sidewall aneurysms only; multivariate analysis
	No	005	among bifurcation aneurysms only; multivariate analysis
Smoking History	NA	012	
	Yes	022	at the time of diagnosis
	Yes	022	as a time-dependent covariate

Systolic BP	No	022	after adjusted for age
TNF-α	Yes	011	
Undulation Index	Yes	004	univariate analysis
	No	004	multivariate analysis with other morphological factors
	No	004	multivariate analysis with other morphological and hemodynamic factors
Wall Shear Stress	Yes	004	average WSS, univariate analysis
	Yes	004	maximum intra-aneurysmal WSS, univariate analysis
	Yes	004	low WSS area, univariate analysis
	No	004	WSS gradient, univariate analysis
	Yes	004	average WSS, multivariate analysis
	No	004	WSS gradient, multivariate analysis
	No	004	maximum intra-aneurysmal WSS, multivariate analysis
	No	004	low WSS area, multivariate analysis
	No	006	time-averaged WSS

In addition to using the *relation extraction* operator to retrieve the risk factors of aneurysm formation, growth, and rupture with associated statistical measurement, I also used *probability retrieval* operator to determine all the probabilities that can be used to estimate the rupture risk for the patient.

Table 4.11 Some conditional probabilities of rupture from PCDM+ to facilitate rupture risk estimation

Patient Characteristics	Observed Value	P(rupture = yes | Patient's observation)
Age	70 yo	15.45%
Gender	female	25%
Clinical presentation	incidental	4.42%
Aneurysm rupture	no	
Aneurysm location	right PCoA	12.34%
Aneurysm size	6.3×6.6×5.3 mm	11.76%
Aneurysm shape	saccular	9.65%
Modality	CTA	
Family history of stroke	yes	24.14%
Family history of hypertension	yes	21.45%
Medical history	head and neck cancer	
Smoking status	former-smoker	12.56%
Alcohol use	none	
Blood pressure	106/65	27.14%
Treatment	No	

(*PCoA: Posterior Communicating Artery; **CTA: Computed Tomography Angiography)

The rupture risk can also be estimated by constructing a BBN with rupture risk as targeted variable. Based on this retrieved information, a belief network was created (Figure 4.10 and 4.11) to answer predictive questions such as, “How likely is it that this aneurysm will rupture?”

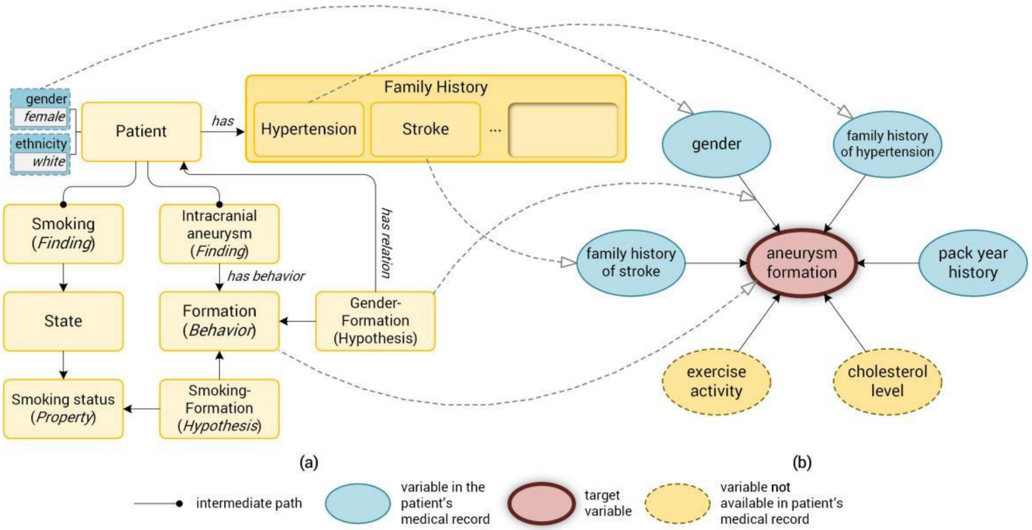


Figure 4.10 The process of creating a BBN to explain and predict aneurysm formation from PCDM+. A solid arrow in the BBN indicates a relation (either an association or a causal relation) between nodes. A dashed arrow indicates a translation from a PCDM+ entity/attribute to a BBN node or a link.

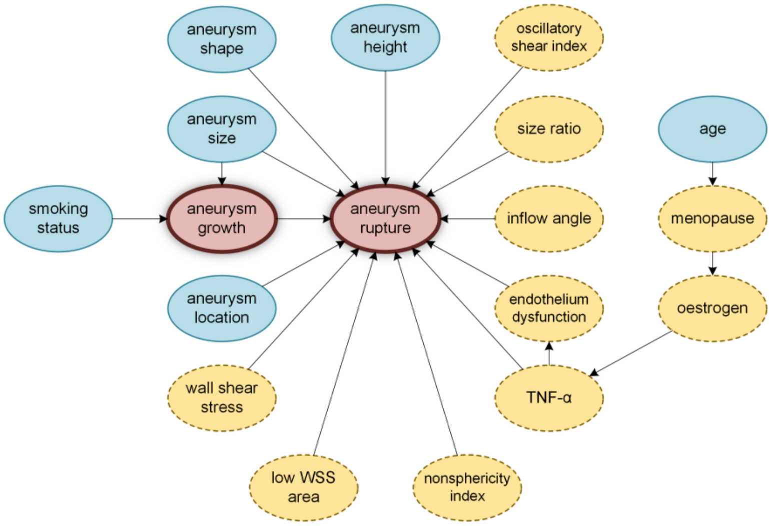


Figure 4.11 Examples of relations associated with aneurysm growth and rupture as reported in matched papers. A solid arrow indicates a relation (either an association or a casual relation) between nodes. The two nodes “aneurysm growth” and “aneurysm rupture” are the target variables of interest. Solid ovals (blue) represent concepts that are observed in the patient, while dashed ovals (yellow) denote concepts that are missing in the patient records.

Here, I demonstrate the potential application of translating PCDM+ to a BBN in order to answer clinical questions. By using the operators previously designed for BBN building (Section 4.2), variables in the networks are translated from PCDM+ entities (e.g., aneurysm formation is a Behavior entity) or attributes (e.g., gender is an attribute of Patient entity); links between variables are translated from relations in PCDM+ (e.g., the arrow from gender to aneurysm formation represents the "gender-formation" relationship in PCDM+), and probabilities captured in PCDM+ are then used to estimate the BBN parameters.

4.3.3 Treatment Planning

I retrieved all the treatment studies discussed in the 22 papers to ascertain the prognostic factors reported in these papers. Study type is recorded as an attribute of a study and the filter "study type = treatment" was used. PCDM+ retrieved five treatment studies from of the set of 22 papers, one of which was excluded as the patient did not satisfy the eligibility criteria of the paper (i.e., the authors only studied the posterior circulation aneurysms and the patient has a posterior communicating aneurysm, which is located in the anterior circulation). Using the *relation extraction* operator, a list of prognostic factors for unfavorable clinical outcomes of SC and EC reported in these papers was retrieved (see Table 4.12).

Table 4.12 Prognostic factors for unfavorable treatment outcome. SC= surgical clipping; EC= endovascular coiling.

Prognostic factor	Treatment	Statistical result	Statistical assessment
Age>70	SC	P=0.039	Negatively significant
Family history of stroke	EC	P=0.004	Significant
Aneurysm Location	SC and EC	Not reported	Not significant
Aneurysm Size	SC and EC	Not reported	Not significant
Multiplicity of aneurysm	SC	P=0.013	Significant
Diabetes	SC	P=0.027	Significant
Hypercholesterolemia	SC	P=0.00	Significant
Smoking	SC	P=0.021	Significant
	EC	P=0.016	Significant

Based on this evidence retrieved from PCDM+ and given that the patient has a family history of stroke, age is ~70, no smoking history, no history of diabetes, and an unruptured aneurysm, endovascular coiling is suggested as a better treatment than surgical clipping.

CHAPTER 5. Conclusion

This chapter summarizes the findings and contributions of this dissertation. Future directions are also presented to further contextualize the impact of this research.

5.1 Contributions

This work addresses the need for a systematic process that can be adopted to synthesize knowledge of a disease across multiple sources (e.g., medical records, published literature, clinical guidelines, expert opinions, and existing models). Knowledge synthesis can facilitate clinical decision making pertaining to individual patients. An intermediate representation was created to consolidate and standardize knowledge fragments and associated context across medical records and scientific literature, and operators were created to translate evidence from this representation to a format useful for answering clinical queries relating to specific patient cases. My research resulted in two key contributions:

1. I created an intermediate representation, called Phenomenon-Centric Data Model Plus (PCDM+), to logically consolidate and standardize knowledge fragments and associated context across medical records and scientific literature. PCDM+ comprises three components: PCDM-Clinic, PCDM-Literature, and Inference Layer. PCDM-Clinic captures individual patient-level observations, PCDM-Literature contains the population-level findings from pertinent scientific literature, and Inference Layer maintains the evidence that was filtered and translated from PCDM+ by operators to answer individual-tailored clinical questions. A requirement analysis revealed that three features are desirable when designing such a standard representation to integrate disease knowledge: (1) classes to encode key findings; (2) classes to capture associated context; and (3) semantic matching across the sources. By adapting a probabilistic entity-relationship model, PCDM+ extended PCDM to satisfy these requirements. The evaluation tasks highlight the completeness and correctness of the representation in its ability to faithfully capture information from patient records and published literature.
2. I developed operators that translate evidence from PCDM+ into aggregated knowledge elements needed to inform clinical decision making. *Patient-population matching*, *relation extraction*, and

probability retrieval operators help to translate PCDM+ into knowledge elements that can assist in answering clinical queries about a specific patient. Based on the representation of eligibility criteria in PCDM+, the patient-population matching operator employed a rule-based algorithm to assess if a patient satisfies the eligibility criteria of sampled populations in research studies. The relation extraction operator allows the users to explore the examined relations and their statistical strength to offer insights on various risk and prognostic factors, and further assist topology specification in disease modeling. The probability retrieval operator returns distributions of features (e.g., demographics, clinical presentation, morphology, treatment) that are observed in the targeted patients as well as probabilities of outcome variables (e.g., rupture, survival) conditioned on those factors. These probabilities can be utilized to estimate the risk in clinical settings, as well as provide constraints to estimate the parameters in disease modeling. In this work, additional operators were designed to facilitate BBN construction in variable selection and discretization, topology specification, and parameter estimation. Use cases demonstrated the functionality of these operators in answering clinical queries of a given patient case with an unruptured brain aneurysm.

5.2 Challenges and Future Work

This dissertation focused on the capacity of PCDM+ for capturing and operationalizing information from medical records and clinical publications. PCDM+ can be effectively used by target users (e.g., clinical translational researchers, practicing physicians) to derive new insights into how a complex disease should be managed in specific patients. I present the challenges, limitations and future work in this section.

- **Generalizability.** I demonstrated PCDM+ functionality through use cases in the domain of intracranial aneurysm but did not explore its application in other domains. However, PCDM+ provides a collection of abstract classes that should be generalizable to other disease domains by creating domain-specific instances.
- **Natural language processing.** Clinical publications are a key source in which domain knowledge is formally captured and subsequently used as evidence to support medical decisions. Publications are

written in natural language and include figures and tables that are challenging to define in a systematic and computer-aided way to appraise and apply the evidence to specific patients. PCDM+ formalizes published research findings (e.g., behaviors, relationships, and observations) in a standardized, machine interpretable manner that supports information retrieval, quality control, and ultimately clinical decision support. In this work, mapping from data sources to PCDM+ was performed manually. However, the standardized classes of PCDM+ provide a structure in which frame-based natural language processing (NLP) representations can serve as input. Future work would include the use of NLP to automate this process to make this work scalable to large corpora of published biomedical literature.

- **Context and provenance.** Often, information sourced from the literature is distilled into a few key points (e.g., individuals with incidental, < 7 mm aneurysms without previous SAH should be observed rather than undergo an intervention), losing most of the context about the study based on which the conclusion is made. Without this surrounding context, data integration for clinical decision support tools is difficult, as the applicability of a recommendation cannot be specifically matched to individuals. Therefore, an important challenge in evidence-based medicine is to define and structure the context that is associated with key findings so that evidence relevant to a specific patient can be easily found and retrieved. Relevant results can be properly interpreted and appraised to provide more targeted treatments. PCDM+ demonstrates the research effort of capturing contextual and provenance metadata from clinical records and published clinical literature, allowing multiple population findings to be extracted and restructured into forms that are applicable to specific patient cases. Based on PCDM+, the operators help achieve the goal of facilitating answering individually-tailored medicine queries. However, in the current dissertation, the contextual metadata in PCDM+ has not been fully utilized and more sophisticated operators need to be developed to achieve the potential of PCDM+ to achieve contextual EBM.
- **Probability estimation.** Physicians often need to estimate the risk of clinical outcomes (e.g., rupture, survival) and explain this information to their patients to convey sound judgment in treatment plan-

ning. A great amount of population evidence provided in the literature (e.g., probabilities, distribution, and statistics) is often overlooked in the research area of EBM. PCDM+ provides a formal structure to maintain marginal, conditional, and joint probabilities with context (e.g., the definition of that probability as a percentage of aneurysm vs. percentage of patient, follow-up period, and annually probability vs. cumulative probability). PCDM+ can, therefore, be used to generate a library of probabilities that can be referenced by physicians. However, the challenge remains in how to consolidate these fragmented probabilities from different studies with a different context in a fully systematic manner. Similarly, additional work would be required to determine how to summarize probabilities from many contexts to generate an overall statistically and clinically meaningful assessment.

- **Probabilistic model building.** Related to the prior point, as a potential extension of this work, the aim should be to improve PCDM+'s ability to create probabilistic models, like BBNs, reported the literature. While prior work has described utilizing information from literature [Antal, 2004], these applications are frequently limited to defining variables and a network topology based on information described in the papers. This work describes a way to transform a PCDM+ instantiation into a BBN. PCDM+ entities/attributes serve as variables, relations encoded in PCDM+ form the network topology, and documented probabilities define the BBN's conditional probabilities. However, combining the partial statistics captured in PCDM+ to form full conditional probability tables and an overall joint probability distribution is still an ongoing task. The mathematical and computational challenges of synthesizing probabilities to generate a joint probability may be solved with optimization algorithms such as MaxEnt, Gibbs sampling, and constraint-based optimization. However, the challenge of reusing these probabilities correctly persists, as they are derived from different sources with different contexts. For instance, the probability of rupture can represent different information across studies: some authors report an annual rupture rate, while others report the cumulative rupture rate over a follow-up period.
- **Knowledge heterogeneity.** During the process of populating PCDM+ with information, several reporting inconsistencies were encountered: (1) the use of different names for the same property across

different sources (e.g., *aneurysm size* and *maximum diameter* are synonymous in many papers); (2) the same name in different sources refers to different properties (e.g., *wall shear stress* can mean *average weighted wall shear stress* or *maximum wall shear stress*); and (3) the same continuous property may be discretized differently in different papers (e.g., *age* is discretized as <70, >=70 in one paper, but <30, 30-50, 51-70, and >70 categories were employed in another study). PCDM+ helps users manage the heterogeneity of information by standardizing the terminology across sources and assigning a preferred name to each variable at the concept level. PCDM+ also explicitly captures properties such as feature definition, discretized states, and synonyms as contextual fragments to enhance the understanding of an observed variable. Given multiple discretization mechanisms, the discretization operator designed in this work can suggest optimal discretization for continuous variables based on maximum entropy.

- **Clinical research versus clinical practice.** A large gap was noted between data that is collected during research and what is reported in clinical practice. Many variables analyzed in research were not routinely observed or collected as part of the standard of care. For ICAs*, aspect ratio*, *undulation index*, *nonsphericity index, ellipticity index,* and *size ratio* are variables often examined in studies that explore morphology as part of rupture risk assessment. Although these parameters are not directly available from clinical data, each can be calculated from clinical variables such as *aneurysm neck diameters, height, vessel angle, aneurysm volume,* and *surface area*. Hemodynamic parameters, such as *wall shear stress, flow pattern, a number of vertices,* and *oscillatory shear index* cannot be sourced from medical records unless computational fluid dynamics simulations are performed to provide estimated values. While many factors exist as to why variables in research have not been translated clinically (e.g., due to computational requirements, issues with reproducibility/standardization), PCDM+ formalizes this information in a way that is more directly sharable. This functionality also indicates that PCDM+ has another potential use to report newly identified significant risk factors and prognostic factors in clinical research and suggest collect observations of those factors in clinical settings.

- **Evidence strength assessment and evidence synthesis.** In this work, PCDM+ attempted to comprehensively represent all available evidence for a phenomenon. It also incorporates context concerning strength of evidence and is capable of providing additional guidance to physicians and statisticians. However, the model and operators do not automatically assess accuracy of certain data sources relative to others. Nor does it provide an explicit mechanism to deal with conflicting evidence. For instance, if multiple papers report differently on a given relationship (non-significant vs. significant vs. inversely significant), PCDM+ will not automatically synthesize multiple observations and suggest a conclusion of such a relation. Rather, PCDM+ provides users with the original information to make such an assessment. Based on PCDM+, separate applications could be developed in the future to synthesize the evidence and assess its strength or detect conflicting evidence.
- **Patient-population matching.** Patient-population matching in this work revealed that determining if a patient satisfied study eligibility criteria was not a necessary or sufficient condition for determining which studies are applicable to individuals. By capturing variables and their distributions in populations, PCDM+ provides a basis for a more robust probabilistic approach for performing patient-population matching in the future. Similarly, patients often have missing data in their medical records for a number of reasons. This work demonstrates a representation of eligibility criteria and a naïve mechanism that can be applied to handle missing data. However, an explicit representation of eligibility criteria capable of handling missing and inconsistent information is desired. To overcome this issue, more structured and comprehensive clinical trial eligibility criteria frameworks [Milian et al., 2012; Dameron et al., 2013] can be integrated to help address missing patient information.
- **Temporality.** Another direction for future work will be to extend PCDM+'s ability to represent changes in knowledge over time via its temporal stream constructs, thus managing evidence propagation. The stream construct has been previously reported [Bui and Taira, 2010]. A stream represents data sequence in the medical records (e.g., sequence of patient states), or the progression of evidence towards a final diagnosis in differential diagnosis. The future efforts should focus on adapting the construct to model: (1) the clinical course of a patient or population (e.g., sequence of events resulting

in a clinical outcome); (2) the experimental flow of a study (e.g., understanding the sequence of interventions given in a clinical trial); and (3) the evolution of our understanding of a given phenomenon (e.g., changes in theories or hypotheses that explain an observation).

- **Matching physicians with recommendations.** Another potential PCDM+ application is to provide a physician "profiling" system. After using PCDM+ to provide information about a patient and the recommendations from the literature, one can further imagine mining the EHR and the past performance for the specific physician to elucidate his/her outcomes pertaining to a particular procedure/treatment. For example, if a patient with certain characteristics comes to the hospital with clinical presentation of a brain aneurysm, PCDM+ can be used to retrieve relevant matching papers and provide the treatment recommendation (e.g., the coiling procedure). Because PCDM+ keeps the past patients' records, the physician's performance regarding coiling procedure will be retrieved (e.g., how many coiling procedures he/she performed and how many were successful).
- **Expressive representation language**. A suitably expressive representation language for PCDM+ is lacking. In this work, I proposed to construct the classes in PCDM+ as a PER model. However, current representations like OWL have no standard to represent probabilities and statistics. Similarly, currently available tools such as Protégé and databases do not fully achieve the functions of PCDM+. Hence, a formal standard needs to be adopted for expressing uncertainty in OWL and supporting tools to specify probabilities.

REFERENCES

Akobeng, A. K. (2005). Principles of evidence-based medicine. *Archives of Disease in Childhood, 90*(8), 837–840. Retrieved from http://doi.org/10.1136/adc.2005.071761

Aliferis, C. F., Tsamardinos, I., & Statnikov, A. (2003). HITON: A Novel Markov Blanket Algorithm for Optimal Variable Selection. *AMIA Annual Symposium Proceedings, 2003,* 21–25.

Antal, P., Fannes, G., Timmerman, D., Moreau, Y., & De Moor, B. (2004). Using literature and data to learn BBNs as clinical models of ovarian tumors. *Artif. Intell. Med., 30,* 257–281. doi:10.1016/j.artmed.2003.11.007

Balaa, Z. E., Strauss, A., & Maximini, K. (2003). Fm-ultranet: a decision support system using case-based reasoning, applied to ultrasonography. In *Workshop on CBR in the Health Sciences* (pp. 0–3).

Bauer, E., Koller, D., & Singer, Y. (1997). *Update rules for parameter estimation in Bayesian networks* (pp. 3–13). Morgan Kaufmann.

Begum, S., Ahmed, M. U., Funk, P., Xiong, N., & Folke, M. (2011). Case-Based Reasoning Systems in the Health Sciences: A Survey of Recent Trends and Developments. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 41*(4), 421–434. doi:10.1109/TSMCC.2010.2071862

Bergmann, R., Kolodner, J., & Plaza, E. (2005). Representation in case-based reasoning. *Knowl. Eng. Rev., 20*(3), 209–213. doi:10.1017/S0269888906000555

Bichindaritz, I., Kansu, E., & Sullivan, K. M. (1998). Case-Based Reasoning in CARE-PARTNER: Gathering Evidence for Evidence-Based Medical Practice. *Proceedings of the 4th European Workshop on Advances in Case-Based Reasoning, EWCBR '98* (pp. 334–345). London, UK: Springer-Verlag. Retrieved from http://dl.acm.org/citation.cfm?id=646178.758707

Boeker, M., Stenzhorn, H., Kumpf, K., Bijlenga, P., Schulz, S., & Hanser, S. (2007). The @neurIST Ontology of Intracranial Aneurysms: Providing Terminological Services for an Integrated IT Infrastructure. *AMIA Annual Symposium Proceedings, 2007,* 56–60.

Bradburn, C., & Zeleznikow, J. (1994). The application of case-based reasoning to the tasks of health care planning. In S. Wess, K.-D. Althoff, & M. Richter (Eds.), *Topics in Case-Based Reasoning, Lecture Notes in Computer Science* (Vol. 837, pp. 365–378). Berlin / Heidelberg: Springer. Retrieved from http://www.springerlink.com/content/y545u58726546104/abstract/

Broughton, R., & Rathbone, B. (2001). What Makes a Good Clinical Guideline? *Hayward Medical Communications*.

Buffart, L. M., Singh, A. S., van Loon, E. C. P., Vermeulen, H. I., Brug, J., & Chinapaw, M. J. M. (2013). Physical activity and the risk of developing lung cancer among smokers: A meta-analysis. *Journal of science and medicine in sport / Sports Medicine Australia*. doi:10.1016/j.jsams.2013.02.015

Bui, A. A. T., & Taira, R. K. (2010). Organizing Observations: Data Models. In A. A. T. Bui & R. K. Taira (Eds.), *Medical Imaging Informatics* (pp. 299–331). US: Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4419-0385-3_7

Cambria, E., & White, B. (2014). Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]. *IEEE Computational Intelligence Magazine, 9,* 48–57. doi:10.1109/MCI.2014.2307227

Cao, Y., Liu, F., Simpson, P., Antieau, L., Bennett, A., Cimino, J. J., . . . Yu, H. (2011). AskHERMES: An online question answering system for complex clinical questions. J. *Biomed. Inform., 44*, 277–288. doi:10.1016/j.jbi.2011.01.004

Chen, H., Quandt, S. A., Grzywacz, J. G., & Arcury, T. A. (2013). A Bayesian Multiple Imputation Method for Handling Longitudinal Pesticide Data with Values below the Limit of Detection. *Environmetrics*, *24*(2), 132–142. doi:10.1002/env.2193

Chen, X., Liu, Y., Røe, O. D., Qian, Y., Guo, R., Zhu, L., . . . Shu, Y. (2013). Gefitinib or erlotinib as maintenance therapy in patients with advanced stage non-small cell lung cancer: a systematic review. *PloS one*, *8*(3), e59314. doi:10.1371/journal.pone.0059314

Chmielewski, M. R., & Grzymala-busse, J. W. (1996). Global discretization of continuous attributes as preprocessing for machine learning. In *International Journal of Approximate Reasoning* (pp. 294–301).

Chou, H.-L., Chen, J.-S., & Cheng, C.-H. (2008). Global Discretization Approach Based on Minimize Entropy in Rough Sets Classification. In *Second International Symposium on Intelligent Information Technology Application, 2008. IITA '08* (Vol. 1, pp. 889–893). Presented at the Second International Symposium on Intelligent Information Technology Application, 2008. IITA '08. doi:10.1109/IITA.2008.485

Clarke, E. J., & Barton, B. A. (2000). Entropy and MDL discretization of continuous variables for Bayesian belief networks. *International Journal of Intelligent Systems*, *15*(1), 61–92. doi:10.1002/(SICI)1098-111X(200001)15:1<61::AID-INT4>3.0.CO;2-O

Cobb, B. R., Rumí, R., & Salmerón, A. (2007). Bayesian Network Models with Discrete and Continuous Variables. In P. L. Dr, J. A. G. Dr, & A. S. Dr (Eds.), A*dvances in Probabilistic Graphical Models* (pp. 81–102). Berlin /Heidelberg: Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-68996-6_4

Cooper, G. F., Aliferis, C. F., Ambrosino, R., Aronis, J., Buchanan, B. G., Caruana, R., . . . Spirtes, P. (1997). An evaluation of machine-learning methods for predicting pneumonia mortality. *Artificial intelligence in medicine*, *9*(2), 107–138.

Correa, E., & Goodacre, R. (2011). A genetic algorithm-Bayesian network approach for the analysis of metabolomics and spectroscopic data: application to the rapid identification of Bacillus spores and classification of Bacillus species. *BMC Bioinformatics*, *12*, 33. doi:10.1186/1471-2105-12-33

D'Amelio, A. M., Cassidy, A., Asomaning, K., Raji, O. Y., Duffy, S. W., Field, J. K., . . . Etzel, C. J. (2010). Comparison of discriminatory power and accuracy of three lung cancer risk models. *British Journal of Cancer*, *103*(3), 423–429. doi:10.1038/sj.bjc.6605759

Dameron, O., Besana, P., Zekri, O., Bourdé, A., Burgun, A., & Cuggia, M. (2013). OWL model of clinical trial eligibility criteria compatible with partially-known information. *Journal of Biomedical Semantics*, *4*, 17. doi:10.1186/2041-1480-4-17

Del Fiol, G., Huser, V., Strasberg, H. R., Maviglia, S. M., Curtis, C., & Cimino, J. J. (2012). Implementations of the HL7 Context-Aware Knowledge Retrieval ("Infobutton") Standard: challenges, strengths, limitations, and uptake. *J. Biomed. Inform.*, *45*, 726–735. doi:10.1016/j.jbi.2011.12.006

Dellaportas, P., Forster, J. J., & Ntzoufras, I. (2000). *Bayesian Variable Selection Using the Gibbs Sampler*.

Dienes, Z. (2011). Bayesian Versus Orthodox Statistics: Which Side Are You On? *Perspectives on Psychological Science*, *6*(3), 274–290. doi:10.1177/1745691611406920

Dimitrova, E. S., Licona, M. P. V., McGee, J., & Laubenbacher, R. (2010). Discretization of Time Series Data. *Journal of Computational Biology*, *17*(6), 853–868. doi:10.1089/cmb.2008.0023

Eekhout, I., de Boer, R. M., Twisk, J. W. R., de Vet, H. C. W., & Heymans, M. W. (2012). Missing data: a systematic review of how they are reported and handled. *Epidemiology (Cambridge, Mass.)*, *23*(5), 729–732. doi:10.1097/EDE.0b013e3182576cdb

Fayyad, U. M., & Irani, K. B. (1993). Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning, *2*(1), 1022–1027.

Friedman, N., & Goldszmidt, M. (1996). Discretizing Continuous Attributes While Learning Bayesian Networks. In *Proc. ICML* (pp. 157–165). Morgan Kaufmann.

Friedman, N., Getoor, L., Koller, D., & Pfeffer, A. (1999). Learning probabilistic relational models. In *IJCAI* (pp. 1300–1309). Springer-Verlag.

Gagnier, J. J., Kienle, G., Altman, D. G., Moher, D., Sox, H., & Riley, D. (2013). The CARE guidelines: consensus-based clinical case reporting guideline development. *Journal of Medical Case Reports*, *7*, 223. doi:10.1186/1752-1947-7-223

Garcia-Gathright, J. I., Matiasz, N. J., Garon, E. B., Aberle, D. R., Taira, R. K., & Bui, A. A. T. (2016). Toward patient-tailored summarization of lung cancer literature (pp. 449–452). IEEE. Retrieved from http://doi.org/10.1109/BHI.2016.7455931

George, E. I., & McCulloch, R. E. (1993). Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association*, *88*(423), 881–889. doi:10.1080/01621459.1993.10476353

Gevaert, O., De Moor, B., & Timmerman, D. (2007). OC157: Optimizing variable selection and cost using a genetic algorithm for modeling adnexal masses with Bayesian networks. *Ultrasound in Obstetrics and Gynecology*, *30*(4), 415–415. doi:10.1002/uog.4263

Goodman, S. N. (1999). Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of internal medicine*, *130*(12), 1005–1013.

Guyatt, G. H., Sackett, D. L., & Cook, D. J. (1994). Users' guides to the medical literature. II. How to use an article about therapy or prevention. B. What were the results and will they help me in caring for my patients? Evidence-Based Medicine Working Group. *JAMA*, *271*, 59–63.

Handels, H., Roß, T., Kreusch, J., Wolff, H. H., & Pöppl, S. J. (1999). Feature selection for optimized skin tumor recognition using genetic algorithms. *Artificial Intelligence in Medicine*, *16*(3), 283–297. doi:10.1016/S0933-3657(99)00005-6

Harrison, R. F., & Kennedy, R. L. (2005). Artificial neural network models for prediction of acute coronary syndromes using clinical data from the time of presentation. *Annals of emergency medicine*, *46*(5), 431–439. doi:10.1016/j.annemergmed.2004.09.012

Heckerman, D. (2008). A Tutorial on Learning with Bayesian Networks. In P. D. E. Holmes & P. L. C. Jain (Eds.), *Innovations in Bayesian Networks* (pp. 33–82). Berlin / Heidelberg: Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-85066-3_3

Heckerman, D., & Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. In *Machine Learning* (pp. 20–197).

Heckerman, D., Meek, C., & Koller, D. (2007). *Probabilistic Entity-Relationship Models, PRMs, and Plate*.

Herrero, J., Díaz-Uriarte, R., & Dopazo, J. (2003). Gene expression data preprocessing. *Bioinformatics*, *19*(5), 655–656. Retrieved from http://doi.org/10.1093/bioinformatics/btg040

Hill, S. M., Neve, R. M., Bayani, N., Kuo, W.-L., Ziyad, S., Spellman, P. T., . . . Mukherjee, S. (2012). Integrating biological knowledge into variable selection: an empirical Bayes approach with an application in cancer biology. *BMC Bioinformatics*, *13*, 94. doi:10.1186/1471-2105-13-94

Hoeting, J., Madigan, D., Raftery, A., & Volinsky, C. (n.d.). Bayesian model averaging: a tutorial. *Statistical Science, 14*(4), 382–417.

Hoot, N., & Aronsky, D. (2005). Using Bayesian Networks to Predict Survival of Liver Transplant Patients. *AMIA Annual Symposium Proceedings*, *2005*, 345–349.

Huang, X., Lin, J., & Demner-Fushman, D. (2006). Evaluation of PICO as a Knowledge Representation for Clinical Questions. *AMIA Annu Symp Proc, 2006,* 359–363.

Ing, C.-K., & Lai, T. L. (2011). A stepwise regression method and consistent model selection for high-dimensional sparse linear models. *Statistica Sinica*, *21*(4). doi:10.5705/ss.2010.081

Jaeschke, R. (2000). Up To Date. *Evidence-Based Medicine, 5*(2), 40–40. Retrieved from http://doi.org/10.1136/ebm.5.2.40

Jaulent, M.-C., Le Bozec, C., Zapletal, E., & Degoulet, P. (1997). A Case-Based Reasoning method for computer-assisted diagnosis in histopathology. In E. Keravnou, C. Garbay, R. Baud, & J. Wyatt (Eds.), *Artificial Intelligence in Medicine, Lecture Notes in Computer Science* (Vol. 1211, pp. 239–242). Berlin / Heidelberg: Springer.

Ji, J., Yan, J., Liu, C., & Zhong, N. (2005). An improved Bayesian networks learning algorithm based on independence test and MDL scoring. In *Proceedings of the 2005 International Conference on Active Media Technology, 2005. (AMT 2005)* (pp. 315–320). Presented at the Proceedings of the 2005 International Conference on Active Media Technology, 2005. (AMT 2005). doi:10.1109/AMT.2005.1505360

John, G. H., & Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence* (pp. 338–345). San Francisco, CA: Morgan Kaufmann Publishers Inc. Retrieved from http://dl.acm.org/citation.cfm?id=2074158.2074196

Juvela, S. (2011). Prevalence of and risk factors for intracranial aneurysms. *Lancet Neurol*, *10*, 595–597. doi:10.1016/S1474-4422(11)70125-9

Juvela, S., Porras, M., & Poussa, K. (2000). Natural history of unruptured intracranial aneurysms: probability of and risk factors for aneurysm rupture. *Journal of neurosurgery*, *93*(3), 379–387. doi:10.3171/jns.2000.93.3.0379

Killeen, T., & Kockro, A. R. (2013). ISUIA / UCAs Aneurysm Calculators. (n.d.). Retrieved from URL:http://www.kockro.com/en/calculators

Kofod-petersen, A. (2006). Challenges in case-based reasoning for context awareness in ambient intelligent systems. *8th European Conference on Case Based Reasoning, Workshop Proceedings, Ölüdeniz* (p. 287).

Koller, D., & Friedman, N. (2009). The Bayesian Network Representation. In *Probabilistic Graphical Models: Principles and Techniques* (1st ed.). The MIT Press.

Koller, D., & Sahami, M. (1996). Toward Optimal Feature Selection (pp. 284–292). Morgan Kaufmann.

Kolodner, J. L. (1983). Towards an understanding of the role of experience in the evolution from novice to expert. *International Journal of Man-Machine Studies, 19*(5), 497–518. doi:10.1016/S0020-7373(83)80068-6

Koton P. (1988). Reasoning about evidence in causal explanations. In T. M. Mitchell & R. G. Smith (Eds.), *Proceedings AAAI-88* (pp. 256-261). Menlo Park, CA: AAAI Press.

Krystyna Milian, A. B. (2012). Building a library of eligibility criteria to support design of clinical trials. doi:10.1007/978-3-642-33876-2_29

Kuo, L., & Mallick, B. (1998). *Variable Selection for Regression Models*.

Kuschner, K. W., Malyarenko, D. I., Cooke, W. E., Cazares, L. H., Semmes, O., & Tracy, E. R. (2010). A Bayesian network approach to feature selection in mass spectrometry data. *BMC Bioinformatics*, *11*, 177. doi:10.1186/1471-2105-11-177

Lam, W., & Bacchus, F. (1994). Learning Bayesian belief networks: An approach based on the MDL principle. *Computational Intelligence*, *10*, 269–293.

Lang, T. A., & Altman, D. G. (2013) Basic Statistical Reporting for Articles Published in Biomedical Journals: The "Statistical Analyses and Methods in the Published Literature" or The SAMPL Guidelines." In P. Smart, H. Maisonneuve, & A. Polderman (Eds.), *Science Editors' Handbook, European Association of Science Editors*, 2013.

Larranaga, P., Sierra, B., Gallego, M. J., Michelena, M. J., & Picaza, J. M. (1997). *Learning Bayesian Networks by Genetic Algorithms. A case study in the prediction of survival in malignant skin melanoma*.

Lauritzen, S. L. (1995). The EM algorithm for graphical association models with missing data. *Computational Statistics & Data Analysis*, *19*(2), 191–201.

Leardi, R., Boggia, R., & Terrile, M. (1992). Genetic algorithms as a strategy for feature selection. *Journal of Chemometrics*, *6*(5), 267–281. doi:10.1002/cem.1180060506

Levin, A. (2001). The Cochrane Collaboration. *Annals of Internal Medicine, 135*(4), 309–312. http://doi.org/10.7326/0003-4819-135-4-200108210-00035

Li, Y., Liu, Y., & Bai, L. (2008). Genetic algorithm based feature selection for mass spectrometry data. In *8th IEEE International Conference on BioInformatics and BioEngineering, 2008. BIBE 2008* (pp. 1–6). Presented at the 8th IEEE International Conference on BioInformatics and BioEngineering, 2008. BIBE 2008. doi:10.1109/BIBE.2008.4696664

Lin, E., & Huang, L.-C. (2008). Identification of significant genes in genomics using Bayesian variable selection methods. *Advances and applications in bioinformatics and chemistry: AABC*, *1*, 13–18.

Liu, H., Hussain, F., Tan, C. L., & Dash, M. (2002). Discretization: An Enabling Technique. *Data Min. Knowl. Discov.*, *6*(4), 393–423. doi:10.1023/A:1016304305535

Luciano, J. S., Andersson, B., Batchelor, C., Bodenreider, O., Clark, T., Denney, C. K., . . . Dumontier, M. (2011). The Translational Medicine Ontology and Knowledge Base: driving personalized medicine by bridging the gap between bench and bedside. *J Biomed Semantics, 2*, S1. doi:10.1186/2041-1480-2-S2-S1

Mao, Q., & Li, X. (2005). Markov Chain Monte Carlo Method of multiple imputation for longitudinal data with missing values in the survey of maternal and children health. *Sichuan da xue xue bao. Yi xue ban (Journal of Sichuan University) Medical science edition*, *36*(3), 422–425.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: *The PRISMA Statement. PLoS Med 6*, e1000097. doi:10.1371/journal.pmed.1000097

Montani, S. (2009). Case-based Reasoning for managing non-compliance with clinical guidelines. *Computational Intelligence, 25*(3), 196-213. doi:10.1111/j.1467-8640.2009.00338.x

Montani, S. (2011). How to use contextual knowledge in medical case-based reasoning systems: a survey on very recent trends. *Artificial intelligence in medicine, 51*(2), 125–131. doi:10.1016/j.artmed.2010.09.004

Montani, S., Bellazzi, R., Portinale, L., Fiocchi, S., & Stefanelli, M. (n.d.). A Case-Based Retrieval System for Diabetic Patients Therapy.

Morita, A., Kirino, T., Hashi, K., Aoki, N., Fukuhara, S., Hashimoto, N., . . . Yoshimoto, T. (2012). The natural course of unruptured cerebral aneurysms in a Japanese cohort. *The New England journal of medicine*, *366*(26), 2474–2482. doi:10.1056/NEJMoa1113260

Newton, P. K., Mason, J., Bethel, K., Bazhenova, L. A., Nieva, J., & Kuhn, P. (2012). A Stochastic Markov Chain Model to Describe Lung Cancer Growth and Metastasis. *PLoS ONE*, *7*(4). doi:10.1371/journal.pone.0034637

Nikovski, D. (2000). Constructing Bayesian Networks for Medical Diagnosis from Incomplete and Partially Correct Statistics. *IEEE Trans. on Knowl. and Data Eng.*, *12*(4), 509–516. doi:10.1109/69.868904

Nilsson, M., (2004). Advancements and Trends in Medical Case-Based Reasoning: An Overview of Systems and System Development.

Nwiabu, N., Allison, I., Holt, P., Lowit, P., & Oyeneyin, B. (2011). Situation awareness in context-aware case-based decision support. *2011 IEEE First International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)* (pp. 9–16). Presented at the 2011 IEEE First International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA). doi:10.1109/COGSIMA.2011.5753761

O'Hara, R. B., & Sillanpää, M. J. (2009). A review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis*, *4*(1), 85–117. doi:10.1214/09-BA403

Patrician, P. A. (2002). Multiple imputation for missing data. *Research in nursing & health*, *25*(1), 76–84.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Perez-Rey, D., Jimenez-Castellanos, A., Garcia-Remesal, M., Crespo, J., & Maojo, V. (2012). CDAPubMed: a browser extension to retrieve EHR-based biomedical literature. *BMC Medical Informatics and Decision Making*, *12*, 29. doi:10.1186/1472-6947-12-29

Royal, L., & Rathbone, B. (2001). What makes a good clinical guideline? *Проблемы стандартизации в здравоохранении*, *1*(11), 1–8. doi:10.1038/295459a0

Sackett, D. L., Strauss, S. E., Richardson, W. S. (2000). *Evidence-based medicine: How to practice and teach EBM*. London, UK: Churchill-Livingstone.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507–2517. doi:10.1093/bioinformatics/btm344

Schardt, C. (n.d.). LibGuides. Introduction to Evidence-Based Practice. Type of Question. Retrieved August 15, 2014, from http://guides.mclibrary.duke.edu/content.php?pid=431451&sid=3530451

Schmidt, R., & Gierl, L. (2003). Case-Based Reasoning for Time Courses Prognosis. In V. Palade, R. Howlett, & L. Jain (Eds.), *Knowledge-Based Intelligent Information and Engineering Systems, Lecture Notes in Computer Science* (Vol. 2773, pp. 984–991). Berlin / Heidelberg: Springer. Retrieved from http://www.springerlink.com/content/94j06tmf0k5yre86/abstract/

Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMJ, 340*, c332. Retrieved from http://doi.org/10.1136/bmj.c332

Shaikh, M., McNicholas, P. D., & Desmond, A. F. (2010). A pseudo-EM algorithm for clustering incomplete longitudinal data. *The international journal of biostatistics*, *6*(1), Article 8.

Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., . . . Rosse, C. (2005). Relations in biomedical ontologies. *Genome Biology 6*, R46. doi:10.1186/gb-2005-6-5-r46

Stajduhar, I., & Dalbelo-Basić, B. (2010). Learning Bayesian networks from survival data using weighting censored instances. *Journal of biomedical informatics*, *43*(4), 613–622. doi:10.1016/j.jbi.2010.03.005

Stajduhar, I., Dalbelo-Basić, B., & Bogunović, N. (2009). Impact of censoring on learning Bayesian networks in survival modelling. *Artificial intelligence in medicine*, *47*(3), 199–217. doi:10.1016/j.artmed.2009.08.001

Stojadinovic, A., Peoples, G. E., Libutti, S. K., Henry, L. R., Eberhardt, J., Howard, R. S., . . . Nissan, A. (2009). Development of a clinical decision model for thyroid nodules. *BMC Surgery, 9,* 12. doi:10.1186/1471-2482-9-12

Suarez, J. I., Tarr, R. W., & Selman, W. R. (2006). Aneurysmal Subarachnoid Hemorrhage. *New England Journal of Medicine, 354*, 387–396. doi:10.1056/NEJMra052732

Tang, Y., & Srihari, S. N. (2012). Efficient and accurate learning of Bayesian networks using chi-squared independence tests. In *2012 21st International Conference on Pattern Recognition (ICPR)* (pp. 2723–2726). Presented at the 2012 21st International Conference on Pattern Recognition (ICPR).

The Many Controversies of Stage IIIA/IIIB Lung Cancer - Cancer Network. (2010, March 22). Retrieved April 10, 2013, from http://www.cancernetwork.com/lung-cancer/content/article/10165/1542458

Tong, S., & Koller, D. (2000). Active Learning for Parameter Estimation in Bayesian Networks (pp. 647–653). Presented at the In NIPS. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.33.8922

Vandenbroucke, J. P., von Elm, E., Altman, D. G., Gøtzsche, P. C., Mulrow, C. D., Pocock, S. J., . . . Egger, M., for the STROBE Initiative. (2007). Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and Elaboration. *PLoS Med, 4,* e297. doi:10.1371/journal.pmed.0040297

Villablanca, J. P., Duckwiler, G. R., Jahan, R., Tateshima, S., Martin, N. A., Frazee, J., Gonzalez, N. R., . . . Vinuela, F. V. (2013). Natural history of asymptomatic unruptured cerebral aneurysms evaluated at CT angiography: growth and rupture incidence and correlation with epidemiologic risk factors. *Radiology, 269*, 258–265. doi:10.1148/radiol.13121188

Virzi, R. A. (1992). Refining the test phase of usability evaluation: how many subjects is enough? *Hum. Factors*, *34*(4), 457–468.

Wasserman, L., & Roeder, K. (2009). High-dimensional variable selection. *The Annals of Statistics*, *37*(5), 2178–2201. doi:10.1214/08-AOS646

Wu, J. A., Hsu, W., & Bui, A. A. T. (2012). Extracting relevant information from clinical records: Towards modeling the evolution of intracranial aneurysms. *Proc AMIA Symp*; 2012. Accepted as poster.

Xuan, P., Guo, M. Z., Wang, J., Wang, C. Y., Liu, X. Y., & Liu, Y. (2011). Genetic algorithm-based efficient feature selection for classification of pre-miRNAs. *Genetics and molecular research: GMR*, *10*(2), 588–603. doi:10.4238/vol10-2gmr969

Zhao, D., & Weng, C. (2011). Combining PubMed knowledge and EHR data to develop a weighted bayesian network for pancreatic cancer prediction. *Journal of biomedical informatics*, *44*(5), 859–868. doi:10.1016/j.jbi.2011.05.004

Zhong, C., Liu, H., Jiang, L., Zhang, W., & Yao, F. (2013). Chemotherapy Plus Best Supportive Care versus Best Supportive Care in Patients with Non-Small Cell Lung Cancer: A Meta-Analysis of Randomized Controlled Trials. *PloS one*, *8*(3), e58466. doi:10.1371/journal.pone.0058466

Zimmermann, A. (2003). Context-Awareness in User Modelling: Requirements Analysis for a Case-Based Reasoning Application. In K. Ashley & D. Bridge (Eds.), *Case-Based Reasoning Research and Development*, *Lecture Notes in Computer Science* (Vol. 2689, pp. 1064–1064). Berlin / Heidelberg: Springer. Retrieved from http://www.springerlink.com/content/797kcp110vjx9whf/abstract/