Bayesian networks from ontological formalisms in radiation oncology

Alan M. Kalet

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

Doctor of Philosophy

University of Washington

2015

Reading Committee:

John H. Gennari, Chair

Mark H. Phillips

Jason N. Doctor

Program Authorized to Offer Degree:
Biomedical Informatics and Medical Education

University of Washington

Abstract

Bayesian networks from ontological formalisms in radiation oncology

Alan M. Kalet

Chair of the Supervisory Committee:
Dr John H. Gennari
Biomedical Informatics and Medical Education

Bayesian networks (BNs) are compact, powerful representations of probabilistic knowledge well suited to applications of reasoning under uncertainty in medical domains. Traditional development of BN topology requires that modeling experts establish relevant dependency links between domain concepts by searching and translating published literature, querying domain experts, or applying machine learning algorithms on data. For initial network development, these methods are time-intensive, and this cost hinders the growth of BN applications in medical decision making. In addition, they result in networks with inconsistent and incompatible topologies, and these characteristics make it difficult for researchers to update old BNs with new knowledge, to merge BNs that share concepts, or to explore the space of possible BN models in any simple intuitive way.

My research alleviates the challenges surrounding BN modeling by leveraging a hub and spoke system for BN construction. I implement the hub and spoke system by developing 1) an ontology of knowledge in radiation oncology (the hub) which includes dependency semantics similar to BN relations and 2) a software tool that operates on ontological semantics using deductive reasoning to create BN topologies. I demonstrate that network topologies built using my software are terminologically consistent and topologically compatible by updating a BN model for prostate cancer prediction with new knowledge, exploring the space of other dependent concepts surrounding prostate cancer radiotherapy, and merging the updated BN

with a different prostate cancer BN containing cross terms with the original model. I also produce a BN to aid in error detection in radiation oncology, showing the extent to which Bayes nets are clinically impactful. Moreover, I show that the methodology developed in this research is applicable to medical domains outside radiation oncology by extracting a BN from a description logic version of the Disease Ontology.

By translating medical domain literature into ontological formalisms and developing a software tool to operate on those formalisms, I establish a novel, feasible, and useful methodology that advances and improves the creation of clinically viable Bayesian network models. In sum, my research represents a foundational component of a larger framework of automation and innovation that contributes to further application of BNs in medical decision support roles.

TABLE OF CONTENTS

Page

List of Figures iv

Chapter 1: Introduction 1
1.1 Background and motivation 1
1.1.1 Bayes net development challenges 1
1.1.2 Model inconsistency 3
1.1.3 Model incompatibility 4
1.2 Solution approach 5
1.3 Contributions of the dissertation 8

Chapter 2: Bayesian networks and biomedicine 11
2.1 Mathematical formalism 12
2.2 BNs in oncology 14
2.3 BNs in radiation oncology 15
2.4 Radiation oncology processes 16
2.5 Summary 19

Chapter 3: Bayesian network based error detection in radiation oncology 21
3.1 Introduction 21
3.2 Methods and materials 24
3.2.1 Network Topology 25
3.2.2 Network Probabilities 27
3.2.3 Evaluation 28
3.3 Results 31
3.3.1 Error detection network 31
3.3.2 Detection performance 35
3.4 Discussion 36

3.5 Conclusion 41

Chapter 4: Automating network topology creation: an ontological approach 42
4.1 Introduction 42
4.2 Biomedical ontologies 42
4.3 Leveraging domain ontologies 43
4.4 Constructing a dependency layered domain ontology 46
4.4.1 Class hierarchy 47
4.4.2 Dependency layer 50
4.4.3 Ontology in full view 54
4.5 Extracting dependency networks from ontological specifications 55
4.5.1 Subsetting ontology via concept selection 55
4.5.2 Subsetting the ontology via dependency layer 57
4.6 Application to other medical domains 57
4.7 Discussion 58
4.7.1 Continued ontology development 59
4.7.2 Algorithmic limitations 60
4.7.3 Scalability 60
4.8 Summary 61

Chapter 5: A software tool for Bayes net development 62
5.1 Goals and architecture 62
5.2 User interface development 66
5.3 Software features 67
5.3.1 Ontology selection 68
5.3.2 Network topology 68
5.3.3 Network edge list 71
5.3.4 Download handling 72
5.3.5 Exploring dependency paths 73
5.3.6 Error handling 74
5.4 Summary 75

Chapter 6: Applications and novel use cases 77
6.1 Introduction 77

6.2 Time savings in initial BN development . . . 79
6.3 Updating networks with new knowledge . . . 80
6.4 Merging network models . . . 83
6.5 Diagnostic networks from the Disease Ontology (DO) . . . 85
6.5.1 Adding description logic to the disease ontology . . . 86
6.5.2 Extracting a diagnostic network . . . 88
6.6 Challenges and insights of the DO as a knowledge hub . . . 91
6.7 Summary . . . 92

Chapter 7: Conclusions and Future Work . . . 93
7.1 A larger framework of automation . . . 93
7.1.1 Semi-automated ontology building . . . 93
7.1.2 Automated queries for conditional probability tables . . . 95
7.1.3 CPT's from big data . . . 97
7.1.4 Integration into clinical workflow . . . 98
7.2 Expansion of ontological scope . . . 99
7.3 Further development of clinically relevant networks . . . 100
7.3.1 Advanced error detection in radiotherapy . . . 100
7.3.2 Other clinical Bayes net use cases . . . 101
7.4 Advanced software developments . . . 101
7.5 Concluding thoughts . . . 102

Bibliography . . . 104

Appendix A: Radiation oncology ontology full class structure . . . 113

Appendix B: BNDE functional dependency structure . . . 118

Appendix C: Copyrights and Permissions . . . 119

LIST OF FIGURES

Figure Number — Page

1.1 Graphic schema of hub-spoke system for network subsetting. Hypothetical models 1, 2, and 3 all share the same set of knowledge concepts {A, B, C, D, E}. Even though the sub-networks are unique, they remain consistent and compatible. 5

1.2 The process of knowledge state transition from Literature to Knowledge base to Bayes net to Decision Support occurs first through manual translation, then through logic operations on an ontology (the knowledge base), then by machine learning on clinical data corresponding to knowledge in Bayes net form. 7

2.1 A simple Bayesian network graph. 12

2.2 A simple Bayes net model for disease diagnosis. 14

2.3 Overview of the radiation oncology process from intake to treatment. 17

2.4 Process map of a clinical radiation oncology process from intake to treatment. Figure adapted by author from Ford et al. [35] 18

2.5 Process components involved in an initial medical physics chart verification. 20

3.1 Initial Bayes net for error detection in treatment plans. Many nodes are highly interconnected (3+ dependencies) to capture the nature of causal knowledge in larger swaths of the radiation oncology domain. Various layers indicate an underlying knowledge structure among concept nodes 26

3.2 Bayesian network topology after parameter learning and node pruning. . . . 32

3.3 Prior distributions (top) of the variables `Total Dose` with no instantiation are contrasted with their respective posterior distributions (bottom) after instantiation and propagation of clinical variables `Histology = Astrocytoma` and `Intent = curative` for the BN 33

3.4 Prior distributions of the variables `Technique` with no instantiation are contrasted with their respective posterior distributions (bottom) after instantiation and propagation of clinical variables `Histology = Astrocytoma` and `Intent = curative` for the BN 34

3.5 Receiver operator characteristic curve for the EBNs for three different tumor sites Lung, Breast, and Brain. The areas under the curves were 0.88, 0.89, 0.98, respectively. These ROC curves can be used to set thresholds for possible error alerts depending on the user's preferences. 36

3.6 ROC curves for brain tumors comparing the performance of the EDN with that of the experts. The expert responses were aggregrated (dashed line) and plotted with the SDOM of responses (shaded grey). AUC for the aggregate human response was 0.90 ± 0.01. 37

4.1 Two-level representation of knowledge based on Ramoni et al. The arrow represents a translation of knowledge from epistemological formalism (ontology) to computational formalism (influence diagrams and Bayes nets) 44

4.2 A screenshot of Protégé showing the top three levels of classes represented in the radonc domain ontology. 49

4.3 A screenshot of Protégé showing part of the bottom level class hierarchy. The annotated term definition "biological No Evidence of Disease" (top right) is given as a citation to the published source which describes the full meaning and context. The set of annotated dependencies to other ontological concepts is also shown (bottom right). 51

4.4 Protégé view of radonc ontology annotation for relation bNED `dependsOn2` Gleason. 53

4.5 Visualization of the radonc ontology using hierarchical edge bundling algorithm. Lines represent dependency between concepts, and are bundled together spatially by the classification scheme. Highlighted here are the dependency paths among terms related to `Plan_Technique`. 55

5.1 Shiny reactivity-initial state with input (`input$obs`) and a valid output (`output$distPlot`). The output `distPlot` gets invalidated, flushed, and re-executed upon changes to `input$obs` 64

5.2 Dynamic information flow between the server, client, and local filesystem in the BNDE 65

5.3 Functional dependency graph among variables in the BNDE used to generate network displays 67

5.4 BNDE upload button and dropdown menu for ontology selection. 69

5.5 BNDE Network Topology tab. The selected ontology's class-subclass structure is recreated (dynamic UI) as a interactive folder-tree appearing on the main sidebar panel (left). 70

5.6 Edge Panel tab 71

5.7 Dowloand Panel in the BNDE 72
5.8 Pathway explorer panel 74

6.1 The hub-spoke system represented with the radonc ontology as the hub. . . . 78
6.2 Prostate utility network, Meyer et al. 2004 81
6.3 Prostate utility network, updated Jan, 2015 81
6.4 Network display of all dependency pathways between "T-stage" and "Disease_Free_Survival" in the radonc ontology. Nodes not in the initially examined network (Figure 6.3) are shaded in green. 83
6.5 Prostate utility network Smith et al. 2009 combined with Meyer et al. 2004 prostate network, updated Jan, 2015 84
6.6 R code for adding classes into the Disease Ontology. 87
6.7 R code for declaring new dependency relations between classes in the Disease Ontology 89
6.8 Diagnostic network topology extracted from the Disease Ontology 90

7.1 Replacing the manual translation step in the knowledge base building process with natural language processing (NLP) methods could greatly increase the portability of an automated network building process to other domains. . . . 94
7.2 Framework for leveraging disparate data sources by mapping their schemas to the ontology for easier queries. 96
7.3 Integration of decision models into clinical workflows can occur by embedding them into the front-end of locally used clinical systems. 98

A.1 Class-subclass hierarchy of the ontology for radaition oncology (version: Jan 2015) 117

B.1 Dependency graph for the Bayesian Network Domain Explorer (BNDE) web-application. Represented here are GUI inputs and outputs, download functions, developer-defined functions, and reactive elements 118

ACKNOWLEDGMENTS

I wish to express my sincere appreciation to my advisors John H. Gennari and Mark H. Phillips for their remarkable insights and support of my research. I also want to thank my parents Ira and Terry Kalet for their endless support and encouragement of me throughout this process. Special thanks to my children Morrigan and Brandan (and their patience when provided). Finally, I also want to give special acknowledgments my GSRs Archis Ghate and Wolf Kohn, and to George Sandison, Eric Ford, and a combined effort of support from the Agency for Healthcare Research and Quality (AHRQ-HS22244-01), the University of Washington Medical Center, and the Department of Radiation Oncology.

DEDICATION

This dissertation is gratefully dedicated to my late father, Ira J. Kalet—a man who passionately nurtured, encouraged, and taught me new perspectives and ways of knowing at every possible turn of life.

Chapter 1

INTRODUCTION

1.1 Background and motivation

Bayesian networks (BNs) are compact and powerful representations of probabilistic knowledge. As such, BNs are particularly well suited to applications of reasoning under uncertainty in medical domains. However, there are challenges in the initial development of viable networks and challenges with updating and combining existing networks. Much of this dissertation research is therefore motivated by the desire to reduce the effort required to construct useful networks, and to increase the compatibility and consistency among network models in clinical medicine.

The domain focus of my research is radiation oncology. Radiation oncology (radonc) is a medical sub-specialty concerned with treating cancer by application of high energy radiation. Like many areas of medicine, it is host to a variety of uncertainty and contains a myriad of complex decision making points. Chapter 2 provides an overview of Bayesian formalism and a historical context of BN applications both in medicine generally and in radiation oncology specifically. In the following sections I present some examples from radonc which show the specific challenges mentioned above in more detail, and the solution approach I take to alleviate those challenges.

1.1.1 Bayes net development challenges

The largest barrier to employing Bayes nets more widely in clinical medicine is both the time and effort it takes to initially create such networks and the follow-on effort to keep them updated and useful. For complicated networks with many variables, the time required by domain experts to create an initial network is often prohibitive. There are two main

reasons for the difficulty; 1) developing an appropriate topology and 2) completing joint probability tables for that topology in order to make the network meaningful and computable. Topology development amounts to establishing the influence (or dependency) of one or more variables on another. For example, deciding that "fever" and "cough" depend on "disease", where there can be many states of disease and many states of symptoms. This network of dependency must be proclaimed in node-arc-node relations and evaluated by independent experts for some measure of accuracy.

Instead of relying on domain experts, one can apply machine learning (ML) methods to learn the network topology, but the computational effort to discover network topology is also time intensive [1, 2, 3]. It has been proven that ML methods for learning network topology is an NP-hard problem [4, 5]. Even with high complexity, computational methods can reduce initial topology generation time, that is, although somewhat costly, the algorithm still finds reasonable topologies in much less time than hand-construction. However, there is no guarantee that resulting topologies will contain meaningful dependency relationships among concepts, and thus domain experts are still required to vet the outcomes of the learning algorithms. If the network does not seem viable to experts, the ML learning process must then be repeated with additional constraints.

Creating conditional probability tables (CPTs) can also be challenging, even when the topology is known. In probability theory, the conditional probability table gives the probability distribution of one variable with respect to another when the other is known. There can be one or even several independent parent variables. For each instantiation of a parent variable, a different distribution for the child variable is required. Thus, the number of probability distributions required to populate a conditional probability table in a Bayesian network grows *exponentially* with the number of parent-nodes associated with that table. Obtaining probability information by querying human experts for every possible case quickly becomes cost prohibitive for anything more than the simplest of networks.

Bayesian networks are also models which make statements about the causality between domain concepts. The way various models make statements about the world suffer from the

"Tower of Babel" problem in that they often use different terminologies to describe the same concepts. This results in inconsistency and a lack of compatibility among models. With inconsistency comes confusion about whether or not separate models in the same domain are describing relationships among the same concepts. Effort is required to investigate this question, and it is not known *a priori* if an answer exists. Incompatibility arises when models describe the same concepts with different causal relationships. This prevents model sharing and integration, and makes computation on model knowledge challenging.

1.1.2 Model inconsistency

Examples of the model inconsistency problem (mismatched terminology) can be readily found in biomedical literature. Meyer et al. and Smith et al. both published Bayesian networks designed to compute probabilities about aspects of prostate radiotherapy [6, 7]. Each network has a different design and overall goal, but both share many of the same knowledge components. This is known because the conditional probability tables for those knowledge components were derived from the same knowledge source, a publication by Horowitz et al. in 2001 [8]. Discovering the shared knowledge source requires critical inspection of each manuscript's text and bibliography in detail. The discovery process is made even more opaque because no naming convention for concept nodes was used–in part because none existed. For example, the nodes "T Stage" in one network and "T class" in the other mean the same thing in both networks[1]. More complicated technical domain terms are less obvious to ascertain. For instance, "ColdVolProstate", is defined in the text of the Meyer et al. article as follows: "The input to ColdVolProstate was defined as $1 - (D_{99.5}/D_{95})$...", but is not clearly defined at all in Smith et al., where the term used is "PTV Cold Spot".

The same problem can be found in other studies. Several research groups have built Bayes nets to model various outcomes in lung cancer radiotherapy, given some clinical findings including the Gross Tumor Volume. In three different networks, three different terms are

[1]This terminological inconsistency occurred across publications in spite of a number of shared co-authors

used for this common volume: “GTV”, “GTV size”, and “lntumorload” in Oh et al., Jayusara et al., and Dekker et al., respectively [9, 10, 11]. In one of these lung networks, the clinical T-stage is termed “T” while in another “tc” is used. It is likely these are the same T-stage tumor descriptor term as in the prostate networks, but since there is no underlying common knowledge source, that statement must remain an assumption.

1.1.3 Model incompatibility

Contributing to the problem of different terminologies being used is the application of different knowledge sources for network construction. The knowledge sources tapped to make networks often varies between domain expert beliefs, published clinical studies, and machine learned data. Sometimes multiple sources are used. The result of this often leads to incompatibility among network topologies. Network incompatibility is particularly acute when topology is learned strictly from data. For example, Stojadinovic et al. produced a set of networks to predict survival in colon cancer patients by applying machine learning algorithms on a large dataset [12]. The dataset was subsetted based on 12, 24, 26, and 60 month follow up times. The resulting topologies have node-arc-node relations which contain opposite causal directions from each other. In the 12 month network, there is a relation “TNM path M” $\rightarrow$ “Mortality”, whereas the in the 60 month network the reverse is present: “TNM path M” $\leftarrow$ “Mortality”. In other words, one network says a pathology causes mortality, but in the other, mortality causes pathology. The algorithm has found a predictive relation between the two domain concepts, but does not have knowledge of the underlying causal nature of disease.

The consequence of modeling networks in isolation and with differing knowledge sources hinders future researchers’ ability to reuse the networks and effectively build on them. A network model is itself a representation of domain knowledge, and knowledge sharing and reuse cannot effectively be accomplished with inconsistent and/or incompatible knowledge representations.

1.2 Solution approach

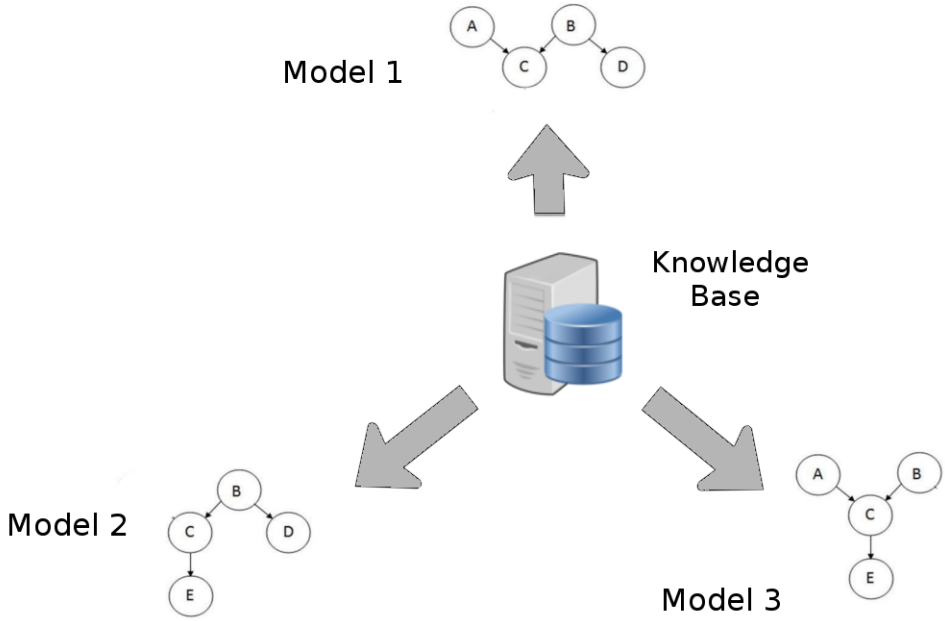


Figure 1.1: Graphic schema of hub-spoke system for network subsetting. Hypothetical models 1, 2, and 3 all share the same set of knowledge concepts {A, B, C, D, E}. Even though the sub-networks are unique, they remain consistent and compatible.

To ease the initial development of probabilistic models as well as their interoperability and reuse among a medical domain, I propose a hub-spoke system where a centralized, standardized knowledge source becomes the hub for constructing consistent and compatible dependency models. Figure 1.1 shows a simple schematic version of how the hub-spoke system works with network models. Unique networks can be built/extracted from the knowledge base independently while retaining consistent and compatible terminology and knowledge relationships.

In this dissertation, I propose a methodology which transforms states of knowledge from

one form to another using algorithms appropriate for each translation. Figure 1.2 broadly outlines the translational steps. To realize the hub-spoke system, a knowledge base must be built and a software system must be constructed to automate the extraction of networks from the knowledge base in simple, intuitive ways. To make extracted network topologies viable as decision support models, population of the network tables must be performed in a way that limits domain expert time. Because the goal is to extract networks consisting of node-arc-node relations, I choose to use an ontology to formalize the knowledge base. Constructing the ontology is a matter of (manually) translating domain knowledge from peer-reviewed literature and other expert sources into sets of subject-predicate-object relations. This manual translation is the top-most step in Figure 1.2. The closely mirrored semantics of ontology triples and Bayesian networks (node-arc-node $\leftrightarrow$ subject-predicate-object) simplify the construction of the ontology and extraction by software methods while adding transparency to the knowledge base and extraction process. In Chapter 4 I discuss my methods for constructing a dependency layered ontology and the algorithms used for extracting and pruning sub-networks from the ontology. An important contribution of this dissertation is a specialized software system (representing the hub-spoke portion of the solution) that applies deductive reasoning (logic) on the ontology triples to create Bayes nets. This is the *Knowledge base* $\rightarrow$ *Bayes net* step of knowledge state translation in Figure 1.2. I detail my development of this specialized software including it's architecture, design, and features in Chapter 5. The probability tables of extracted network topologies are completed by applying a machine learning algorithm to clinical data sets in the *Clinical Data*/*Bayes net* + *ML* $\rightarrow$ *Decision Support* step at the bottom right of Figure 1.2. As a part of my work, I show that when a topology is given, using ML methods to compute CPTs from medical data stores is an efficient means of completing probability tables. I describe my specific methods for CPT construction with data extracted from a medical relational database in Chapter 3.

The methodology described in this dissertation represents a novel application of encoding and retrieving medical domain knowledge from ontologies. By encoding the relevant dependency semantics, I form a network of meaningful computable relationships among medical

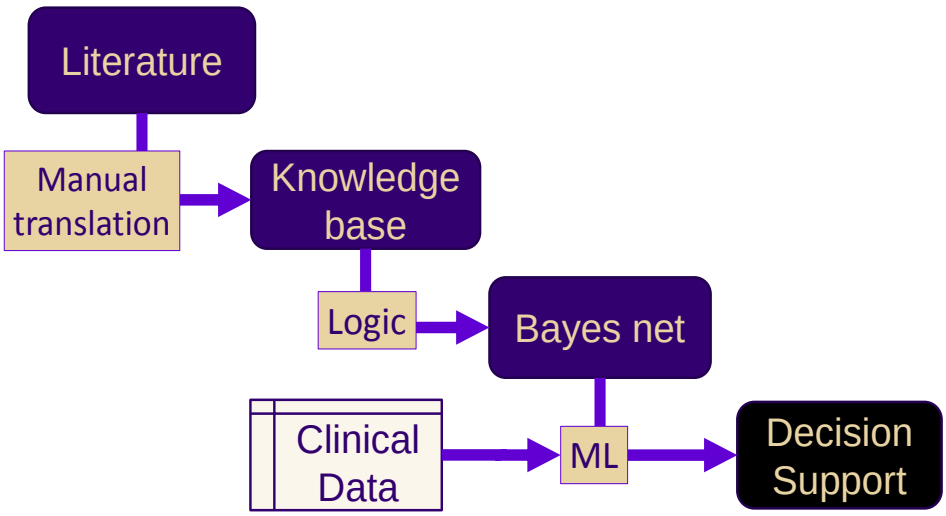


Figure 1.2: The process of knowledge state transition from Literature to Knowledge base to Bayes net to Decision Support occurs first through manual translation, then through logic operations on an ontology (the knowledge base), then by machine learning on clinical data corresponding to knowledge in Bayes net form.

concepts which can be used to build a variety of Bayes nets that address causality and uncertainty for a variety of different decision making settings and stakeholders. My goal is to make building these Bayes nets easier and to make the knowledge in those networks more consistent. For example, creating a generic class for "complications" and then listing disease-specific complication (esophagitis, pneumonitis, rectal complications, etc.) as subclasses, we can better organize the knowledge and potentially perform some completeness and correction tests of that knowledge.

An ontology also provides a foundational structure that can be more easily used to augment and change network topology as new knowledge and causal dependencies are added. A demonstration of how an ontological formalism can be used for updating models with new domain knowledge is given in Chapter 6. Methods of ontology use for Bayes net construction

have been explored previously as a viable means of auto-generating Bayes nets for telecommunications applications[13] and for use in developing medical diagnostic networks[14], but the flexibility to create networks capable of answering specific questions about topics in the domain goes well beyond a re-application of a technique. In the ontology, I build in semantics that describe the directed dependency links, having translated the dependency from the original knowledge source. Many of the ontology links therefore represent clinical judgments or causalities which are not found directly in the literature. Anyone can subset this ontology without restriction, which means they can build any kind of network to answer any question of interest, regardless of whether there is some pre-existing model for that network. The dependency layer sets a groundwork protocol for establishing dependency across medical domains.

As the knowledge base grows, entering new knowledge in the hierarchical class-subclass form becomes more tractable than adding nodes to a large Bays net. Ontology knowledge entry is manageable on scales up to around a thousand concepts, beyond which ontologies themselves may become difficult to manage. However, Bayes nets are only manageable on the order of 10-20 nodes, beyond which it becomes problematic to read, navigate, and understand the network. Moreover, the number of different networks that can be explored grows exponentially with the number of new concept nodes. This exponential scaling opens up possibilities for generating previously undiscovered networks and concept dependencies.

1.3 Contributions of the dissertation

In this dissertation I develop a novel methodology for BN construction demonstrated in but not limited to a specific medical domain (radiation oncology). I show that my methodology, a software-based, semi-automated BN extraction from ontological knowledge sources creates consistent and compatible network topologies which solves both the terminological problem and the knowledge source problem. I also show by specific example (with the Disease Ontology) that the method is portable to other areas of clinical medicine.

The software tool was developed to aid researchers in navigating and selecting various

domain concepts such that they can generate networks designed to answer new questions of interest in the domain. Class-subclass and strength of dependency type relations are leveraged by the software to allow for researchers to select from or add constraints to groups of highly interconnected domain terms. I demonstrated a prototype system for BN topology exploration and qualitatively evaluate its ability to perform updating of old networks with new knowledge, merging of networks with cross terms, and potential applications outside the radiation oncology domain. In particular, I created a description logic version of the Disease Ontology [15] to give it a formal dependency layer from its existing (but alternative) dependency semantics and showed that my software extracts diagnostic type Bayes nets from the modified ontology.

As a part of this research, I also built and verified a BN for error detection applicable to a crucial step in the radiotherapy plan verification process. I show practical utility of BN models by constructing a network model aimed at identifying high risk, low occurrence errors within radiotherapy treatment plans in the radiation oncology domain. I created a) a suitable concept topology from domain expert input and b) extracted clinically derived data from a relational database model and c) applied machine learning algorithms on these data in order to create the conditional probability tables in the network. This probabilistic model is designed to capture types of information which provide *likelihood* of error, even in cases where all hard constraints (rules) are met. I introduced planning errors derived from a local "clinical continuing safety improvement" database into a subset of offline clinical data. The error cases were run through the BN propagation to test the strength of the network's ability to flag errors. I also presented these error test cases to human expert for chart review, and compared the results of human experts' assessments of plan correctness to the network results. This proof of concept network shows potentially high clinical impact in its ability to detect probabilistic classes of errors with sensitivity and specificity comparable to human counterparts performing the same task [**?**]. The BN performance results exemplify the utility that BNs can have on the safety and quality of complex modern healthcare.

The concepts making up the topology of this network served as an initial knowledge set for

expansion into a larger ontology. That ontology became the knowledge base, (the hub in the hub spoke model shown in Figure 1.1) and forms the foundation from which new or existing Bayes net models can be built and/or changed. I constructed the ontology initially from the BN relations derived for the error detection network, and then expanded it by adding concepts from various other knowledge sources such as published peer-reviewed literature and domain experts. I identified the canonical semantic relationships in this domain in order to formalize radiation therapy plan and other concepts into a common standard using Web Ontology Language (OWL) description logic. Generation and validation of this initial ontology was performed by domain experts. The ontology not only necessarily includes BN-specific relations such as dependency, but also a practical class-subclass structuring of radiation oncology concepts which has not yet been performed in the domain.

In developing the BN for radiotherapy error detection, I make some discoveries regarding the significant mismatch between clinical knowledge terminology and clinical data source terminology. Additionally, I find there is a looming "Big Data" problem for BNs which hope to leverage large clinical data repositories for CPT construction. These findings, among others, dictate the important directions of future work in this area. I outline this and other challenges related to further automating steps of knowledge state translation in Chapter 7.

Chapter 2

BAYESIAN NETWORKS AND BIOMEDICINE

The complex nature of healthcare and biomedicine has been approaching a state where decision support tools are becoming increasingly useful—in particular—adaptable models for decision making. Modeling medical decisions computationally requires solid fundamental principles regarding medical decision making and cognition itself; an excellent systematic review of which can be found here[16]. This kind of work, in conjunction with early prototype systems resulted in many novel systems. Though the utility of these expert systems and decision support was recognized over twenty years ago[17], few, if any, of these methodologies have been incorporated into standard practice in the time since, in large part because the overhead required to make these systems scalable. Research has been progressing, however, and the core ideas are still valid. Because so much of medical decision making is based on likelihood or level of belief in one variable given some clinical evidence, research into probabilistic models like Bayesian networks has been growing. Bayes nets are probabilistic directed graphical models with nodes that represent variables and arcs that represent conditional independence assumptions. They provide a compact representation of joint probability distributions which are effective at modeling uncertainty in various systems or processes and have been explored in many areas of biomedicine[18]. In this chapter, I describe the mathematical formalism behind Bayesian networks and discuss how Bayesian networks are used in oncology. To provide additional context to this dissertation, I give an overview of the radiation oncology sub-specialty, how Bayes nets have been previously applied in radiation oncology specifically, and why the domain is well suited for further applications of Bayes nets.

2.1 Mathematical formalism

Bayesian Networks (BNs) are directed acyclic graph networks represented formally by a set of joint probabilities distributions described by probability calculus. Graph networks themselves are simple visual models, containing *nodes* and *arcs* (sometimes referred to as edges) that connect two nodes to indicate a relation between them. In directed graphs, arcs specify a `from` and `to` in the relation. For example, the arc $a = (X, Y)$ in Figure 2.1 has direction `from` X, `to` Y. X is said to be the parent node and Y is the child node. In acyclic graphs, there are no directed paths among any of the nodes and arcs that leads back to the start of the path. Bayesian networks satisfy both directionality and acyclic constraints because BNs are mathematically rigorous representations of causality. Thus, these constraints force avoidance of circular reasoning in models.

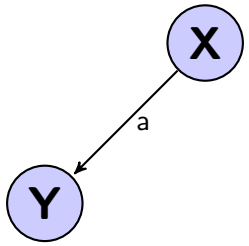


Figure 2.1: A simple Bayesian network graph.

The basic concept in Bayesian statistics is *conditional probability.* Conditional probability is a statement of the form: "given event X, the probability of event Y is l". This statement is formally expressed as $P(Y \mid X) = l$. Generally, $P(Y \mid X) \neq P(X \mid Y)$ because they are conditioned on differing events. For instance, the probability of testing positive for a rare disease, given you actually have the disease is far greater than the probability of having that rare disease, given that your test result is positive. In order to compute actual values to prove this claim we must use Bayes Theorem, which can be derived from the fundamental rule for

probability calculus. The formal definition of conditional probability in given in (2.1)[1], where $P(X, Y)$ is the probability of joint event $X \wedge Y$, and $P(Y)$ is the independent probability of observing event Y.

$$P(X \mid Y) = \frac{P(X, Y)}{P(Y)} \tag{2.1}$$

$$P(Y \mid X) = \frac{P(Y, X)}{P(X)}, \tag{2.2}$$

Substituting (2.2) into (2.1) we get (2.3).

$$P(X, Y) = P(X \mid Y), P(Y) = P(Y \mid X) \tag{2.3}$$

Putting this result back into (2.1) produces Bayes Theorem:

$$P(X \mid Y) = \frac{P(Y \mid X)\, P(X)}{P(Y)} \tag{2.4}$$

With this we can compute the probability from the "rare disease" example above. For a test sensitivity of 99% and specificity of 99%, (that is, the test will produce 99% true positive results for those who have the disease and 99% true negative results for who don't) and a rare disease with prevalence of say, 0.5%, we compute the following probability of having the rare disease, given a positive test result:

$$P(D_+ \mid T_+) = \frac{P(T_+ \mid D_+)\, P(D_+)}{P(T_+)} = \frac{P(T_+ \mid D_+)\, P(D_+)}{\sum_j P(T_+ \mid D_j) P(D_j)}, \tag{2.5}$$

$$\Rightarrow \frac{0.99 \times 0.005}{0.99 \times 0.005 + 0.01 \times 0.995} = 0.332 \tag{2.6}$$

This result is indeed quite different from the probability of 0.99 for known positive disease carriers. The probability is conditioned on the low prevalence of disease. Therefore, our confidence in the test should be low.

---

[1]The converse statement for $P(Y, X)$ is given in (2.2)

By inserting a test result in this example, we are instantiating evidence. Evidence changes the probability distribution of the dependent variable. The resulting distribution is known as the *posterior* probability. In networks with chained dependency (e.g. $X \leftarrow Y \leftarrow Z$), evidence instantiated can propagate between variables. For evidence in Z, the posterior is computed for Y. Given the new distribution in Y, the posterior can be computed for X. Likewise, if evidence is instantiated in Y, it propagates to X and Z. In a fully computable BN, each node requires a conditional probability table (CPT) representing all the joint probabilities between that node and it's parent nodes. Leveraging CPTs and propagation among variable distributions, we could make our example into a more accurate diagnostic model by including a second independent test to give more information to our disease probability, forming a topology shown in Figure 2.1. A thorough treatise of this and other traditional methods for building up network model topologies can be found in Jensen [19].

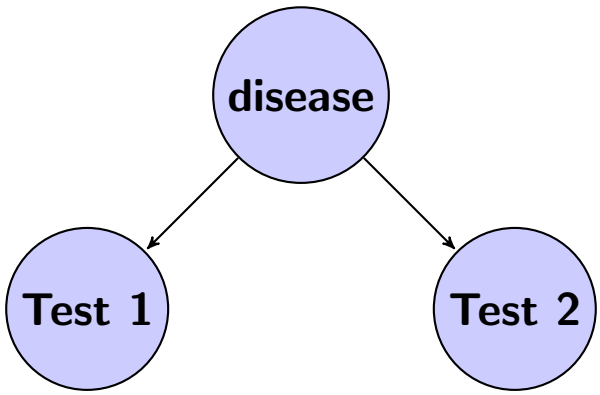


Figure 2.2: A simple Bayes net model for disease diagnosis.

2.2 BNs in oncology

The properties of BNs allow computational reasoning tasks to be performed in a way that mimics components of human intelligence and understanding of causality. This has led researchers to build BNs for various tasks including diagnosis, prognosis, error detection, and other decision making in a variety of medical specialties.

Much has been accomplished building networks for cancer diagnosis. For example, an impressive series of progressively more extensive and successful models for breast cancer have been built by several researchers over the last few decades[20, 21, 22, 23]. More recently, BN based decision models are being pursued for prognosis and disease progression when diagnosis is already known. Verduijna et al.[24] give a good overall description of medical prognostic models in general. A good deal of progress has been made in cancer prognosis specifically. Nissan et al.[25], for example, have used BNs to predict nodal involvement for colorectal cancer staging, while others, like the NasoNet project[26] and Exarchos et al.[27] developed Bayesian networks to model the progression of certain classes of head and neck cancers (oral and nasopharyngeal). Others have used BNs to predict the likelihood of survival for bone metastases in order to choose whether surgical intervention should be recommended[28]. A learned BN for lung cancer survival was shown to outperform support vector machines by Dekker et al.[11]. Networks which inform decisions about modifying treatment mid-course (a very bleeding edge change to modern techniques termed "adaptive therapy") have been recently explored[29]. Despite these efforts, at present, there are far fewer predictive models than situations arising in real clinical environments. However, these prognostic models in particular (if effective) have the potential to be of clinical impact and are also in line with national incentive programs outlined in meaningful use phase 3, e.g. improving quality, safety, and efficiency, leading to improved health outcomes and decision support for national high-priority conditions[30].

2.3 BNs in radiation oncology

The radiation oncology (radonc) domain is the sub-specialty of oncology primarily concerned with treatment of (mostly) solid tumors with electromagnetic radiation. Radiation oncology is a dynamic field in which technological and computer developments are continually providing more versatile tools for the planning and delivery of radiation. In addition, advances in genomics, imaging and other medical fields provide more information about responses of individuals and their diseases to radiation treatments. However, even given an advancing

knowledge of cancer biology, the outcomes of the cancers and their treatments are characterized by probabilities with high variances—far from deterministic. In many cases, even the diagnoses and locations of the tumors are uncertain. The role that uncertainty plays in decision making in this field makes Bayes networks a natural choice for modeling the process of radiation oncology. As noted, previously, Bayes nets are a compact and versatile approach, and have the ability to provide some degree of transparency—that is, the ability to query the network to explain decisions and determine which nodes are the most critical[19, 31].

The use of Bayes networks to represent the integrated radiation therapy process—from implementation of technological solutions to health state outcomes—provides the opportunity to use them in many different contexts beyond prognosis[32]. For example, networks have been constructed to (a) to help in making individualized cancer treatment decisions[6, 7] (b) to provide guidelines for when to ask for diagnostic tests and which therapy methods to use in given clinical situations[33], (c) to aid in identifying potential diagnostic errors [34], and (d) to help in allocating healthcare resources. The previous research in this field illustrates the versatility and flexibility of Bayes networks in many different situations of decision making under uncertainty.

2.4 Radiation oncology processes

The fundamental process by which radiation oncology operates is a differential sensitivity to high energy radiation between cancerous and normal tissues. High energy radiation damages cell DNA in a way that causes apoptosis and other cell death mechanisms. Many cancerous tissues grow at a faster rate than normal tissues, making them less able to recover from radiation damage than normal tissues. Given this knowledge, radiation can be delivered to cancerous tumors at a specified rate which kills tumor cells while allowing some level of normal tissue recovery. Radiation sensitivity varies among tumor types which leads to several ranges of applying radiation dosage among the total delivered dose, the dose given for every treatment, the timing frequency of delivered dose, and other factors. However, he high toxicity of radiation to both cancerous and normal tissues combined with the limitation

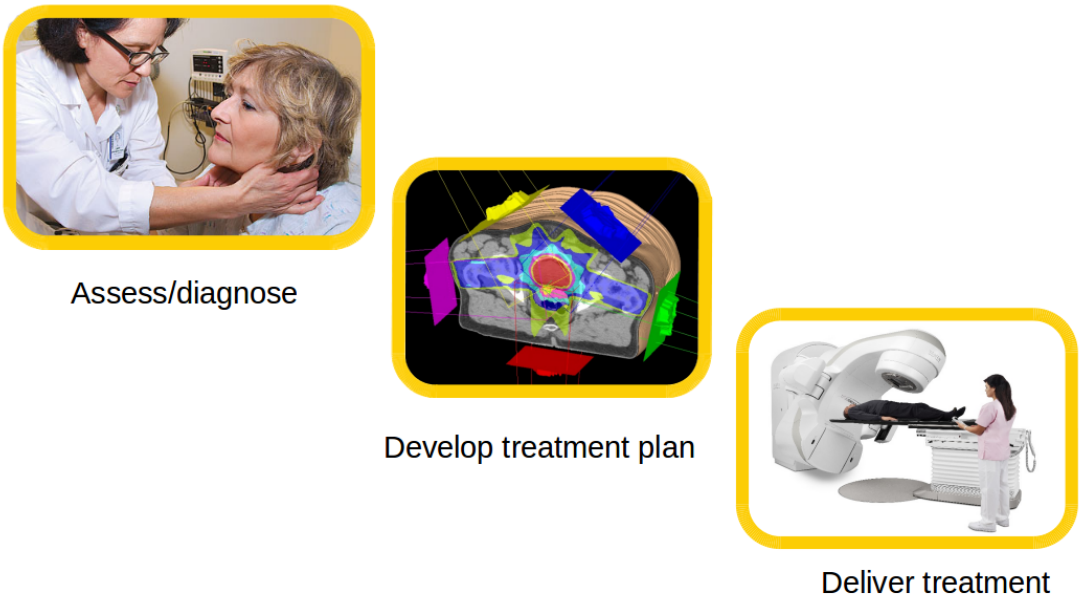


Figure 2.3: Overview of the radiation oncology process from intake to treatment.

of radiation delivery methods make this a potentially very dangerous treatment method. Therefore, the process of effectively delivering therapeutic radiation to patients in a safe and useful manner is quite complex.

Radiation oncology is a tertiary specialty that most patients are referred to from outside oncologists, and typically patients have had some initial diagnostic testing performed indicating cancerous growth. However, this initial diagnostic information is typically insufficient for further radiation treatment. Patients first come into radonc and undergo additional thorough diagnostic assessment often including imaging such at CT, PET-CT, or MRI. With extensive imaging information, physicians draw outlines (on 2-D image sets) using specialized software of where they believe the border of the tumor lies inside the patient. Other organs and physical regions of interest are outlined in this step as well. These regions are used to develop a radiation treatment plan in a treatment planning system (TPS). The TPS software contains a model of the radiation delivery device and an algorithm to compute radiation distribution inside a patient's imaging scan. Specialized clinical personnel (dosimetrists) apply various

methods and planning techniques to achieve the prescribed radiation dose distribution to the tumor while minimizing dose distribution to the surrounding normal tissues and structures outlined in the previous stage. Once the plan has been checked and approved, the plan is implemented by therapists on the radiation delivery device. This process is shown broadly in Figure 2.3.

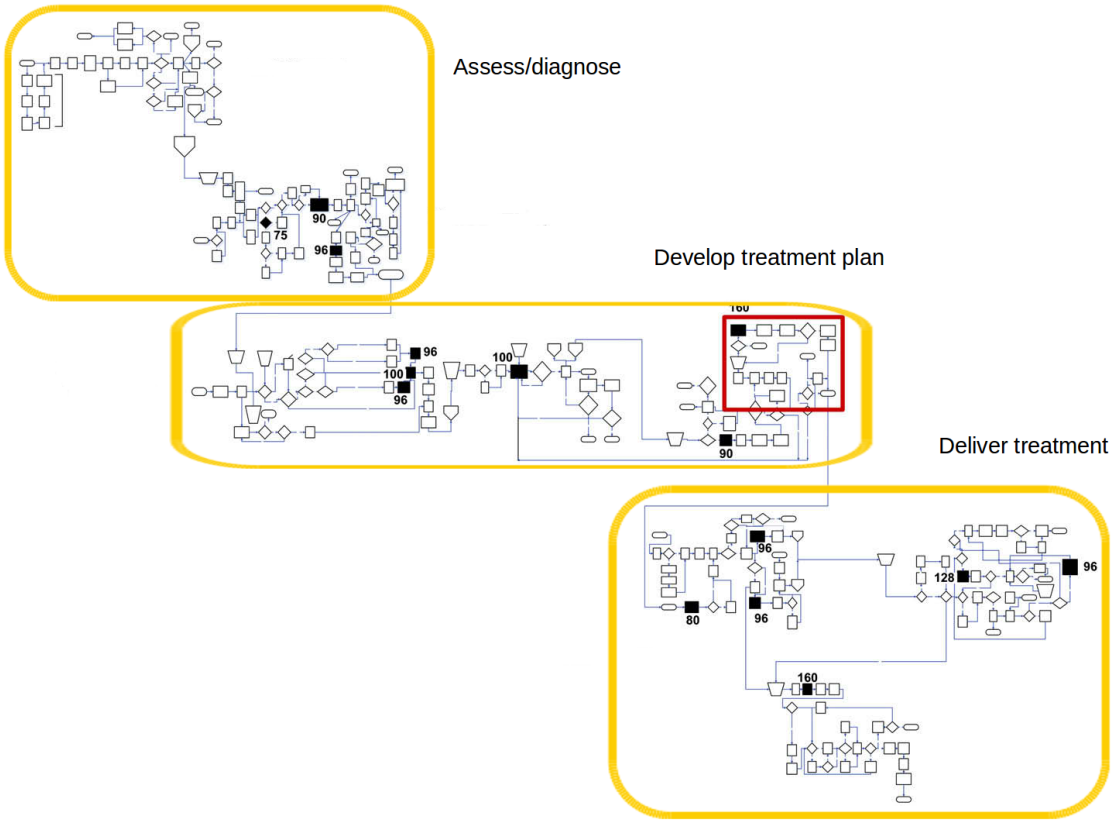


Figure 2.4: Process map of a clinical radiation oncology process from intake to treatment. Figure adapted by author from Ford et al. [35]

Though the overall process of radiation therapy workflow can be roughly divided into three stages, the actual processes, personnel, and steps taken within those stages is complicated and non-linear. Ford et al. developed a process map for radiation oncology clinical workflow at Johns Hopkins Department of Radiation Oncology and Molecular Radiation Sci-

ences (shown in Figure 2.4) to visualize the extent of process complexity and non-linearity. The process map outlined by the study represents a clinic which had yet switched to an electronic medical record and verify system, though many of the steps involved do not change with this switch. If anything, the introduction of multiple (sometimes poorly) integrated software systems since the publication of this workflow model has only made the process more complex and increased opacity regarding information transfer between systems and between systems and personnel.

Examine more closely the part of the workflow process for verifying a digitally reconstructed radiograph (DRR) before the patient is given radiation. Shown in Figure 2.5, this process represents only a small part of a clinical physicist's task of initial chart verification[2]. Much of the cognitive work performed in any individual part of this process is itself hidden in each step. The top left step "Export DRR to R&V" may be a matter of a few button clicks by a dosimetrist, while the bottom right step, "Physics check vs plan - is it okay?" typically involves dozens of rule based comparisons and several professional judgment considerations across multiple independent software systems. The requirement for task specialization within a complex domain such as radonc is such that evaluation of an entire treatment plan by any individual is more challenging than this workflow model suggests. Many areas of the full process map can be broken down into components this way—where judgment type decisions are made which effect downstream processes in a non-obvious manner (i.e. they occur outside the process map).

2.5 Summary

In the survey presented here, I identify many examples of BN usefulness in biomedical applications, particularly in oncology and radiation oncology. Within the formalism described above, it becomes clear why it is useful to apply Bayes networks to answer questions of conditional dependence occurring in medicine. Within the radiation oncology sub-specialty,

[2]Figure reproduced with permission from Elsevier–see Appendix C for copyright transfer agreement

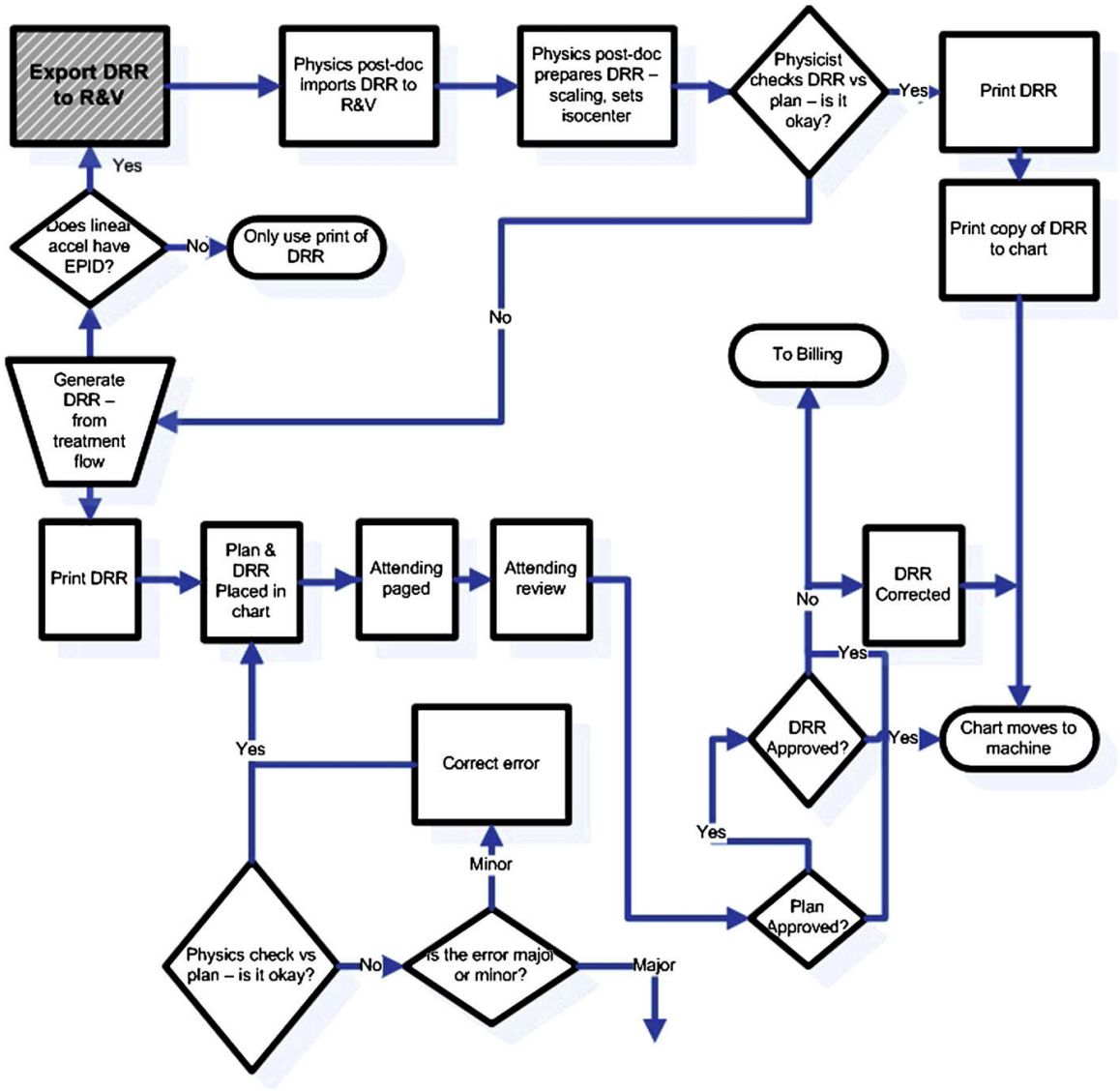


Figure 2.5: Process components involved in an initial medical physics chart verification.

I show that the extent of uncertainty and complexity among workflow tasks and the nature of some of those tasks harbors significant potential for further applications of BNs.

Chapter 3

BAYESIAN NETWORK BASED ERROR DETECTION IN RADIATION ONCOLOGY

As a part of this dissertation, I designed and developed a probabilistic network for detecting errors in radiotherapy plans for use at the time of initial plan verification. I have initiated a multi-pronged approach to reduce these errors. To build my networks I first interviewed medical physicists and other domain experts to identify the relevant radiotherapy concepts and their associated interdependencies and to construct a network topology. Next, to populate the network's conditional probability tables, I used the Hugin Expert software to learn parameter distributions from a subset of de-identified data derived from a radiation oncology based clinical information database system. These data represent 4990 unique prescription cases over a 5 year period. Under test case scenarios with approximately 1.5% introduced error rates, network performance produced areas under the ROC curve of 0.88, 0.98, and 0.89 for the lung, brain and female breast error detection networks, respectively. Comparison of the brain network to human experts performance (AUC of 0.90 $\pm$ 0.01) shows the Bayes network model performs better than domain experts under the same test conditions. These results demonstrate the feasibility and effectiveness of a comprehensive probabilistic models as part of decision support systems for improved detection of errors in initial radiotherapy plan verification procedures.

3.1 Introduction

The effects of errors in the delivery of radiation therapy can range from inconsequential to disastrous. In-depth data collection and evaluation has provided both insight and hard data

on potential and actual errors in radiation oncology, including a variety of studies among different institutions which report error frequency in radiation oncology on the order of 1% per course of treatment [36, 37, 38, 39]. The volume of treatments (over one million patients annually) at this rate results in approximately 5000-6000 errors per year and suggests that the existing QA procedures are insufficient [40]. In addition to the frequency of errors, we must consider their consequences as well. Overdoses can lead to catastrophic side-effects, while underdoses and geographic misses can result in a significant reduction in the probability of tumor control—errors that are nearly impossible to detect after the fact.

Current error-detection systems involve redundant manual reviews of treatment variables at different steps in the therapy planning and delivery processes [41]. One of the most critical steps in the chain is a review of the radiotherapy plan by the medical physicist. Ford et al. have shown that this pretreatment plan and chart review by the medical physicist is the most effective at capturing high severity incidents, however, this method is still prone to human error and biases [42]. The requirement for task specialization within a complex domain such as radiation oncology is such that evaluation of an entire treatment plan by any individual is challenging. Though some of this challenge is alleviated by instituting multiple checks both before and during treatment by different individuals, many errors can and do still slip through [43].

The last several decades have seen rapid improvements in the localization and delivery of radiation with a concomitant explosion in the complexity of the planning and treatment processes and in the number and types of variables that define a treatment. Thus, the process of conducting a plan and chart verification by examining the variables has become much more difficult. Software to aid in the detection of planning errors has been addressed by a number of groups [44, 45, 46] by means of rule-based systems. One limitation of such an approach is the system's inability to alert the user to errors of judgment, that is, many plans may and do meet all rule based criteria for acceptability, yet still contain errors or suboptimal treatment choices. Rules and checklists can verify the existence of hard constraint violations such as monitor unit matching of electronic plan transfer from radiation treatment planning

(RTP) system to radiation delivery device, but they cannot reliably capture error classes of such as a misinterpretation of prescription or inappropriate planning technique for a given tumor type/location.

In addition, an expert plan review must account for the complexity of the medical decision making process. The relationships between many variables and the magnitudes of the variables cannot be encapsulated easily into rules since they depend on details of the disease, its location, and prior treatments, none of which are apparent in the treatment plan itself. In addition, physician preference can be a contributing factor, for example, in the decision to use a certain fractionation scheme. All of these factors lead to the conclusion that in many cases probabilistic relationships are the most appropriate way of characterizing plan variables. Furhang et al. have explored the performance of an automated initial chart checking processes that employs case-based reasoning to measure similarity between the current plan parameters and historic plan parameters in a probabilistic way [47], however, the probabilities of these models are independent of each other and static. One way to encapsulate more dynamic probability distributions which represent interdependency between variables is to employ probabilistic networks such as Bayesian networks.

Bayesian networks (BN) are probabilistic directed graphical models with nodes that represent variables, and arcs that represent conditional independence assumptions. They provide a compact representation of joint probability distributions which are effective at modeling uncertainty in various systems and have been explored in many areas of biomedicine [18] including applications in radiation oncology [6, 33, 7]. BNs also have the ability to provide transparency—that is, the ability to query the network to explain decisions and determine which of its nodes are the most critical [19, 48]. BNs allow one to ask (and receive an answer for) queries of the form "given a clinical evidence set X, what is the probability of observing therapy planning set Y?" even when the set Y is conditionally dependent on the set X in a complicated way. More detail regarding the mathematical formalism of BNs is given in chapter 2. Unlike other statistical models which answer similar questions, BNs can be built on a transparent underlying formalism establishing real-world dependency between

concepts in a domain, thereby eliminating spurious correlations among parameters which are not causally related. Such networks have been shown to aid in error detection both in clinical trial data [49] and in clinical laboratory settings [50, 51].

In this chapter, I describe my initial results in developing a Bayesian network for identification of potential errors in radiotherapy plans. Ultimately, these models will be coupled with a rules-based approach to create a two-tiered software model for error detection. As I show below, I have demonstrated feasibility with results that show high accuracy on test cases. Overall, the goal is not to replace the physicist treatment plan review; rather, it is to provide decision support: to highlight those parts of the plan that are most likely to be in error and to provide guidance in subsequent inquiries. In other words, it is a decision aid to relieve the tedium of finding a "needle in a haystack" and to reduce the time needed to identify the source of errors.

3.2 Methods and materials

Bayesian networks are directed acyclic graphs representing conditional dependencies among a set of variables. The classical example of a Bayesian network for diagnosis would represent the probabilistic relationships among diseases and symptoms; such a network can be used to predict the probability of a given disease given a number of symptoms. However, in this case, I use the formalism to represent relationships among clinical characteristics, radiation dosing prescriptions, and radiation treatment plans. In a BN, the probability of any particular variable achieving a given value in its domain is conditioned on the value of its parent variables in the directed graph. Construction of a fully computable network requires both a suitable topology—the set of nodes and arcs making up the structure of dependency relations—and also sets of conditional probability tables (CPT) for each node. Our overall strategy for creating the error detection Bayes network involves three sequential steps:

1. Manually construct a suitable concept topology from expert knowledge

2. Extract data from a clinical database for each network concept node

3. Automatically populate the network's CPTs from extracted clinical data

3.2.1 Network Topology

Using a semi structured interview technique [52] I queried domain experts (medical physicists) about their knowledge of dependency relationships in order to capture an initial topology for the BN. The network is designed to model the relevant parts of a radiotherapy (RT) plan examined during the initial plan and chart verification step performed by a medical physicist in the radiation oncology workflow process. As previously noted, this check is a critical step that comes after the completion of a treatment plan, but prior to the actual delivery of the radiation plan to the patient. I created a set of interview questions to identify the concepts most commonly employed by clinicians to evaluate the probabilistic types of questions regarding the accuracy and completeness of RT plans at this point in the workflow. In addition to enumerating the concepts, the interview questions also sought to identify the dependency relationships among those concepts. Many of the interview questions involved asking about a particular concept and what additional information (if any) was needed to evaluate the correctness or accuracy of that concept's state. Answers which involved language predications similar to "depends on" were translated into arcs between concept nodes in a BN formalism. The semi-structured interview method also allows for an additional round of follow up with modified or reworked questions based on responses from the initial interview with the goal of achieving a consensus on the most relevant concepts and their appropriate dependencies. The topology of this network was then encoded into a BN using the Hugin Expert software[1] [53].

The initial error detection Bayes network contains nodes for a broad range of clinical variables ranging from the primary tumor location (`Laterality`) to the number of beams (`N_Beams`) for a given RT plan. During the interview process, I found that many domain concept sets settled out into a few overarching superclasses, noted in Figure 3.1: a clinical

[1]Hugin Expert A/S, Aalborg, Denmark

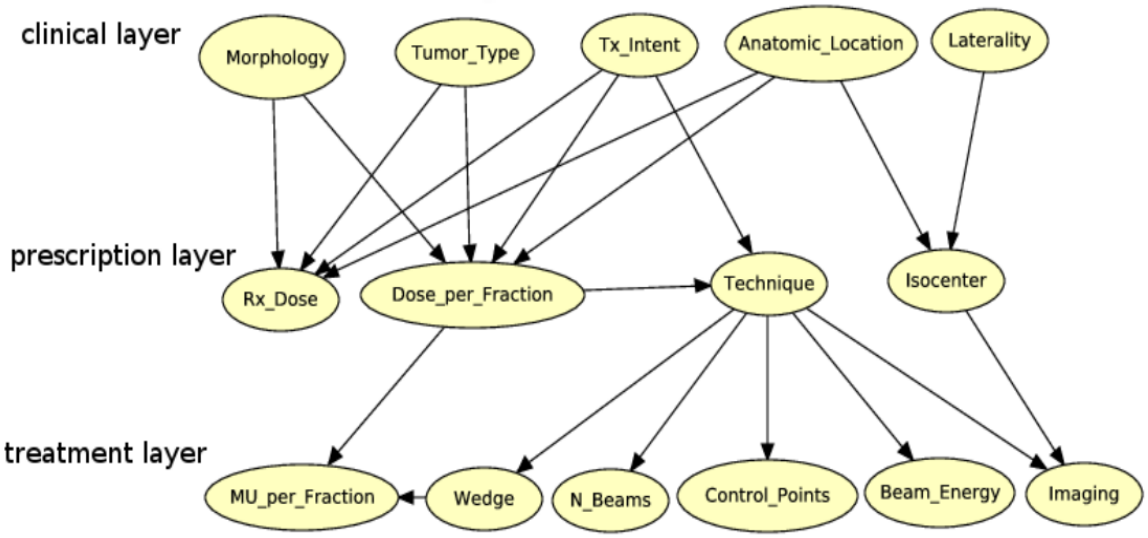


Figure 3.1: Initial Bayes net for error detection in treatment plans. Many nodes are highly interconnected (3+ dependencies) to capture the nature of causal knowledge in larger swaths of the radiation oncology domain. Various layers indicate an underlying knowledge structure among concept nodes

layer, a prescription layer, and a treatment layer. These layers make the network model more transparent to the flow of knowledge and choices made in a clinical environment: for example, a patient presents with clinical findings, which causes a decision to be made about the prescription, which is then implemented on the radiation delivery device. Carefully formalizing domain knowledge using a variety of semantic formalisms over and above the BN formalism has benefits towards future use of these networks. In particular, it enhances the potential for the networks to be more easily revisited, augmented, and updated without repeating the entire interview processes as new knowledge enters the domain or as the state of the art of clinical practice and treatment changes. Further discussion of this advantage is provided in Chapter 6.

3.2.2 Network Probabilities

To derive the probability tables, I extracted data from a Mosaiq oncology information system (OIS)[2] using a standard structured query language (SQL). I queried for all states of identifiable topological concepts from the initial network. These anonymous data were preprocessed for quality and consistency and saved into a flat file to be read by the Hugin system's machine learning (Expectation Maximization) algorithm [54] for auto-population of the CPTs. Due to disease recurrence and cases of metastasis, several patients had multiple prescriptions. Some patients also had an "initial" prescription followed by a "boost" prescription typically given to a smaller region. Given that there is no clear internal distinction in the relational database (RDB) schema itself which distinguishes these prescriptions, they were treated as separate instances.

In the queried clinical test data subset, there were 73 anatomic categories, 329 different dose totals, and 251 International Classification of Diseases (ICD-9) codes. Given the vast amount of data collected for this set (approx 5 years of full time clinical practice in a university radiation oncology department running multiple radiation treatment units), I found that memory and computational constraints were insufficient for creating one large network from all the data. For example, a joint probability table for "Rx_Dose", given parent nodes anatomic categories (ICD-O) and cancer histology (ICD-9) would have required over *6 million unique entries*. To reduce this constraint, we split the data into several subsets by common anatomic region, resulting in sets of 2780, 1195, 1015 unique prescription cases for breast, lung, and brain sites, respectively. For each of these sites, a network was generated using the same topology but populated with separate respective data subsets.

---

[2]Elekta AB, Stockholm, Sweden.

3.2.3 Evaluation

Potential errors and testing data

For each network I took a sample of 100 cases from the initial dataset and introduced known potential errors[3] at a nominal rate of 1.5% error per set. Each case contained 9 parameters of which only 6 were evaluated giving 600 tests per network and approximately 9 errors per set of 600 tests. The introduced errors varied by type and severity based on consensus recommendations for assigning severity to actual and near miss events [55]. A few example cases are presented in Table 3.1. In the brain test set, the example in the first row of Table 3.1 is an unusually high dose when the intent is palliative given the increased possibility of complications. However, a total dose of 54 Gy dose is not uncommon for *curative* brain tumor treatments, which can result in this case having low visibility to manual plan checkers. In the second case (row 2) I have introduced a more subtle error in the form of a suboptimal choice of beam energy. In clinics with multiple beam energies available, it is not uncommon to find individual use of IMRT techniques *or* 18MV beam energy used alone, but the combination of IMRT *and* 18MV can produce suboptimal dose distributions for this anatomic region. This is a type of conjoined error for which joint conditional probability tables might be well suited to find. The last example, from the breast test case set (row 3) gives an unusually low dose per fraction. This error would result in a non-curative dose and carries a high severity. Other cases of error types include instances where the technique prescribed is "4 field box" whereas the plan has 7 fields. Because the technique is a character string and the number of fields is a countable integer, this kind of mismatch is challenging to flag using a rule-based system which typically only compares like data types.

[3]I use the term "potential errors" here to reflect the fact that in some instances, there are factors outside the scope of this network that may justify the parameter value. In our applications, the intent is to flag such cases for further investigation, rather than define the parameter explicitly as an error. For simplicity, in the rest of this manuscript we will simply use "error", with the understanding that all of these are actually potential errors.

Table 3.1: Example potential error cases. Introduced errors are indicated in bold

Tx_Intent (site)	Technique	Modality	Dose_Ttl	Dose_Tx	numFields	Severity
Palliative (Brain)	Conformal Plan	x06	**5400**	200	6	7
Curative (Brain)	IMRT	**x18**	6000	200	12	2
Palliative (Breast)	PA	x18	2268	**81**	2	8

Network analysis

In a BN, one can instantiate any variables and propagate the effect of that information to the rest of the network. To obtain individual probability values for comparison with the known error set, I instantiated three "clinical finding" parameters in the network: intent, histology, and laterality (paired organ) for each case. By instantiating these variables, I are stating something about one part of the plan parameter set in order to evaluate it's effect on other parts of the set. The instantiation and propagation could in fact be run in the opposite direction by instantiating plan parameters, but the initial topology of the network and clinical layering were instructive in our choice. The clinical finding nodes are "dictatorial" in the sense that they influence many other nodes whom are its dependents, as described by the flow of information from one layer to the next in section 3.2.1. This process also mirrors that used by many clinical medical physicists where clinical findings are often assumed correct until plan parameters indicate otherwise.

After instantiating evidence, we ran the BN propagation to compute each of the six remaining parameter probabilities and then compared each resulting probability for the variable state present in the chart to an error detection threshold. Probabilities below the threshold were flagged as errors. The error detection threshold was incremented in steps of 0.001 from 0 to 1 and comparisons recomputed to produce a receiver operating characteristic curve (ROC) for each network. The ROC curve plots sensitivity versus 1-specificity to pro-

duce a measure of each network's ability to discriminate errors. I compute the area under the ROC curve (AUC) to evaluate the predictive performance of the models. An area under the curve of 0.5 indicates no discriminating ability (displayed in ROC graphs as the 1:1 slope line, i.e. true positive rate = false positive rate) [56]. Analysis was performed using the R statistical language and the RHugin package to interface the Hugin decision engine via the available API [57, 58].

Validation and comparison with human observers

To compare system performance against human domain experts, I recruited participants for an IRB-approved study assessing their ability at detecting potential radiation plan errors via a web-based survey. The survey consisted of ten radiation treatment plan cases taken from the same error infused test set used to evaluate the Bayes net model. Much like the evaluation for the Bayesian network detection model, participants were given three initial "clinical finding" parameters, and then asked to evaluate the remaining six parameters of the plan and respond by checking a box regarding the likelihood of correctness of each parameter in each case (four-point Likert type scale of likelihood). The adverbs used in the responses translated answers into probabilities with a mapping based on the normal cumulative distribution and standard deviations, where "somewhat" indicated one standard deviation, "very" indicated two deviations and the likely/unlikely indicated direction of the deviation from the mean, e.g. "very unlikely" mapped to $p = 0.023$; "very likely" to $p = 0.97$. Quantification done in this way allowed the participants to express their belief in correctness using their own statistical heuristics without the potentially confusing requirement to assign a probability of correctness to every parameter[59]. The mapping created a distribution of probabilistic belief for each parameter based on human decisions, and could therefore be analyzed in the same way as the Bayes nets—by varying a threshold over the probabilities to produce an ROC curve.

3.3 Results

3.3.1 Error detection network

Upon querying the clinical RDB for populating the CPTs, I found several factors which forced changes to aspects of the network topology. In the RDB schema, there exists no digitally readable record of the plan parameters "isocenter" or "imaging", therefore no data were available to populate such a node. Nodes which required some counting operations in the query were successful in obtaining a count for the number of beams for a given prescription, but obtaining monitor units (MU) per fraction was problematic due to the splitting of MU over hundreds of control points in IMRT and VMAT techniques. Differences in node terminology between the initial and final network are caused by back-end schema mismatching the software front-end labeling, for example, `Laterality` in the initial network is equivalent to `Paired_Organ` in the final network. Similarly, `Morphology` is mapped to `Histology` in the schema, as the terminologies are used synonymously in the ICD system. Figure 3.2 shows the final network after pruning nodes that could not be populated from our data.

The Bayes net contains a set of initial probability distributions (priors). Once a variable is instantiated in the network, the change is propagated throughout all other nodes resulting in new (posterior) probability distributions. Figures 3.3 and 3.4 illustrate prior and posterior distributions of the Brain tumor BN for two different variables, `Total Dose`, and `Technique`, respectively. It can be seen that by instantiating `Tx_Intent = curative` and `Histology = Astrocytoma`, the broad prior distribution transformed into a very narrow posterior distribution. The opposite behavior is illustrated in Figure 3.4 in which the prior distribution is quite peaked, and the instantiation of the same two variables (`Tx_Intent` and `Histology`) does little to change it.

These histograms of joint prior probability distributions convey information differences found in the different parameter sets, which can be formalized using the concept of *entropy* [48]. Computing the entropy (H) for these distributions explicitly using Eq. 3.1 gives a

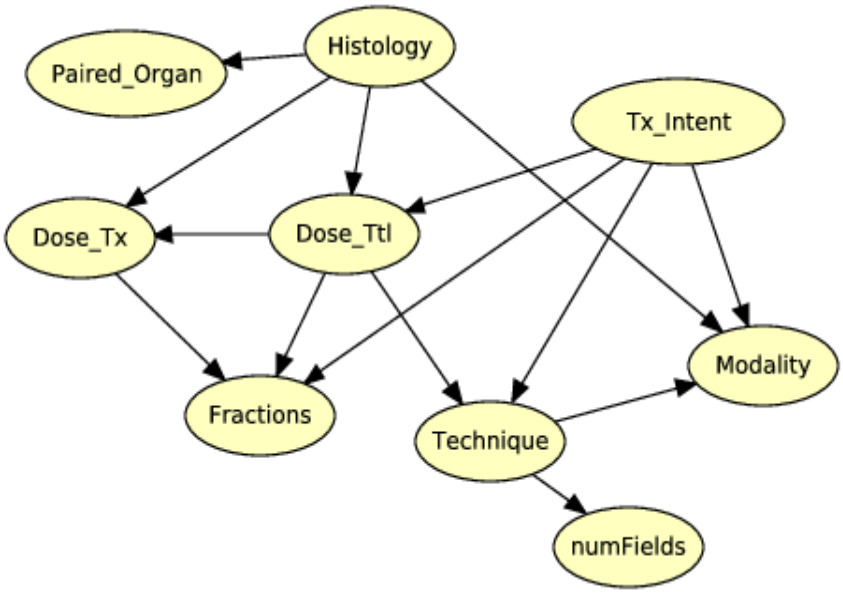


Figure 3.2: Bayesian network topology after parameter learning and node pruning.

specific metric for comparison of prior and posterior distributions, where p_i is the probability of a given variable being in the i^{th} state, summed over all possible states. Entropy values for the examples in Figures 3.3 and 3.4 are shown in Table 3.2.

$$H = -\sum_i p_i \log p_i \tag{3.1}$$

Table 3.2: Information entropy (H) before and after instantiation of clinical variables Histology (Astrocytoma) and Treatment Intent (curative)

	Prior entropy (H)	Posterior entropy (H)	Δ H
Total Dose	3.387	0.950	2.437
Technique	1.687	1.258	0.429

In addition to computing single node entropy, I also perform a Value of Information analysis (VOI) on all observation network nodes with respect to the clinical findings nodes.

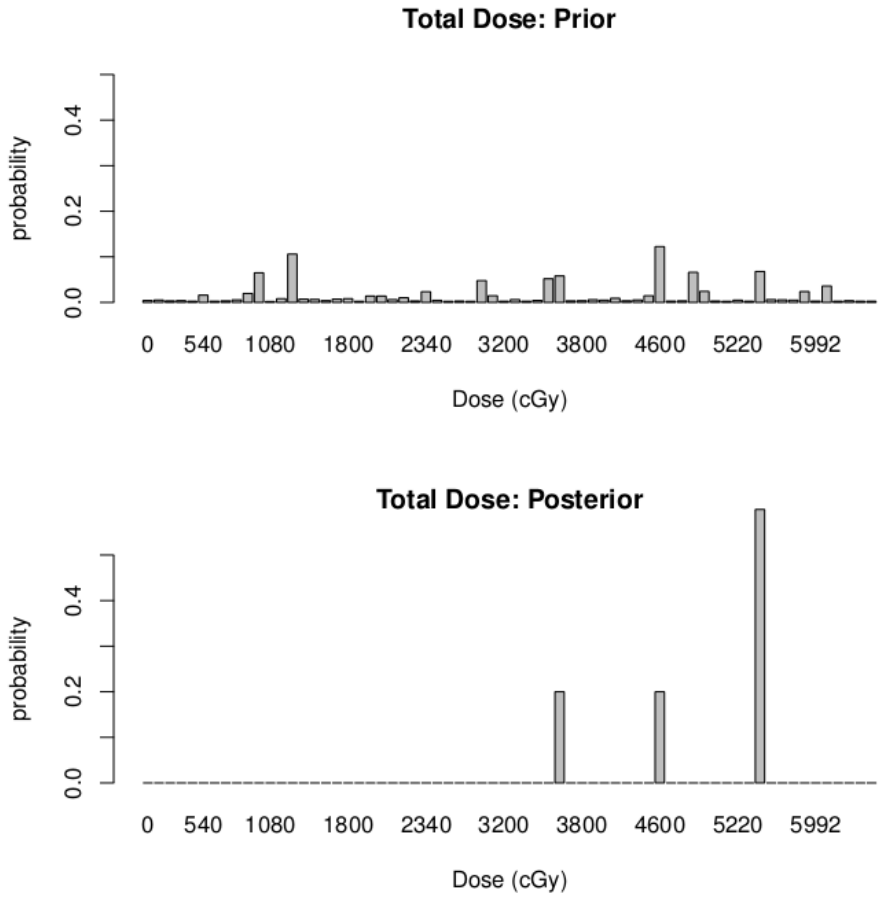


Figure 3.3: Prior distributions (top) of the variables `Total Dose` with no instantiation are contrasted with their respective posterior distributions (bottom) after instantiation and propagation of clinical variables `Histology = Astrocytoma` and `Intent = curative` for the BN

The VOI analysis uses the mutual information I in Eq. 3.2 to produce a value for relative cost-benefit comparison of obtaining and instantiating evidence [48]. Here, p_i and p_j are prior independent probabilities for the i^{th} and j^{th} states of each node distribution and p_{ij} is the joint probability distribution between two nodes. The summation is over all possible states of i and j.

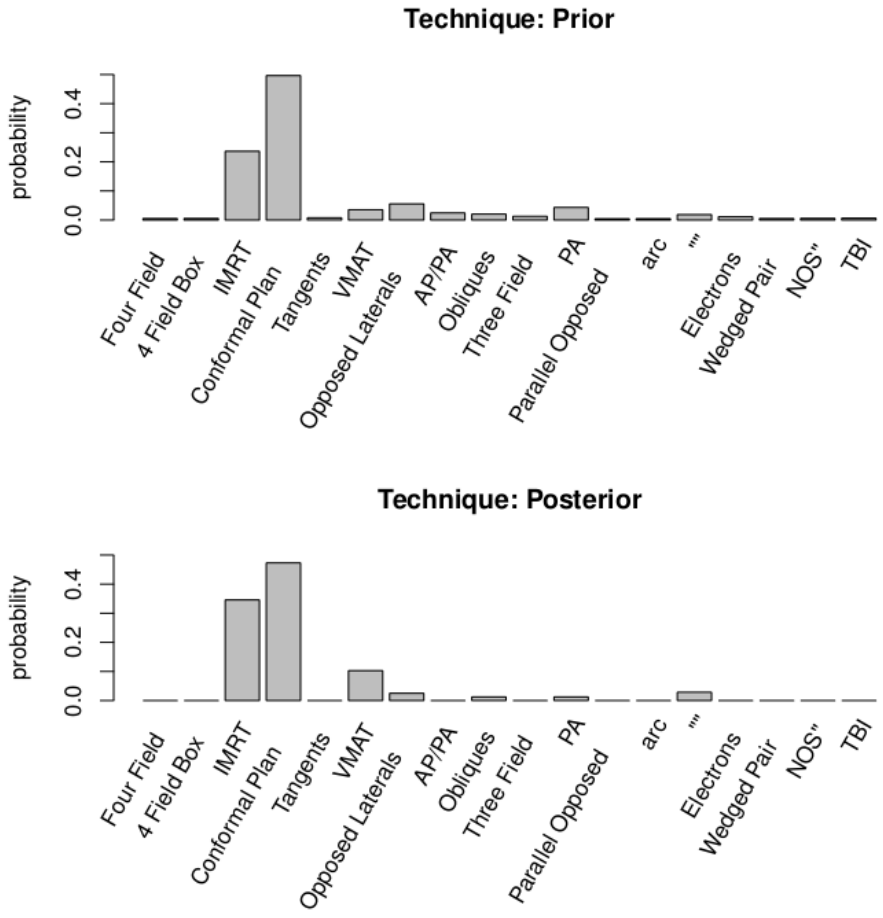


Figure 3.4: Prior distributions of the variables `Technique` with no instantiation are contrasted with their respective posterior distributions (bottom) after instantiation and propagation of clinical variables `Histology = Astrocytoma` and `Intent = curative` for the BN

$$I = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{p_i p_j} \tag{3.2}$$

Table 3.3 shows the mutual information (I) between the evidence nodes and each of the selected information variables. In this model, the Histology evidence node's effect outweighs the other two clinical finding variables by a factor of 4 or more, depending on the specific node examined, and in no cases is it less than a factor of 1. This result indicates that

Histology is one of the most impactful nodes instantiated in the model.

Table 3.3: Value of information metric for the observation network variables based on clinical evidence nodes

	Total Dose	Technique	Modality	N Fields	Fractions	Dose per Fraction
Histology	0.53	0.1	0.21	0.02	0.43	0.31
Tx Intent	0.18	0.08	0.06	0.02	0.11	0.09
Laterality	0.12	0.03	0.05	0.005	0.09	0.09

3.3.2 Detection performance

Figure 3.5 illustrates the results of the Bayesian network performance under the test conditions described in Section 3.2.3. AUCs for the Lung, Breast, and Brain sites were found to be 0.88, 0.89 and 0.98, respectively. The ROC curve for the Brain network shows that an error detection threshold of $p = 0.177$ results in 100% detection at a false positive rate of only 6%. The ROC curves for the Breast and Lung networks indicate that no reasonable threshold setting could be used to find 100% of the errors. A threshold of $p = 0.157$ would detect 91% of the errors with a false positive rate of 21% for the lung network, and a threshold of $p = 0.136$ can be set to detect 89% of errors in the breast network with a false positive rate of 13%.

I obtained a total of six (6) respondents to the human subjects survey for the brain site cases and applied ROC analysis amongst individuals. The error detection rates were aggregated to produce a mean true positive rate, false positive rate, and standard deviation of the mean(SDOM). This ROC curve is shown in Figure 3.6 along with the Bayes net results for the same brain site cases. The AUC for domain experts was found to be 0.90 ± 0.01, compared to 0.98 for the Bayes net.

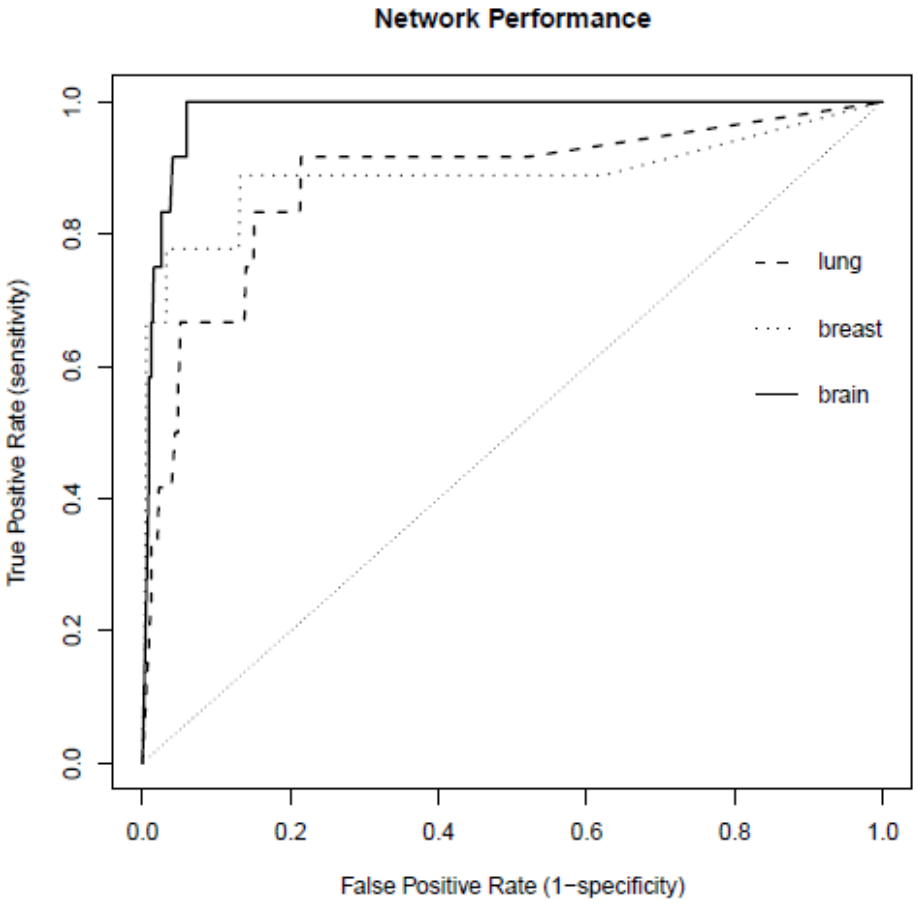


Figure 3.5: Receiver operator characteristic curve for the EBNs for three different tumor sites Lung, Breast, and Brain. The areas under the curves were 0.88, 0.89, 0.98, respectively. These ROC curves can be used to set thresholds for possible error alerts depending on the user's preferences.

3.4 Discussion

The results presented above provide support for my hypothesis that a BN can be used to detect errors in radiation therapy plans. In addition, preliminary results indicate that it can do so more accurately than can experts. This work has highlighted a number of strengths of this approach. One strength is the nature of the conditional probabilities such that for a

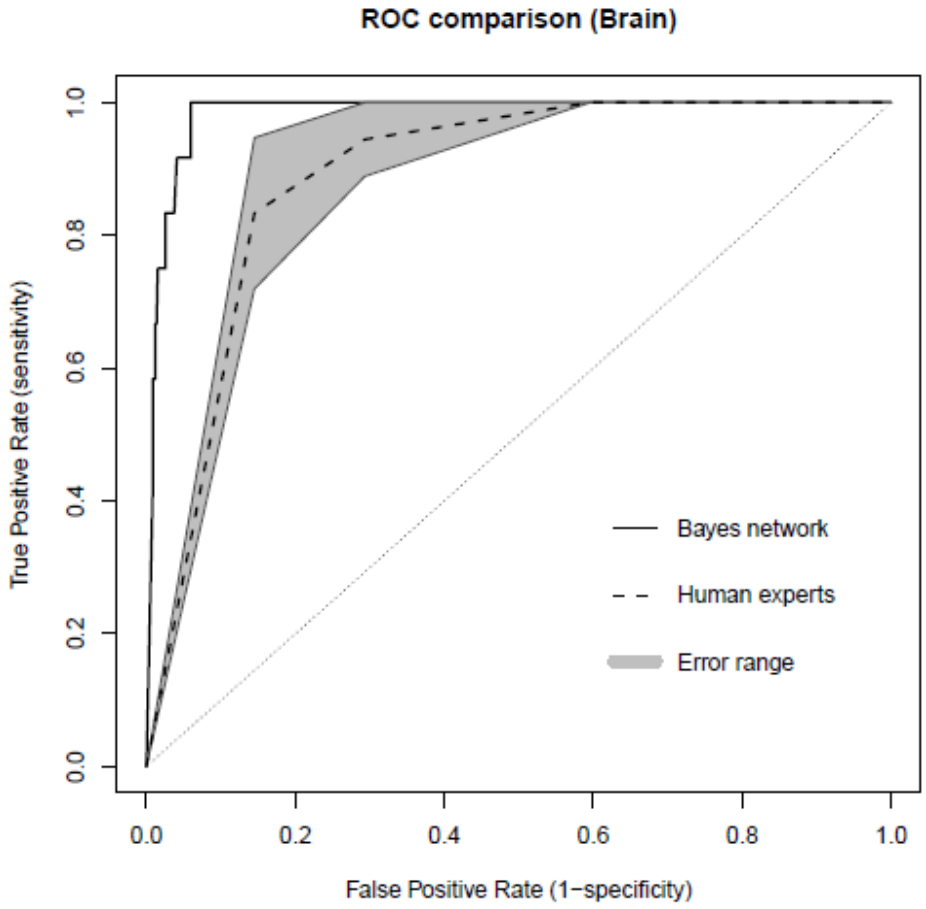


Figure 3.6: ROC curves for brain tumors comparing the performance of the EDN with that of the experts. The expert responses were aggregrated (dashed line) and plotted with the SDOM of responses (shaded grey). AUC for the aggregate human response was 0.90 ± 0.01.

given set of plan variables certain small collections of values can result in large probabilities, whereas the much larger set of possible combinations result in low probabilities. Thus the algorithm does an efficient job of identifying the few likely combinations in this high noise data set. Another strength is the use of clinical data to determine the probabilities. Since the CPTs are created in a data-driven manner, the resulting performance of the Bayesian network reflects the policies and practice of the particular institution or group of physicians. Additionally, as new standards of practice are implemented clinically, the probabilities will

shift to reflect these changes. Perhaps most importantly, the method described stems from, is dependent on, and ultimately encapsulates the knowledge and judgment of the experts. It also reflects the uncertainties and range of practice inherent in real clinical environments.

Moving forward, a significant challenge is scaling up my approach to include more variables. The current model included nine variables chosen, to some degree, by a limited ability to extract data from our OIS. Constructing BNs comes with challenges both in developing topology and populating the CPTs. There are both manual and automated methods for addressing these challenges. Here, I will discuss the CPT's, whereas the network topology building problem is the topic of chapters 4 and 5.

Manual methods for elucidating probabilities from experts, given an existing topology can produce fairly accurate probabilities, but are well known for being time consuming for developers and domain experts [60]. In cases where network variables have sets upwards of 100 possible states, it becomes impractical to query human experts, making automated methods for populating the network tables more attractive. Though the number of probability distributions required to populate a CPT in a BN grows exponentially with the number of parent-nodes associated with each particular table (thus increasing the computational power needed to run the expectation maximization algorithm), I do not expect the performance of the system to necessarily suffer. It is intuitive to anticipate that adding more arcs to the graph could reduce the likelihood that certain combinations will cause posterior probabilities to increase significantly. However, we anticipate that future work will increase the number of variables rather than increasing the number of arcs per variable. One way to reduce dimensionality is by building separate BN's for different disease sites. Moreover, we can leverage the transparency of variable distributions as a strength of the BN formalism to perform VOI analysis and evaluate the entropy changes as a function of each variable in the network. In this way, we retain only the parameters of the model which deliver high utility to decision making. This will also make the calculation of probabilities a more tractable problem. In the VOI presented in this work, the relatively low influence of `Laterality` may be due in part to

the node's d-separation[4] in the topology, but this result also suggests that laterality of tumor location simply is not as clinically relevant to the planning factors. In future iterations of these network topologies, less impactful nodes can be removed to save computational effort. The more critical nodes also indicate areas where clinician's information entry into clinical database systems is most important in terms of predicting outcomes and capturing errors. This information can be used to guide workflows towards improving the effect.

Just as for the CPTs, some machine learning algorithms can be used to derive network topological structure, given a very limited number of network node connections and states [1, 2, 3, 10]. A disadvantage of automatically learning structure directly from data is that correlations can be construed into causality by the learning algorithm, leading to dependency relationships which are nonsensical in the real world. These learning algorithms also suffer from having NP hard complexity [4, 5]. Due to this algorithmic complexity, large data sets become computationally infeasible to employ as scale increases. For these reasons, I used a manual method for generating our initial topology. This also informed my decision to formalize our topology into semantic structures as mentioned in section 3.2.1. Adding new sets of dependent concepts to a semantic formalism amounts to adding one item to a larger list, independent from all other concepts in the list. Our list structure was defined in the ontology building software Protégé [61]. Future networks can then be automatically extracted from this ontological list of dependent concepts—a much simplified task compared to the process of adding arcs and nodes to a large directed graphical network.

The comparison of the algorithm's performance relative to expert performance is not intended to be a conclusive result. As discussed above, the network itself is limited so that it is difficult to judge whether the relative paucity of information in the test hindered human judgment. The purpose of the comparison was to determine if my method was representative of realistic conditions. My results suggest that it is, particularly given the fact that humans performed similarly, but not identically. Despite the relatively low number of participants,

[4]See [19] for more details.

the magnitude of variations in human performance gives some justification that the test cases were neither too simple nor too hard.

Figure 3.5 shows ROC curves for three different tumor sites. The curve for brain cancers reaches 100% sensitivity at a relatively high value of specificity, whereas the other two curves never reach that level. Whether this has to do with the nature of the disease, the state of the art for treating these tumors, or just reflect the practice of the physicians involved is unknown.

The ROC curves illustrate an important point in this decision making scenario. Making the threshold for an alert to an error too sensitive will lead to a reduced number of missed errors but may become burdensome due to the effort needed to weed out the false positives. Conversely, a higher threshold will increase the specificity while simultaneously increasing the false negative rate, and hence, reducing the chances of detecting errors. Ultimately, this aspect of the algorithm will need to be integrated with software to allow the user to tune the system to their personal preference which may vary depending on the tumor site.

As stated above, this software is intended to be part of an entire suite of error-detection processes, one of which is rule-based. The BN model's transparency and VOI analysis are advantages which complement the rule-based approach. With the ability to compute mutual information among node pairs, I not only identify errors, I can also identify investigative clues to the possible sources of errors when they are flagged. In that sense, the BN approach to modeling radiotherapy planning also has implications beyond error detection. The topology of the network contains knowledge about treatment planning and processes as discussed in section 3.2.1. This level of knowledge can be employed to guide forward planning of radiation therapies by modifying the initial query of the decision model from "given a clinical evidence set X, what is the probability of observing therapy planning set Y?", to "given a clinical evidence set X, what therapy planning set Y should one create?". Though much more detail would be required for the BN to answer that question with sufficient confidence, the model development for both applications remains the same.

My plan for future work is to implement some variation on a rules-based system to single

out plans with errors that violate hardware or departmental (protocol) constraints. Plans that pass this step will then be examined by the BN. As I have it designed now, the user would be given a choice as to which BN is appropriate. Other options, such as thresholds and sensitivity, could be chosen as well. The output of the algorithm still needs to be designed to make it most useful. Finally,the network should be tested using data from other centers to explore the extent that the structure and probability tables are compatible across different clinical practices.

3.5 Conclusion

This chapter presents one aspect of a planned error-detection system. I described the construction and use of a Bayesian network model of the radiation therapy plan. By populating the networks entirely with data captured from a clinical oncology information management system acquired over the course of several years of normal practice, I was able to create accurate conditional probability tables with no additional time spent by experts or clinicians. These probabilistic descriptions of treatment planning allow verification of treatment plan parameters to be within the normal scope of practice, given some initial set of clinical evidence and thereby detect for potential outliers to be flagged for further investigation. I have presented a proof of concept that probabilistic networks can identify a large proportion of potential clinical errors while minimizing false-alarms. Future efforts in this area include expansion of the network models to include additional treatment parameters not considered in this work, e.g. extracting dosimetric data from treatment planning systems, as well as integration of the error flagging algorithm into a usable interface for error checking and evaluation of treatment plans in real-time. The success of this work and the the topology generated for the error detection network informed the design of the ontology and the ontology based approach to the topology problem presented in the next chapter.

Chapter 4

AUTOMATING NETWORK TOPOLOGY CREATION: AN ONTOLOGICAL APPROACH

4.1 Introduction

In this chapter I describe the methodology and specifications for creating a dependency layered domain ontology and the computational methods that leverage this type of ontology to generate network topologies of interest. I demonstrate each component of the methodology with an example. First, I present a dependency layered ontology which encompasses a portion of the radiation oncology (radonc) domain. Then, I present a Bayesian network extracted from the radonc ontology. These tasks comprise a proof of concept that the methodology presented here is a viable means of auto-generating dependency networks from complex medical domain ontologies.

4.2 Biomedical ontologies

Formally speaking, ontologies are "explicit specifications of a conceptualization [where] definitions associate the names of entities in the universe of discourse (e.g., classes, relations, functions, or other objects) with human-readable text describing what the names are meant to denote, and formal axioms that constrain the interpretation and well-formed use of these terms" [62]. Essentially, ontologies can be used to compute on, act as storage/retrieval for, and serve as a standardization for various types of knowledge.

There are various types of ontologies and their uses inform their structure. For instance, the Foundational Model of Anatomy (FMA) strictly contains anatomic structural entities and partonomic relations [63]. The FMA object properties consist of spatial locations and relative locations to other anatomic objects. Other ontologies simply try to organize terminologies

or vocabularies into hierarchies (SNOMED-CT, ICD-9/10, etc), and have few relational properties beyond definition slots or mappings to other terminologies [64, 65]. There are ontologies which are designed to handle meta-data mapping to aid in case based reasoning [66, 67]. These meta-data ontologies take semantic similarities in one problem and apply them to new, unknown cases of classification.

Ramoni et al. provides a breakdown of how ontologies fit into medical knowledge with regards to Knowledge Based Systems (KBS) in terms of two levels: epistemological and computational[68]. Though one could argue this classification scheme of this breakdown isn't in itself very well formalized, what is clear (with respect to this thesis) is the separation between ontology domain knowledge and the translation of that knowledge into computational systems that can operate on data. In this organization of knowledge representations, what is not shown is how to move from one level to the next. The arrow in Figure 4.1 represents a translational movement between levels, in particular, between ontology and influence diagrams (Bayesian networks).

4.3 Leveraging domain ontologies

To leverage the benefits of ontologizing medical domain knowledge for research and reuse in Bayes nets, one must encode the kinds of concept dependency relations found in Bayes nets into the ontology, and then translate them into networks, e.g. one must formally define the arrow in Figure 4.1. The topology (the nodes and edges) of any Bayes net can be represented in an ontology—nodes become classes, and dependency arcs become relationships between these classes. Traditionally, a BN has no formal representation of knowledge at the epistemological level whatsoever, that is, they have no ontology. In this work, I supply both the ontology and the means to translate from the ontology to the computational level of a particular Bayes net.

Some work has been done in this area to support the idea that this is possible. Devitt et al. for example, created an ontology of Bayesian concepts and merged it with an ontology of domain concepts in order to create networks of interest[13]. In particular, it uses an

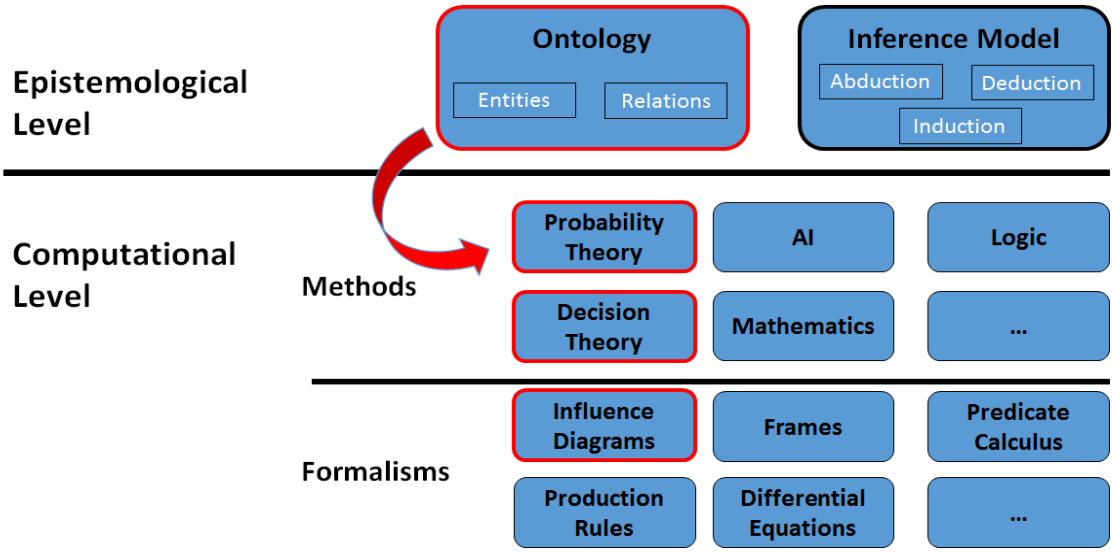


Figure 4.1: Two-level representation of knowledge based on Ramoni et al. The arrow represents a translation of knowledge from epistemological formalism (ontology) to computational formalism (influence diagrams and Bayes nets)

additional ontological layer for the Bayes net models, and applies concept instances from concept sets in a telecommunication network. This additional layer of modeling might be considered restrictive in a medical domain, where the structure of a network model itself can vary significantly depending on the application. Others have utilized a "security ontology" for the Bayesian threat probability determination[69]. In the medical domain, Helsper and van der Gaag have built BNs by creating an ontology for esophageal cancer staging[70]. These applications are relatively specific, however, they validate the theoretical aspects of the method. Among the remaining questions then is how to scale and expand the method towards large, complex medical domains such as radiation oncology.

It is reasonable to ask if there are existing viable knowledge sources or ontologies from which to construct dependency models. In fact, there are ontologies in oncology, notably the National Cancer Institute thesaurus (NCIT) and the ACGT master ontology for can-

cer research, though these are mainly terminologies with somewhat complex class-subclass hierarchies[71, 72]. A. Miller developed an ontology for radiation oncology specifically[73]. The ontology developed by Miller relies on a unique clinical markup language and hierarchy most of which is designed for patient management goals in clinical trials/research. Another radiation oncology ontology, the Radiation Treatment Guideline Ontology (RTOG) attempts to help clinicians more efficiently use clinical publications by representing them in an ontological framework. In this manner, data from clinical publications are then extracted using the framework to produce graphical representation of treatment recommendations [74]. Some of the concepts and terms in these contemporary ontologies might be reusable, (Clinical Finding, Complication, Dose Volume, etc.) but none of the ontologies quite have the semantics required to create meaningful dependency statements among those concepts.

One can also consider deriving an ontology or set of relational networks from metadata itself. Methods for making a Bayesian network directly from relational models[75] and for mapping relational databases to RDF are available[76]. The challenges with applying these methods to EMR data are multifaceted. The schema used in many clinical database systems is far from aligned with an appropriate class-subclass structure, the terminology is not standardized, and the datasets often are too large to be handled by these algorithms in a reasonable time frame.

In order to extract dependency network knowledge from an ontology, and thereby automated topology extraction from the ontology while minimizing deviation from existing ontology efforts to date. In addition, what the ontology needs to represent more than the current slate of knowledge bases (i.e beyond partonomic and declarative statements about what things are) is the kind of *process knowledge* found in published literature and clinical workflows regarding concepts and decisions made in clinical environments. The statement in Chapter 3, "a patient presents with clinical findings, which causes a decision to be made about the prescription, which is then implemented on the radiation delivery device", is a description of a causal process. Representing these types of processes in an ontology by using dependency semantics connects the epistemological level to the computational level, e.g. it

formally defines the arrow in Figure 4.1. In this sense, my work results in the establishment of a new standard. By collating knowledge across these sources into a single ontology, I standardize terminology as well as organize the knowledge into a class-subclass hierarchy.

4.4 Constructing a dependency layered domain ontology

I demonstrate these ontology construction methods to represent knowledge in the radiation oncology domain. Ontology construction requires defining classes in the ontology, arranging the classes in a class-subclass hierarchy, defining relationships and then filling in the relationships among classes. As an iterative process, it can take some definition cycles to bring the ontology to a useful form. The ontology is constructed with two operational goals: 1) to organize domain knowledge into a navigable class-subclass structure, and 2) to establish dependency among domain concepts. Tertiary goals of the domain ontology include 1.a) annotation of classes to existing ontological terms to maintain the consistency this work with those efforts where possible, and 2.a) annotation of the dependency relations to their original knowledge source. With respect to the above goals, I also specify the level of knowledge this ontology is required to cover, i.e. what depth of class structure is most useful to declare dependency. Much of this depth leveling is defined by the current knowledge sources available.

For my work, I have chosen to use Protégé-Owl, for ease of development and for expressivity. However, my work to date does not leverage description logic inference, so other simpler formalisms could also have been used. I resolve any knowledge inconsistencies by consulting with local domain experts. In the case of circular dependency, such structures are strictly disallowed in the Bays net formalism. It is possible that they truly exist in the domain in some time-dependent form. For example, metastasis can be an outcome dependent on a treatment which depends on the original tumor, but later in time the metastasis could be considered the tumor to be treated, thus completing a dependency loop. To resolve this type of issue I separate certain concepts in the ontology by time (to continue the example case) by creating two classes of metastasis: one for clinical presentation and one as a post-treatment

outcome—recognizing that the ontology is designed to be static.

4.4.1 Class hierarchy

Before establishing any specific sets of dependency arcs, I first define what domain concepts exist and how to categorize them. To accomplish this I take a combination of the"bottom-up" and "top-down" approaches. Determining the bottom level requires consideration of dependency, however. Bayes networks compute on states, and each set of states belongs to a concept. The state parameters are used to derive conditional probability tables in the network and concepts make up the topology, as described in chapter 2. The goal of the ontology is to capture the topology of concepts, not states themselves, as these may vary for each data set source or application of the network. The state level therefore defines a floor.

The top level is organized by greater concept categories commonly used in clinical radiation oncology. Generally, a template dependency network is going to have specific concepts which fall under these broader categories of medical practice. For example, outcomes depend on treatment, treatments depend on diagnosis, diagnosis depend on symptoms, and so forth. This structure dictates the top level category types required, i.e. diagnostic factors, treatments, outcomes, etc. that are used in the domain.

In this ontology, I opt to emphasize breadth over depth in the class-subclass hierarchy due to the focus on building networks from leaf node concepts (bottom level). I seek to keep the hierarchy to a scale reasonable for the user navigation task of finding leaf nodes sets of interest from which to build a network topology. Thus, new branches of the internal hierarchical structure are only created when delineation amongst subclasses is relevant (distinct), have common usage, or aided navigational tractability.

To obtain the specific ontological terms and their relations I examine several knowledge sources including domain experts, published peer-reviewed experiments from well known domain journals, published treatment guidelines, and clinical trials. The knowledge source defines the set of concept parameters. For example, a publication from the journal *Oral Oncology* gives the results of a 62 patient trial of head and neck cancer patients post ra-

diotherapy and analyzes trismus versus a TGF β1 genotype [77]. The hypothesis, from the abstract: "Radiation can trigger an intense fibrosis within the masticatory muscles and transforming growth factor beta 1 (TGF β1) is involved in this process." Trismus is a complication defined by a reduction in the ability to open the mouth. In a network model, the state value possibilities would be "yes/no" for trismus and TGF β1, and therefore the bottm level concepts are `Trismus` and `TGF` β`1 polymorphism`.

Continuing the *Oral Oncology* example, we can say generally that trismus is defined as a type of `Complication` and complications are a type of `Clinical_Outcome`. However, the set of complications is too broad to be reasonably navigable. There are thousands of complications, and typically only a smaller subset of them are of interest for any given treatment. For this reason, the complication class is redefined as `Complication_by_site`, and subclasses are defined per anatomic region: `Brain_comp`, `Pelvis_comp`, etc.

The decision to split certain categories into anatomic subcategories makes sense when considering that radonc is primarily concerned with solid tumors located in a known position of the patient body, yet there are cases where other usage has equivalent weight and suggests added distinction. The practical clinical concern of a complication like trismus, for instance, is not just trismus itself, it is also the subsquent (non-treatment related) complications of that condition, i.e. diffuluty eating—leading to sigificant weight loss and/or further deterioration of health state. For these clinical concerns, I created another subclass of Clinical_Outcome: `Complication_by_system`, subclassed by systematic categories `Endocrine_disorder`, `Gastrointestinal_disorder`, and so on. Most of the bottom level complications are subclasses of both of these superclasses, making them easier to find regardless of the clinical concern (site vs. sytsem). The top few levels of the ontology are shown in Figure 4.2

Some of the challenges building this ontology include differentiation of complications from symptoms, such as a "headaches". To alleviate this conflict, I include only complications which are well defined states or conditions, or that coincide closely with RTOG endpoints/toxicity since those are also are well defined for purposes of outcome reporting in

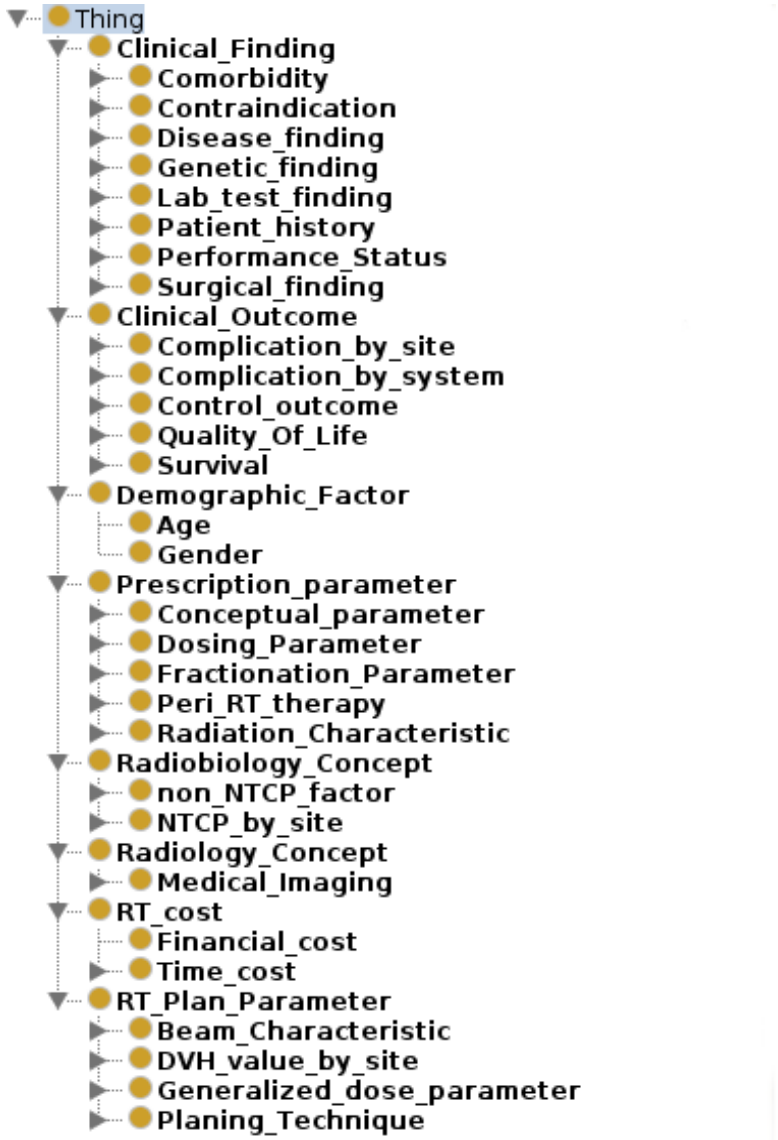


Figure 4.2: A screenshot of Protégé showing the top three levels of classes represented in the radonc domain ontology.

experimental clinical trials. Other issues arise with the specificity of terms. Clinical publications will often use a general term when they mean a specific one. For example "nodal involvement". Because the literature often studies a specific cancer class (say, prostate) the terms must be formalized in context. This means other site specific terms are created in the

ontology building process, i.e. `Pelvic_nodal_involvement`.

A consequence of the building choices made is the lack of any direct mapping or adherence to an existing biomedical ontologys class structure. Mapping of individual terms do at least allow one to connect the ontologies at a local point, and are described in more detail below.

Concept annotation

Annotations for classes having similar or exact matching terms in an existing ontology or standard terminology are added to the ontology as "comments" (annotation property) on the class itself. The most commonly used annotations map to SMOMED-CT, ICD-10, and the NCIT. Each annotation is identified with a specific annotation property under the convention `NCI_id`, `SNOMED_id`, `NCI_defintion`, `SNOMED_defintion`. I map to these sets of existing defined terms to retain some terminological standards. If a concept term cannot be found in an existing ontology, then a comment is used to create the defintion and/or citation to source which describes the term. For example, "Sticky Saliva" is a term commonly used in the radiation oncology literature to indicate a thickening of the saliva. It is the product of glandular response to radiation delivered to structures in the mouth and throat. Though it has no proper medical definition in the existing ontologies, and is closer to a symptom than a disorder or disease, it is used often enough in the literature to qualify as a complication, or complicating factor-similar to xerostomia or dysphagia. Figure 4.3 (top right) shows how annotations appear for the `bNED` term. At times, inconsistent use of terminology between journal publications and existing ontologies, or between ontologies themselves arise. I choose to resolve ontology-ontology conflict by favoring the NCIT, given that radonc is a sub-specialty of oncology in general.

4.4.2 Dependency layer

To represent the dependency arcs, I define a transitive property,"dependsOn". Thus, for knowledge that could be expressed in natural language as "the likelihood of symptoms such as fever and nausea depend on a diagnosis of influenza", I would define three classes {fever,

nausea, influenza}, and two "dependsOn" relationships. In the Bayesian formalism, influenza becomes the parent node to the other two, which are causally dependent on the presence or absence of that disease. Thus, I translate the node-arc-node framework of the Bayes net into an object-relation-object triple formalism that has become standard for semantic web representations.

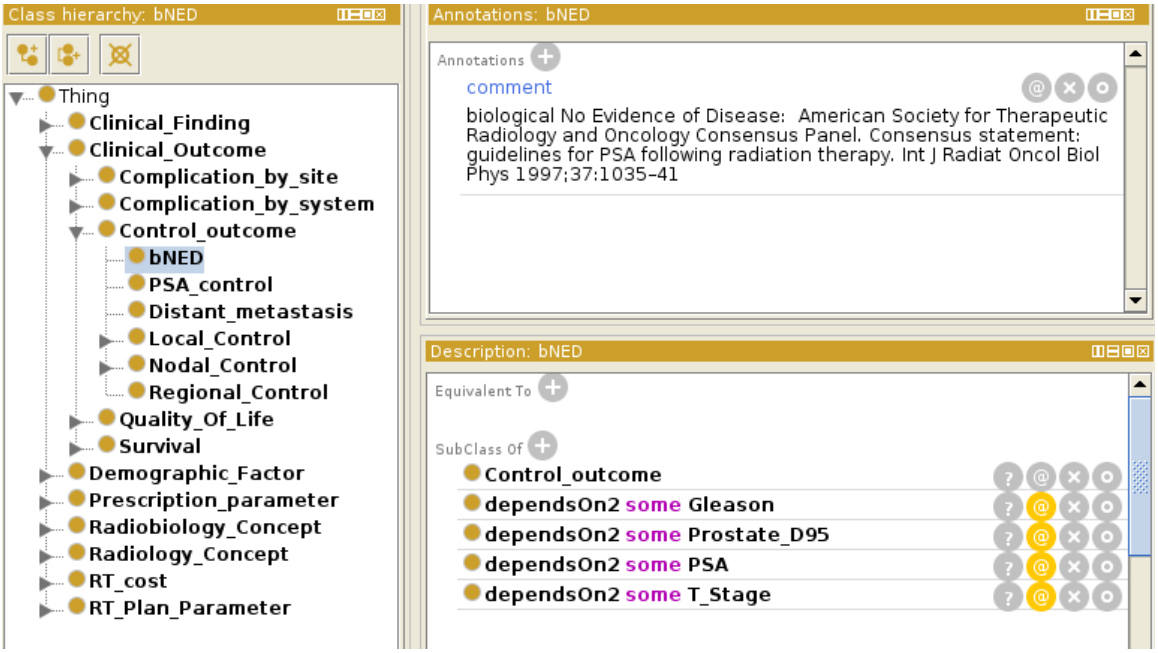


Figure 4.3: A screenshot of Protégé showing part of the bottom level class hierarchy. The annotated term definition "biological No Evidence of Disease" (top right) is given as a citation to the published source which describes the full meaning and context. The set of annotated dependencies to other ontological concepts is also shown (bottom right).

It is recognized that although the Bayes net structure only allows for one kind of dependency, this constraint does not limit the actual dependency type which may be chosen by the researcher to include in a network. Some dependencies are strictly mathematical, like calculable machine parameters based on first principle radiation physics and tissue structure while others are correlated evidentially as observational results of various kinds of clinical trials. Another goal of this ontology is to capture this stratification of dependency. This is

an important feature of the ontology which gives researchers the flexibility to explore the knowledge base at variable levels of causal belief to account for differing goals. In one case, a researcher searching for an extremely robust network for practical clinical use desires less uncertainty in the topology, and may want to only include *strongly dependent* concepts, while other researcher goals such as examination of the knowledge base towards constructing further studies in the domain may want to include (or restrict to) the *weakly dependent* terms. To accomplish this I establish a threshold for dividing dependency in the previously mentioned sources based on level of evidence factors [78]. Formally, this results in a set of object sub-properties, `dependsOn1`, `dependsOn2`, `dependsOn3`, and `dependsOn4`; the numeric tag value corresponding to the evidence level of that dependency.

Most of the dependency between concepts is established by examining studies that control for specific concept states, as described in section 4.4.1. In addition to manually extracting the concepts from the literature, I also evaluate the study for the dependency level, based on level of evidence standards. Conclusions from the published study, guideline, or source are coded into the ontology using the proper "dependsOn" term.

The set of concepts which `bNED` depends, extracted from Horwitz et al. [8] are shown in the bottom right of Figure 4.4.

The nature of classification allows for inheritance of properties by subclasses. In some areas of the domain, it both saves time and is theoretically correct to declare dependency of a superclass rather than enumerate for every specific subclass. For instance, one can declare that the superclass `DVH_value_by_site` `dependsOn4` some `PTV_dose_general`. The PTV dose effects the surrounding doses which constitute the Dose Volume Histogram (DVH). This is a well known physical property of the radiation beam scattering with significant experimental evidence and is also derivable from first principles. Thus, all subclasses of the DVH type (now and in any entered in the future) will inherit this property of dependence. Unfortunately, because the object property applies only to the declared class, it's inverse cannot be inferred, i.e. the ontology does not inherit the dependency to any subclasses of `PTV_dose_general`. For this reason, an additional object property was created, `casualFactorOf`, for declaring

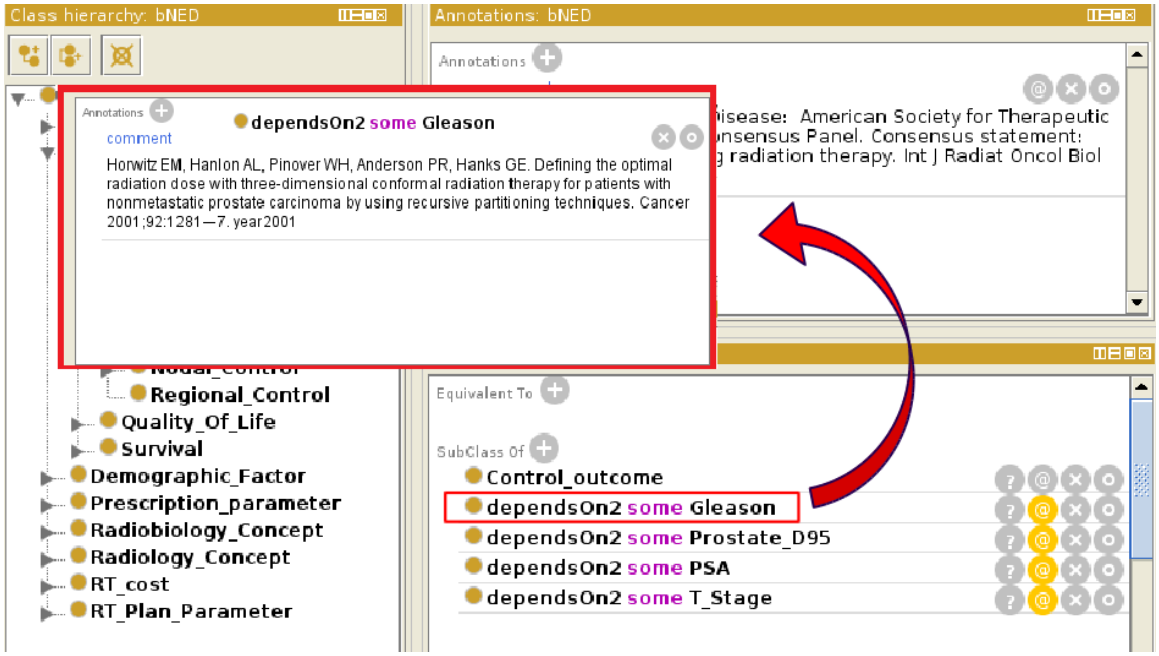


Figure 4.4: Protégé view of radonc ontology annotation for relation bNED `dependsOn2` Gleason.

dependency (in a causal sense) which can be inherited by subclasses and converted in processing to a "dependsOn" type outside the ontology. In this ontology a declaration is made that `Complication_by_site` and `Complication_by_system` are each a `causalFactorOf` the "quality of life" (`QOL`, as quality of life depends on any/all complications.

Dependency annotation

To give legitimacy to the dependency declarations in this ontology, dependency declarations are annotated with the knowledge that describes the dependency. As in the `bNED` example shown in figure 4.3, `bNED DependsOn2` some `Gleason` is annotated with the citation for Horwitz et al.(cited above), where the peer reviewed experimental evidence was published. Currently only one citation per dependency is provided, though there is no restriction to the number of possible sources that can be attached.

4.4.3 Ontology in full view

As ontologies become large, the full set of classes and relationships cannot be viewed well in standard ontology building tools, making it challenging to see the overall picture. For reference, the full radonc ontology class-subclass hierarchy (as of Jan, 2015) is presented in appendix A. Though this dissertation focuses on extraction of subsets from the larger ontology, it is still useful to have a visualization of the whole. Visualizations are intended to organize and display information, relationships, and patterns, and by doing so hopefully help people make choices regarding that information. One way of visualizing an ontology with two orthogonal relations—class hierarchy and dependency relationships—in a way that can be examined all at once is hierarchical edge bundling (HEB) [79]. HEB presents bottom level concepts in a radial display, with dependency paths between concepts as lines. The lines are bundled together according to their location in the hierarchy The hierarchy itself is not explicitly shown but underlies the paths in a radial tree format (the trunk of the tree being at the center of the circle).

Figure 4.5 shows the result of implementing HEB using the data-driven documents object model [80] on an intermediate version of the radonc ontology (transformed to JSON data format). This visualization shows some of the overarching patterns present in the ontology, in particular the inherited sets of dependency among complications, doses to specific organs, and quality of life (`QOL`). Additionally, there are highly bundled paths between the set of complications and `Plan_Technique` as well as `PTV_dose_general`. Despite much of the dependency layer being built at the bottom layer, these emergent patterns suggest that the the higher level medical template discussed in section 4.4.1 is reasonably produced from the knowledge base. Analysis of this sort was performed viewing the HEB form ontology in a browser (as intended). One of the limitations of the HEB is that it cannot handle multiple inheritance, however, in this case, bundling one dependency path along two different hierarchies doesn't aid the visualization since retaining the split hierarchy isn't of critical importance to conveying information about the larger ontology.

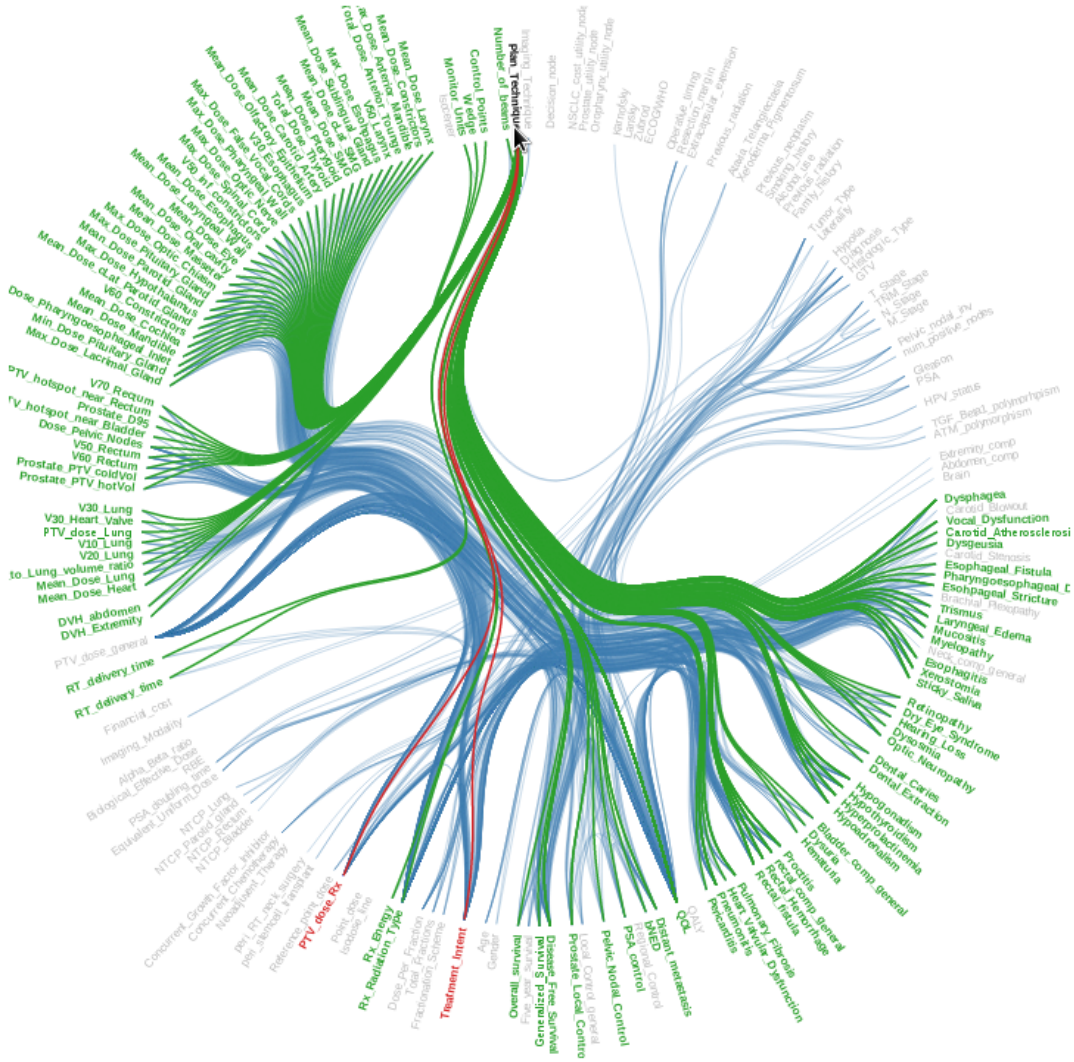


Figure 4.5: Visualization of the radonc ontology using hierarchical edge bundling algorithm. Lines represent dependency between concepts, and are bundled together spatially by the classification scheme. Highlighted here are the dependency paths among terms related to `Plan_Technique`.

4.5 Extracting dependency networks from ontological specifications

4.5.1 Subsetting ontology via concept selection

Auto-generation of Bayes nets from an ontology requires three critical operations: 1) extracting the ontology into a Bayesian form 2) finding all the dependency pathways between

a subset of nodes of interest and 3) modifying the network to remove nodes which are outside the scope of interest.

To accomplish the initial extraction, I employ a sparql query using the rrdf and rhugin libraries available in the R statistical programming language[57, 81] to pull all concept nodes and dependency relationships from the ontology that contain "dependsOn" type relationships. Supplementary to the initial extraction, interpretation is performed on any concepts containing the `causalFactorOf` relation—transforming them into inverse triples with "dependsOn" relations—to keep the set homogeneous with respect to relational properties. To construct a narrow BN from this larger network, a set of nodes of interest $\{ X_1, X_2, X_3 ...\}$ and the extracted ontology is passed to a path finding function which identifies and returns all unique dependency pathways between nodes X_i and X_j $(i \neq j)$. This is performed using an exhaustive level order non-binary search method. It is not assumed that a user would know the topology or directionality of dependence (dependency being non- commutative) *a priori*. However, this knowledge is specified in the ontology and so accounted for in the code. Paths are always returned with correct directional dependence.

To construct a narrow BN from this larger network, methods of network pruning based on an adjustable user input selection of individual nodes are devised. First, I develop pathway discovery functionality by applying deductive reasoning along paths of dependent nodes. In a pruning stage, I apply a node removal process which retains the dependency between edge nodes in a layered dependency. Because dependency relations are by definition transitive, all pathway diagrams in the set are commutative. For the d-separated nodes of interest (A, C) in a relationship $A \rightarrow B \rightarrow C$, given no information about the node B, dependency is retained between A and C by way of deductive reasoning. Removal of B means generation of new dependency relationships between all parents of B and all children of B, and then deletion of B from the set along with preceding dependencies in the local network resulting in $A \rightarrow C$. The task of removing multiple nodes, i.e. removing (B, C) from pathway $A \rightarrow B \rightarrow C \rightarrow D$ is nothing more than a matter of applying the one parent node pruning method iteratively. As I describe in Chapter 5, I have developed software that implements this approach.

4.5.2 Subsetting the ontology via dependency layer

In a self consistent knowledge base, the pruning principles are applicable in pruning networks based on any selection criteria including the dependency relations outlined in section 4.4. With a stratified dependency layer, a conflict arises as to what the resulting strength of the final dependency arc ought to be when, for example, a set of two arcs `dependsOn2` and `dependsOn4` are to be combined. Without resolving this interesting epistemological question in deductive reasoning, a conservative approach is taken whereby the "dependsOn" of weakest evidence level is used. The approach claims that a (logic) chain is only as strong as its weakest link.

Since a variety of object properties, annotations, and other information can be included in an ontology, subsetting and pruning based on these factors can generally also be done in a similar fashion. Finally, I note that the functional design is not domain-unique, that is, these functions are operable on any OWL ontology formalized using the object properties outlined in this dissertation.

4.6 Application to other medical domains

An important way to demonstrate the possible breadth of this methodology is to consider whether it can be applied in other medical spaces outside the radiation oncology domain. We ask, how would one add dependency layer to some other ontological formalism, or use another similar formalism to construct dependency networks? Consider the disease ontology. The disease ontology (DO) is designed to connect disease concepts, genes contributing to disease, symptoms, findings and signs using well defined, standardized resources [15]. Looking closely at a specific disease from the DO, say `hepatitis B`, a definition can be found which incorporates some unique properties:

> Hepatitis B:
>
> A viral infectious disease that results_in inflammation located_in liver, has_material_basis_in Hepatitis B virus, which is transmitted_by sexual contact, transmitted_by blood

transfusions, and transmitted_by fomites like needles or syringes. The infection has_symptom fever, has_symptom fatigue, has_symptom loss of appetite, has_symptom nausea, has_symptom vomiting, has_symptom abdominal pain, has_symptom clay-colored bowel movements, has_symptom joint pain, and has_symptom jaundice.

The kinds of relations used here could be interpreted for possible diagnostic applications. The terms `transmitted_by` and `has_symptom` might be considered a "dependsOn" type, as hepatitis B is causal of those particular symptoms. This would constitute a small directed network graph for hepatitis B. A translation of the disease ontology at large would allow a more useful network to be constructed that represents more closely the diagnostic tasks expressed in clinical medicine. That is, one could construct a network of various diagnostic patient factors (symptoms, patient history, recent exposures, etc), and a range of possible causal diseases. Other properties in the disease ontology that might be useful to develop causal networks include `results_in, results_in_formation_of`, and `effects`. The disease ontology is not currently in a format where these relations have been translated to description logic object properties, so some preprocessing of disease definitions would need to be performed. Nonetheless, porting the dependency network extraction methods to operate on a version of the disease ontology and aid in creating probabilistic diagnostic models is entirely plausible. I provide a more thorough treatment of this idea in Chapter 6.

4.7 Discussion

The methodologies outlined here are combined with a software user interface tool I built to make the task of developing Bayes nets or other dependency network topologies easier. The software tool is described in the next chapter. Compared to traditional development methods, computational methods of extracting networks reduces the domain expert time required of the process. With this knowledge base format, extracting a Bayes net and it's associated knowledge source citations can be condensed to the simplified task of selecting relevant

concepts from a larger list. Additionally, the ontological format provides the flexibility to add new object properties and classes at will. This leaves the ontology open to including new potentially useful relations such as time-dependencies, functional (deterministic) relations, or personnel associations describing work-flow processes, to give a few examples.

Despite the fact that the initial goal of this knowledge base is to generate Bayes networks among medical concepts, there is nothing in the ontology itself which specifies *probabilistic* dependency. Therefore, deductive reasoning operations can be performed on the ontology knowledge itself. But to what end? One thing users can do is use the ontology itself to learn about concept relationships in the domain without having to scour the literature base. It is possible for a user to quickly construct a set of literature and informational sources regarding a topic of interest, say, a review paper on the complications of head and neck radiotherapy. In that sense, the ontology acts as a meta-level knowledge source. The rich semantics in a dependency layered ontology connect domain concepts in a way not found in search engines and thus give it some advantage over those types of resources. As such, a learning tool could be built on top of this ontological knowledge base to aid in training medical residents or clinical researchers.

4.7.1 Continued ontology development

Though there are clear benefits to using an ontological formalism, its use also forces one to consider its technical and/or practical limitations and explore the possible space of solutions. For instance, the sheer volume of literature available currently in addition to the ongoing advances of domain knowledge (i.e. studies which advance the level of understanding of a particular dependency, or which increase/decrease the dependency level) may render manual curation of a domain ontology insufficient—or at the very least a task which must be supported in some meaningful way. In the future, it may be possible to address this by employing natural language processing (NLP) tools to create some kind of supervised agent to extract the appropriate semantics (class and dependency) from domain literature. This is likely to be a challenging task in and of itself (and is currently at the forefront of medical

informatics research).

4.7.2 Algorithmic limitations

When considering alternative state levels from which to construct a network, a few limitations on network subsetting arise. The algorithm in its current form does not allow construction of a network where the data is a level above the bottom state level chosen here. For example, if I desired a Bays node `Comorbidity` where specific subclasses such as `HPV status` are my desired network states, not my network nodes. This level of concept selection is not currently supported because that level of generality is not often studied/published. As previously noted in section 4.4.1, existing studies focus on specific variables. However, this is not *always* the case, and since data can be generated from EHRs, etc. it is possible this could be a useful feature. Future improvements to software tools could include this level of concept extraction, requiring some technical differentiation between user selection of `concept A` to mean *only* `concept A` versus the current *all leaf nodes descended from* `concept A`.

4.7.3 Scalability

Its is not known what level of scalability can be achieved. As a domain ontology grows with respect to it's number of dependency connections, so does the pathway algorithm's search time. At what point does traversing the domain ontology hierarchy becomes unwieldy? Both breath first (used here) and depth first search algorithms scale similarly, thus, solutions beyond the current search-on-demand type might be required. One possible solution to this problem might rely on the latency between ontology updates. If updates to the ontology are not frequent (with respect to user demand frequency), then it might make sense to construct and tabulate all possible network models after a substantive ontology update. The software could then simply select the stored model corresponding to the selection input. Pruning and other changes would then occur on the sub-model with minimized computational time. The overall computation time would obviously be increased, but with most of it being ported to an update cycle instead of it being done on the user's time.

4.8 Summary

I have described in this chapter the methodologies and technical specifications for developing and extracting dependency networks from ontological specifications. These specifications can be utilized to perform the specified computing tasks outlined in section 1.2, namely, applying deductive reasoning on ontology triples to subset consistent and compatible Bayesian models. While some limitations have been identified, most of them appear to be solvable using technological tweaks without impacting the broader advantages of using an ontology based knowledge hub for storing domain knowledge and dependency semantics. In the next chapter, I describe the software tool that I built which applies these methodologies to perform network extraction from dependency layered ontologies.

Chapter 5

A SOFTWARE TOOL FOR BAYES NET DEVELOPMENT

In this chapter I outline the details of a software tool designed to give users a simple graphical interface for extracting dependency network topologies from dependency layered ontologies (as described in Chapter 4). What is important about this tool and how it is different from other software like Hugin or Netica is that it derives knowledge directly from a dependency layered ontological knowledge base and therefore does not require significant user input towards developing a topological structure. There are many features which this tool does *not* have, such as the ability to create conditional probability tables or perform any sort of learning algorithms on data. However, topologies created with this tool can be exported and read into other systems such as Hugin for those tasks, making this tool complementary to other software systems for Bayes net development.

5.1 Goals and architecture

I developed a user interface prototype, the "Bayesian Network Domain Explorer" (BNDE), with a few key goals. The first goal is to allow for users to reasonably navigate an ontology class-subclass hierarchy and select concepts of interest. Secondly, the software needs to generate and display the resulting networks and information about those networks, as built from the ontological knowledge base and user selections. Lastly, the software has to provide user controlled network pruning and downloading of resulting network topologies in useful formats for further use. In accordance with these goals, I also seek *platform independence*, that is, to limit the software's dependence on operating system, hardware, or other device types.

To meet these goals I use the `shiny` package for the `R` statistical programming language.

There are several advantages to using `shiny` for this software. `Shiny` is a web application framework for `R` and can be run on a local server tied to `R` processes [82]. This allows for fairly simple development of web-based user interface tools while retaining access to powerful computational `R` libraries. Web-based also means that many widgets, buttons, layouts, graphical methods, javascript and css stylesheets already available via open source projects such as bootstrap [83] and other webtools can be utilized. Because shiny applications are web-based, they are also fairly platform independent. Shiny applications work on almost all common contemporary web browsers, allowing users to apply their local browser settings, configurations, and schemes independent of the application. Shiny also uses a reactive programming model which effectively performs real-time updating of graphical output (or any other reactive functionality) based on user input.

The essential model of reactivity in `shiny` applications is shown in Figure 5.1. A thorough overview of reactivity is given in RStudio's article on reactivity [84]. I present a shortened version for context. Client-side inputs are tied to server-side outputs (or any other descendant server-side reactive functions). When the input changes, the descendant functions are invalidated. After all invalidations are finished, the outputs (graphical display, dynamic UI, etc) are flushed and redrawn using the new user inputs. This event scheduling occurs on the order of milliseconds. The shiny server checks for input changes frequently—effectively updating changes with high responsiveness. If an underlying server-side function needs to run for some time, there can be delays in re-drawing the display. Fortunately, there are many available features in the development set to provide client-side user feedback during server-side computing processes.

In order to meet the desired specifications above I employed the `rrdf`, `RHugin`, `shinysky`, `shinyBS`, and `shinyIncubator` packages as well as a set of developer defined functions. The core functionality relies on a few computational tasks:

1. parse description logic ontology
2. search for dependency paths between nodes

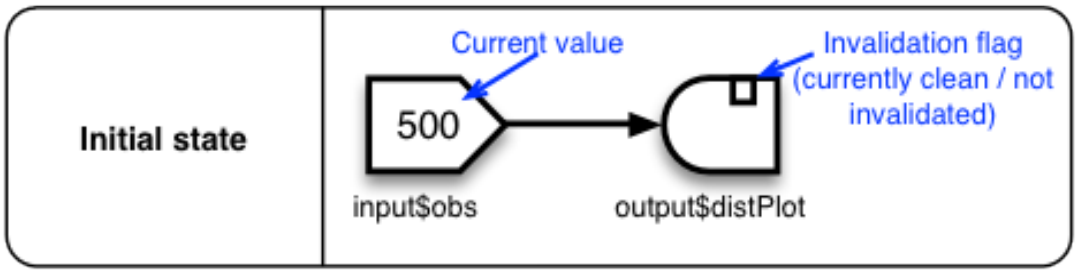


Figure 5.1: Shiny reactivity-initial state with input (`input$obs`) and a valid output (`output$distPlot`). The output `distPlot` gets invalidated, flushed, and re-executed upon changes to `input$obs`

3. prune the network nodes

Description logic ontologies can be read into `R` as a set of rdf triples, however, they are represented as a Java class. To parse the ontology into a more easily searched matrix of dependency triples, sparql queries of the form below are employed to extract all the terms in the ontology which have dependence on other terms. All terms in the new triple sets (for each type of dependency) are `SELECT`ed and collated to form a 3xN matrix.

```
CONSTRUCT {?object2 ro:dependsOn3 ?object1}
WHERE {?object1 rdfs:subClassOf ?restriction.
?restriction owl:onProperty ro:dependsOn3.
?restriction owl:someValuesFrom ?object2.
?restriction ?restrictionPredicate ?object2.}
```

From the larger ontology, the software now has a subset of triples to operate on. Here, a level-order non-binary search method (breadth-first search) is used to traverse the set of triples to discover pathways between user defined subset of terms. The minutia of both this method and network pruning operations are outlined with more detail in chapter 4.

Marrying these core operations to the user interface requires assigning a series of reactive elements. A broad overview of the architecture which links UI inputs/outputs between the server, client, and client local filesystem is shown in Figure 5.2. The user interface (UI) is tied to R processes running on the server—gathering inputs and sending back outputs for displays or requests for uploads/downloads. There are ontologies loaded on the server filesystem, but options are available for users to upload their own ontology to the server and access it using the UI running in the client's local browser.

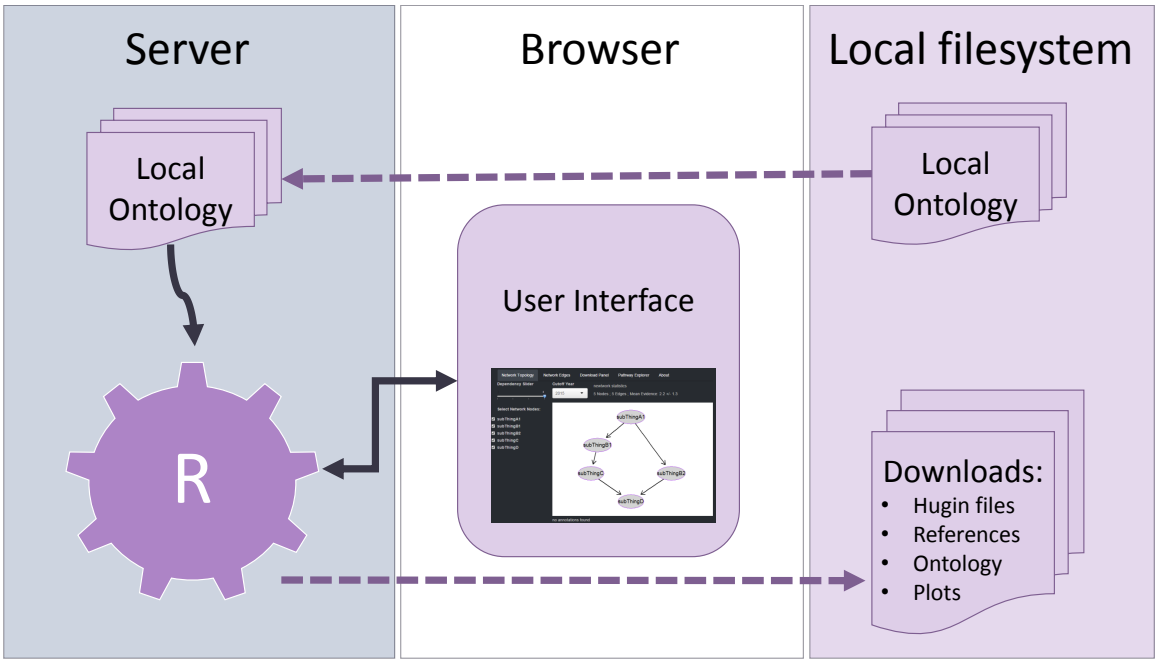


Figure 5.2: Dynamic information flow between the server, client, and local filesystem in the BNDE

Because the core functionality of this tool is to perform logical deductive reasoning among dependent terminologies and express their topological connections, it can technically be applied among non-probabilistic variables as well, such as those found in computer programs. Therefore, if the application is given an ontology of its own functional dependencies among variables and reactive elements, it can describe itself! Figure 5.3 is a directed graph network representing all chained dependencies among the various inputs, outputs, and functions used

in the BNDE to generate a network display. This was generated by uploading a dependency layered ontology representing the BNDE. Concepts are prefixed to indicate their superclass (user interface input/out `GUI_in:/GUI_out:`, or developer/reactive functions `DFN:/RFN:`). The `RFN:Z` function generates the network graph object used by several components of the software including a graph display function (`GUI_out:bnPlot`) at the bottom which generates the graph display output. A dependency graph of the entire BNDE can be found in Appendix B. At the time of writing, the BNDE ontology is available in the server-side local ontology selection list, that any brave of heart transparency seeking enthusiast might peruse to understand better how the software application works.

5.2 User interface development

Much of the design for the BNDE software was guided by the fact that the author of the software is also a potential future user. The layout of the tool is limited to some extent by screen space in general, regardless of device. The basic design is intended meet the goals of network development, i.e. the information to be displayed is mainly classes, networks, and some widgets for network pruning. However, this leaves little screen real-estate for other planned features. Therefore, this software could not be programmed as a flat-page site. This informed the choice to use a tabbed environment, where additional features could be placed into other tabs. Having a tabbed page also lets users stay on one site while working and not have to reload any pages or refresh any information.

While some parameters can be static, others are functionally tied to network properties in a dynamic way. For example, the displayed node names can be longer than the size of the node circle and become difficult to read. So the node `width` parameter is set as a function of the length (in characters) of the node name. Css stylesheets are used to maintain legibility of text, as well as minimal, gentle color schemes. Contrast is used to improve understanding of presented information. I found it important to keep required user actions as simple as possible, and provide *useful* feedback wherever issues arise.

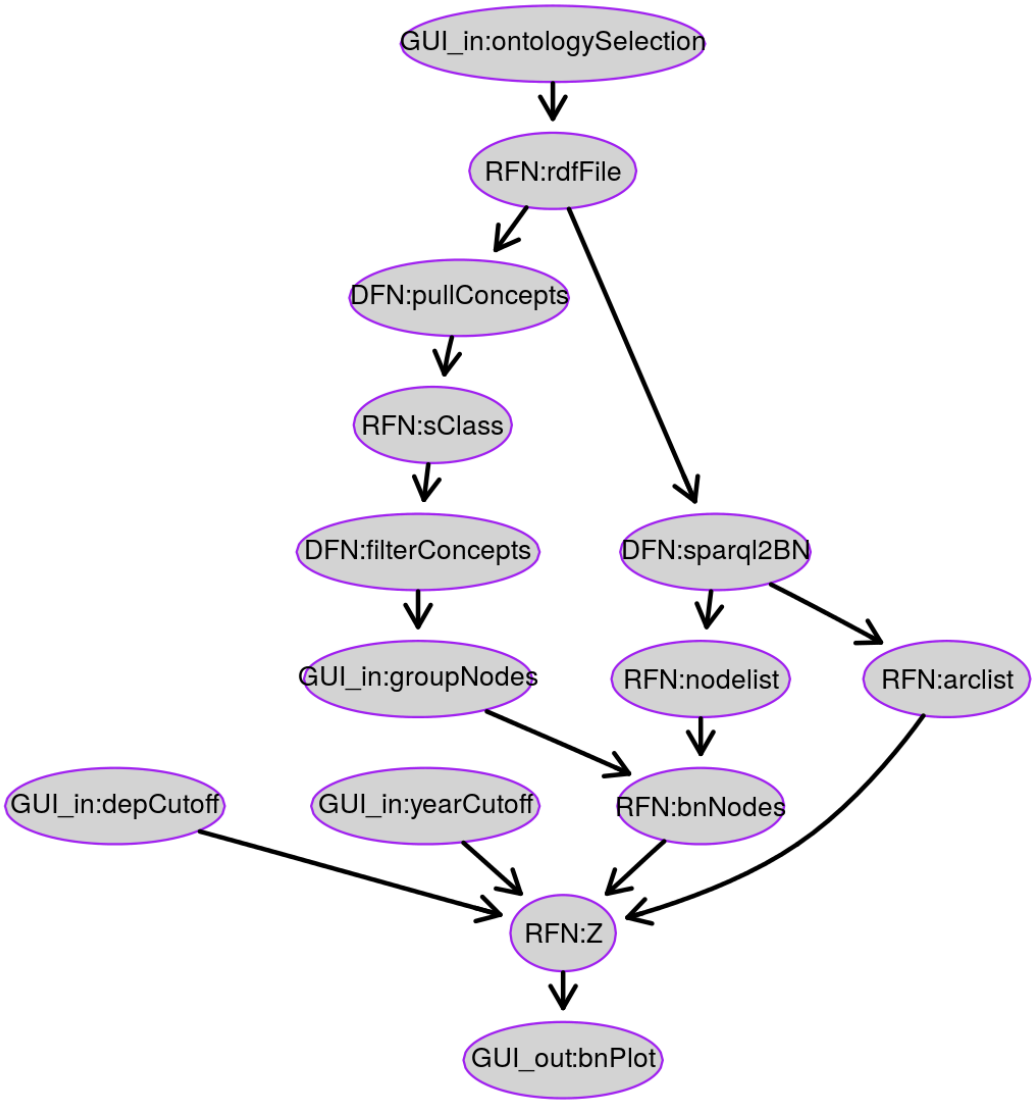


Figure 5.3: Functional dependency graph among variables in the BNDE used to generate network displays

5.3 Software features

The primary use of this software is to semi-automate the construction of dependency networks (Bayesian networks) from a domain ontology knowledge base. The domain ontology specifies

what concepts are dependent on others, so it is not necessary to have *a priori* knowledge of what they are. When the user selects a set of concepts of interest, the tool will automatically create the network nodes and arcs in a graph object and display the structure of the network, among other things (described in more detail below). This eases the development of Bayesian networks in that the user need not reconstruct a network from scratch for every use case, nor search the literature or interview domain experts to establish an appropriate network structure. The interface is also interactive, thus, changes in concept selection, dependency level, and other parameters result in real-time updating of the graphics and other reactive features.

5.3.1 Ontology selection

In order to use this tool the user must select or upload an ontology. The main sidebar panel has options to do both. Here, a dropdown menu has been provided and a few preloaded ontologies can be selected from. Uploaded ontologies will appear in this selectable list of available ontologies after successful upload. A simple ontology ("testontology.owl") is available for experimenting with software features. These ontologies contain object properties that define the dependencies between classes.

Once an ontology is selected, the software will automatically read the class-subclass hierarchy and recreate it in the main sidebar panel as a folder-tree. It is generated from the ontology at the time of loading by using a recursive function that traverses the ontology's class-subclass tree to recreate the class hierarchy. This folder tree structure is an instance of (javascript) jstree and is therefore interactive and selectable. Though a few ontologies are available locally, the uploading feature gives users the flexibility to use these tools on any ontology constructed with the dependency layer protocol e.g. object properties of "dependsOn" type.

5.3.2 Network topology

With an ontology loaded and a folder-tree built, a user can select among the various concepts shown in the folder tree. Holding the 'ctrl' key selects for multiple concepts. Selection of a concept or category of concepts will add it and any sub-concepts to the checkbox list just to the right of the folder tree. Some ontologies can be too large to view easily in a folder display, so this list is here to keep track of all the things are have selected so far. The checkbox group is also selectable, so network nodes can be removed and added from the graph without re-navigating the folder tree.

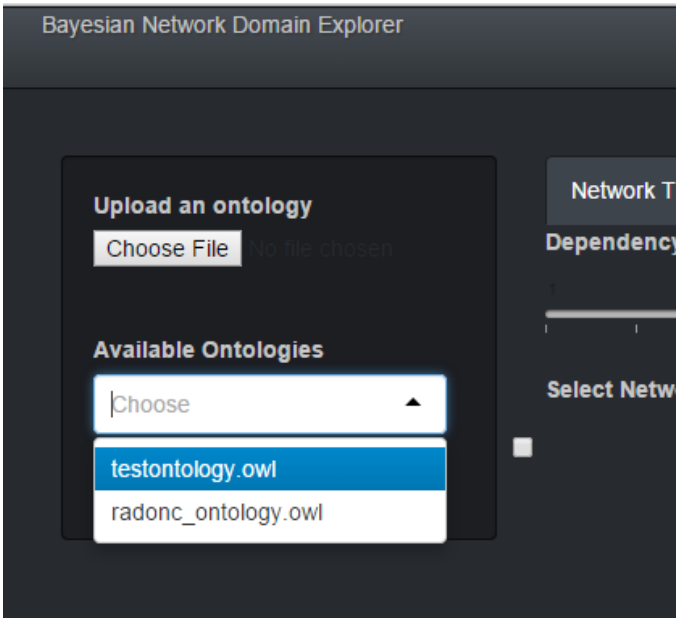


Figure 5.4: BNDE upload button and dropdown menu for ontology selection.

If dependencies among the selected concepts exist, then the network graph belonging to this set of concepts will be immediately computed and displayed in this panel. If there are no dependencies among the selected terms, nothing will be graphed or displayed, except an error message (see section 5.3.6). Similarly, if a selected node has no edge connection to any other terms in the set, it will not be displayed or included in the graph structure. If the dependency properties in the ontology have numeric tags specifying the strength of dependency, ("dependsOn3", for example) then some statistical metrics describing the network (nodes and edges) will be computed and displayed above the graph.

The "Mean Evidence" metric is the mean value of the set of dependency arcs strengths. The "Evidence" term used comes from a medical context and represents a ranking system used to describe the strength of the results measured in a clinical trial or research study,

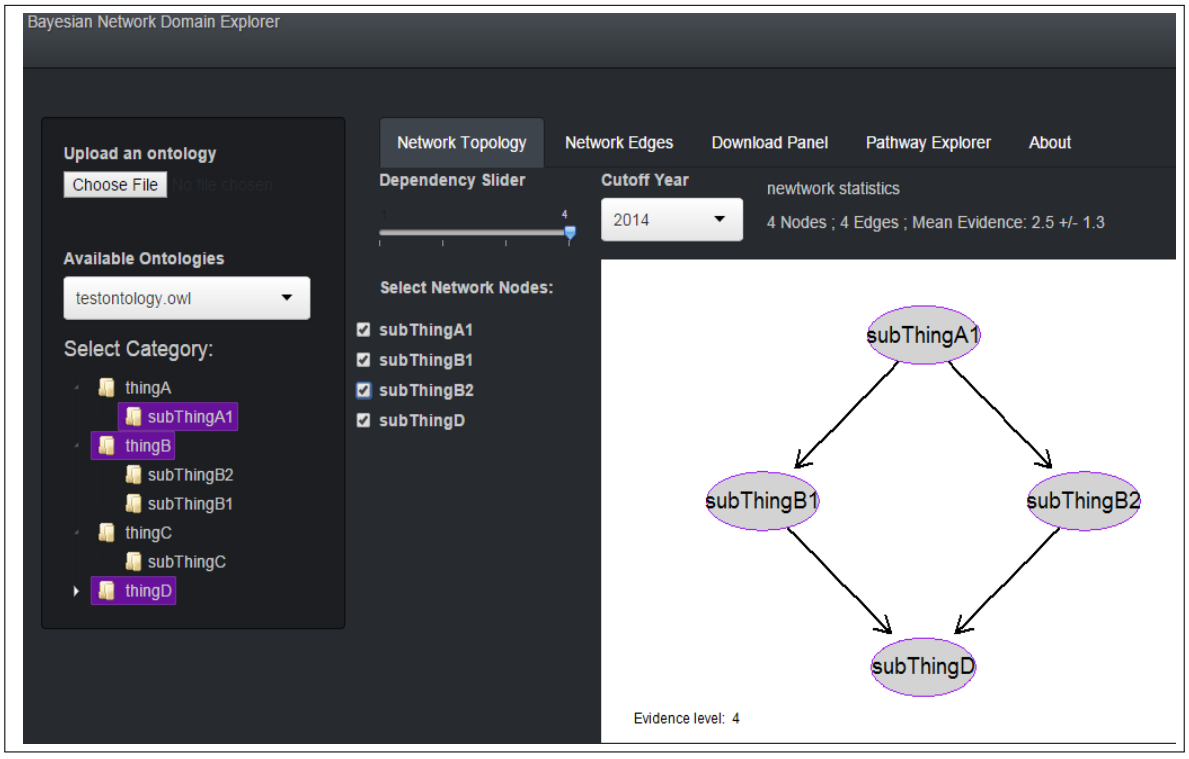


Figure 5.5: BNDE Network Topology tab. The selected ontology's class-subclass structure is recreated (dynamic UI) as a interactive folder-tree appearing on the main sidebar panel (left).

though this value could represent some other type of dependency more generally. For an ontology that contains a layered dependency such as this, the "Dependency Slider" can be changed to exclude nodes below the selected level. If no specific strength is provided by the ontology, all arcs are considered level 1 strength. Networks are updated in real-time as this value is changed. With respect to levels of evidence, one can use this slider to adjust or examine the strength of the overall network. As previously mentioned, because the software is reactive, changes in inputs (ontology choice, selected nodes, dependency slider, etc) will automatically result in re-computation and redrawing of the network display and any other dependent factors on other tab panels.

Because often the concept names used in ontologies are acronyms or jargon, it can be

hard to read the network plot. For this reason, this software make the nodes of a displayed Bayesian network clickable. Clicking on nodes will produce their annotated definitions, NCIT definition, or SNOMED-CT definition, or all three if they exist. The node definition is printed just below the graph. If there are no definitions then a message is printed below to indicate that nothing was found[1].

5.3.3 Network edge list

On the second tab in the tabset, a searchable, sortable, datatable is provided which displays the user constructed network's edge connections and each edge's dependency type. A screenshot of this panel is shown in Figure 5.6.

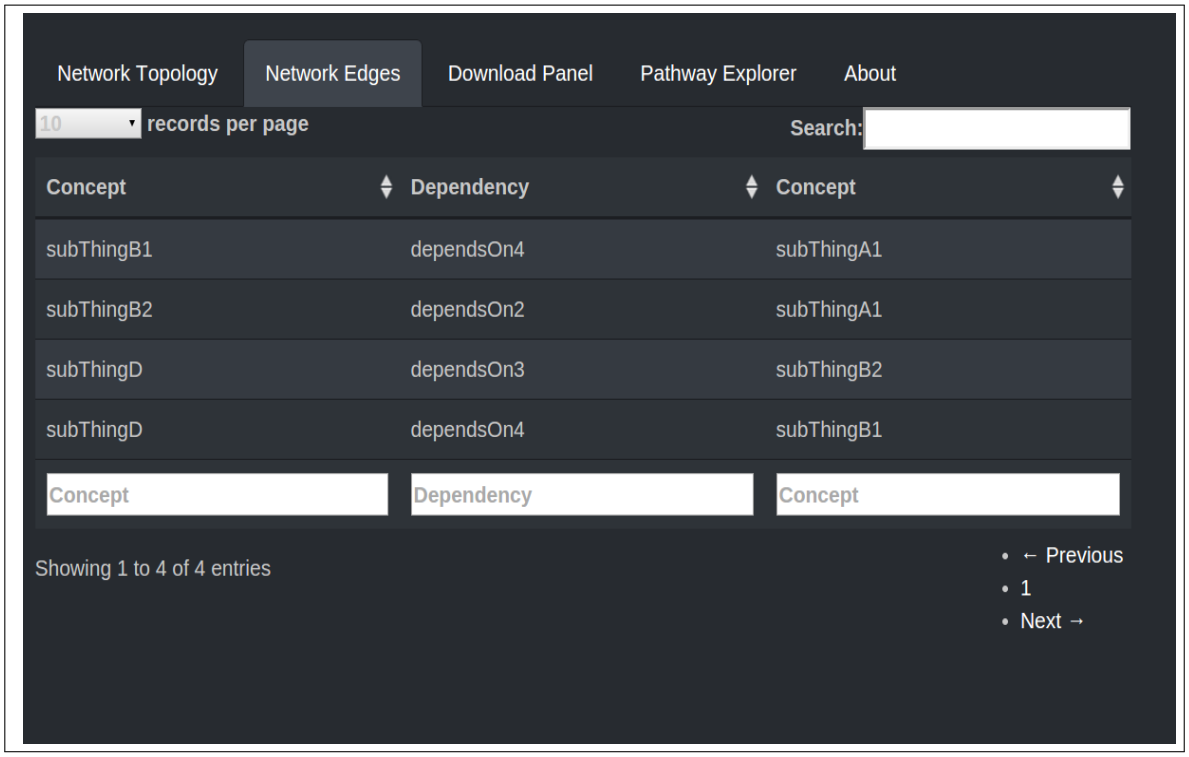


Figure 5.6: Edge Panel tab

[1]Due to unresolved mismatching between server/client click-identification points, this is an unstable feature and remains in-progress

5.3.4 Download handling

The download panel provides several options for saving user generated networks and obtaining more detailed information from the networks:

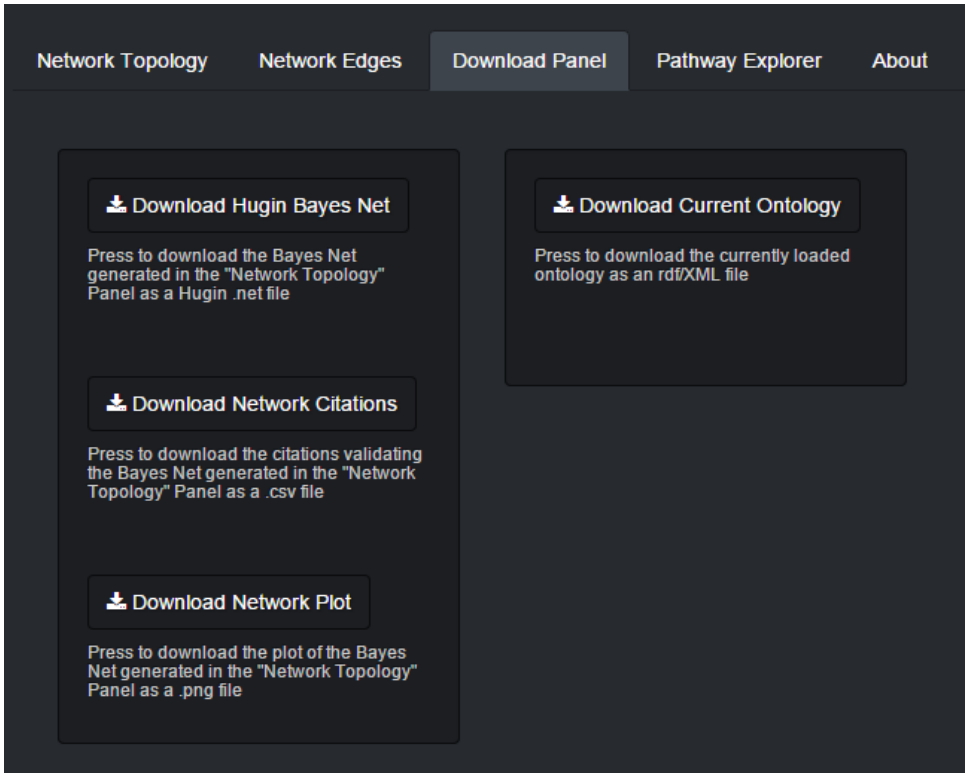


Figure 5.7: Dowloand Panel in the BNDE

- Download a .net file that is compatible with the Hugin Expert[2] software. The Hugin software can accept a topology generated from this tool, and its facilities allow one to populate the conditional probability tables with user provided data or belief values to make a network fully computable.

- Download the citations belonging to the network edges. This feature allows downloading of the entire set of references associated with the network edges in the network

[2]Hugin Expert A/S, Aalborg, Denmark

created. Ontologies whose dependencies contain annotations (e.g. references to a journal publication or other set of information which supports the dependency evidence between two concepts) are extractable via this option. In this way, users can obtain the list of references (read: justifications) for the structure of any particular network. As part of the way the algorithm creates networks in this system, some edges/references will appear in this list that might not appear in the network itself. This is because dependency pathways might be logically chained from distant nodes via deductive reasoning methods. In the citation list, the chained dependency is included.

- Download the graphical plot of the user generated network in portable network graphics format.

- Download the current ontology in RDF/XML format. These files can be viewed or edited using the free, open source ontology editor tool Protégé or other similar software.

5.3.5 Exploring dependency paths

The pathway explorer panel includes a feature which allows for examination of the full dependency paths between any two nodes in a network. This exploratory feature may be useful to understand further some of the underlying knowledge within a set of domain concepts. For instance, in the figure below, examining the paths between "subThingA1" and "subThingD" reveals a dependency path which includes a concept node from outside the currently selected network, "subThingC".

Using this feature is straightforward. Users select one node from each of the dropdown lists (these are populated from the selected network concepts in the main panel) and if there are dependency pathways between these nodes, their graphical representation will be computed and shown in this panel, as well as an edge list below the graph which includes any reference annotations. There is a download button here as well which provides a spreadsheet file of the dependencies and their annotations, for the case that one is interested in exploring

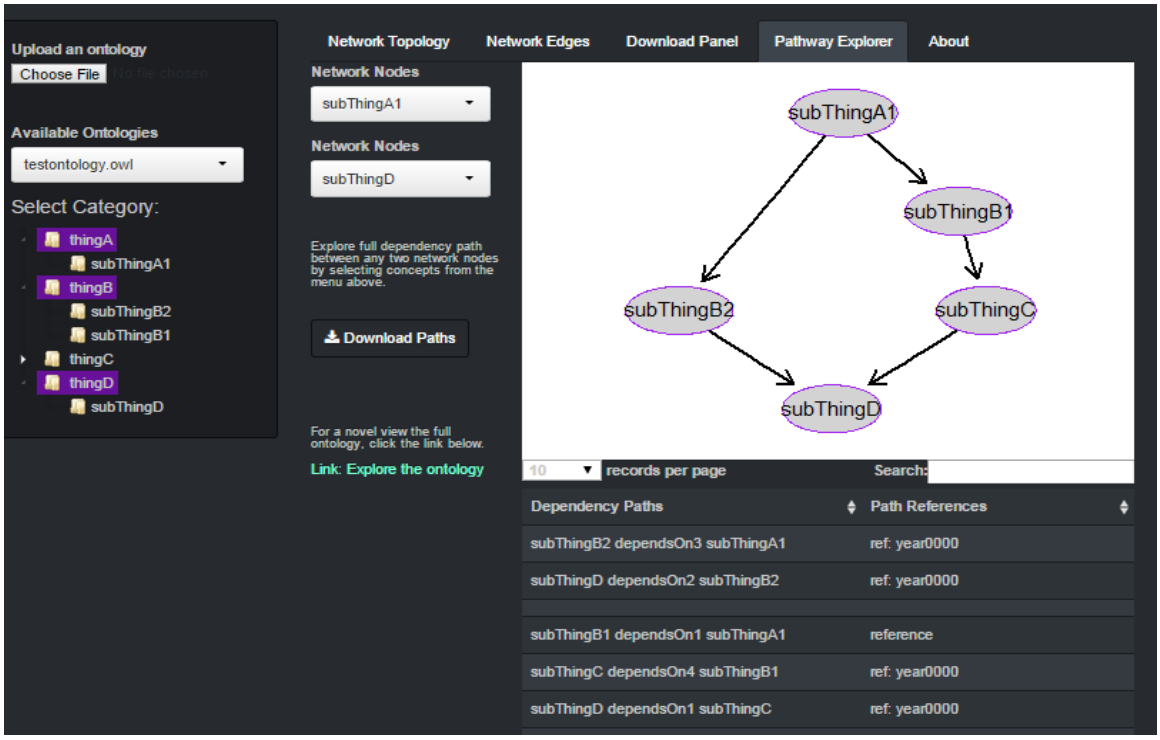


Figure 5.8: Pathway explorer panel

a particular set of dependency paths and their reference sources outside this software.

5.3.6 Error handling

The BNDE software has been tested in a host of modern browsers including Chrome, Iceweasel, Firefox, and Opera. The software also works on mobile browsers, though screen size limits much of what can be performed with this application as it is not optimized for mobile devices. When known errors occur during a network request or other operations, user feedback is given in the form of custom error messages. These errors don't crash the system, they simply inform the user. The application runs as usual even after encountering these errors. The following is a list of some messages currently output and what they indicate:

Error messages:

`Error: please select additional concepts to create a network`

Explanation: This notice occurs when not enough dependent concepts are selected to generate a network

`Error: no paths exist for these concepts`

Explanation: This notice occurs when no dependency paths exist between any of the selected concepts, regardless of external constraints

`Error: no paths exist for these conditions`

Explanation: This notice occurs when dependency paths exist among the selected concepts, but due to constraints put on by the user (for example, the dependency slider setting) all the available nodes are excluded from the network.

`Error: no paths exist among these concepts`

Explanation: This notice occurs when no dependency paths can be found to exist between the two selected nodes

5.4 Summary

The software application described here is a valuable tool for leveraging dependency layered domain ontologies for building directed graphical networks in a fairly straightforward way. With this application, researchers can take advantage of computational elements to subset ontologies and extract networks and network information. One limitation of this software (in its current beta form) becomes obvious when attempting to select for highly interconnected networks containing many nodes (30+). The path search algorithm is not optimized for performance, and it can take several minutes of waiting for a larger network to be constructed. Given real-time updating, any small change to input parameters of a network restarts the entire construction process. Despite this limitation, it is expected that network development time is still significantly reduced compared to the task of manually researching and justifying dependency among a similarly large set of domain concepts. Additionally, both breath-first

and depth-first search algorithms process on a time proportional to the order of the number of edges plus the number of nodes, which sets a fundamental lower computational bound.

Chapter 6

APPLICATIONS AND NOVEL USE CASES

In this chapter I describe the benefits a domain ontology with dependency semantics can give researchers in several distinct use cases. I provide some examples of what can be accomplished with software tools that compute on ontological knowledge by updating Bayesian network models and performing model merging within the radiation oncology domain. I demonstrate how the ontology-based method can be expanded to other medical domains by applying it to an existing medical ontology (the Disease Ontology) which contains dependency semantics that can be interpreted to imply causality and therefore "depends on" relations.

6.1 Introduction

My approach to the problem of BN development and increase BN model consistency and compatibility is to formalize the domain knowledge into an ontology. Ontologies not only offer a method for standardizing terminologies and concepts, they also offer a method for encoding relationships among those knowledge concepts. By formalizing the domain knowledge, we get the benefits of having well defined, cited terminology for common understanding of domain concepts. With well defined terminology comes ease of building, merging, reuse and updating of existing network models. This is the crux of the hub-spoke system I describe in Chapter 1.

To this end, I constructed a domain ontology for radiation oncology (radonc). Methods of ontology construction and encoding are described with more detail in chapter 4. While the ontology does not describe all of radonc, it is large enough to be able to extract meaningful network topologies for clinical use. Networks which share terms using this centralized knowledge source will agree on both terminology and underlying fundamental topology.

Traditionally, a researcher wanting to build a slightly different or updated Bayes network to compute probabilities on aspects of a particular cancer would have to start from scratch—searching the literature, interviewing domain experts, or possibly using machine learning on data to find a suitable topology. With a domain ontology and software operating on the ontology, researchers can simply look for the terms of interest, examine, merge, explore, and extract centrally sourced and verified knowledge quickly and directly. I show in Figure 6.1 a representation of the radonc ontology as the knowledge base ('hub'), with various BN models constructed on the spokes.

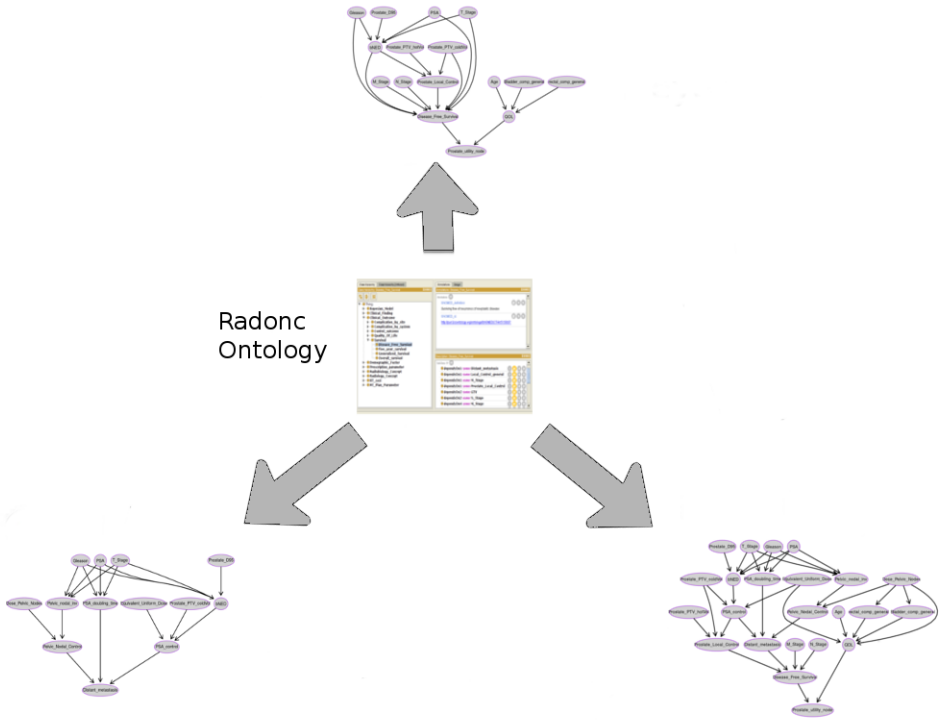


Figure 6.1: The hub-spoke system represented with the radonc ontology as the hub.

6.2 Time savings in initial BN development

While not measured in any formal way there is an informal argument to be made for impact of the hub-spoke system to reduce initial development time. It stands to reason that if N people independently want to make Bayes nets from the same knowledge base, then total development time is of order N. On the other hand, if only one person puts the found knowledge into the ontology, then N other people can use it via the hub-spoke model and the BNDE software tool[1] and (at most) the initial development time is reduced by a factor of N-1. Put more atomically, each relation in the ontology "A dependsOn B" can be used in more than one Bayes net. Each time that term is used is one less time someone had to search the literature or other knowledge source to discover it.

Recall from Chapter 2, that Bayes net topology only represents the abstract causal relations between concepts—the outer shell of the network. The conditional probability tables (CPTs)—wrought from data obtained from various places e.g. published trials, EHRs, clinical surveys, etc.—constitute the inner workings inside the shell which give the network its computational characteristics. Therefore, it is entirely possible that the same BN topology would be built by more than one individual for use in their particular area because they desire the network to compute probabilities more closely matching their own data sources. In fact, one of the next steps for the error detection model project outlined in Chapter 3 is to populate the same topology with data from another clinical facility and compare performance. The fundamentally reusable nature of BN topology lends itself well to the centralized knowledge source model. It is here that the *other* network effect (i.e. the exponential growth in the number of possible subnetworks to be built as the number of new relations grows) can be leveraged to expand the scope of modelers options and time savings.

---

[1] the BNDE software tool is described in Chapter5

6.3 Updating networks with new knowledge

With software tools that operate by accessing an ontological knowledge source, updating Bayesian networks with the state of the art knowledge is a process similar to any other software system which updates via centralized knowledge sources, e.g. spoke and hub models. Where new arcs appear in the updated network, additional knowledge is present between concepts that wasn't considered before. New concepts can quickly be sought and added if desired.

Using the software tool described in Chapter 5, I built the network using terms corresponding to the prostate model published by Meyer et al. in 2004 with a cutoff on the knowledge base for year 2004 [6]. This model was designed to capture the tradeoffs between quality of life (QOL) and disease control in prostate cancer radiotherapy treatments. Figure 6.2 shows the resulting model topology at the time of the model's publishing. By adjusting the network pruning to a knowledge cutoff of 2015 (via a simple dropdown menu selection in the UI), the network is instantly rebuilt using the full knowledge base up to the year 2015. The updated model is shown in Figure 6.3. Compared to the previous model, this new model contains additional arcs, indicating some new dependencies have been added to the knowledge base in the decade or so between their construction.

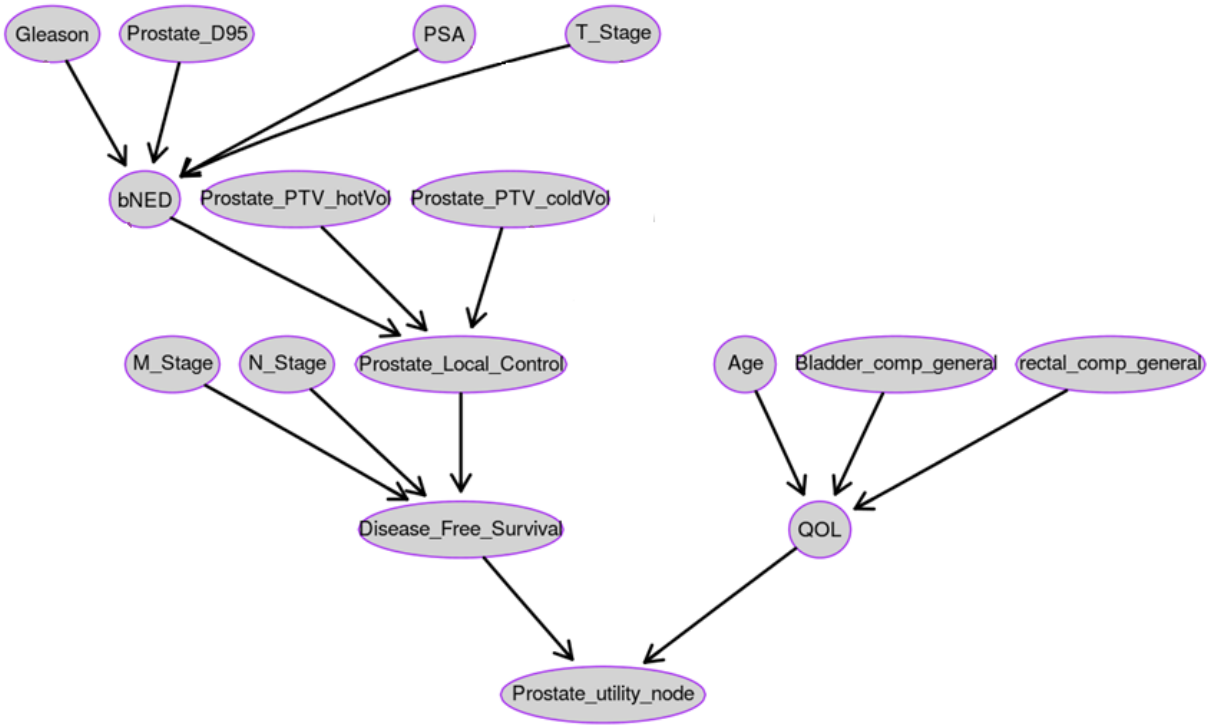


Figure 6.2: Prostate utility network, Meyer et al. 2004

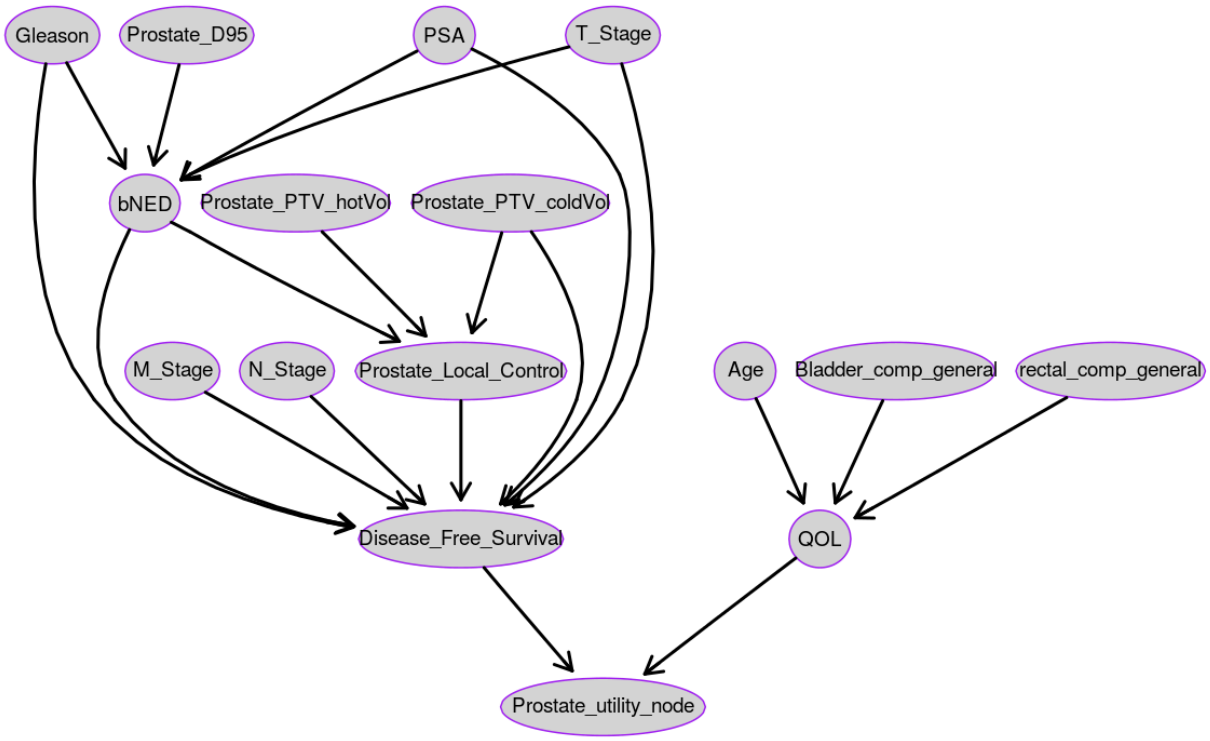


Figure 6.3: Prostate utility network, updated Jan, 2015

subject `predicate` **object**	**knowledge citation**
Pelvic_nodal_inv `dependsOn2` T_Stage	Makarov, DV et al. Urology 69.6 (2007):1095-1101
Pelvic_Nodal_Control `dependsOn2` Pelvic_nodal_inv	Makarov, DV et al. Urology 69.6 (2007):1095-1101
Distant_metastasis `dependsOn2` Pelvic_Nodal_Control	Makarov, DV et al. Urology 69.6 (2007):1095-1101
Disease_Free_Survival `dependsOn1` Distant_metastasis	axiom

Table 6.1: Extracted list of citations (right) for additional arcs (left) found in the updated prostate BN.

Exploring one of the new arcs, (between "T-stage" and "Disease_Free_Survival") using software features uncovers exactly what new knowledge was added to the domain. There are several resulting knowledge pathways, one of which is shown in Table 6.3. The extracted citation in the table shows a study published in 2007 which enumerated several predictors of distant metastasis based on pelvic nodal involvement, pelvic nodal control, and T-stage. The bottom dependency link is an axiom[2] added to the ontology which required no citation, since disease free survival is defined in part by the presence of distant metastasis. The full set of pathways between "T-stage" and "Disease_Free_Survival" can be displayed graphically. In Figure 6.4, all the paths found in the ontology are shown in a directed acyclic graph. The nodes shaded in green are concepts not present in the initial set of concepts selected by the user to build the network, but are still present in the knowledge base.

To a model builder, having this knowledge available opens up some options. One choice would be to include some or all of these new concepts to add richness to the model's description of a system and possibly improve it's accuracy. Though a modeler may not possess

[2]We use 'axiom' in annotations to indicate domain assumptions as opposed to its subsumption usage in OWL

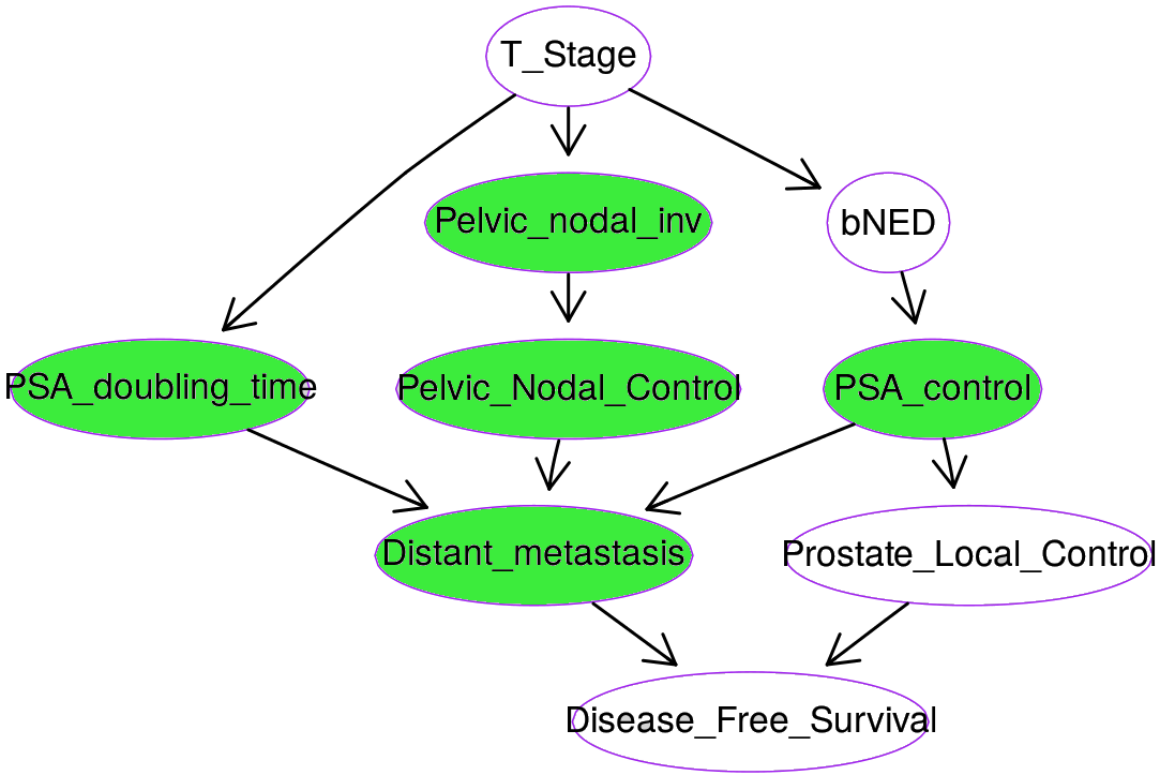


Figure 6.4: Network display of all dependency pathways between "T-stage" and "Disease_Free_Survival" in the radonc ontology. Nodes not in the initially examined network (Figure 6.3) are shaded in green.

the data to populate CPTs for a larger network nor introduce the added complexity, my methodology and software demonstrates the ease at which new information can be brought to a user's attention and makes topology choices more readily available.

6.4 Merging network models

A dependency layered domain ontology that standardizes terminology also gives modelers the flexibility to perform network construction tasks that were previously very challenging. Continuing with the prostate model example; instead of only adding in the missing in-between nodes found in the pathway search, one could choose to add a set of terms from

another published model—in effect—merging the two models to form an updated, more comprehensive model. This task is easily demonstrated because there have been other Bayes net models published about prostate cancer radiotherapy. Smith et al. have produced a model to compute the probability of distant metastasis, given some information about the radiation treatment plan, pelvic nodal involvement, and other prognostic factors including many of the same factors from the 2004 prostate Bayes net [7]. The cross-terms in each model provide a connection point (or points) for model merging. Combining these models (using the 2015 knowledge base) results in the larger, more highly interconnected network seen in Figure 6.5.

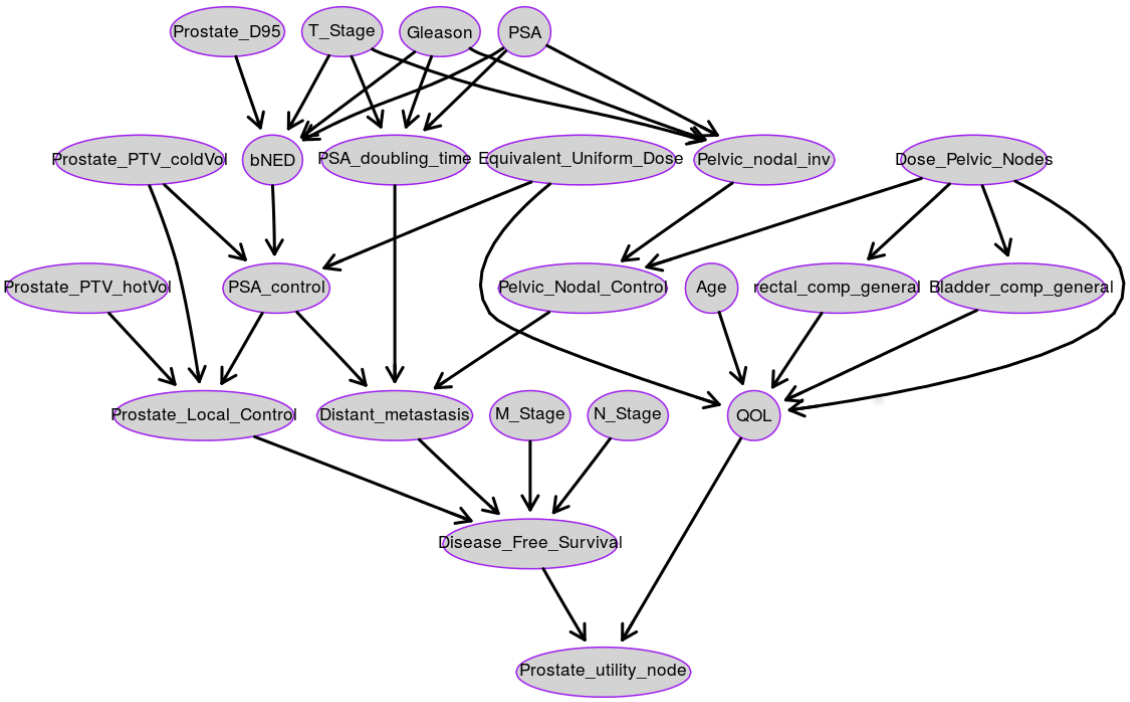


Figure 6.5: Prostate utility network Smith et al. 2009 combined with Meyer et al. 2004 prostate network, updated Jan, 2015

This new model more closely represents the complex interplay between the clinical complications caused by treatment choices (radiation dose to pelvic nodes) and disease free sur-

vival, whereas in the 2004 model these were topologically independent factors. The additional arcs in this combined model between "Equivalent_Uniform_Dose", "Dose_Pelvic_Nodes" and "QOL" can be explored in the same manner as in the previous example. Found within those arcs are seven (7) specific subclasses of bladder and rectal complications. Specific complications like these could be added to the model if that level of granularity were so desired, again, at the expense of complexity.

6.5 Diagnostic networks from the Disease Ontology (DO)

One of the more important aspects of this thesis is to demonstrate the utility of an ontology-based network development methodology beyond one specialized domain. The core claim is that leveraging dependency type semantics in description logics to construct Bayes networks is broadly applicable. In this example, I show how the methods can be applied to an existing medical ontology (the Disease Ontology) to construct networks for diagnostic purposes. Here, the Disease Ontology acts as the 'hub', in the hub-spoke system.

The Disease Ontology (DO) is a knowledge base of human diseases which semantically integrates disease and medical vocabularies via mapping to other standardized medical terminology such as MeSH, ICD, NCI's thesaurus, SNOMED CT and OMIM [15]. Though the DO is available in OWL, it is not in a full description logic (DL) form. Within the annotated definitions for each disease, however, there are relations used to describe various components of the disease. In the DO, the four relations, `has_symptom`, `results_in`, `caused_by`, and `transmitted_by` are of particular interest in that they relate symptoms, environmental factors, and diseases in causal way. These relations are defined in the ontology as object properties. Unfortunately, they are not used to make any logic expressions. If the relations were used in a DL form, then they could be easily translated in external software and used to build networks. As part of this example case, I convert the disease ontology into description logic format using the following steps:

`1. parse disease definitions by relation type to capture relation terms`

2. declare new class for each relation type
3. declare new subclasses of relation terms under its corresponding relation type
4. declare logic expressions for each disease from relations and new subclasses

Once converted to DL format, I upload the disease ontology into the Bayesian Network Domain Explorer to create networks from diseases, symptoms, causes, and transmissions.

6.5.1 Adding description logic to the disease ontology

Since the relations are fairly well defined, a special purpose parser was written to extract portions of each definition that follow the predefined relations. Using SPARQL queries to pull each disease's definition and string operations on the definition paragraphs, individual terms and relations can be isolated. Take for example, the term for chickenpox. Its definition is shown below:

Definition:
"A viral infectious disease that results_in infection located_in skin, has_material_basis_in Human herpesvirus 3, which is transmitted_by direct contact with secretions from the rash, or transmitted_by droplet spread of respiratory secretions. The infection has_symptom anorexia, has_symptom myalgia, has_symptom nausea, has_symptom fever, has_symptom headache, has_symptom sore throat, and has_symptom blisters."

From this definition, the parser extracts triples of the form:

chickenpox `results_in` infection located_in skin
chickenpox `transmitted_by` direct contact with secretions from the rash
chickenpox `has_symptom` anorexia
chickenpox `has_symptom` myalgia
chickenpox `has_symptom` nausea

```
1   # create XML string with new class label and DOID
2
3 symp<-c("<owl:Class",
4 paste("rdf:about=","\"","http://purl.obolibrary.org/obo/DOID_1s","\"
    ",">",sep=""),
5         "<rdfs:label",
6 paste("rdf:datatype=","\"","http://www.w3.org/2001/XMLSchema#string
    ","\"",
7 ">Symptom</rdfs:label>", sep=""),
8  "</owl:Class>")
9
10  # add new class into the ontology
11
12 DO<-append(DO, symp, after=grep("</rdf:RDF>", DO)-1)
```

Figure 6.6: R code for adding classes into the Disease Ontology.

chickenpox `has_symptom` fever
chickenpox `has_symptom` headache
chickenpox `has_symptom` sore throat
chickenpox `has_symptom` blisters

In DL formalism, all the terms on the right-hand-side of this list need to become classes. Therefore, a set of new classes is defined back into the ontology: `Symptom`, `Transmission`, `Result`, and `Cause`. Under each of these classes, the newly parsed terms above can be added into the ontology as subclasses. The code (using the `R` language) for this processes is shown in Figure 6.6.

New description logic statements are added into the ontology in much the same manner that new classes for relation terms are added. From a list of diseases and associated relation terms, strings are constructed to declare each term using corresponding object properties. The disease term is searched for in the ontology and the new logic statements are added to the file. Technically, in order to make the statement "chickenpox `has_symptom` fever", *chickenpox* is required to have appropriate XML syntax and the declarations "subClassOf" "Restriction" "onProperty" "someValuesFrom" applied to the resource *fever*, as shown in Figure 6.7.

6.5.2 Extracting a diagnostic network

With a DL version of the disease ontology, queries can now be run against the object properties and translated into appropriate dependency types with directionality of dependence assumed from the terminological meaning of the relations. For some generic classes A and B, the statement "A `has_symptom` B" becomes "B `dependsOn` A", "A `caused_by` B" becomes "A `dependsOn` B", "A `transmitted_by` B" becomes "A `dependsOn` B", and "A `results_in` B" becomes "B `dependsOn` A". This translation is accomplished using queries of the form below:

```
CONSTRUCT {?object2 ro:dependsOn4 ?object1}
WHERE {?object1 rdfs:subClassOf ?restriction.
?restriction owl:onProperty obo:has_symptom.
?restriction owl:someValuesFrom ?object2.
?restriction ?restrictionPredicate ?object2.}
```

The disease ontology is fairly large (over 185,000 triples before the addition of four new class sets), so in order to keep the computational time minimal, diseases without any description logic relationships were removed, resulting in 37,875 triples. This abridged version of the DO in DL was read into the Bayes net development software, and classes were selected using

```
1
2   # starting with 2 x N matrix of disease-id's and assoiciated
        symptoms
3
4   # 1. find location of each disease class in the DO and make unique
        DOID
5
6 classPos<-grep(sympList[i,2], DO)
7 genDOID<-paste("DOID_", which(sympClass==sympList[i,1]),"s", sep="")
8
9   # 2. declare semantic statement
10
11 newDL<-c(
12 "<rdfs:subClassOf>",
13 "<owl:Restriction>",
14 "<owl:onProperty",
15 paste("rdf:resource=",
16 "\"","http://purl.obolibrary.org/obo/doid#has_symptom","\"","/>",
       sep=""),
17
18 "<owl:someValuesFrom",
19 paste("rdf:resource=",
20 "\"","http://purl.obolibrary.org/obo/", genDOID ,"\"","/>", sep=""),
21 "</owl:Restriction>",
22 "</rdfs:subClassOf>"
23 )
24   # 3. add statement to the ontology
25
26 DO<-append(DO, newDL, after=classPos)
```

Figure 6.7: R code for declaring new dependency relations between classes in the Disease Ontology

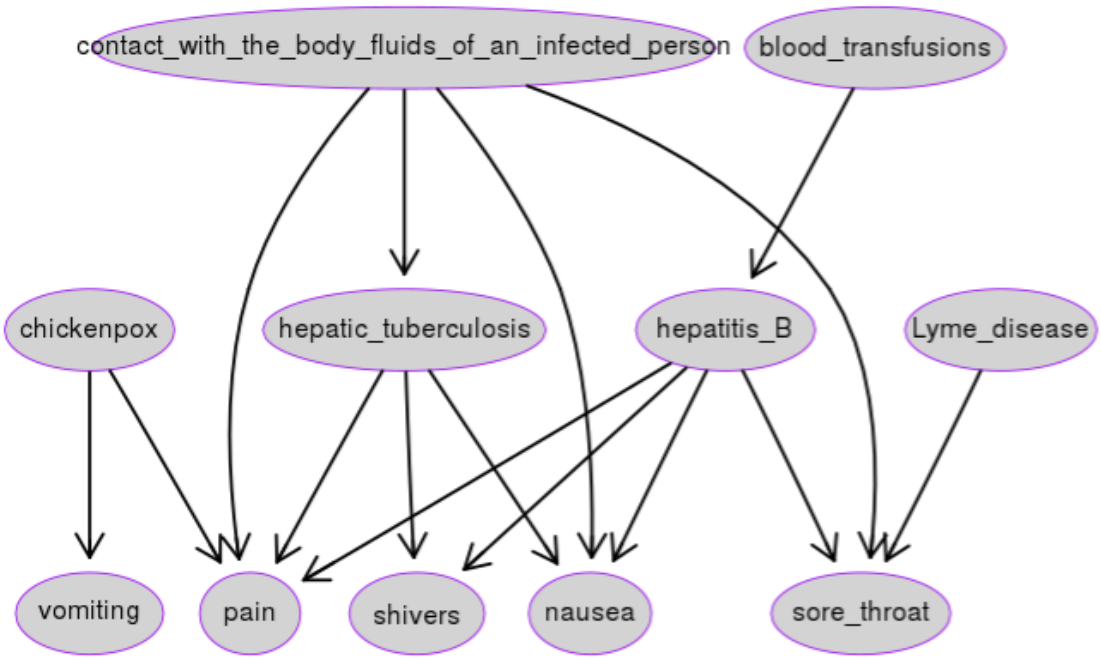


Figure 6.8: Diagnostic network topology extracted from the Disease Ontology

the GUI. Given the size of the ontology, the possible choices of network to create are nearly endless. For a diagnostic network, is expected that the choice of network topology would be informed by a particular set of symptoms, exposure factors, and a range of diseases. Figure 6.8 shows an example diagnostic network containing a set of fairly common symptoms, some diseases that can cause those symptoms, and transmission types for some of the diseases.

Initial inspection of this network, although it is much simplified from real world diagnostic practices, shows that the dependency directionality and ontological concepts agree with a common understanding of symptoms, diseases, and causes. What stands out in this topology are the arcs from "contact_with_body_fluids_of_an_infected_person" to "pain", "sore throat", and "nausea". This arc bypasses the listed diseases, which indicates that there is an additional pathway between that exposure parameter and resulting symptoms. Unlike

the radonc ontology[3], the structure of the DO has fewer levels. That is, we can say that exposures like body fluid contact don't cause symptoms directly, and so most likely there is another disease that has not been selected for (and was therefore pruned out of the final network). In terms of modifying the network for more effective diagnostic tasks, this suggests that additional diseases might be considered.

6.6 Challenges and insights of the DO as a knowledge hub

Even with nearly 38,000 declared triples, some of the descriptions appear incomplete. For the example in Figure 6.8, one would expect to see "contact_with_body_fluids_of_an_infected_person" to be related to "hepatitis_B", perhaps from a `transmitted_by` relation (and hence an arc). This relation is not present in the ontology, though it is a well known fact in the medical community. Thus, the methodology not only allows us to just build diagnostic networks—we also gain a unique way of visually finding weaknesses and missing links in the DO.

A more thorough inspection comparing the DL version of the DO with the original DO reveals that there are some symptoms which do not appear to have been correctly identified or declared in a correct dependency, suggesting a few bugs in the implementation. Future endeavors into creating a description logic transformation of the disease ontology should note to take a careful look at using methods to verify parsed results and reinsertion of classes.

One challenge to navigating the DO class lists using my software is that there is no defined structure to the new class sets. This is an expected result, as there was no such code written for including additional class structure to the ontology beyond the four overarching relation superclasses. Despite this fact, there are few surprising exception cases. As it happens, a few *Symptom* and *Result* terms had previously been defined in the ontology as diseases, for example, "arthritis". By simply declaring "arthritis" as a subclass of *Symptom* without knowledge of it being a subclass of *Disease*, it becomes interpreted a subclass of both, and acquires some hierarchical structure through multiple inheritance.

---

[3]see chapter 4, section 4.4

Though the size of the ontology makes the BNDE software somewhat impractical to use, it does provide for a proof of concept that the functionality works in more general domains which contain dependency semantics. The functionality could be put to use more efficiently on the back-end of an efficient logic system to create diagnostic networks for some kind of decision aid, given a more comprehensive (or more narrowly scoped) disease ontology.

6.7 Summary

In this chapter I have demonstrated how a user of my software tool can quickly and easily create, update, and/or merge Bayesian network models from dependency layered description logic ontologies in two different medical domains. In addition, I have shown that there is potential time savings for network modelers using the software which grows with the number of users (N). The examples described here do two things: 1) they validate the functionality of the software tool, i.e. it's ability to subset based on multiple aspects of the knowledge base (terminological selection and year cutoff), and, more importantly, 2) **give proof of concept that the central repository model and its associated extraction methods meet the initial goals set forth by this research proposal**, namely, to provide the means to create consistent and compatible probabilistic topological models with real-world foundational knowledge.

Chapter 7
CONCLUSIONS AND FUTURE WORK

In this chapter I address the broader process of automating domain knowledge into decision support systems, how my research fits into that process, and some solutions to the challenges in automating those parts of the process which are not covered in this dissertation. I give an overview of possible extensions of the radiation oncology domain ontology and what those extensions mean. Finally, I outline development directions for other clinical use cases and potential software features.

7.1 A larger framework of automation

As discussed in Chapter 1, my research involves a process of knowledge capture and translation. Though only a few stages of translation process are automated by my work, they solve a important core part of the larger challenge of automating and integrating probabilistic knowledge and reasoning into clinical systems. There are several areas of the process which can still benefit from further automation, and many other interesting research topics and questions arise.

7.1.1 Semi-automated ontology building

As mentioned in section 4.7.1 the process of manual ontology building is a task requiring specialized knowledge and domain expertise. The radonc ontology developed in my research, while fairly comprehensive at the top level, is still incomplete. In fact, the ongoing curation of a domain ontology is technically never complete, as domain knowledge is constantly entering the public sphere. To reproduce my method of ontology construction in domains where terminology is extensive could be too time consuming to be effective. An alterna-

tive method for ontology construction might employ natural language processing (NLP) to discover and translate relationships between variables into an ontological formalism. This semi-automation would replace the manual translation step as shown in Figure 7.1.

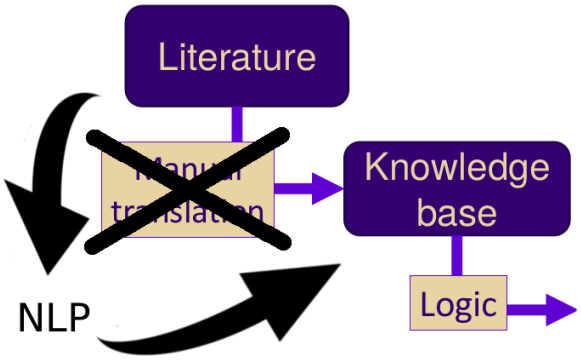


Figure 7.1: Replacing the manual translation step in the knowledge base building process with natural language processing (NLP) methods could greatly increase the portability of an automated network building process to other domains.

One of the challenges posed in this problem is the lack of common semantics and language structure used among scientific literature. For example, we examine a portion of the *Conclusions* section of an abstract in a recent paper published by E.C. Osmundson et al. [85]:

$V_{BED10}72$, $V_{BED10}66$, and $Dmean_{BED10}$ to cHBT are associated with HB toxicity.

This language would need to be translated and reduced to a dependency triple (or set of triples, since there are three nouns relating to one term), and the associated classes assigned in the hierarchy. For this example, a system would need to decide what $V_{BED10}72$ is. The abstract further states: "...doses were converted to biologically effective doses (BED) by using the standard linear quadratic model $\alpha/\beta = 10$ (BED10)". An NLP parser needs to know (or be told) the difference between a DVH (VN) versus a radiobiological parameter (BED). Also, a determination needs to be made about what "toxicity" amounts to, possibly requiring parsing of other portions of the abstract, e.g. "To identify dosimetric predictors of

hepatobiliary (HB) toxicity associated with stereotactic body radiation therapy (SBRT) for liver tumors.". This abstract parsed and translated would hopefully result in the triples of the form, "HB toxicity `dependsOn` V72". Training an NLP system to capture all the various ways medical knowledge is encoded into published literature is clearly non-trivial. However, a successful application of NLP to building domain ontologies would be of tremendous benefit for dependency layered ontology construction.

7.1.2 Automated queries for conditional probability tables

Electronic medical data repositories have become a very large, uncontrolled set of rich information. One way of leveraging this information retrospectively is with probabilistic networks. Networks of joint probability distributions capture the convolution of influence among many parameters without having to keep any of them constant across a cohort. In fact, a probabilistic network can be more information rich if homogeneity does *not* occur in the cohort. Given this, BNs are an attractive way of generating predictions from these growing uncontrolled information sets. However, the connection between network influence directions and the data which populates each network table is rarely established. Using the ontological model for extracting causal links alleviates the directionality problem in the topology, but in order to populate the topology with the data, a file needs to be associated with each node and it's terminology.

For small networks, manually writing SQL queries to generate flat data files for machine learning CPTs can be a reasonable task to connect network diagrams with the data appropriate to each network node, as was the case in Chapter 3, but it is a task which necessitates trained staff who understand database (DB) queries, the database schema, and how the schema maps to real-world terminology (for every query!). Moving forward, it would be more practical to engage the DB with an appropriate mapping to ontological terminologies, such that for every network generated from the ontology, a mapping exists which can be employed to query the DB for populating the CPT's. Figure 7.2 shows the part of the knowledge translation process representing a mapping of ontological terminologies to data

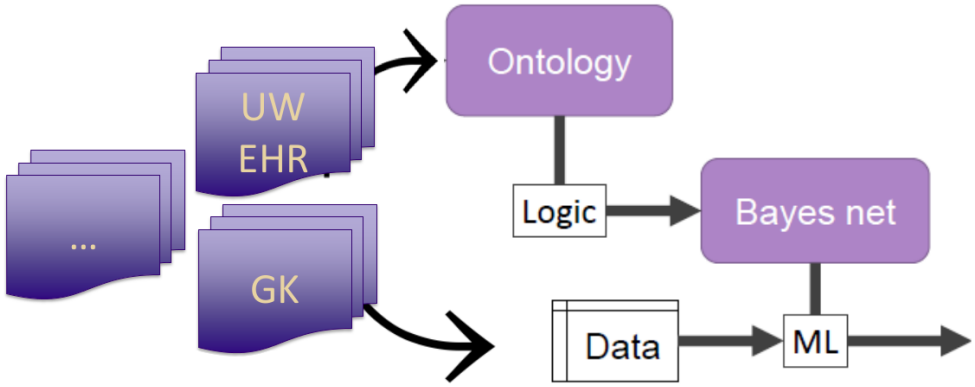


Figure 7.2: Framework for leveraging disparate data sources by mapping their schemas to the ontology for easier queries.

sources.

The current implementation of a Mosaiq clinical information system (CIS) in radiation oncology uses a DB schema over 900 pages including thousands of terms. These terms have no standardized mapping to any other set of known terminology. A mapping could also inform the process of mosaiq RDB construction in future versions of the CIS as well as being hugely beneficial to the network development process, and making the radonc ontology a practical knowledge source to the many other clinics using Mosaiq. Mapping an ontology to one DB schema only solves the CPT problem for that unique DB. Leveraging other sources of knowledge and information from disparate systems (a more common way of storing patient data than not) requires a systematic mapping of each DB. In the UWMC system, the local mosaiq DB mostly stores information about patient's linac based radiotherapy. The Harborview Gamma Knife (GK) center operates a different kind of radiation delivery but maintains a separate data repository. Similarly, the UW EHR contains yet another separate

patient database. Future work in the direction of automating CPT construction should consider a host of mappings to these repositories and technologies. The mappings would allow for cross-source queries of data in order to build CPT's for more thorough probabilistic models that aid decisions, lower costs, or improve patient care.

7.1.3 CPT's from big data

One of the problems populating CPTs from large data sources is the fact that table size grows like $\prod N_i$ where N is the number of states in the ith connected node. The computational requirements needed to process the number of states often found in medical data (sometimes hundreds to thousands of states per node) makes machine learning CPTs unachievable with desktop computing systems (as they exist at the time of writing). Any future effort to automate CPT construction processes by extracting and learning experience tables from mapped datasets needs to address the big-data nature of this problem.

Many of the methods for machine learning CPTs employ the expectation maximization (EM) algorithm. There are some approaches to parallel computing the EM algorithm in distributed ways which could be applied to BN learning to make it more tractable [86, 87]. Since the network topology is fixed and known, we might be better served by using tree structured learning [88]. In this method, each CPT is learned by separately from the larger network based only on it's conditional dependence relationships (e.g. parent nodes). A similar method which could provide more accurate probability distributions involves breaking a larger network into sets of subnetworks based on the Markov blanket corresponding to each node. The CPT for that node is computed and the tables recombined to reproduce the larger network with full CPTs. Given the BNDE software described in Chapters 4 and 5, which prunes networks into smaller models, this latter method (using Markov blankets) seems quite promising.

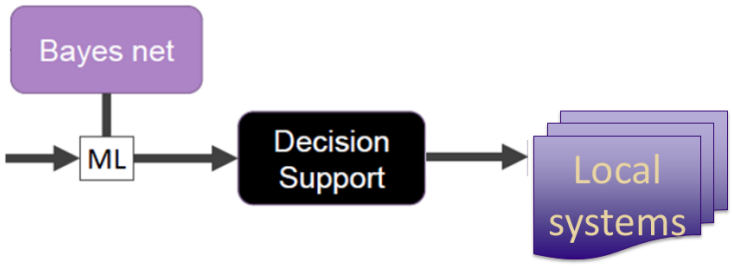


Figure 7.3: Integration of decision models into clinical workflows can occur by embedding them into the front-end of locally used clinical systems.

7.1.4 Integration into clinical workflow

One of the remaining potentials of this research is integration of probabilistic clinical decision support models into everyday clinical workflows. Consider the error detection network, for example. The format of the error detection model presented in Chapter 3 is not a practical decision aid for medical physicists use in checking radiotherapy charts on a daily basis. This is largely because instantiating evidence and propagating information changes would require searching for and collecting certain patient data from the mosaiq user interface, individually entering data manually into a separate software system, computing results and comparing each result with a tolerance threshold for error. Few clinical physicists know how to run R much less have the time it takes to perform this task on every chart. Thus, to make these decision models more *clinically impactful*, some measure of tool building needs to be performed. Specifications for such a tool include patient selection, automated extraction of

relevant clinical data from that patient chart, error detection analysis, and display of metrics and results in a clear and simple format. Specification for other network models will be dictated by the varying details of their use cases and how they fit into a particular workflow.

7.2 Expansion of ontological scope

Beyond adding to the set of concepts and "dependsOn" relations in the ontology, some consideration should be given to other types of useful object properties and annotations. Some of this is discussed in section 4.7. There has been some research describing methods of combining probabilistic and deterministic components into one Bayes net [89]. It is possible that one would desire a network with components that are explicitly functionally dependent as well as probabilistically dependent. To create that network, we could still use the "dependsOn" ontological properties describing the relationship, but would need specific annotations describing the exact functional relationship (i.e. $y = x^2 + 1$), and the software would have to be changed to incorporate the new functions into the network object.

Another consideration would be to include a set of relations describing a workflow and/or personnel layer. Given this layer, one could extract, design, analyze and develop workflow models regarding diagnostic, planning, QA, and/or treatment processes. Integrating workflow models from the ontology into incident reporting systems could help clinicians and administrators better understand what components of a clinical process are related to medical incidents and incident types. Error detection networks could flag errors in certain planning parameters which correspond to process steps in the workflow and clinical personnel classes responsible for those processes. This extra layer of knowledge in the ontology could then lead to opportunities for process or personnel improvement as opposed to just detecting errors.

As briefly mentioned in the Chapter 1, the dependency layer actually sets a groundwork protocol for establishing dependency across medical domains. What does this mean? It means that ontologies can be connected via dependency semantics, providing the terminology is reasonably consistent between them. This allows cross-domain networks to be constructed. No medical domain stands completely in isolation. Relations between radiation oncology

concepts could have implications to other medical care choices. Late complications from radiation therapy, for example, are often treated by a patient's primary care physician and not the radonc practitioners. In that sense, complications in the radonc ontology would constitute a subclass of clinical findings in another ontology, along with a host of associated treatment options and other concepts in that domain. Connecting ontologies in this way would allow for construction of more comprehensive, multi-domain models.

It is possible that inconsistent ontological class-subclass relations would pose issues with respect to things like inheritance of causal object properties, and these minutia would need to be considered in any effort to expand the dependency layer from one ontology to another. Of course, one solution would be to perform ontology merging before hand, however, this comes at the cost of increasing the search space. In this sense, decentralized but consistent ontologies would be preferred.

7.3 Further development of clinically relevant networks

7.3.1 Advanced error detection in radiotherapy

With the error detection networks described in Chapter 3, I showed beneficial results over existing paradigms with network models. The data used in the project were entered into the medical record system during the normal course of treatment, meaning exactly zero additional time was required by domain experts other than extraction of the data tables from a relational model to a flat file—a task reasonably accomplished with well known SQL type queries. In radiation oncology, only three vendors account for 90% of the radiation oncology clinical database software, leaving significant room for portability of this technology in the selected domain. With automated CPT creation methods, clinics can instantiate networks with their own data representing the distributions within their local practice.

Natural next steps to further developing error detection networks involve 1) creating advanced topology with additional variables, and 2) verifying the portability of the decision model to other institutions by collaborating and applying the ML methods on different data.

The choice of new variables can be informed by comparing the current network concepts to a set of polled data gleaned from medical physicist questionnaires.

7.3.2 Other clinical Bayes net use cases

Beyond error detection, there are other questions of clinical interest in the radiation oncology domain. Given the financial cost and complication rate differences between photon and proton radiotherapy, a cost-benefit analysis for competing modalities is a Bayes network which can answer the question: "which treatment modality gives the best outcome (quality of life) for the lowest price?". Similarly, a administrative or financial stakeholder could use this same cost-benefit network to ask the inverse question and get answers regarding price points, e.g. for a fixed outcome, "what would therapy X's price have to be in order to be cost-competitive to therapy Y?". Generally speaking, sensitivity analysis on the network would be applicable in answering these questions.

Of equal importance is a decision model to answer the "next test" question. That is, given a set of tests, what test should be ordered next to give the most information to put towards a clinician's next treatment decision. In oncology, a physician might have to decide between biopsy, blood test, imaging, or no test, with each test have different levels of invasiveness, cost, and giving different amounts and types of information related to the tumor not known *a priori*. With sets of prior distributions in a BN, the physician could weigh these many possibilities in an empirical way before making a decision.

7.4 Advanced software developments

One of the limitations of the BNDE software for network development is that it could become challenging for users to navigate the ontology class-subclass structure manually. This is related to the size of the ontology, which is only expected to grow. For less technically savvy users (or even users who might understand what a BN is and what it can tell us but aren't versed in building them), the process of building a BN from *any* set of concepts at all might be daunting. A set of meta-level BN models could make the BNDE more approachable to

these users.

As I stated in section 4.4.1 regarding the top-level of the ontology: "outcomes depend on treatment, treatments depend on diagnosis, diagnosis depend on symptoms...". This statement implies that a higher level network model—a template—could be coded into the software to build complicated networks from very simple requirements. Template models could be accessed by a simple two parameter function displayed as dropdown menu items in the UI. For the radiation oncology domain, the template could be presented as two simple parameters *site* and *goal*, where the site is a general anatomic location e.g. "lung" and the goal is the class of question to be answered by the network (the template) such as "cost-effectiveness", "error detection", " or "quality of life". The software can then extract the set of concepts corresponding to the selected site and which have some superclass belonging to the template model. The current software already automatically builds BN topology from selected concepts, thus, additional browsing of the ontology would not be necessary.

I also believe adding some form of ontological model repository would be useful. This would allow researchers to save their models, merge or connect them to other available models. It is not clear if adding models to the existing working ontology (as a set of repeated classes) would bloat the ontology size and/or impact the search time. It should be investigated whether or not the added utility of a model repository is outweighed by technical constraints.

7.5 Concluding thoughts

In this dissertation, I addressed the challenges of manually building Bayes networks by automating parts of the process and establishing a consistent ontology knowledge base for future use in radiation oncology. In the process of carrying out this research, I learned that although the primary motivation of this work was to ease Bayes net development, the ontological approach that I investigated to achieve that task actually has quite a bit more potential. By formalizing relations between concepts in a domain, we get to ask "what does dependency/causality mean?" In the context of this dissertation, it mostly means

"conditionally dependent" in statistical terms. However, it could mean (as discussed above) deterministically dependent in terms of some well defined continuous mathematical function. In other forms, it could mean something discontinuous, or perhaps even rule-based, e.g. for some dependent concepts A and B: if $state(A) = a_i$, then $state(B) = b_j$, and so on. Bayes nets are most useful in domains where uncertainty is relatively high. Other domains might find rules or determinism more valuable. The implication of this broad scope of potential is the ability to *reuse* the knowledge base for a variety of applications.

In the end, my research represents an important proof-of-concept, but the real value of this work lies in operating on more comprehensive representations of the radonc domain, the disease ontology, and other currently unrepresented domains. Thus, the long term impact of this work depends on the continued development of the approach and the tools that operate on top of the knowledge base to perform useful reasoning.

BIBLIOGRAPHY

[1] Gal Elidan, Iftach Nachman, and Nir Friedman. Ideal parent structure learning for continuous variable bayesian networks. *J Mach Learn Res*, 8:1799–1833, 2007.

[2] Rónán Daly, Qiang Shen, and Stuart Aitken. Learning bayesian networks: approaches and issues. *The Knowledge Engineering Review*, 26:99–157, 4 2011.

[3] Ai-Ling Zhu, Jian Li, and Tze-Yun Leong. Automated knowledge extraction for decision model construction: a data mining approach. In *AMIA Annual Symposium Proceedings*, volume 2003, page 758. American Medical Informatics Association, 2003.

[4] David Chickering, Dan Geiger, and David Heckermn. Learning bayesian networks is np hard. Technical Report MSR-TR-94-17, Microsoft Research, 1994.

[5] MJ Druzdel and Linda C Van Der Gaag. Building probabilistic networks:" where do the numbers come from?". *IEEE Transactions on knowledge and data engineering*, 12(4):481–486, 2000.

[6] J Meyer, M H Phillips, P S Cho, I Kalet, and J N Doctor. Application of influence diagrams to prostate intensity-modulated radiation therapy plan selection. *Phys Med Biol*, 49:1637–1653, 2004.

[7] WP Smith, J Doctor, J Meyer, IJ Kalet, and MH Phillips. A decision aid for intensity-modulated radiation-therapy plan selection in prostate cancer based on a prognostic bayesian network and a markov model. *Artif Intell Med.*, pages 119–130, 2009.

[8] Eric M Horwitz, Alexandra L Hanlon, Wayne H Pinover, Penny R Anderson, and Gerald E Hanks. Defining the optimal radiation dose with three-dimensional conformal radiation therapy for patients with nonmetastatic prostate carcinoma by using recursive partitioning techniques. *Cancer*, 92(5):1281–1287, 2001.

[9] Jung Hun Oh, Jeffrey Craft, Rawan Al Lozi, Manushka Vaidy, Yifan Meng, Joseph O Deasy, Jeffrey D Bradley, and Issam El Naqa. A bayesian network approach for modeling local failure in lung cancer. *Phys Med Biol.*, 56, 2011.

[10] K. Jayasurya, G. Fung, S. Yu, C. Dehing-Oberije, D. De Ruysscher, A. Hope, W. De Neve, Y. Lievens, P. Lambin, and A. L. A. J. Dekker. Comparison of bayesian network and support vector machine models for two-year survival prediction in lung cancer patients treated with radiotherapy. *Med Phys.*, 37:1401–7, 2010.

[11] A. Dekker, C. Dehing-Oberije, D. De Ruysscher, P. Lambin, A. Hope, K. Komati, G. Fung, Shipeng Yu, W. De Neve, and Y. Lievens. Survival prediction in lung cancer treated with radiotherapy: Bayesian networks vs. support vector machines in handling missing data. In *Machine Learning and Applications, 2009. ICMLA '09. International Conference on*, pages 494–497, 2009.

[12] Alexander Stojadinovic, Anton Bilchik, David Smith, John S Eberhardt, Elizabeth Ben Ward, Aviram Nissan, Eric K Johnson, Mladjan Protic, George E Peoples, Itzhak Avital, et al. Clinical decision support and individualized prediction of survival in colon cancer: Bayesian belief network model. *Annals of surgical oncology*, 20(1):161–174, 2013.

[13] Ann Devitt, Boris Danev, and Katarina Matusikova. Constructing bayesian networks automatically using ontologies. 2006.

[14] Peter J Haug, Jeffrey P Ferraro, John Holmen, Xinzi Wu, Kumar Mynam, Matthew Ebert, Nathan Dean, and Jason Jones. An ontology-driven, diagnostic modeling system. *Journal of the American Medical Informatics Association*, 20(e1):e102–e110, 2013.

[15] Lynn Marie Schriml, Cesar Arze, Suvarna Nadendla, Yu-Wei Wayne Chang, Mark Mazaitis, Victor Felix, Gang Feng, and Warren Alden Kibbe. Disease ontology: a backbone for disease semantic integration. *Nucleic acids research*, 40(D1):D940–D946, 2012.

[16] Vimla Patel, David R. Kaufman, and Jose F Arocha. Emerging paradigms of cognition in medical decision-making. *J Biomed Inf*, 35:52–75, 2002.

[17] Nilesh Jain and Michael Kahn. Clinical decision-support systems in radiation therapy. In *First International Symposium on 3D Radiation Treatment Planning and Conformal Therapy*. Citeseer, 1993.

[18] Peter Lucas, Linda C. van der Gaag, and Ameen Abu-Hanna. Bayesian networks in biomedicine and health-care. *Artif Intell Med*, 30(3):201–214, 2004.

[19] FV Jensen. *Bayesian Networks and Decision Graphs*. Springer-Verlag, 2001.

[20] Charles E. Kahn Jr, Linda M. Robertsb, Katherine A. Shaffera, and Peter Haddawya. Construction of a bayesian network for mammographic diagnosis of breast cancer. *Computers in Biology and Medicine*, 27:19–29, 1997.

[21] Xiao-Hui Wang, Bin Zheng, Walter F. Good, Jill L. King, and Yuan-Hsiang Chang. Computer-assisted diagnosis of breast cancer using a data-driven bayesian belief network. *International Journal of Medical Informatics*, 54(2):115 – 126, 1999.

[22] E. Burnside, D. Rubin, and R. Shachter. A bayesian network for mammography. pages 106–110, 2000.

[23] Nicandro Cruz-Ramirez, Hector Gabriel Acosta-Mesa, Humberto Carrillo-Calvet, Luis Alonso Nava-Fernndez, and Rocio Erandi Barrientos-Martinez. Diagnosis of breast cancer using bayesian networks: A case study. *Computers in Biology and Medicine*, 37:1553–1564, 2007.

[24] Marion Verduijna, Niels Peeka, Peter M.J. Rosseeld, Evert de Jongeb, and Bas A.J.M. de Molc. Prognostic bayesian networks: I: Rationale, learning procedure, and clinical use. *Journal of Biomedical Informatics*, 40:609–618, 2007.

[25] A. Nissan, M. Protic, A. Bilchik, J. Eberhardt, G.E. Peoples, and A. Stojadinovic A. Predictive model of outcome of targeted nodal assessment in colorectal cancer. *Ann Surg.*, 251, 2010.

[26] S.F. Galan, F. Aguado, and F.J. Dez andJ. Mira. Nasonet, modeling the spread of nasopharyngeal cancer with networks of probabilistic events in discrete time. *Artificial Intelligence in Medicine*, 25:247–264, 2002.

[27] KP Exarchos, G Rigas, Y Goletsis, and DI Fotiadis. Towards building a dynamic bayesian network for monitoring oral cancer progression using time-course gene expression data. In *Information Technology and Applications in Biomedicine (ITAB), 2010 10th IEEE International Conference on*, pages 1–4. IEEE, 2010.

[28] Jonathan Agner Forsberg, John Eberhardt, Patrick J. Boland, Rikard Wedin, and John H. Healey. Estimating survival in patients with operable skeletal metastases: an application of a bayesian belief network. *PLoS One*, 6, 2011.

[29] Daniel McShan, Yi Luo, Matt Schipper, and Randall TenHaken. Bayesian decision support for adaptive lung treatments. *Journal of Physics: Conference Series*, 489(1):012053, 2014.

[30] US Congress. American recovery and reinvestment act of 2009. *Public Law*, (111-5):111, 2009.

[31] UB Kjaerulff and AL Madsen. 2010.

[32] DJ Spiegelhalter, KR Abrams, and JP Myles. John Wiley and Sons, 2004.

[33] MH Phillips, WP Smith, U Parvathaneni, and G Laramore. The role of pet in the treatment of occult disease in head and neck cancer: A modelling approach. *Int J Radiat Oncol Biol Phys*, pages 1089–1095, 2010.

[34] Robert C Lee, Edidiong Ekaette, Karie-Lynn Kelly, Peter Craighead, Chris Newcomb, and Peter Dunscombe. Implications of cancer staging uncertainties in radiation therapy decisions. *Medical decision making*, 26(3):226–238, 2006.

[35] Eric C. Ford, Ray Gaudette, Lee Myers, Bruce Vanderver, Lilly Engineer, Richard Zellars, Danny Y. Song, John Wong, and Theodore L. DeWeese. Evaluation of safety in a radiation oncology setting using failure mode and effects analysis. *International journal of radiation oncology, biology, physics*, 74(3):852–858, 2009.

[36] Brenda G. Clark, Robert J. Brown, Jodi L. Ploquin, Anneke L. Kind, and Laval Grimard. The management of radiation treatment error through incident learning. *Radiotherapy and Oncology*, 95(3):344 – 349, 2010.

[37] Margie A. Hunt, Gerri Pastrana, Howard I. Amols, Aileen Killen, and Kaled Alektiar. The Impact of New Technologies on Radiation Oncology Events and Trends in the Past Decade: An Institutional Experience. *International Journal of Radiation Oncology*Biology*Physics*, 84(4):925–931, 2012.

[38] Gregory A Patton, David K Gaffney, and John H Moeller. Facilitation of radiotherapeutic error by computerized record and verify systems. *International Journal of Radiation Oncology* Biology* Physics*, 56(1):50–57, 2003.

[39] Jean-Pierre Bissonnette and Gaylene Medlam. Trend analysis of radiation therapy incidents over seven years. *Radiotherapy and Oncology*, 96(1):139–144, 2010.

[40] Jesmin Shafiq, Michael Barton, Douglas Noble, Claire Lemer, and Liam J. Donaldson. An international review of patient safety measures in radiotherapy practice. *Radiotherapy and Oncology*, 92(1):15–21, 2009.

[41] Grace Huang, Gaylene Medlam, Justin Lee, Susan Billingsley, Jean-Pierre Bissonnette, Jolie Ringash, Gabrielle Kane, and David C Hodgson. Error in the delivery of radiation therapy: results of a quality assurance review. *International Journal of Radiation Oncology* Biology* Physics*, 61(5):1590–1595, 2005.

[42] Eric C Ford, Stephanie Terezakis, Annette Souranis, Kendra Harris, Hiram Gay, and Sasa Mutic. Quality control quantification (qcq): A tool to measure the value of quality

control checks in radiation oncology. *International Journal of Radiation Oncology* Biology* Physics*, 84(3):e263–e269, 2012.

[43] Benedick A. Fraass. Errors in radiotherapy: Motivation for development of new radiotherapy quality assurance paradigms. *International Journal of Radiation Oncology*Biology*Physics*, 71(1, Supplement):S162 – S165, 2008.

[44] Fatemeh Azmandian, David Kaeli, Jennifer G Dy, Elizabeth Hutchinson, Marek Ancukiewicz, Andrzej Niemierko, and Steve B Jiang. Towards the development of an error checker for radiotherapy treatment plans: a preliminary study. *Physics in Medicine and Biology*, 52(21):6511, 2007.

[45] Martin A Ebert, Annette Haworth, Rachel Kearvell, Ben Hooton, Rhonda Coleman, Nigel Spry, Sean Bydder, and David Joseph. Detailed review and analysis of complex radiotherapy clinical trial planning data: evaluation and initial experience with the swan software system. *Radiotherapy and Oncology*, 86(2):200–210, 2008.

[46] Ramon Alfredo Siochi, Edward Pennington, Timothy Waldron, and John Bayouth. Radiation therapy plan checks in a paperless clinic. *Journal of Applied Clinical Medical Physics*, 10(1), 2009.

[47] Eli Furhang, James Dolan, Jussi Sillanpaa, and Louis Harrison. Automating the initial physics chart checking process. *Journal of Applied Clinical Medical Physics*, 10(1), 2009.

[48] Uffe B Kjaerulff and Anders L Madsen. *Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis: A Guide to Construction and Analysis*, volume 22. Springer, 2012.

[49] Gregory Strylewicz and Jason Doctor. Evaluation of an automated method to assist with error detection in the accord central laboratory. *Clinical Trials*, 7(4):380–389, 2010.

[50] Jason N. Doctor and Greg Strylewicz. Detecting wrong blood in tube errors: Evaluation of a bayesian network approach. *Artificial Intelligence in Medicine*, 50:75–82, 2010.

[51] Q.A. Le, G. Strylewicz, and J.N. Doctor. Detecting blood laboratory errors using a bayesian network: an evaluation on liver enzyme tests. *Med. Decis. Making*, 31:325–337, 2011.

[52] Tom Wengraf. *Qualitative research interviewing: Biographic narrative and semi-structured methods*. Sage, 2001.

[53] Stig K Andersen, Kristian G Olesen, Finn Verner Jensen, and Frank Jensen. Hugin-a shell for building bayesian belief universes for expert systems. In *IJCAI*, volume 89, pages 1080–1085, 1989.

[54] Leonard E Baum. An equality and associated maximization technique in statistical estimation for probabilistic functions of markov processes. *Inequalities*, 3:1–8, 1972.

[55] EC Ford, L Fong de Los Santos, T Pawlicki, S Sutlief, and P Dunscombe. Consensus recommendations for incident learning database structures in radiation oncology. *Medical physics*, 39(12):7272–7290, 2012.

[56] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

[57] Ross Ihaka and Robert Gentleman. R: a language for data analysis and graphics. *Journal of computational and graphical statistics*, 5(3):299–314, 1996.

[58] K Konis. Rhugin. *R package version*, pages 7–5, 2011.

[59] Richard E Nisbett, David H Krantz, Christopher Jepson, and Ziva Kunda. The use of statistical heuristics in everyday inductive reasoning. *Psychological Review*, 90(4):339, 1983.

[60] L.C. Van der Gaag, S. Renooij, C.L.M. Witteman, B.M.P. Aleman, and B.G. Taal. Probabilities for a probabilistic network: a case study in oesophageal cancer. *Artificial Intelligence in Medicine*, 25:123–148, 2002.

[61] Holger Knublauch, Ray W Fergerson, Natalya F Noy, and Mark A Musen. The protégé owl plugin: An open development environment for semantic web applications. In *The Semantic Web–ISWC 2004*, pages 229–243. Springer, 2004.

[62] Thomas R Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220, 1993.

[63] Cornelius Rosse and José LV Mejino Jr. A reference ontology for biomedical informatics: the foundational model of anatomy. *Journal of biomedical informatics*, 36(6):478–500, 2003.

[64] "World Health Organization". *ICD-O: International classification of diseases for oncology*. World Health Organization, 1976.

[65] Michael Q Stearns, Colin Price, Kent A Spackman, and Amy Y Wang. Snomed clinical terms: overview of the development process and project status. In *Proceedings of the AMIA Symposium*, page 662. American Medical Informatics Association, 2001.

[66] Ralph Bergmann and Martin Schaaf. Structural case-based reasoning and ontology-based knowledge management: A perfect match? *J. UCS*, 9(7):608–626, 2003.

[67] EAM Lotfy Abdrabou and A Salem. A breast cancer classifier based on a combination of case-based reasoning and ontology approach. In *Computer Science and Information Technology (IMCSIT), Proceedings of the 2010 International Multiconference on*, pages 3–10. IEEE, 2010.

[68] M. Ramoni, M. Stefanelli, L. Magnani, and G. Barosi. An epistemological framework for medical knowledge-based systems. *Systems, Man and Cybernetics, IEEE Transactions on*, 22(6):1361–1375, 1992.

[69] Stefan Fenz, A Min Tjoa, and Marcus Hudec. Ontology-based generation of bayesian networks. In *Complex, Intelligent and Software Intensive Systems, 2009. CISIS'09. International Conference on*, pages 712–717. IEEE, 2009.

[70] Eveline M Helsper and Linda C van der Gaag. Building bayesian networks through ontologies. In *ECAI*, volume 2002, page 15th, 2002.

[71] Mathias Brochhausen, Andrew D. Spear, Cristian Cocos, Gabriele Weiler, Luis Martn, Alberto Anguita, Holger Stenzhorn, Evangelia Daskalaki, Fatima Schera, Ulf Schwarz, Stelios Sfakianakis, Stephan Kiefer, Martin Drr, Norbert Graf, and Manolis Tsiknakis. The {ACGT} master ontology and its applications towards an ontology-driven cancer research and management system. *Journal of Biomedical Informatics*, 44(1):8 – 25, 2011.

[72] Gilberto Fragoso, Sherri de Coronado, Margaret Haber, Frank Hartel, and Larry Wright. Overview and utilization of the nci thesaurus. *Comparative and Functional Genomics*, 5(8):648–654, 2004.

[73] Alexis Miller. Developing an ontology for radiation oncology. Master's thesis, University of Wollongong, 2012.

[74] Thomas M Minta. *Ontological Representation of Radiation Treatment Data*. PhD thesis, Wake Forest University, 2011.

[75] Joe Suzuki. A construction of bayesian networks from databases based on an mdl principle. In *Proceedings of the Ninth international conference on Uncertainty in artificial*

intelligence, UAI'93, pages 266–273, San Francisco, CA, USA, 1993. Morgan Kaufmann Publishers Inc.

[76] Christian Bizer and Richard Cyganiak. D2r server-publishing relational databases on the semantic web. In *5th international Semantic Web conference*, page 26, 2006.

[77] Andrew J Lyons, Siobhan Crichton, and Thomas Pezier. Trismus following radiotherapy to the head and neck is likely to have distinct genotype dependent cause. *Oral oncology*, 49(9):932–936, 2013.

[78] OCEBM Levels of Evidence Working Group et al. The oxford 2011 levels of evidence, 2011.

[79] Danny Holten. Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. *Visualization and Computer Graphics, IEEE Transactions on*, 12(5):741–748, 2006.

[80] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D^3 data-driven documents. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2301–2309, 2011.

[81] Egon Willighagen. Accessing biological data with semantic web technologies. http://dx.doi.org/10.7287/peerj.preprints.185v1, 2013.

[82] R Development Core Team. shiny: Web application framework for r, 2014.

[83] Mark Otto and Jacob Thornton, 2011.

[84] RStudio, 2014. Accessed: 2014-09-30.

[85] Evan C Osmundson, Yufan Wu, Gary Luxton, Jose G Bazan, Albert C Koong, and Daniel T Chang. Predictors of toxicity associated with stereotactic body radiation therapy to the central hepatobiliary tract. *International Journal of Radiation Oncology* Biology* Physics*, 91(5):986–994, 2015.

[86] Jason Wolfe, Aria Haghighi, and Dan Klein. Fully distributed em for very large datasets. In *Proceedings of the 25th international conference on Machine learning*, pages 1184–1191. ACM, 2008.

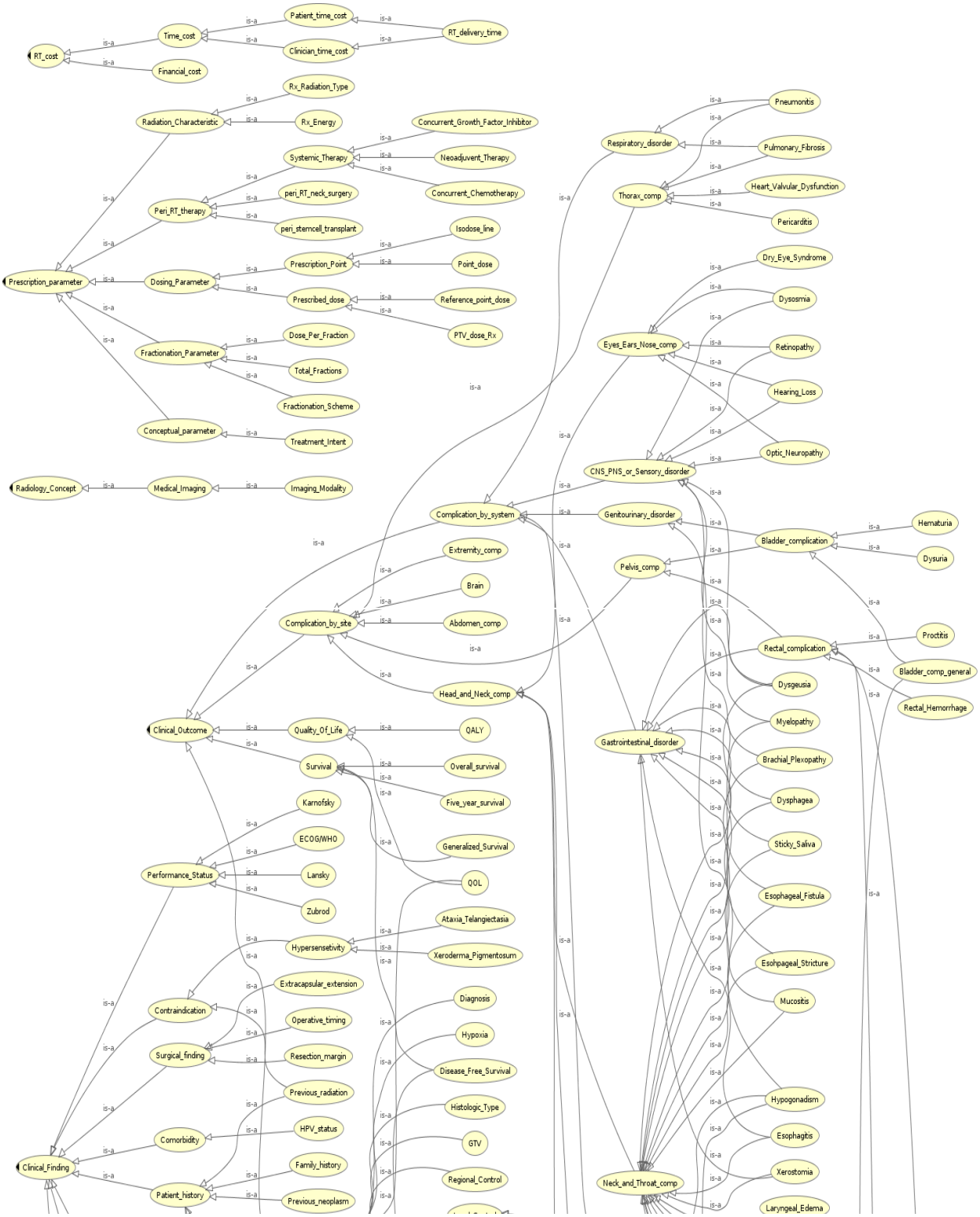
[87] Jun Zhu, Jianfei Chen, and Wenbo Hu. Big learning with bayesian methods. *CoRR*, abs/1411.6370, 2014.

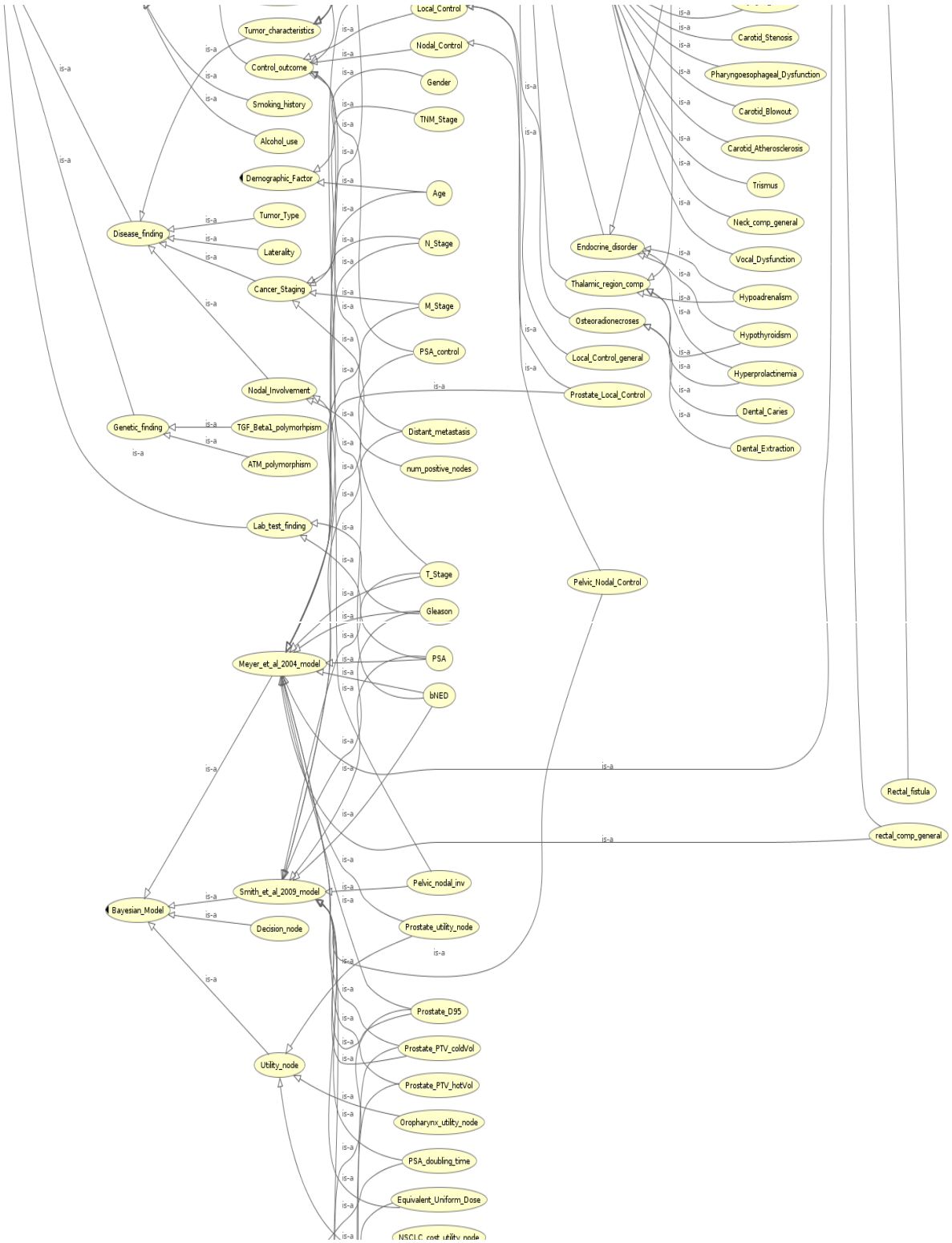
[88] MARIE DESJARDINS, PRIYANG RATHOD, and LISE GETOOR. Learning structured bayesian networks: Combining abstraction hierarchies and tree-structured conditional probability tables. *Computational Intelligence*, 24(1), 2008.

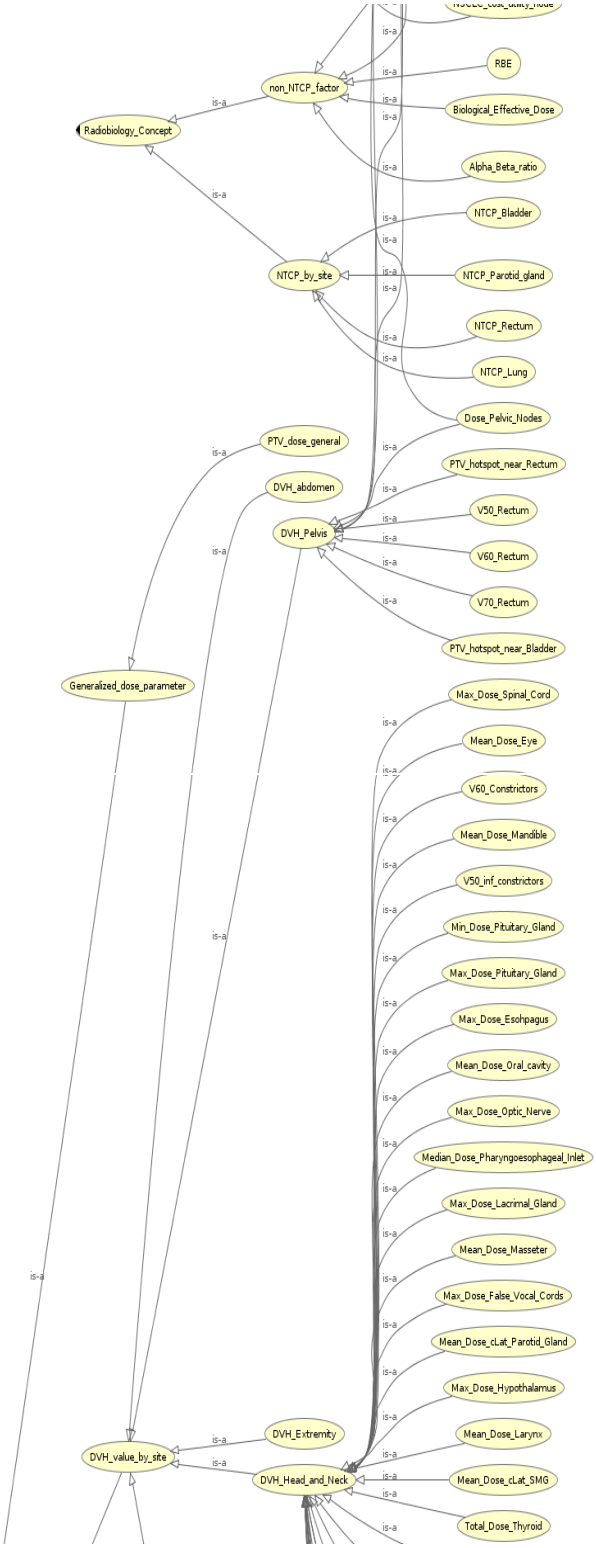
[89] Robert Mateescu and Rina Dechter. Mixed deterministic and probabilistic networks. *Annals of mathematics and artificial intelligence*, 54(1-3):3–51, 2008.

Appendix A

RADIATION ONCOLOGY ONTOLOGY FULL CLASS STRUCTURE


RT_cost
Time_cost
Financial_cost
Patient_time_cost
Clinician_time_cost
RT_delivery_time
Rx_Radiation_Type
Radiation_Characteristic
Rx_Energy
Concurrent_Growth_Factor_Inhibitor
Systemic_Therapy
Neoadjuvent_Therapy
Concurrent_Chemotherapy
peri_RT_neck_surgery
Peri_RT_therapy
peri_sterncell_transplant
Isodose_line
Prescription_Point
Point_dose
Prescription_parameter
Dosing_Parameter
Prescribed_dose
Reference_point_dose
PTV_dose_Rx
Dose_Per_Fraction
Fractionation_Parameter
Total_Fractions
Fractionation_Scheme
Conceptual_parameter
Treatment_Intent
Radiology_Concept
Medical_Imaging
Imaging_Modality
Respiratory_disorder
Pneumonitis
Pulmonary_Fibrosis
Thorax_comp
Heart_Valvular_Dysfunction
Pericarditis
Dry_Eye_Syndrome
Dysosmia
Eyes_Ears_Nose_comp
Retinopathy
Hearing_Loss
Optic_Neuropathy
CNS_PNS_or_Sensory_disorder
Complication_by_system
Genitourinary_disorder
Hematuria
Bladder_complication
Dysuria
Pelvis_comp
Extremity_comp
Brain
Complication_by_site
Abdomen_comp
Head_and_Neck_comp
Proctitis
Rectal_complication
Bladder_comp_general
Dysgeusia
Rectal_Hemorrhage
Myelopathy
Gastrointestinal_disorder
Brachial_Plexopathy
Clinical_Outcome
Quality_Of_Life
QALY
Survival
Overall_survival
Five_year_survival
Karnofsky
ECOG/WHO
Generalized_Survival
Performance_Status
Lansky
QOL
Zubrod
Dysphagea
Sticky_Saliva
Esophageal_Fistula
Ataxia_Telangiectasia
Hypersensetivity
Xeroderma_Pigmentosum
Esohpageal_Stricture
Extracapsular_extension
Mucositis
Contraindication
Diagnosis
Operative_timing
Hypoxia
Surgical_finding
Resection_margin
Disease_Free_Survival
Previous_radiation
Hypogonadism
Histologic_Type
Comorbidity
HPV_status
Esophagitis
GTV
Clinical_Finding
Family_history
Xerostomia
Regional_Control
Patient_history
Previous_neoplasm
Neck_and_Throat_comp
Laryngeal_Edema
is-a


Local_Control
Tumor_characteristics
Nodal_Control
Control_outcome
Gender
Smoking_history
TNM_Stage
Alcohol_use
Demographic_Factor
Age
Tumor_Type
Disease_finding
Laterality
N_Stage
Cancer_Staging
M_Stage
PSA_control
Nodal_Involvement
Genetic_finding
TGF_Beta1_polymorhpism
ATM_polymorphism
Distant_metastasis
num_positive_nodes
Lab_test_finding
T_Stage
Gleason
Meyer_et_al_2004_model
PSA
bNED
Smith_et_al_2009_model
Pelvic_nodal_inv
Bayesian_Model
Decision_node
Prostate_utility_node
Prostate_D95
Prostate_PTV_coldVol
Utility_node
Prostate_PTV_hotVol
Oropharynx_utility_node
PSA_doubling_time
Equivalent_Uniform_Dose
NSCLC_cost_utility_node
Endocrine_disorder
Thalamic_region_comp
Osteoradionecroses
Local_Control_general
Prostate_Local_Control
Pelvic_Nodal_Control
Carotid_Stenosis
Pharyngoesophageal_Dysfunction
Carotid_Blowout
Carotid_Atherosclerosis
Trismus
Neck_comp_general
Vocal_Dysfunction
Hypoadrenalism
Hypothyroidism
Hyperprolactinemia
Dental_Caries
Dental_Extraction
Rectal_fistula
rectal_comp_general
is-a


RBE
non_NTCP_factor
Biological_Effective_Dose
Radiobiology_Concept
Alpha_Beta_ratio
NTCP_Bladder
NTCP_by_site
NTCP_Parotid_gland
NTCP_Rectum
NTCP_Lung
Dose_Pelvic_Nodes
PTV_dose_general
PTV_hotspot_near_Rectum
DVH_abdomen
V50_Rectum
DVH_Pelvis
V60_Rectum
V70_Rectum
PTV_hotspot_near_Bladder
Generalized_dose_parameter
Max_Dose_Spinal_Cord
Mean_Dose_Eye
V60_Constrictors
Mean_Dose_Mandible
V50_inf_constrictors
Min_Dose_Pituitary_Gland
Max_Dose_Pituitary_Gland
Max_Dose_Esohpagus
Mean_Dose_Oral_cavity
Max_Dose_Optic_Nerve
Median_Dose_Pharyngoesophageal_Inlet
Max_Dose_Lacrimal_Gland
Mean_Dose_Masseter
Max_Dose_False_Vocal_Cords
Mean_Dose_cLat_Parotid_Gland
Max_Dose_Hypothalamus
DVH_Extremity
Mean_Dose_Larynx
DVH_value_by_site
DVH_Head_and_Neck
Mean_Dose_cLat_SMG
Total_Dose_Thyroid
is-a

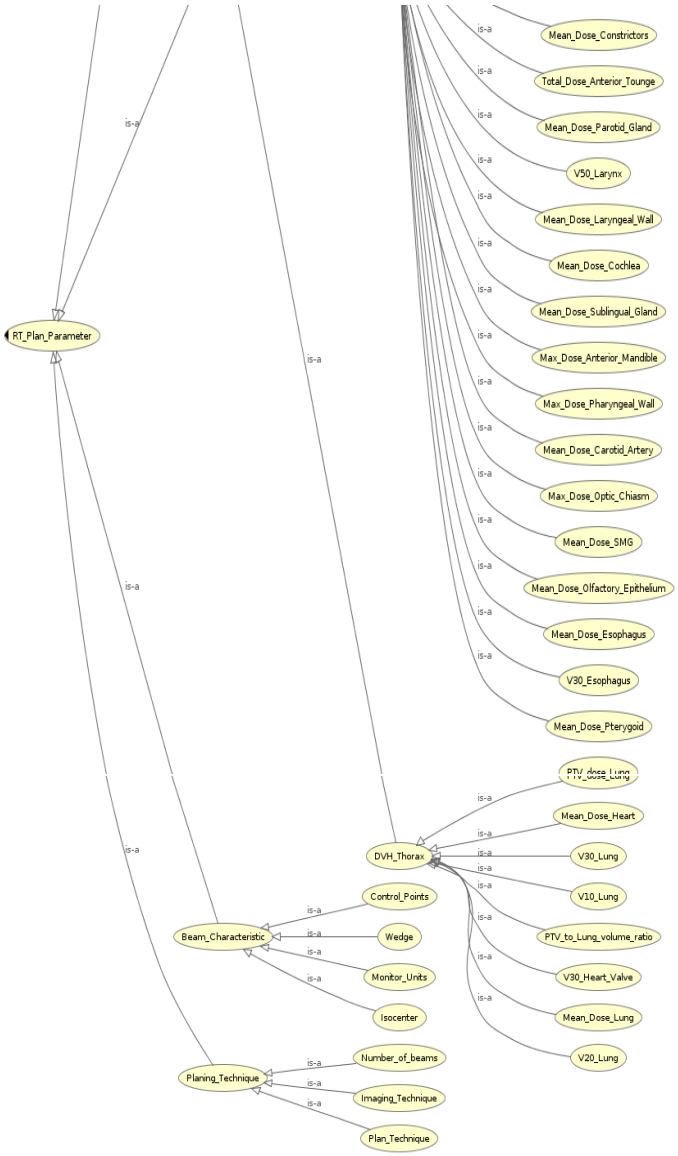


Figure A.1: Class-subclass hierarchy of the ontology for radaition oncology (version: Jan 2015)

Appendix B

BNDE FUNCTIONAL DEPENDENCY STRUCTURE

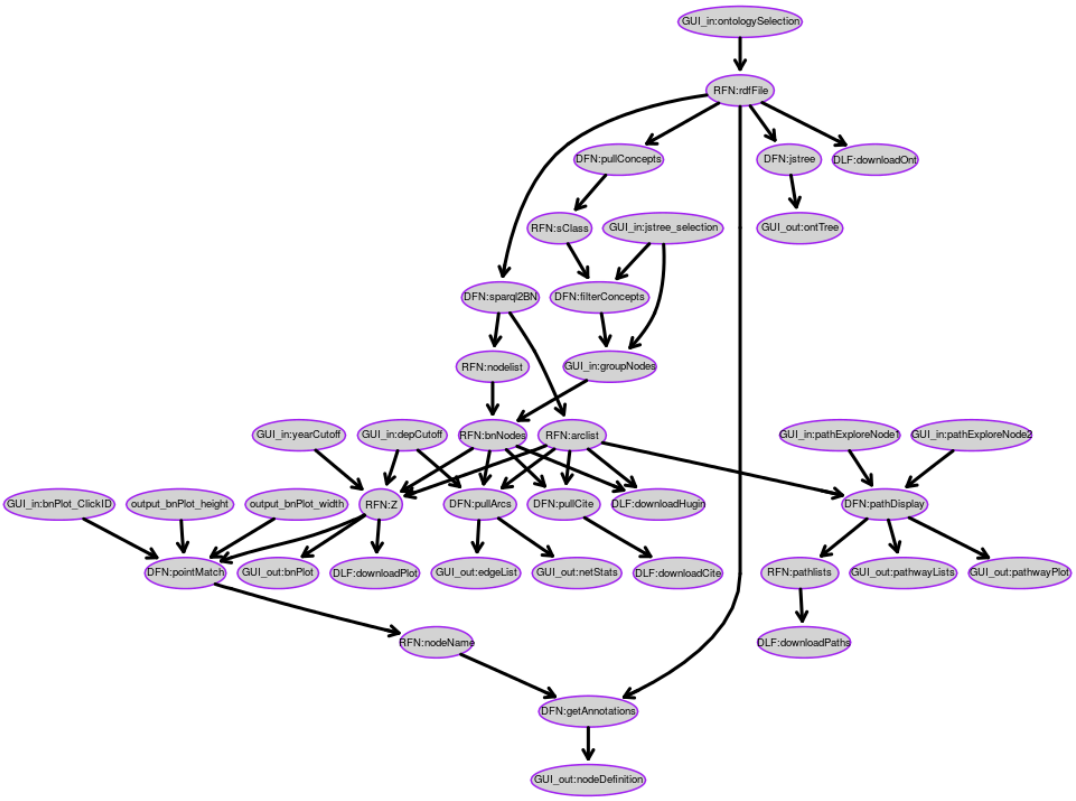


Figure B.1: Dependency graph for the Bayesian Network Domain Explorer (BNDE) web-application. Represented here are GUI inputs and outputs, download functions, developer-defined functions, and reactive elements

Appendix C

COPYRIGHTS AND PERMISSIONS

ELSEVIER LICENSE
TERMS AND CONDITIONS

Apr 27, 2015

This is a License Agreement between Alan Kalet ("You") and Elsevier ("Elsevier") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Elsevier, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

Supplier	Elsevier Limited The Boulevard,Langford Lane Kidlington,Oxford,OX5 1GB,UK
Registered Company Number	1982084
Customer name	Alan Kalet
Customer address	1959 pacific st SEATTLE, WA 98195
License number	3617231213591
License date	Apr 27, 2015
Licensed content publisher	Elsevier
Licensed content publication	International Journal of Radiation Oncology*Biology*Physics
Licensed content title	Evaluation of Safety in a Radiation Oncology Setting Using Failure Mode and Effects Analysis
Licensed content author	Eric C. Ford,Ray Gaudette,Lee Myers,Bruce Vanderver,Lilly Engineer,Richard Zellars,Danny Y. Song,John Wong,Theodore L. DeWeese
Licensed content date	1 July 2009
Licensed content volume number	74
Licensed content issue number	3
Number of pages	7
Start Page	852
End Page	858
Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	2
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No

Original figure numbers	Fig. 1 and Fig. 2
Title of your thesis/dissertation	Bayesian network models from ontological formalisms in radiation oncology
Expected completion date	Jun 2015
Estimated size (number of pages)	130
Elsevier VAT number	GB 494 6272 12
Price	0.00 USD
VAT/Local Sales Tax	0.00 USD / 0.00 GBP
Total	**0.00 USD**

Terms and Conditions

INTRODUCTION

1. The publisher for this copyrighted material is Elsevier. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).

GENERAL TERMS

2. Elsevier hereby grants you permission to reproduce the aforementioned material subject to the terms and conditions indicated.
3. Acknowledgement: If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies. Suitable acknowledgement to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:
"Reprinted from Publication title, Vol /edition number, Author(s), Title of article / title of chapter, Pages No., Copyright (Year), with permission from Elsevier [OR APPLICABLE SOCIETY COPYRIGHT OWNER]." Also Lancet special credit - "Reprinted from The Lancet, Vol. number, Author(s), Title of article, Pages No., Copyright (Year), with permission from Elsevier."
4. Reproduction of this material is confined to the purpose and/or media for which permission is hereby given.
5. Altering/Modifying Material: Not Permitted. However figures and illustrations may be altered/adapted minimally to serve your work. Any other abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of Elsevier Ltd. (Please contact Elsevier at permissions@elsevier.com)
6. If the permission fee for the requested use of our material is waived in this instance, please be advised that your future requests for Elsevier materials may attract a fee.
7. Reservation of Rights: Publisher reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.
8. License Contingent Upon Payment: While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by publisher or by CCC) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.
9. Warranties: Publisher makes no representations or warranties with respect to the licensed

material.
10. Indemnity: You hereby indemnify and agree to hold harmless publisher and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.
11. No Transfer of License: This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without publisher's written permission.
12. No Amendment Except in Writing: This license may not be amended except in a writing signed by both parties (or, in the case of publisher, by CCC on publisher's behalf).
13. Objection to Contrary Terms: Publisher hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and publisher (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.
14. Revocation: Elsevier or Copyright Clearance Center may deny the permissions described in this License at their sole discretion, for any reason or no reason, with a full refund payable to you. Notice of such denial will be made using the contact information provided by you. Failure to receive such notice will not alter or invalidate the denial. In no event will Elsevier or Copyright Clearance Center be responsible or liable for any costs, expenses or damage incurred by you as a result of a denial of your permission request, other than a refund of the amount(s) paid by you to Elsevier and/or Copyright Clearance Center for denied permissions.

LIMITED LICENSE

The following terms and conditions apply only to specific license types:
15. **Translation**: This permission is granted for non-exclusive world **English** rights only unless your license was granted for translation rights. If you licensed translation rights you may only translate this content into the languages you requested. A professional translator must perform all translations and reproduce the content word for word preserving the integrity of the article. If this license is to re-use 1 or 2 figures then permission is granted for non-exclusive world rights in all languages.
16. **Posting licensed content on any Website**: The following terms and conditions apply as follows: Licensing material from an Elsevier journal: All content posted to the web site must maintain the copyright information line on the bottom of each image; A hyper-text must be included to the Homepage of the journal from which you are licensing at http://www.sciencedirect.com/science/journal/xxxxx or the Elsevier homepage for books at http://www.elsevier.com; Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu. Licensing material from an Elsevier book: A hyper-text link must be included to the Elsevier homepage at http://www.elsevier.com . All content posted to the web site must maintain the copyright information line on the bottom of each image.

Posting licensed content on Electronic reserve: In addition to the above the following clauses are applicable: The web site must be password-protected and made available only to bona fide students registered on a relevant course. This permission is granted for 1 year only. You may obtain a new license for future website posting.
17. **For journal authors:** the following clauses are applicable in addition to the above:
Preprints:
A preprint is an author's own write-up of research results and analysis, it has not been peer-reviewed, nor has it had any other value added to it by a publisher (such as formatting, copyright, technical enhancement etc.).
Authors can share their preprints anywhere at any time. Preprints should not be added to or enhanced in any way in order to appear more like, or to substitute for, the final versions of articles however authors can update their preprints on arXiv or RePEc with their Accepted Author Manuscript (see below).

If accepted for publication, we encourage authors to link from the preprint to their formal publication via its DOI. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help users to find, access, cite and use the best available version. Please note that Cell Press, The Lancet and some society-owned have different preprint policies. Information on these policies is available on the journal homepage.
Accepted Author Manuscripts: An accepted author manuscript is the manuscript of an article that has been accepted for publication and which typically includes author-incorporated changes suggested during submission, peer review and editor-author communications.
Authors can share their accepted author manuscript:

- immediately
  - via their non-commercial person homepage or blog
  - by updating a preprint in arXiv or RePEc with the accepted manuscript
  - via their research institute or institutional repository for internal institutional uses or as part of an invitation-only research collaboration work-group
  - directly by providing copies to their students or to research collaborators for their personal use
  - for private scholarly sharing as part of an invitation-only work group on commercial sites with which Elsevier has an agreement
- after the embargo period
  - via non-commercial hosting platforms such as their institutional repository
  - via commercial sites with which Elsevier has an agreement

In all cases accepted manuscripts should:

- link to the formal publication via its DOI
- bear a CC-BY-NC-ND license - this is easy to do
- if aggregated with other manuscripts, for example in a repository or other site, be shared in alignment with our hosting policy not be added to or enhanced in any way to appear more like, or to substitute for, the published journal article.

Published journal article (JPA): A published journal article (PJA) is the definitive final record of published research that appears or will appear in the journal and embodies all value-adding publishing activities including peer review co-ordination, copy-editing, formatting, (if relevant) pagination and online enrichment.
Policies for sharing publishing journal articles differ for subscription and gold open access articles:
Subscription Articles: If you are an author, please share a link to your article rather than the full-text. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help your users to find, access, cite, and use the best available version.
Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.
If you are affiliated with a library that subscribes to ScienceDirect you have additional private sharing rights for others' research accessed under that agreement. This includes use for classroom teaching and internal training at the institution (including use in course packs and courseware programs), and inclusion of the article for grant funding purposes.
Gold Open Access Articles: May be shared according to the author-selected end-user license and should contain a CrossMark logo, the end user license, and a DOI link to the formal publication on ScienceDirect.
Please refer to Elsevier's posting policy for further information.
18. **For book authors** the following clauses are applicable in addition to the above: Authors are permitted to place a brief summary of their work online only. You are not allowed to download and post the published electronic version of your chapter, nor may you scan the printed edition to create an electronic version. **Posting to a repository:** Authors are permitted to post a summary of their chapter only in their institution's repository.
19. **Thesis/Dissertation**: If your license is for use in a thesis/dissertation your thesis may be

submitted to your institution in either print or electronic form. Should your thesis be published commercially, please reapply for permission. These requirements include permission for the Library and Archives of Canada to supply single copies, on demand, of the complete thesis and include permission for Proquest/UMI to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission. Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.

Elsevier Open Access Terms and Conditions

You can publish open access with Elsevier in hundreds of open access journals or in nearly 2000 established subscription journals that support open access publishing. Permitted third party re-use of these open access articles is defined by the author's choice of Creative Commons user license. See our open access license policy for more information.

Terms & Conditions applicable to all Open Access articles published with Elsevier:

Any reuse of the article must not represent the author as endorsing the adaptation of the article nor should the article be modified in such a way as to damage the author's honour or reputation. If any changes have been made, such changes must be clearly indicated.

The author(s) must be appropriately credited and we ask that you include the end user license and a DOI link to the formal publication on ScienceDirect.

If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source it is the responsibility of the user to ensure their reuse complies with the terms and conditions determined by the rights holder.

Additional Terms & Conditions applicable to each Creative Commons user license:

CC BY: The CC-BY license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article and to make commercial use of the Article (including reuse and/or resale of the Article by commercial entities), provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. The full details of the license are available at http://creativecommons.org/licenses/by/4.0.

CC BY NC SA: The CC BY-NC-SA license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article, provided this is not done for commercial purposes, and that the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. Further, any new works must be made available on the same conditions. The full details of the license are available at http://creativecommons.org/licenses/by-nc-sa/4.0.

CC BY NC ND: The CC BY-NC-ND license allows users to copy and distribute the Article, provided this is not done for commercial purposes and further does not permit distribution of the Article if it is changed or edited in any way, and provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, and that the licensor is not represented as endorsing the use made of the work. The full details of the license are available at http://creativecommons.org/licenses/by-nc-nd/4.0. Any commercial reuse of Open Access articles published with a CC BY NC SA or CC BY NC ND license requires permission from Elsevier and will be subject to a fee.

Commercial reuse includes:

- Associating advertising with the full text of the Article
- Charging fees for document delivery or access
- Article aggregation
- Systematic distribution via e-mail lists or share buttons

Posting or linking by commercial companies for use by customers of those companies.

20. **Other Conditions**:

v1.7

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

Gratis licenses (referencing $0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.

VITA

Alan Michael Kalet was born in Philadelphia, Pennsylvania and studied at the University of Washington where he earned a bachelor's degree in Physics. He went on to pursue graduate level Physics at the University of California, Riverside and earned a masters degree there before returning to Seattle, WA for clinical medical physics training at the University of Washington Medical Center. Post residency training, he continued on at the University of Washington Medical Center as a medical physicist where he decided to pursue his interests in decision support tool development in radiation oncology. In 2015, he earned his Doctor of Philosophy from the University of Washington's Biomedical and Health Informatics program.