

REPRESENTING AND REASONING WITH CLINICAL KNOWLEDGE IN
RADIATION THERAPY PUBLICATIONS: A STEP TOWARDS
EVIDENCED-BASED MEDICINE

by

Yi Zhen

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Computing and Information Systems

Charlotte

2019

Approved by:

Dr. Yaorong Ge

Dr. Mirsad Hadzikadic

Dr. Jackie Wu

Dr. Wlodek Zadrozny

ABSTRACT

YI ZHEN. Representing and Reasoning with Clinical Knowledge in Radiation Therapy Publications: A step towards evidenced-based medicine. (Under the direction of DR. YAORONG GE)

To improve the quality and consistency of health care, evidence-based medicine (EBM) was proposed to promote the wide adoption of current best evidence to make decisions about care of individual patients. The practice of EBM in Radiation Oncology is a process of integrating clinical expertise, patient's expectation, and research evidence to support decision-making in Radiation Therapy (RT). One of the goals of the RT decision-making aims to design an ideal RT plan that achieves most damage to target treatment site and least harm to surrounding healthy organs and tissues. This aim requires radiation oncologists to understand and maintain up-to-date Radiation Oncology knowledge, also known as the clinical evidence published in clinical guidelines and clinical research studies, such as radiation-induced adverse events, dosimetric criteria recommendation, and meta-analysis of randomized controlled clinical trials.

As the amount of clinical evidence increases, it is becoming increasingly difficult for clinicians to maintain and adopt the best and most up-to-date clinical evidence in their clinical practices. This demands the development of effective systems for automated and intelligent clinical decision support (CDS), which relies on knowledge engineering methods to translate narrative clinical knowledge into computable forms and enable reasoning of the computerized knowledge. In the domain of RT, we believe that computerized Radiation Oncology knowledge will improve the ability and quality of intelligent decision-making in an efficient way.

This dissertation aimed to advance the state-of-the-art of research towards evidence-based medicine in general and with a specific focus on intelligent decision support for radiation therapy. First, we explored radiation-induced adverse events and their

grading standards used in clinical research studies using ontological modeling and text mining methods. Second, we investigated the challenges and developed a framework for extracting Radiation Oncology knowledge from clinical guidelines and clinical research studies. Third, we focused on the specific and challenging problem of uncertainty nature of human biological systems and biomedical research approaches. Toward this end, we investigated the feasibility of probabilistic models for representing extracted RT knowledge and the ability of performing reasoning. Specifically, we developed novel methods to encode uncertain Radiation Oncology knowledge using Markov Logic Networks and conducted a study of quantifying uncertainties in Radiation Oncology clinical evidence. We demonstrated the feasibility of using the proposed methods as a general knowledge engineering framework for representing complex and uncertain knowledge in Radiation Oncology for decision support.

ACKNOWLEDGEMENTS

First, I would especially like to thank my advisor Dr. Yaorong Ge, for his helpful advice and support during my seven years' PhD life. As my teacher and mentor, he has shown me by example what a good scientist and person should be. Thanks to his patience and encouragement, I learned how to manage time, pressure and emotions when I encountered failures and rejections. More importantly, I learned how to be happy and maintain a balance between scientific research and my personal life through each conversation with him.

Second, I would like to thank the committee members for providing me extensive professional guidance and academic feedbacks to help me improve this dissertation from its proposal to the final writing. Dr. Mirsad Hadzikadic suggested me to elaborate my contribution to the academic community at a PhD level. Dr. Jackie Wu and her team in Department of Radiation Oncology provided me professional advice and research support as domain experts. And Dr. Wlodek Zadrozny taught me general text mining methods and introduced me the current progress of question-answer systems in medical field.

I am grateful to all of those with whom I have had the pleasure to work during this and other related projects, including my lab mates in Health Informatics Lab at UNC Charlotte, colleagues from Department of Radiation Oncology, Duke Medical Center, and my fellow internship friends at Philips Research. I also would like to thank Dr. Albert Park, Dr. Charles Yee and Dr. Yong Mao as honest friends whom encouraged me to pursue my career goal.

It is never too late to thank my family for their best education and support. I would like to thank my parents and my younger brother, whose love and tolerance are always with me in whatever I pursue.

TABLE OF CONTENTS

LIST OF TABLES	ix
LIST OF FIGURES	x
LIST OF ABBREVIATIONS AND SYMBOLS	xiii
CHAPTER 1: INTRODUCTION	1
1.1. Radiation Therapy	3
1.2. Clinical Knowledge in Radiation Therapy	6
1.3. Knowledge Representation Formalisms	10
1.4. Summary of Contributions	13
CHAPTER 2: REPRESENTING STANDARD FOR ADVERSE EVENTS WITH ONTOLOGICAL MODELING	16
2.1. Background	16
2.2. Related Work	19
2.3. Methods	21
2.3.1. Five Steps of Generating An Ontology for CTCAE	21
2.3.2. Mapping Rules	26
2.4. Experiments	28
2.5. Discussion	32
2.6. Summary	32
CHAPTER 3: UNDERSTANDING THE GRADING STANDARDS FOR RADIATION-INDUCED ADVERSE EVENTS IN RT STUDIES	34
3.1. Grading Standards for Radiation-Induced Adverse Events	34

3.2. Methods	35
3.2.1. Data Preprocessing	36
3.2.2. Feature Extraction	36
3.2.3. Classifier Modeling	36
3.2.4. Cancer Type Identification	39
3.3. Experiments	40
3.3.1. Materials	40
3.3.2. Training and Validation of Classifiers	41
3.3.3. Usage Trends of Grading Standards for Radiation-Induced Adverse Events During 2010-2015	44
3.4. Discussion	45
3.4.1. Error Analysis	46
3.4.2. Limitations	47
3.4.3. Strategies to Improve CTCAE	47
CHAPTER 4: CLINICAL KNOWLEDGE REPRESENTATION AND REASONING FOR RT TREATMENT PLANNING WITH MARKOV LOGIC NETWORKS	51
4.1. Knowledge Engineering with Markov Logic Networks	51
4.2. Methods	53
4.2.1. Materials	53
4.2.2. Overview of Proposed Framework	53
4.2.3. Structure Learning	54
4.2.4. Weight Learning	57
4.2.5. Inference	58

4.3. A Knowledge Representation Example of Generated MLNs	59
4.4. Discussion	62
CHAPTER 5: Quantifying Uncertainty of Hedging Terms in Radiation Oncology Knowledge	63
5.1. Related Works	64
5.2. Study Design	65
5.2.1. Select 14 Hedging Terms as Research Objects	65
5.2.2. Experiments: Context-Domain V.S Context-Specific	66
5.2.3. Survey Questions	67
5.3. Results	70
5.3.1. Demographic Information	70
5.3.2. The Ranking of Hedging Terms	70
5.3.3. The Probablistic Scores of Hedging Terms	70
5.4. Conclusion	83
CHAPTER 6: CONCLUSIONS	84
6.1. Limitations	84
6.2. Future Direction of Research Work	85
REFERENCES	86

LIST OF TABLES

TABLE 2.1: The comparison between our generated CTCAE ontology and CTCAE OWL using the NCBO Biportal Ontology Metrics.	29
TABLE 3.1: The definitions and example of text component used in the regular expressions discovery process.	37
TABLE 3.2: Confusion matrix of the RE based classifier.	42
TABLE 3.3: Confusion matrix of the Naïve Bayes classifier.	42
TABLE 3.4: The top adverse event types reported with occurrence $> 55\%$ for the four major cancer types in RT clinical studies.	50
TABLE 4.1: The table of variables for constructing MLNs in RT	56
TABLE 4.2: The table of predicates for constructing MLNs in RT	57
TABLE 4.3: The results of making inference	61

LIST OF FIGURES

FIGURE 1.1: The principles of Radiation therapy.	5
FIGURE 1.2: The pyramid of clinical knowledge in radiation therapy.	7
FIGURE 1.3: The outline of the QUANTEC reviews.	9
FIGURE 2.1: The example of AEs and their grading definitions in CTCAE with tabular format.	18
FIGURE 2.2: The process of ontology generation.	22
FIGURE 2.3: The step of converting CTCAE table into an XML.	23
FIGURE 2.4: An example of extracted XML schema defined with XSD language.	24
FIGURE 2.5: An example of generated XML-Schema Graph(XSG).	25
FIGURE 2.6: An example of generating OWL entities from XSG.	26
FIGURE 2.7: The comparison between CTCAE OWL and our CTCAE ontology viewed under Protégé.	31
FIGURE 3.1: The evolution of the F-measures during the iterative training process of RE based classifier	41
FIGURE 3.2: The usage trends of the three grading standards during 2010-2012	43
FIGURE 3.3: The usage trends of the three grading standards from 2010 to mid 2015	44
FIGURE 3.4: The usage proportion of the three grading standard grouped by cancer types	45
FIGURE 3.5: Total numbers of adverse event types reported for four major cancer types and average, minimal, and maximal numbers of adverse event types per article.	49
FIGURE 4.1: The workflow of representing RT planning knowledge with Markov Logic Networks (MLNs) framework.	54

FIGURE 4.2: The example of a knowledge statement, its first-order logic translation, and the corresponding MLNs script	60
FIGURE 5.1: The top 30 frequently occurred hedging terms in QUANTEC.	66
FIGURE 5.2: The demographic question in the survey.	68
FIGURE 5.3: The ranking question in the survey.	68
FIGURE 5.4: The context-domain questions in the survey.	69
FIGURE 5.5: The context-specific question in the survey.	69
FIGURE 5.6: The ranking distribution of the 14 hedging terms.	70
FIGURE 5.7: The Box Plot of probabilistic scores for the 14 hedging terms on point estimate.	71
FIGURE 5.8: The Box Plot of probabilistic scores for the 14 hedging terms on upper bound.	72
FIGURE 5.9: The Box Plot of probabilistic scores for the 14 hedging terms on lower bound.	73
FIGURE 5.10: The mean values of probabilistic score for each hedging term	74
FIGURE 5.11: The Box Plot of probabilistic scores for the hedging term <i>'common'</i> .	75
FIGURE 5.12: The Box Plot of probabilistic scores for the hedging term <i>'could be'</i> .	75
FIGURE 5.13: The Box Plot of probabilistic scores for the hedging term <i>'frequently'</i> .	76
FIGURE 5.14: The Box Plot of probabilistic scores for the hedging term <i>'high risk'</i> .	76
FIGURE 5.15: The Box Plot of probabilistic scores for the hedging term <i>'likely'</i> .	77

FIGURE 5.16: The Box Plot of probablistic scores for the hedging term ' <i>may</i> '.	77
FIGURE 5.17: The Box Plot of probablistic scores for the hedging term ' <i>might</i> '.	78
FIGURE 5.18: The Box Plot of probablistic scores for the hedging term ' <i>possible</i> '.	78
FIGURE 5.19: The Box Plot of probablistic scores for the hedging term ' <i>potential</i> '.	79
FIGURE 5.20: The Box Plot of probablistic scores for the hedging term ' <i>probably</i> '.	79
FIGURE 5.21: The Box Plot of probablistic scores for the hedging term ' <i>rarely</i> '.	80
FIGURE 5.22: The Box Plot of probablistic scores for the hedging term ' <i>risk of</i> '.	80
FIGURE 5.23: The Box Plot of probablistic scores for the hedging term ' <i>suggest</i> '.	81
FIGURE 5.24: The Box Plot of probablistic scores for the hedging term ' <i>usually</i> '.	81
FIGURE 5.25: The One-Way ANOVA of mean probablistic scores on five groups.	82
FIGURE 5.26: The One-Way ANOVA of mean probablistic scores on two groups.	82

LIST OF ABBREVIATIONS AND SYMBOLS

Abbreviations

AE	An acronym for Adverse Event
CDS	An acronym for clinical decision support.
CTCAE	An acronym for Common Terminology Criteria for Adverse Events
EBM	An acronym for Evidence-Based Medicine.
LENT-SOMA	An acronym for Late Effects Normal Tissue Task Force-Subjective, Objective, Management and Analytic
MLNs	An acronym for Markov Logic Networks.
OWL	An acronym for Web Ontology Language
PSL	An acronym for Probabilistic Soft Logic.
RE	An acronym for Web Regular Expression
RT	An acronym for Radiation Therapy.
RTOG	An acronym for Radiation Therapy Oncology Group
XML	An acronym for eXtensible Markup Language
XSG	An acronym for XML schema graph

Symbols

F	The first-order-logic formula.
w	The associated weight with its formula.

CHAPTER 1: INTRODUCTION

Evidence-based medicine (EBM) has been proposed as one of the most significant developments in the practice of medicine as a result of the need to cope with information overload and scientific decision-making over the last three decades. EBM is defined as an integration of the best clinical evidence that is currently available from systematic research with individual clinical expertise and patient values to support clinical decision-making [1]. To improve the quality and consistency of health care, EBM has evolved as a promising tool to promote the broad adoption of the current best evidence to make decisions about the care of individual patients [2]. With the rapid growth of clinical evidence in published research and the explosive increase of information in clinical environments worldwide, it became difficult and, in some cases, no longer possible for individual medical professionals to adopt, manage and maintain up-to-date clinical evidence or knowledge upon which to base their decisions. In particular, the practice of EBM is critical and challenging for radiation oncologists, medical physicists and dosimetrists to support effective and scientific decision-making in radiation therapy (RT), a cancer treatment that incorporates the best clinical evidence of radiation oncology into treatment decisions. This situation demands the development of effective and efficient systems for automated and intelligent clinical decision support in order to implement the practice of EBM in RT.

To promote EBM, the health care community developed models, protocols, and tools for clinical practitioners to train, educate, guide, and facilitate the practice of EBM [3]. The fundamental model of the practice of EBM comprises five steps: (1) formulate the need into an answerable question; (2) acquire the best research evidence to answer the formulated question; (3) critically appraise that evidence for its

validity, impact, and applicability; (4) apply the appraised evidence; and (5) evaluate the effectiveness and efficiency in executing these steps. Since informatics has been accepted and utilized in health care, the practice of EBM has become more practical in transferring the best evidence into clinical practice. Recent efforts have been made in the development of intelligent systems for each step of the EBM practice model [4, 5, 6, 7, 8]. More specifically, various systems and methods have been developed for automated clinical evidence retrieval [9, 10, 11, 12], automatic summarization of clinical research studies [13, 14], clinical evidence grading [15, 16], computerized clinical guidelines [17, 18, 19, 20, 21], and intelligent decision support in health care [22, 23, 24, 25]. While these methods and systems have improved the practice of EBM, few efforts have tackled the limitations and expanded the benefits of implementing the practice of EBM in RT because of the following major challenges and knowledge gaps. Firstly, the search and identification of up-to-date clinical knowledge on RT, with its different formats and heterogeneous resources, is labor-intensive and prone to error. Secondly, clinical knowledge in RT is essentially a form of statistical knowledge that captures the generalities of classes of patients rather than the peculiarities of a specific patient [26]. Radiation oncologists are required to manually specialize such highly generalized knowledge based on patients' individual characteristics [27]. Thirdly, the uncertainty of clinical knowledge in RT is challenging to understand and interpret using computer agents when determining the strength of recommendation for clinical decision-making.

Therefore, a critical step toward the practice of EBM in RT is the development of a system or a knowledge base to allow practitioners to acquire up-to-date clinical knowledge in RT publications, thus enabling them to utilize and specialize clinical knowledge for RT decision-making with uncertainty. This dissertation investigates approaches to representing clinical knowledge in RT publications and proposes a framework based on knowledge engineering methods to bridge the gap between sci-

entific research and clinical practices in RT in an efficient and intelligent way.

In the following sections, we provide some background information related to the domain of RT, clinical knowledge in RT publications, and knowledge representation formalisms. In the next two chapters, we use text mining methods and ontological modeling to explore radiation-induced adverse events and their grading standards as used in RT research studies. Chapter 3 also presents the results of the extraction of clinical knowledge from RT publications with term identification methods. In Chapter 4, we describe the proposed method of representing clinical knowledge in RT and the results of knowledge reasoning with uncertainty. Chapter 5 illustrates a study on quantifying the uncertainty conveyed by hedging terms in RT publications. In the final chapter, we conclude the dissertation and make recommendations for future work.

1.1 Radiation Therapy

Radiation Therapy (RT) is currently one of the most effective cancer treatment modalities by which to cure cancer, reduce cancer symptoms, and prevent cancer from spreading. According to the American Cancer Society and the National Cancer Institute, from 2012 to 2019, more than half of cancer patients received RT either as a sole cancer treatment or in combination with other treatments [28]. Unlike chemotherapy, which usually exposes the whole human system to anti-cancer drugs, RT is usually a local treatment that aims to do the most damage to cancer cells with the least possible harm to surrounding healthy cells. RT uses high-energy radiation to kill cancer cells and shrink tumors by destroying the DNA of targeted cancer cells, thereby preventing them from growing and dividing. Two primary types of RT are frequently adopted in clinical practices: external beam RT and internal RT. External beam RT applies a machine outside the patient’s body to deliver high-energy radiation beams from many directions to a specific part of the body that contains the tumor, whereas internal RT involves the implanting of a small amount of radioactive material

inside the patient’s body within or near the tumor. The type of RT that is adopted depends on the type of cancer and the location of the tumor. External beam RT is a widely used treatment for most cancer types and has different treatment techniques, such as three-dimensional conformal RT (3D-CRT), intensity-modulated RT (IMRT), and image-guided RT (IGRT). Internal RT or brachytherapy is usually used to treat several types of cancers, including head and neck cancer, breast cancer, cervix cancer, and prostate cancer.

RT not only destroys cancer cells with radiation but also causes side effects by exposing normal cells to radiation, as illustrated in Figure 1.1. In this dissertation, we refer to the side effects that follow RT as radiation-induced adverse events. Radiation-induced adverse events - also known as radiation toxicity on normal tissues and organs - are defined as the inevitable damage caused by RT. For instance, fatigue and hair loss are two commonly reported radiation-induced adverse events after RT. Like any cancer treatment, the therapeutic benefit of RT is balanced against potential radiation-induced adverse events. Therefore, radiation-induced adverse events are critically important factors for RT treatment planning, RT outcome evaluation, and treatment safety and quality control. Over the past decades, many clinical studies and trials have explored the balance between tumor control performance and radiation-induced adverse events [29, 30, 31]. A number of standard scoring criteria have been proposed and used in RT clinical studies for reporting and grading radiation-induced adverse events, such as the Common Terminology Criteria for Adverse Events (CTCAE) [32], the Radiation Therapy Oncology Group (RTOG) [33], and the Late Effects Normal Tissue Task Force-Subjective, Objective, Management and Analytic (LENT-SOMA) [34].

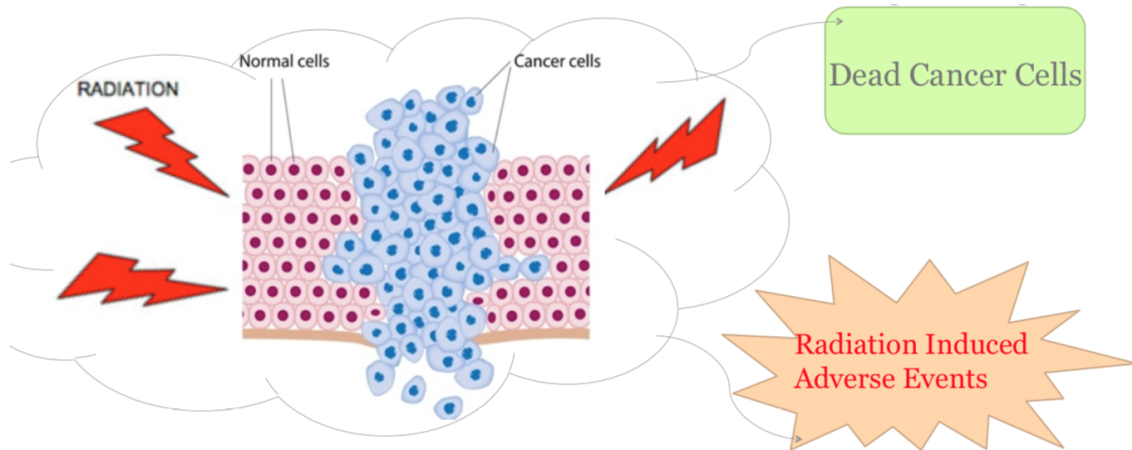


Figure 1.1: RT not only kills cancer cells with high-energy radiation but also causes radiation-induced adverse events by damaging surrounding normal cells

RT is a highly sophisticated technology that requires a team of radiation oncologists, radiation therapists, medical physicists, and dosimetrists to spend significant time and effort in the care of cancer patients. The RT process usually involves the following five steps: (1) in the consultation step, the radiation oncologist diagnoses the cancer type and stage, identifies the tumor with radiology imaging and determines the RT treatment techniques; (2) once the treatment technique determined, a radiation simulation is performed on the patient to outline the exact treatment area, determine the best angle and location for radiation treatment, and ensure immobilization if needed by the RT team; (3) after simulation, the RT team develops a unique optimal treatment plan for the patient and evaluates the plan by checking its quality and safety iteratively in the RT treatment planning step; (4) the radiation therapists deliver the treatment to the patient by following the generated plan; and (5) the radiation oncologist reports and manages the side effects after treatment delivery in the follow-up stage. The practice of radiation therapy requires the RT team members to combine their expertise and knowledge of radiation therapy with patients' values. For the recent thirty years, each step of the RT process has involved computer modeling and information techniques to assist decision-making. Among these steps, treatment

planning forms the core of RT and is the field with the most potential for the application of intelligent systems. In this dissertation, we focus on the improvement of treatment planning by representing, managing and reasoning clinical knowledge from RT publications.

An optimal RT plan maximizes damage to the cancer cells with high-energy radiation while minimizing damage to the normal cells surrounding the cancer cells. In order to develop an effective and accurate cancer treatment plan with the fewest radiation-induced adverse events, radiation oncologists rely on a number of radiation therapy publications during RT planning by following clinical guidelines such as the Quantitative Analyses of Normal Tissue Effects in the Clinic (QUANTEC) and referencing large numbers of peer-reviewed clinical research papers. During the process of RT decision-making, radiation oncologists must take complex clinical knowledge into account, including the prescribed dosage, the volume of tissue affected by radiation, and the risk of radiation-induced adverse events. Ultimately, the dissertation aims to assist radiation oncologists in making intelligent decisions about treatment plans with less risk of radiation-induced adverse events by representing and reasoning clinical knowledge in RT publications in computerized form.

1.2 Clinical Knowledge in Radiation Therapy

In order to make scientific clinical decisions in RT planning, medical professionals must adopt the best and most up-to-date clinical knowledge provided in RT publications. Clinical knowledge in RT publications exists in two types of resources: primary original research and secondary pre-appraised research. Primary original research contains original research results and synopses without a critical evaluation process and includes observational studies and randomized controlled trials. Secondary pre-appraised research is filtered information research that has passed critical appraisal and evaluation by experts and authorities; it includes meta-analysis, systematic review, and clinical practice guidelines. The pyramid diagram in Figure 1.2 visualizes

the ranking of clinical knowledge resources based on their strength of recommendation and reliability in clinical decision-making. The reliability level of clinical knowledge becomes higher and more robust from expert opinions at the bottom to clinical practice guidelines at the top. In the practice of EBM in RT, the RT practitioners should select as high a level of evidence as possible.

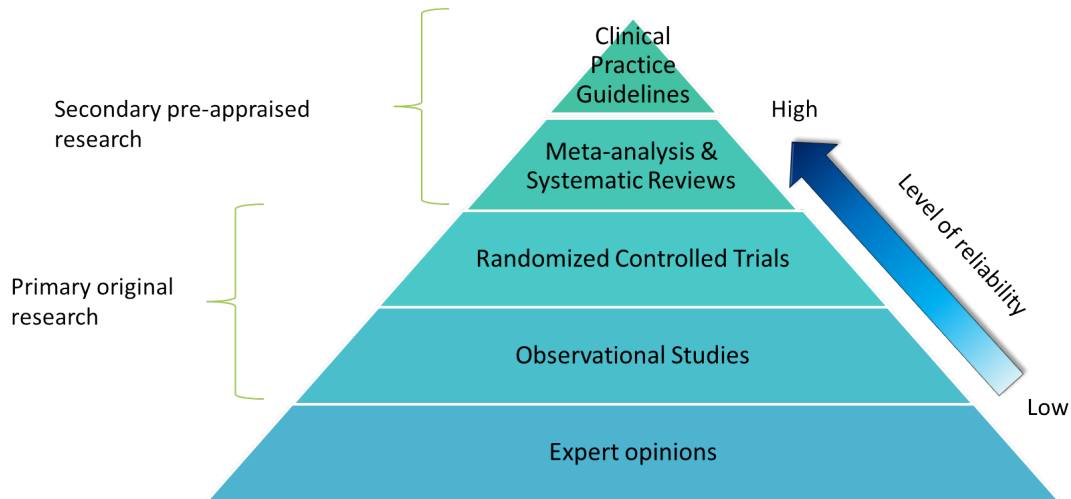


Figure 1.2: The pyramid diagram ranks the reliability of different levels of clinical knowledge in RT, usually from expert opinions at the bottom (least reliable) to meta-analyses or systematic reviews and clinical practice guidelines at the top (most reliable).

The focus of this dissertation is on clinical knowledge in RT publications from two primary resources: (1) Quantitative Analyses of Normal Tissue Effects in the Clinic (QUANTEC) [35], which are the clinical guidelines frequently used in RT treatment planning and contain meta-analysis of RT plans and radiation-induced adverse events, as well as recommendations for acceptable dose/volume constraints; and (2) clinical studies referenced by QUANTEC reviews. The QUANTEC reviews play an important role of clinical knowledge resources on assisting RT professional in determining acceptable dose/volume constraints to protect normal tissues and organs from radiation-induced adverse events [36]. Figure 1.3 displays the outline of the QUANTEC reviews which consist of three sections: 1) introductory papers; 2) organ-specific papers; and 3) vision papers. The clinical knowledge that we are extracting from QUANTEC

reviews is primarily from the section of the sixteen organ-specific papers. Each of the organ-specific papers is organized with a consistent structure of ten topic sections. The following example shows a clinical knowledge statement coming from the section of 'Recommended Dose/Volume Limites' in the organ-specific paper about rectum in the QUANTEC reviews. The clinical knowledge statement carries clinical knowledge of our interest for RT treatment planning. It lists the dose/volume constraints on rectum for patients with prostate cancer, and the potential risk rate of having 'Grade 2 late rectal toxicity' if following these constraints while receiving 3D conventional RT. The example is given as:

'For patients with prostate cancer, the following dose-volume constraints for conventional fractionation up to 78 Gy are provided as a conservative starting point for 3D treatment planning: $V_{50} < 50\%$, $V_{60} < 35\%$, $V_{65} < 25\%$, $V_{70} < 20\%$, and $V_{75} < 15\%$. And the NTCP models predict that following these constraints should limit Grade 2 rectal toxicity below about 10%. '

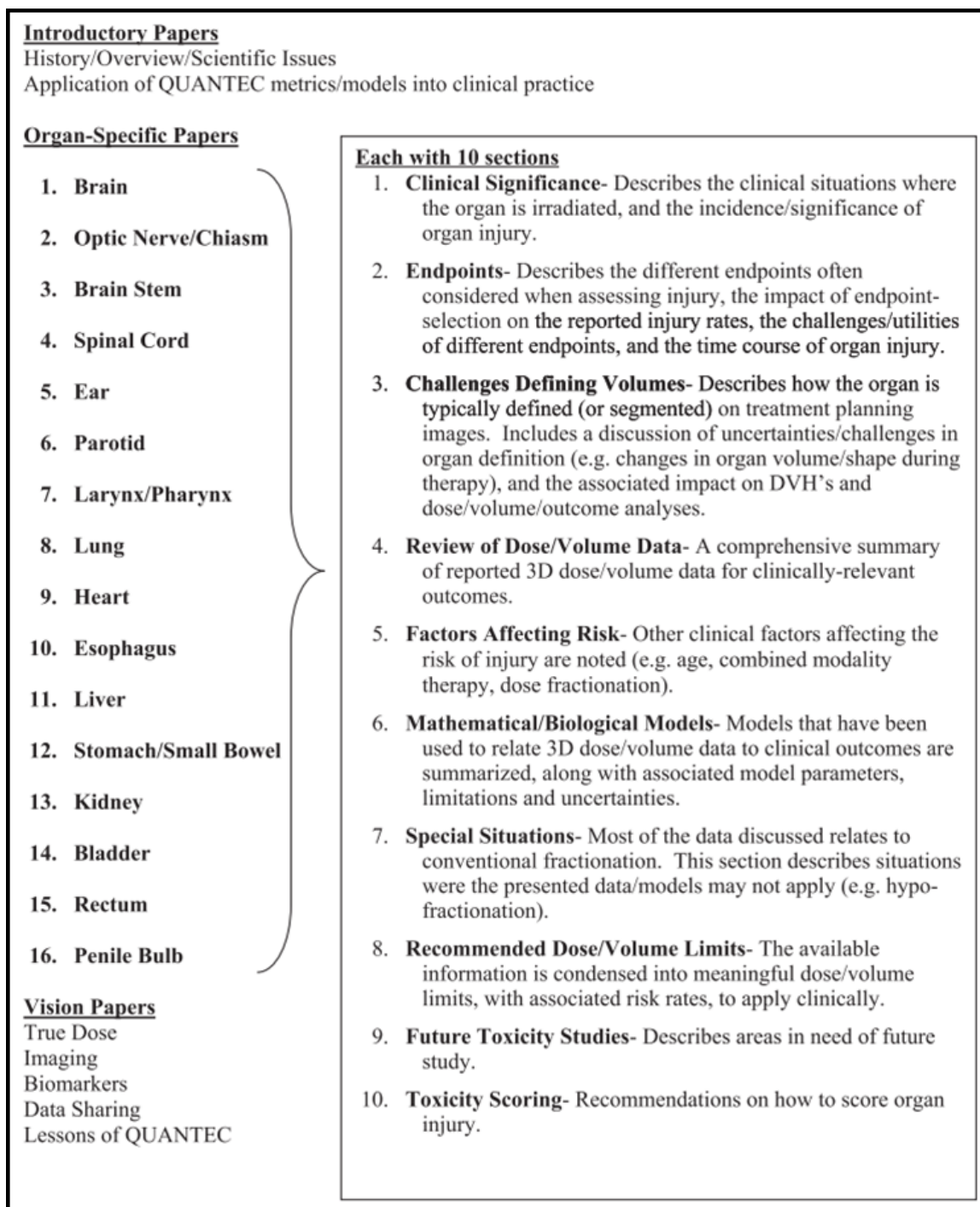


Figure 1.3: The outline of the QUANTEC reviews consisting of three sections: 1) two introductory papers; 2) sixteen organ-specific papers, each paper containing ten topic sections; and 3) five vision papers. (From: 'Guest editor's introduction to QUANTEC: a users guide')

With regard to content, we investigate clinical knowledge in RT by focusing on the following classes: radiation-induced adverse events, diagnoses, dose prescription, dosimetric criteria, treatment technique, and target site [37]. As a crucial measure in evaluating an RT plan, radiation-induced adverse events are the primary RT area of interest. The knowledge related to radiation-induced adverse events includes the name of adverse events, the duration, the grade, and the risk of radiation-induced adverse events. The knowledge related to diagnoses in RT consists of patient clinical diagnosis for RT, including patient information (e.g., age group and gender), primary cancer type for RT, and relevant medical conditions (e.g., previous medical interventions on the target site and medical history). Dose prescription is the radiation dosage prescribed on the target by the radiation oncologist. Dosimetric criteria refer to the dose-volume constraint protocols followed by radiation oncologists to limit the effects of radiation on organs at risk. Dosimetric criteria consist of specified planning parameters, such as maximum dose, mean dose, and the volume percentage receiving a specific dose. A treatment technique is a commonly used technique for RT to cure cancers and includes IMRT, 3D-CRT, VMAT, and so on. A target site is defined as an anatomical site with cancer that receives a prescribed dose.

1.3 Knowledge Representation Formalisms

Knowledge representation and reasoning is a field of artificial intelligence that concerns how a computer agent uses data, information, and knowledge to make decisions [38]. In order to guide the problem-solving process, a computer agent first requires a formal representation of domain knowledge in computer-interpretable language, then inference answers to given problems or queries based on knowledge representation formalisms. Knowledge representation formalisms provide computer agents with a structure of hypothesis space consisting of key entities, semantic relationships, and associated contexts from different resources for different purposes [39]. However, it is still difficult to represent knowledge with clear logic, compact description, and prob-

abilistic and temporal properties. Furthermore, it is a challenging problem for the agent to capture and extract knowledge from traditional resources, such as narrative text in literature, datasets and even figures in studies. Knowledge reasoning then uses formalized knowledge and logical reasoning to make inferences and solve problems.

Knowledge-based systems usually offer several different ways of representing knowledge in a domain and reasoning with this knowledge automatically to derive conclusions. In terms of a logical representation, logical language acts as a basis for knowledge representation, such as first-order logic (e.g., Horn-clause logic), description logic (e.g., OWL, the Web Ontology Language), and temporal logic (e.g., Linear Temporal Logic) [40, 41, 42]. Existing representations such as Arden Syntax [43], EON method [17] and ATHENA-CDS project [44] are insufficient for representing clinical knowledge in RT. Among these examples, first-order logic is a robust representation, but it is also problematic: firstly, representing knowledge in first-order logic as a list of propositions does not clarify the relationships and structures among propositions; secondly, inference in first-order logic is only semi-decidable; moreover, the world represented by first-order logic is absolute, while the clinical knowledge in RT is highly uncertain. Ontology is a general knowledge formalism in the biomedical field and is defined as a formal representation that captures concepts and relationships in a specific domain to allow the sharing of knowledge. However, ontology is unable to represent clinical knowledge in RT with uncertainty.

There has also been significant progress in the development of knowledge representation and reasoning with uncertainty. Probabilistic graphic models have been accepted as a natural formalism to represent and reason with uncertain biomedical knowledge. Bayesian Networks (BNs) - also termed causal probabilistic networks - are directed acyclic graph networks represented formally by a set of joint probability distributions that are described by probability calculus. In directed graphs, edges specify the from-and-to direction as relation or causal relationship. The properties of

BNs have led to progress in using them to handle the uncertainty of knowledge representation and reasoning in the biomedical domain [45, 46]. For example, researchers have used BNs to build networks for cancer diagnosis in the early 2000s [47, 48, 49]. BN-based decision models have been proposed for prognosis and disease progression when a diagnosis has already been made [50] and even for predicting cancer progression [51]. In RT, BNs have been used to represent the integrated RT process to make individualized cancer treatment decisions, to provide guidelines for therapy, and to allocate healthcare resources. Kalet [52] constructed BNs from ontology in radiation oncology and implemented a software tool to create a BN topology based on ontological semantics.

However, the relationships between entities in RT planning are more complex than that in the above cases; they are not simple causal relationships. Recent efforts to represent and reason knowledge have led to the development of statistical relational learning methods to handle both complexity and uncertainty in the real world [53, 54, 55, 56]. Knowledge representation and reasoning in the RT domain require not only the description of knowledge with logical language but also the handling uncertainty with probability theory. Markov logic networks (MLNs) have been proposed in knowledge representation and reasoning as a single formalism that generalizes both first-order logic and Markov networks by attaching weights to first-order logical formulas. The idea is to use predicate logic to generate Markov networks, i.e., joint probability distributions that have an associated undirected graph. It is also a probabilistic relational model that has the potential to address problems in health informatics, such as by providing better estimates for clinical diagnostic criteria and integrating information across health records [57]. MLN has also been used in ontology matching to identify corresponding semantics between entities of different ontologies in the biomedical field [58]. Several approaches and systems have been proposed to extract knowledge from biomedical literature with MLNs [59, 60, 61].

In addition, several projects have supported the implementation of MLNs, including the Alchemy project [62, 63], and the TUFFY project [64]. Moreover, outside the biomedical field, Snidaro showed an example by applying MLNs to fuse uncertain knowledge and evidence for maritime situational awareness [65].

Much recently, Probabilistic Soft Logic (PSL) was proposed as another formalism in statistical relational learning [66, 67]. Both MLNs and PSL combine logic and probabilistic graphical model in a single representation, whereby each formula is associated with a weight and the probability distribution over possible worlds is derived from the weights of the formulas that are satisfied by the possible worlds. However, PSL is more general in interpreting knowledge since it is based on fuzzy logic [68] and the suitability comparison between the two formalisms depends on their specific applications [69]. Considering the uncertainty and complexity inherent in RT planning, we believe that an MLN is a more suitable representation formalism.

In simple terms, an MLN is a first-order logic knowledge base with a weight attached to each formula [70, 71]. First-order logic enables an agent to compactly represent RT clinical knowledge; a Markov network enables an agent to efficiently handle uncertainty in RT decision-making. However, in addition to the challenges of knowledge representation and reasoning in RT, MLNs have several challenges to overcome: (1) how to translate knowledge from narrative text or figures into first-order logic in accurate and efficient ways; (2) how to construct the probabilistic graphic model with weights; and (3) how to assess the MLN built for RT planning.

1.4 Summary of Contributions

This dissertation proposes a framework based on knowledge engineering methods to bridge the gap between scientific research and clinical practices in RT in an efficient and intelligent way. In particular, the problem we are interested in is to represent, manage, and utilize clinical knowledge in RT publications for clinical decision-making in RT treatment planning. First, we explored radiation-induced adverse events and

their grading standards used in clinical research studies by applying ontological modeling and text mining methods. We developed an ontology for adverse events grading standard with an accurate and efficient way to the radiation oncology community as an initial step towards knowledge engineering. Second, we investigated the methods and developed a framework for extracting radiation oncology knowledge from clinical guidelines and clinical research studies. We suggest the RT community to improve the grading standards for radiation-induced adverse events with three strategies: sub-setting, re-examination, and ontological modeling. Third, we focused on the specific and challenging problem of the uncertain nature of human biological systems and biomedical research approaches. We found that MLNs can encode uncertain and complex knowledge, and support reasoning to answer queries, and takes patient individual specialization into account. To this end, we investigated the feasibility of probabilistic models for representing extracted RT knowledge and their ability to reason. We found that learning weights from experts knowledge should be cautious, and a data-driven weight learning method is more reliable in construction MLNs.

The dissertation comprises the following previous work I contributed to publications or will be published. We investigated automatic methods to generate an ontology for Common Terminology Criteria for Adverse Events (CTCAE), a standard for reporting and grading adverse events in cancer care, in

Y. Zhen, J. Wu, L. Yuan, Y. Ge. A semi-automatic method for generating ontology of the Common Terminology Criteria for Adverse Events. *AMIA Joint Summit on Translation Sciences*, 2015.

We then analyzed the use of the three most prevalent standards for grading radiation-induced adverse events in RT clinical studies by applying text mining methods in

Y. Zhen, Y. Jiang, L. Yuan, J. Kirkpatrick, J. Wu, Y. Ge. Understanding the use of adverse event scoring criteria in radiation therapy: a literature

mining approach. *American Medical Informatics Association (AMIA) Symposium*, 2015.

An extension of this work that discussed the factors of adopting radiation-induced adverse events grading standards in RT clinical studies appeared in

Y. Zhen, Y. Jiang, L. Yuan, J. Kirkpatrick, J. Wu, and Y. Ge. Analyzing the usage of standards in radiation therapy clinical studies. *IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, 2017.

We presented results of extracting radiation-induced adverse events from RT studies with Named Entity Recognition methods in

Y. Zhen, S. Karki, L. Yuan, J. Wu, Y. Ge. Radiation Induced Adverse Events Reported in Radiation Therapy Studies, *AMIA Informatics Summit*, 2018.

We first proposed the idea of using Markov Logic Networks to represent clinical knowledge in RT publications concerning both complexity and uncertainty in

Y. Zhen, S. Karki, L. Yuan, J. Wu, Y. Ge. Representation of Radiation Oncology Knowledge Using Markov Logic Network. *IEEE BHI*, 2018

This work later appeared as a conference paper in

Y. Zhen, T. Xie, S. Karki, L. Yuan, J. Wu, and Y. Ge. Representing Knowledge for Radiation Therapy Planning with Markov Logic Networks. *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2018.

We plan to submit our study of quantifying uncertainty of clinical knowledge in RT studies to the 2020 AMIA informatics Summit.

CHAPTER 2: REPRESENTING STANDARD FOR ADVERSE EVENTS WITH ONTOLOGICAL MODELING

In this chapter, we present a semi-automatic approach to developing an ontological representation for the Common Terminology Criteria for Adverse Events (CTCAE) to enable accurate and efficient reporting of adverse events in cancer care. CTCAE is a terminology standard for Adverse Event (AE) reporting in cancer treatment. The ontological modeling of CTCAE converts the standard from the current narrative and tabular format into an explicit, formal and computerized representation formalism that captures important concepts and semantic relationships specified in CTCAE. In addition, the proposed approach is general and can be extended to generate ontological representations for other clinical terminology standards in narrative and tabular format.

2.1 Background

Accurate and standardized reporting of side effects or toxicity for patients receiving cancer treatments are critically important for treatment outcome evaluation and its safety and quality control. CTCAE is a descriptive terminology standard developed by the National Cancer Institute (NCI) for Adverse Event (AE) reporting in cancer research and clinical care. An AE is any unfavorable and unintended sign, symptom, or disease temporally associated with the use of a medical treatment or procedure. In short terms, AE is a term referring to side effects or toxicity associated with cancer treatment in medical documentation and scientific analysis. A growing number of clinical trials and research studies rely on this standard to capture AEs of cancer treatment, such as radiation therapy, chemotherapy, and other systemic cancer ther-

apy treatments [72]. Not only do the medical professionals use CTCAE to standardize the process of AE reporting in the follow-up cancer care, but also clinical decision support systems must reference the standard as a consistent knowledge resource of AEs.

CTCAE is written in a tabular format that includes all AEs and their grading scales. In CTCAE, AEs are grouped by 26 System Organ Classes, which are identified by the anatomical or physiological system, etiology, or purpose. Thereby, CTCAE contains 26 tables of AEs organized by System Organ Class. Each row in the tables is an AE and its grading definitions that describe the severity of AE after cancer treatment. Figure 2.1 shows an example of several AEs in the table of gastrointestinal disorders in CTCAE v 4.03. With the narrative and tabular format, the concepts and relationships in CTCAE can hardly be retrieved and reasoned by computer agents for clinical decision support. While CTCAE in its current tabular form may be sufficient for manual entry of adverse events data in clinical trials, automatic reporting of adverse events in Electronic Medical Records (EMRs) and other advance use of CTCAE in decision support systems will require an explicit and formal representation for this standard. The representation must not only capture essential concepts and semantic relationships specified in CTCAE, but it also must allow medical professionals and clinical decision support systems in cancer care to access, query, and manage clinical knowledge related to AEs efficiently and intelligently.

Gastrointestinal disorders					
Adverse Event	Grade				
	1	2	3	4	5
Rectal fistula	Asymptomatic; clinical or diagnostic observations only; intervention not indicated	Symptomatic; altered GI function	Severely altered GI function; TPN or hospitalization indicated; elective operative intervention indicated	Life-threatening consequences; urgent intervention indicated	Death
Definition: A disorder characterized by an abnormal communication between the rectum and another organ or anatomic site.					
Rectal hemorrhage	Mild; intervention not indicated	Moderate symptoms; medical intervention or minor cauterization indicated	Transfusion, radiologic, endoscopic, or elective operative intervention indicated	Life-threatening consequences; urgent intervention indicated	Death
Definition: A disorder characterized by bleeding from the rectal wall and discharged from the anus.					
Rectal mucositis	Asymptomatic or mild symptoms; intervention not indicated	Symptomatic; medical intervention indicated; limiting instrumental ADL	Severe symptoms; limiting self care ADL	Life-threatening consequences; urgent operative intervention indicated	Death
Definition: A disorder characterized by inflammation of the mucous membrane of the rectum.					
Rectal necrosis	-	-	Tube feeding or TPN indicated; radiologic, endoscopic, or operative intervention indicated	Life-threatening consequences; urgent operative intervention indicated	Death
Definition: A disorder characterized by a necrotic process occurring in the rectal wall.					

Figure 2.1: The example shows a table including four AEs and their grading definitions under the class of gastrointestinal disorders from CTCAE v 4.03.

Trying to improve the efficiency of using CTCAE, several computerized tools for CTCAE have been developed so far [73]. However, they are based on simple representations that ignore the structure and semantic relationships in its definitions, which still limits integrating and exploring it within other clinical information. We note that an OWL version of CTCAE [74] in NCBO BioPortal has been released and maintained by the Cancer Therapy Evaluation Program (CTEP) at the National Cancer Institute. It is not a comprehensive ontology because it is a direct translation of the tabular format into an OWL representation without careful definitions of important concepts and relationships pertaining to AE.

In this chapter, we describe a semi-automatic approach to convert CTCAE into an ontological representation. According to Gruber’s definition, Ontology is the term referring to the understanding sharing of a specific domain, which is often conceived as a set of concepts or class relations, functions, axioms, and instances [75]. It is a data model to formally represent defined concepts and their relationship for both computing systems and humans. It also provides a shared framework of the com-

mon knowledge of specific domains for computer agents’ communication and heterogeneous information integration. We aim to develop a comprehensive ontological representation of the CTCAE that includes important concepts and semantic relationships about AE to enable knowledge sharing and semantic reasoning. Manually constructing ontologies is time-consuming, labor-intensive and error-prone. Moreover, manually updating ontologies when new standards are published could also introduce significant delays and hinder the development and application of the ontologies. Thus, we propose an approach to semi-automatically converting standards in tabular representations, such as CTCAE, into ontologies that maintain concept-relationship consistency and enable semantic reasoning.

2.2 Related Work

Several studies have developed general ontology design principles as guidelines for ontology generation [76]. Guarino [77] proposed a methodology for ontology design known as ‘Formal Ontology’. The design principles included the need to: (1) be clear about the domain; (2) take identity seriously; (3) isolate a basic taxonomic structure; and (4) identify roles explicitly. Uschold and Gruninger [78] proposed a skeletal methodology for building ontologies via a purely manual process. Ontological design patterns (ODPs) were proposed by Reich [79], which could be used to abstract and identify ontological design structures, terms, larger expressions, and semantic contexts. Moreover, the ontology design method was successfully applied in the integration of molecular biological information. Hwang [80] proposed a series of desirable criteria for the final generated ontology. The generated ontology should be (1) open and dynamic (both algorithmically and structurally for easy construction and modification), (2) scalable and interoperable, (3) easily maintained, (4) context-independent.

Based on the above guidelines, several approaches have emerged for generating ontologies from heterogeneous data sources, such as from scratch, from structured in-

formation, from existing ontologies, or any combination of the three sources [81]. According to the different sources, ontology generation methods were grouped into four main categories [82]: (1) conversion or translation methods prove that the ontological representation is more comprehensive than other structured knowledge representation, such as XML or UML; (2) mining-based methods implement mining techniques in order to retrieve enough information to generate an ontology mostly from unstructured sources; (3) external knowledge-based methods build or enrich a domain ontology by integrating external dictionaries, existing ontologies or other knowledge resources; (4) frameworks based methods generate ontologies by utilizing predefined modules and libraries in framework tools like Protégé [83].

In conversion or translation methods, experiences show that this approach presents a high degree of automation and a simple solution to ontology generation from another representation format. However, simple conversion does not address the whole problem of the ontology generation, especially in a specific task. Hannes Bohring [84] has developed a tool that converts given XML files to OWL format. He hypothesizes that items in XML schema can be converted to ontology's classes, properties, and instances without any other intervention on semantics and structures during the transformation. This transformation is implemented by XML style-sheet language transformation (XSLT) and has been also applied to the OntoWiki platform. Ghawi [85] and Yahia [86] improved Bohring's method to generate OWL ontology from multiple XML sources based on XML schema and XML schema graph (XSG). Their methods enable ontology generation to deal with more complex cases and possible design patterns. Hence, our proposed method is based on the conversion or translation approach to convert highly structured CTCAE into OWL ontology. We also applied Protégé knowledge-based framework to refine generated ontology.

2.3 Methods

2.3.1 Five Steps of Generating An Ontology for CTCAE

We developed a flexible approach to generating an ontology from a terminology standard that is specified in tabular document format and applied this process to convert CTCAE into an ontology. The first step is a manual process that converts the CTCAE document into eXtensible Markup Language (XML) using an existing tool. Then, in the ontology generation step, the XML representation is automatically transformed into an OWL representation. Finally, the generated ontology is manually refined by mapping important concepts and relationships to existing ontologies. The ontology generation step makes use of four representation schemes, the XML, XML schema, XML schema graph (XSG) and the Web Ontology Language (OWL).

We use XML to store CTCAE metadata and maintain consistency with the table-layer structure. CTCAE is composed of a list of adverse event terms and their grading scale definitions. The AEs are grouped by the System Organ Classes (SOCs). Thus, CTCAE contains many organ-based disorders tables. Each organ-based disorder table contains many adverse events that are associated with the specific organ. These relationships are captured in XML as metadata. An XML Schema describes the structure of an XML document. XML schema graph captures the graph structure of the XML schema, which matches our expected OWL graph structure.

Figure 2.2 depicts the ontology generation process. The overall method consists of five steps:

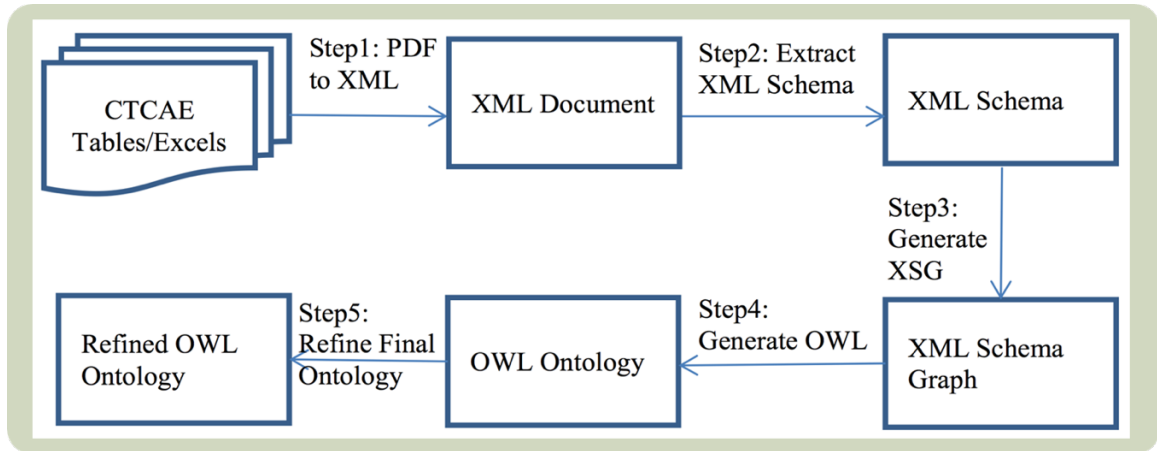


Figure 2.2: The ontology generation consists of five steps: 1) the step of converting CTCAE tables in PDF document into XML format; 2) the step of extracting XML Schema for each XML document; 3) the step of generating XSG based on XML Schema; 4) the step of generating OWL entities based on mapping rules; 5) the step of refining ontology.

Step 1: Convert each CTCAE table into an XML format. In this step shown as Figure 2.3, CTCAE tables are exported into Microsoft Excel to clear irrelevant data. Then, the clean excel document is converted into XML file by a java program. In the resulting XML file, the XML root element is CTCAE; a root's child element refers to an organ-based disorders table; each XML grandchild element of root represents an adverse event in an organ-based disorders table; and the XML leaf element is a grading scale definition of an adverse event.

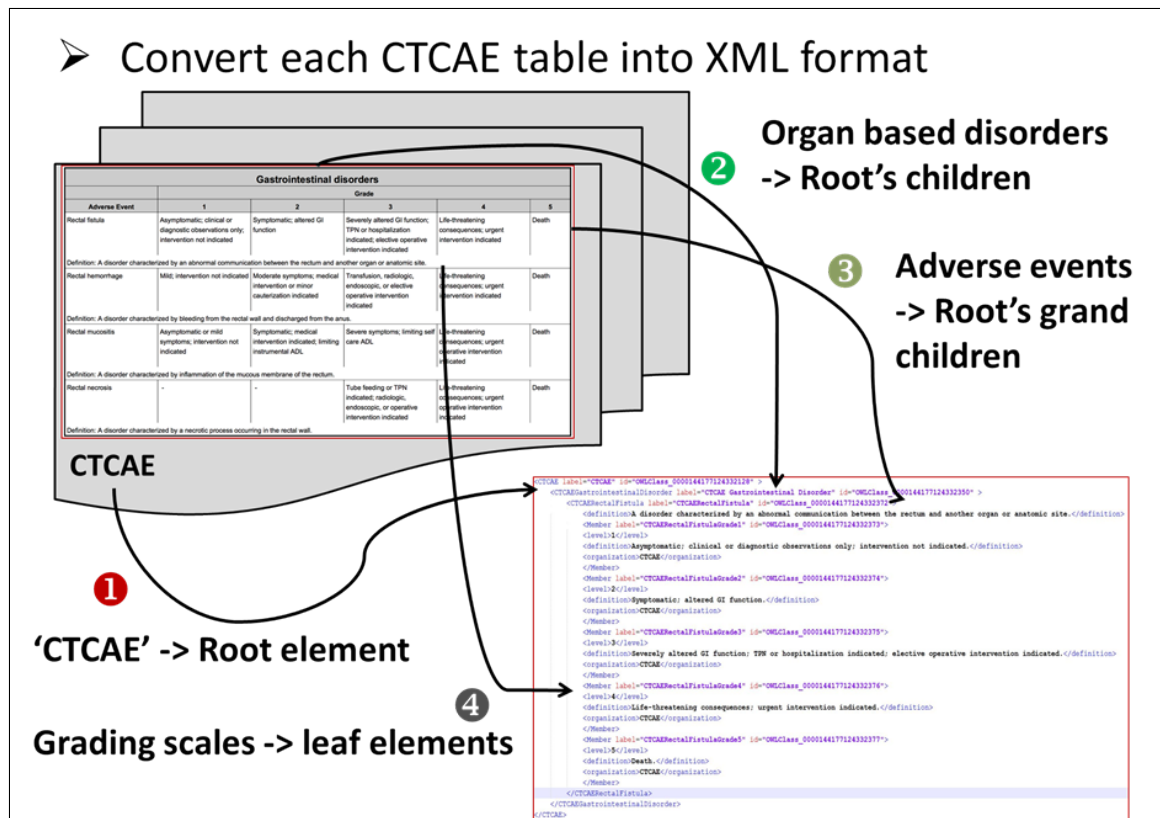


Figure 2.3: An example shows the step of converting each CTCAE table into an XML format. 'CTCAE' is mapped into the root element in XML; organ-based disorders are mapped into the root element's children; adverse events are mapped into the root element's grandchildren; and the grading scales definitions of each adverse event are mapped into the leaf elements in XML.

Step 2: Extract XML schema out of the XML document by parsing general elements in XML to specific structures such as complex types, elements, and attributes. XML-Schema Object Model (XSOM) processes this step. XML schema is in XML Schema Definition (XSD) language. Figure 2.4 gives an example of extracted XML schema from its XML document by using XSD language.

```

<xs:element name="CTCAE">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="GastrointestinalDisorder">
        <xs:complexType>
          <xs:sequence>
            <xs:element name="RectalFistula">
              <xs:complexType>
                <xs:sequence>
                  <xs:element type="xs:string" name="definition"/>
                  <xs:element name="Member" maxOccurs="5" minOccurs="0">
                    <xs:complexType>
                      <xs:sequence>
                        <xs:element type="xs:integer" name="level"/>
                        <xs:element type="xs:string" name="definition"/>
                      </xs:sequence>
                      <xs:attribute type="xs:string" name="title" use="optional"/>
                      <xs:attribute type="xs:string" use="optional"/>
                    </xs:complexType>
                  </xs:element>
                </xs:sequence>
                <xs:attribute type="xs:string" name="title"/>
                <xs:attribute type="xs:string" />
              </xs:complexType>
            </xs:element>
          </xs:sequence>
          <xs:attribute type="xs:string" name="title"/>
          <xs:attribute type="xs:string"/>
        </xs:complexType>
      </xs:element>
    </xs:sequence>
    <xs:attribute type="xs:string" name="title"/>
    <xs:attribute type="xs:string"/>
  </xs:complexType>
</xs:element>

```

Figure 2.4: An example shows the XML schema extracted from its XML document by using XSD language.

Step 3: Generate an XML-Schema Graph (XSG) by analyzing the XML-Schema. An XSG is a directed acyclic graph that has a unique root vertex that is the vertex of XML root element. It is composed of a vertex set and an edge set, seen as Figure 2.5. The vertex set contains all elements, attributes, non-

primitive types, element groups, and attribute groups. The edge set contains the edges that link (1) each element to its type, if not primitive type, and (2) each type, element group or attribute group to their contained elements and attributes.

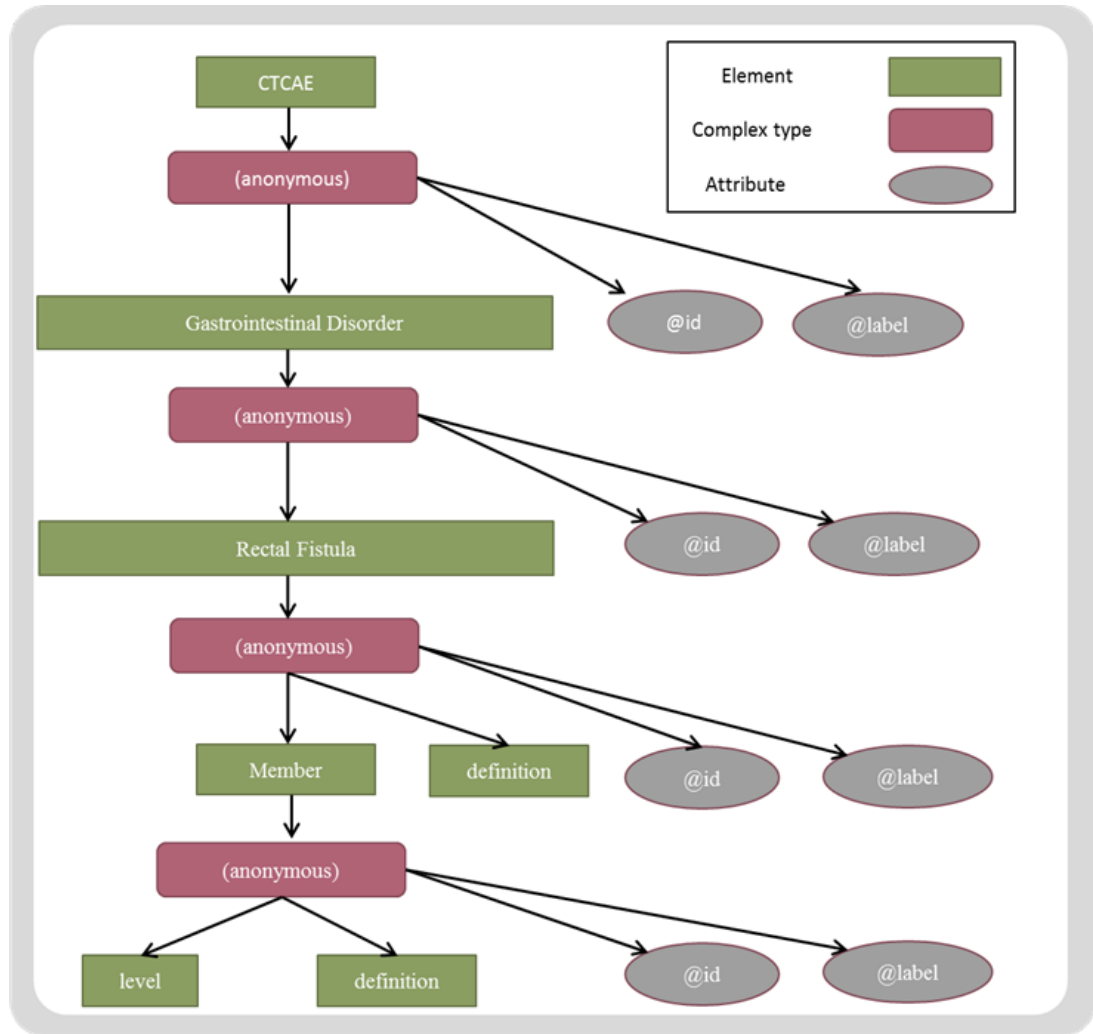


Figure 2.5: An example shows the XML-Schema graph generated from its XML schema, which consisting of vertex sets and edge sets. In the XML-Schema graph, the vertex set contains all elements, attributes and complex types.

Step 4: Generate OWL entities from XSG based on a set of mapping rules which will be elaborated in the later subsection. OWL classes and individuals emerge from complex types, and element-group declarations; object prop-

erties emerge from element-subelement relationships; datatype properties emerge from attributes and simple types.

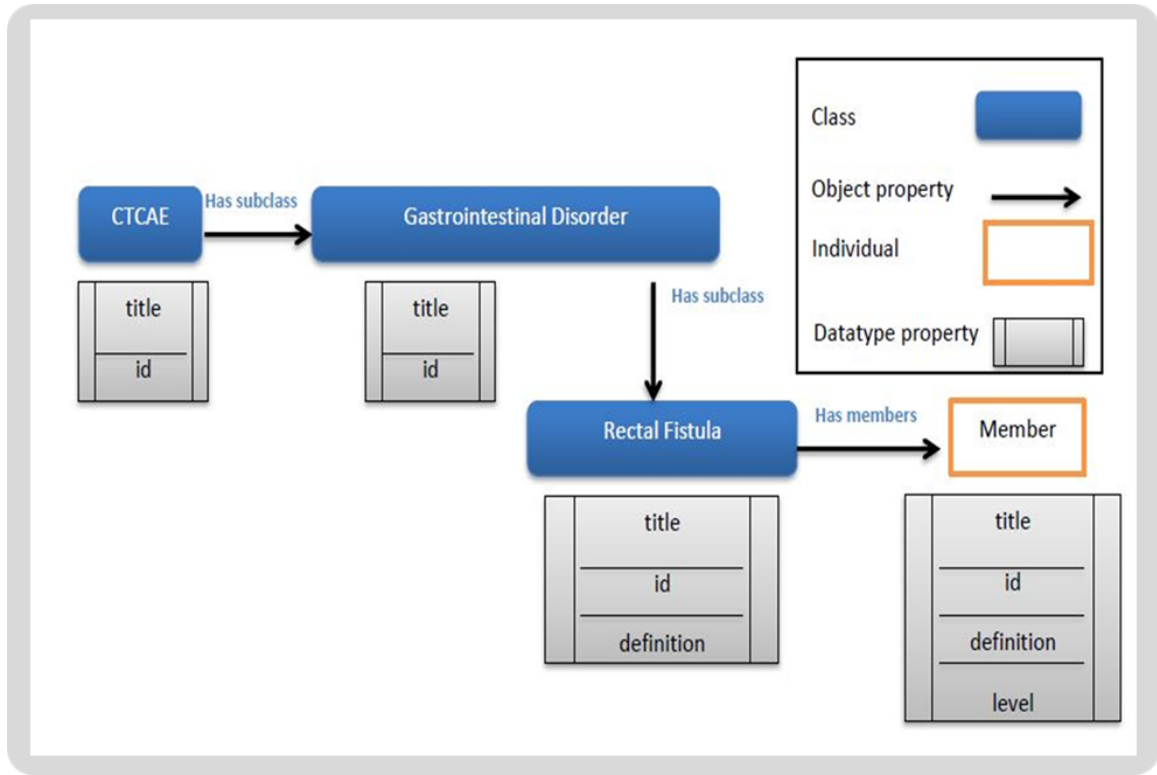


Figure 2.6: An example shows the generation of OWL entities from XSG based on a set of mapping rules.

Step 5: Manually refine the generated ontology by mapping important concepts to existing ontologies, such as Disease Ontology (DO) [87] for disease concepts and Foundational Anatomy Model (FMA) ontology [88] for anatomical concepts.

2.3.2 Mapping Rules

Our method constructs an OWL ontology by transforming XML schema entities into OWL model entities based on a set of mapping rules. In our CTCAE ontology model, each adverse event’s grading scale definitions are instances of the adverse event class. For example, Anemia Grade1 is a member of the class Anemia because it denotes a single anemia grade. Our CTCAE ontology model contains four types

of OWL entities: OWL classes, OWL individuals, object properties and datatype properties. Thus, the transformation from XML schema entities into OWL entities is based on the following four types of mapping rules:

Rule 1: OWL class mapping: maps an XML node (except leaf nodes) to an OWL concept. XML node is defined as complex types and named elements in XML schema. An OWL class generated from complex types has the name of the complex element. For example, a complex type named 'CTCAE' maps into OWL class 'CTCAE'.

Rule 2: OWL individuals mapping: maps an XML leaf node to an OWL instance. All leaf nodes in XML document are defined as complex elements named 'Member' in XML schema. These kinds of complex elements named 'Member' are mapped into OWL individuals of classes.

Rule 3: Object property mapping: maps a relationship between two XML nodes to an OWL object property. More specifically, the object properties are mapped from the element-subelement relationship in XML schema. An object property is added to the ontology when a complex element is already mapped to an OWL class and relates to its surrounding complex elements. The object property has a domain of corresponding classes and a range of its surrounding classes. For instance, rectal bleeding indicates bleeding. An object property 'indicates' existed between class 'RectalBleeding' and class 'Bleeding'.

Rule 4: Datatype property mapping: maps an attribute of XML nodes to an OWL datatype property. A simple element in XML schema is mapped to the datatype property. The datatype property has a domain of the OWL class corresponding to its surrounding complex types and a range of its XSD data type. Using the complex element 'Member' mentioned in the second rule, the

datatype properties are 'level' and 'definition'; their data types respectively are 'xs: integer' and 'xs:string'.

2.4 Experiments

The ontology generation method is applied to the tables of CTCAE v.4.03 in PDF format. The automatic process of CTCAE ontology generation takes about 8 minutes. The generated ontology covers all the terms, definitions, grades of adverse events in CTCAE v.4.03 accurately. The refined ontology includes important semantic relationships that map CTCAE entities to concepts in existing ontologies, such as the Disease Ontology (DO) and the Foundational Model of Anatomy ontology (FMA).

Table 2.1 presents the ontology metrics for the generated CTCAE ontology based on BioPortal Ontology Metrics. The table is composed of two parts: 1) Statistical metrics, and 2) Quality-control and quality-assurance metrics. The ontology contains 817 classes, including one root class 'CTCAE', 26 organ-based disorders classes, and 790 adverse events classes. The ontology has 3050 individuals, which denote to all grading scales of adverse events. The number of properties in the ontology is 7. The properties include '*is - a*', '*hasSite*', '*indicates*', '*hasDefinition*', '*located_in*', '*definition*', and '*level*'. For OWL ontology, the maximum depth only counts the '*is - a*' relationship as a hierarchical relationship. So, the maximum depth of our class hierarchical tree is 3. Maximum number of siblings in our ontology is 117. The class '*GastrointestinalDisorders*' has the maximum number of children. The average number of siblings at one lever in the class hierarchical tree is 4.

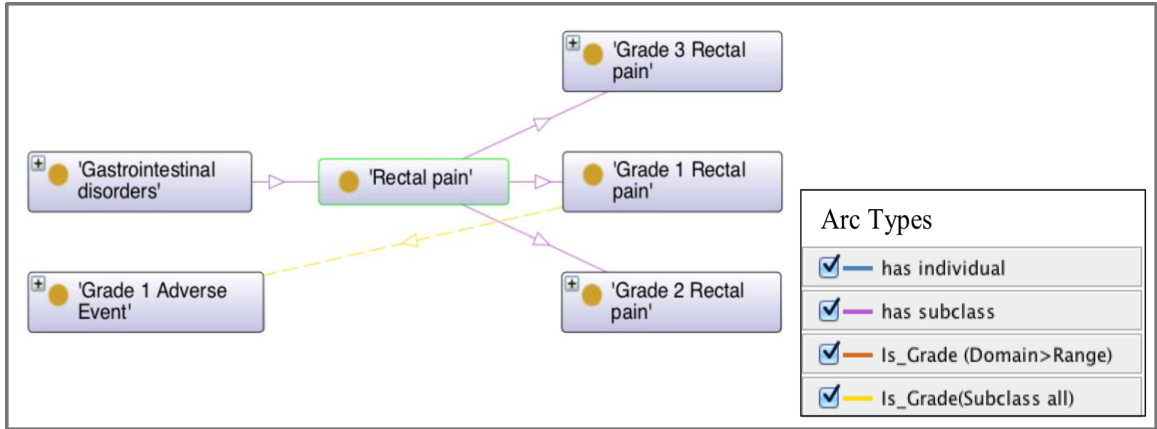
The second part of Table 2.1 shows Quality-Control and Quality-Assurance Metrics, which indicates the quality of the ontology and helps to improve the quality. The number of classes with only one subclass is one since only one adverse event term exists in 'Congenital, familial and genetic disorders'. This number often indicates that either the hierarchy is under-specified, or the distinction between the class and the subclass is not appropriate. The number of classes with more than twenty-five

subclasses is 12 since twelve organ-based disorders tables in CTCAE contain over twenty-five adverse events. A class that has more than twenty-five subclasses is a candidate for additional distinctions and categorization is needed. The number of classes with no definition in our ontology is 53 since there is no definition in CTCAE for root class 'CTCAE', organ-based disorders classes, and adverse events class named with 'other'. For comparison, we list the metrics of NCI CTEP's OWL version of CTCAE (referred as CTCAE OWL) in the last column. These metrics indicate that the automatically generated part of our CTCAE ontology is comparable to CTCAE OWL. The difference in the number of classes and individuals is due to difference in treating grading scales as individuals in our ontology and subclasses in CTCAE OWL.

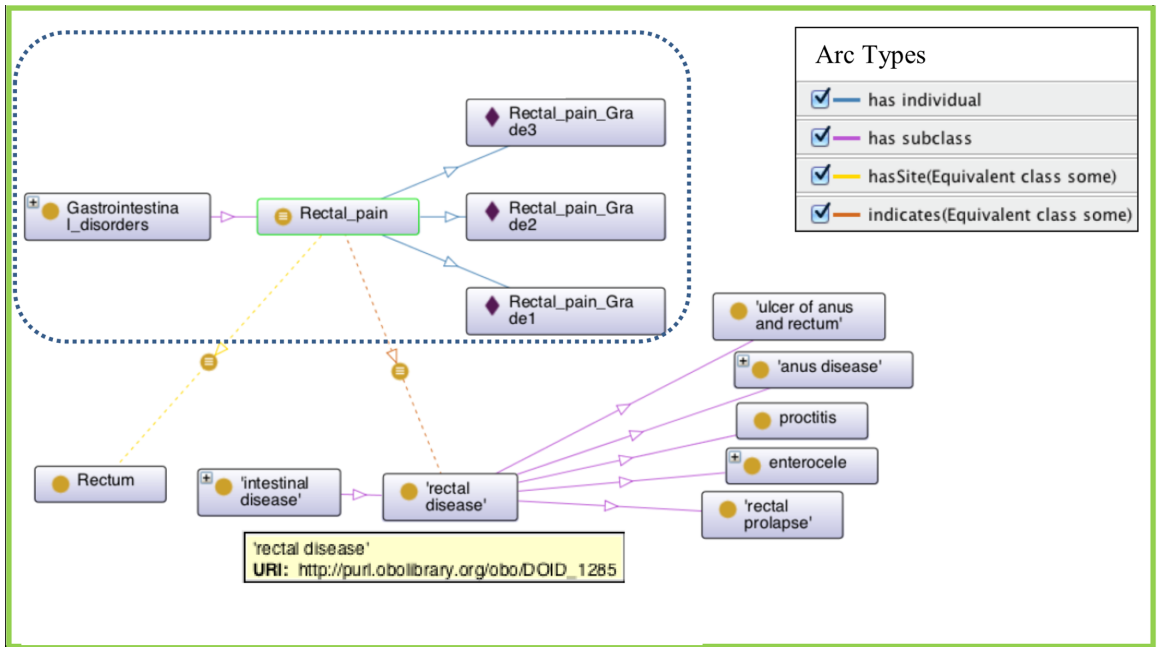
Table 2.1: The comparison between our generated CTCAE ontology and CTCAE OWL using the NCBO Bioportal Ontology Metrics.

Statistical Metrics	Our generated CTCAE ontology	CTCAE OWL
# of Classes	817	3874
# of Individuals	3050	0
# of Properties	7	7
Maximum Depth	3	4
Maximum Number of Siblings	117	117
Average Number of Siblings	4	4
Quality-Control and Quality-Assurance Metrics		
# of Classes with only one subclass	1	24
# of Classes with more than 25 subclass	12	12
# of Classes with no definition	53	3874

A careful comparison of the generated ontology with the original standard indicates that our method generates CTCAE ontology correctly. The generated CTCAE ontology contains all concepts, instances, and relationships as consistent with CTCAE v.4.03. Consistency and accuracy of generated ontology are comparable to CTCAE OWL. Our generated ontology improves logic and clarity of CTCAE ontology by using class-individuals to represent an AE- its grade scales instead of class-subclasses. Considering large numbers of OWL entities, we compare CTCAE OWL to our proposed CTCAE ontology by using an example of the AE named '*Rectal Pain*'. Figure 2.7 displays a comparison by viewing ontology graphs that focus on '*Rectal Pain*' using Protégé 4.319 - an ontology development environment. In the graph view of CTCAE OWL shown in Figure 2.7(a), the three grade entities are subclasses of adverse event class '*Rectal pain*'. However, in our generated ontology displayed in Figure 2.7(b), the three grades of rectal pain are individuals of class '*Rectal_pain*'.



(a) CTCAE OWL viewed under Protégé



(b) Our resulting ontology after refining viewed under Protégé

Figure 2.7: The comparison between CTCAE OWL and our CTCAE ontology viewed under protégé. (a) The ontology graph view focusing on the class '*Rectal pain*' in CTCAE OWL; (b) The ontology graph view focusing on the class '*Rectal_pain*' in our resulting ontology.

Moreover, our resulting ontology promotes usability by extending and reusing existing ontologies in the final refining step. Figure 2.7(b) shows that reusing the Disease Ontology (DO) and the Foundational Anatomy Model (FMA) ontology extends our generated class '*Rectal_pain*'. The adverse event '*Rectal_pain*' in CTCAE indicates

a '*rectal disease*' in DO and has site of '*Rectum*' in FMA. Hence, we can obtain more concepts and relationships about such AEs from resulting ontology in oncology adverse event domain.

2.5 Discussion

Our resulting CTCAE ontology provides a computerized representation for CTCAE. The ontological representation expresses contents in CTCAE as structured classes, instances, and properties. Reusing existing ontologies expands entities and enriches semantic relationships for CTCAE. Compared with current tabular CTCAE or CTCAE in OWL version (referred as CTCAE OWL), our proposed CTCAE ontology supports knowledge sharing, reasoning, and querying in a domain of grading adverse events. In our CTCAE ontology model, each adverse event's grading scales are instances of the adverse event class, rather than subclasses of the adverse event class stated in existing CTCAE OWL. The method developed in this work is applicable to other ontology generation tasks. Compared to manual ontology generation, the proposed approach requires less time and labor, especially for large-scale terminology standards. It will also reduce errors associated manual processes.

The proposed method has a couple of limitations as it stands. One is that the refining step is a manual task. Future work would extend the generation method to automatic reusing existing ontologies and other domain knowledge within adverse events grading and reporting. The second is that the method lacks inter-operations in annotating definitions. The definitions of adverse events and grades are rich text for semantic reasoning. We would use Natural Language Processing and Text Mining to annotate definitions automatically in future work.

2.6 Summary

In conclusion, this chapter provides an efficient approach for generating ontological representations from terminology standards that are written in narrative and tabu-

lar formats. Moreover, we developed an ontological representation for CTCAE, the common standard for AE reporting in cancer care. As the first step of this dissertation, the proposed CTCAE ontology provides a knowledge base of radiation-induced adverse events for clinical decision support in RT treatment planning.

CHAPTER 3: UNDERSTANDING THE GRADING STANDARDS FOR RADIATION-INDUCED ADVERSE EVENTS IN RT STUDIES

Grading standards for radiation-induced adverse events after radiation therapy (RT) is crucial for integrated, consistent, and accurate analysis of toxicity results at large scale and across multiple studies. This chapter discusses the trends of using the three most commonly used adverse event criteria in RT studies with text mining methods. The adoption of standard for grading radiation-induced adverse events has significant impact on the assessment and improvement of RT treatment. We also investigate the factors affecting the adoption of these grading standards.

3.1 Grading Standards for Radiation-Induced Adverse Events

Several grading standards have been proposed and utilized for reporting radiation-induced adverse events in RT studies. Three of the most commonly used standards are the Common Terminology Criteria for Adverse Events (CTCAE) [32], the Radiation Therapy Oncology Group (RTOG) [33], and the Late Effects Normal Tissue Task Force-Subjective, Objective, Management and Analytic (LENT-SOMA) [34]. These adverse event scoring criteria have been revised multiple times in recent years. In particular, CTCAE is strongly promoted as the comprehensive standard for adverse event reporting in all cancer care, with a set of criteria for the standardized classification of adverse effects of drugs or procedures [89, 90]. The initial development of CTCAE referenced the standards by the Radiation Therapy Oncology Group (RTOG)/European Organization for Research and Treatment of Cancer (EORTC) Late Morbidity System created in 1984. The RTOG standards have been updated for many versions and are still in use, containing grading standards for late radiation

morbidity, acute radiation morbidity, and common toxicity standards. To overcome the shortcomings of the RTOG/EORTC Late Effects System, the LENT-SOMA was published by the joint efforts of the RTOG and EORTC in 1995 to produce a universal system for measuring and recording late effects of radiation therapy [91]. A number of efforts have been reported to study and improve the scoring standards for grading radiation-induced adverse events [92, 93, 94, 95, 96, 97]. However, there is a lack of studies that attempts to understand how the various standards have been used in RT studies, and therefore a lack of understanding as to how these standards should be adopted, harmonized, or improved.

3.2 Methods

In this chapter, we are interested in addressing two specific questions: (1) the portion of clinical studies that use each standard by year and by cancer type; (2) the trend of usage in recent years. In order to answer these two questions, we developed a text mining method for automatically categorizing articles based on grading standards, and identifying cancer types of interest in the articles. While manual analysis by human experts is not prohibitive for individual questions, text mining techniques enable highly efficient and automatic investigation of multiple comprehensive questions using large-scale biomedical literature. Automated text analysis also allows continuous updates of trends analysis as newly published articles are added. Numerous literature exists for text mining techniques used in clinical medicine [98, 99, 100, 101]. We develop and compare two text mining approaches, one based on regular expression and one based on machine learning.

The text mining methods presented in this chapter focuses on two main tasks, categorizing articles and identifying cancer types. To understand the use of the three adverse event grading standards in RT clinical articles, we need to categorize articles based on the standards used. Then we would like to identify the type of cancer the clinical articles addressed. Our approach consists of four basic steps: data

preprocessing, feature extraction, classifier training, and cancer type identification.

3.2.1 Data Preprocessing

Each article was first converted into a simple text document without figures, tables, or references. Second, we applied tokenizer to each document to remove numerals and punctuations transforming each document into a list of sentences, and each sentence tokenized into a list of words. Then, we removed stop words in each document based on the stop-words list. Finally, we applied lemmatization, a WordNet's built-in function, to group different inflected forms of a word.

3.2.2 Feature Extraction

We extracted features for statistical analysis and classifier modeling. Major features include n-gram frequency, term frequency and inverse document frequency. Based on experimental analysis, we use 4-gram frequency in modeling Naïve Bayes Classifier in the article categorization task. Term frequency is calculated by counting the occurrence of phrases after applying n-gram model to each document. To avoid bias caused by different length of documents, we use the total number of terms in a document to normalize raw term frequency in such document.

3.2.3 Classifier Modeling

The goal of the document categorization task is to classify all documents based on grading standards used. The task categorizes each document with one of the three standards, '*CTCAE/CTC*', '*RTOG*', and '*LENT-SOMA*'. We separately used two classification methods to categorize all documents, one is based on regular expression (RE), and the other is based on Naïve Bayes classifier. Both methods have been shown to work well in text mining tasks [102, 103].

3.2.3.1 Regular Expression Based Classifier

The regular expressions enable a rule-based classifier. It assumes that the basic feature to differentiate documents is a specific pattern, such as particular characters,

words, or patterns of characters. The patterns determine which document uses which standards. If a document contains strings that match RE patterns for one of the three standard standards, the document is categorized as a sample of that criterion. Some documents may be labeled with more than one standards class if they contain strings that match more than one set of RE patterns.

Table 3.1: The definitions and example of text component used in the regular expressions discovery process.

Text Components	Definition	Example
Snippet	A sequence of characters that provide semantic information for our categorization task	<i>'All symptoms were scored according to the Common Terminology Criteria for Adverse Events.'</i>
Token	Words, numbers, or symbols in snippet	<i>'All', 'symptoms', 'were', 'scored', 'according', 'to', 'the', 'Common', 'Terminology', 'Criteria', 'for', 'Adverse Events'.</i>
Phrase	A sequence of consecutive tokens	<i>'All symptoms', 'were', 'scored', 'according to', 'Common Terminology Criteria for Adverse Events'.</i>
Key	An ordered list of phrases	<i>['were', 'scored', 'according to'] ['scored', 'according to', 'Common Terminology Criteria for Adverse Events'].</i>
Regular Expression	A sequence of characters that define a search pattern	<code>[scored\s+.*\s+(ctcae)\s+(v version)?\s+\d?</code>

We use an iterative process to discover REs in three basic steps: 1) Extract text snippets from labeled object articles; 2) Extract keys from snippets; 3) Generate REs. A text snippet is defined as a sequence of characters that provide semantic information for our categorization task. A snippet contains tokens that are defined as words, numbers, or symbols. A sequence of consecutive tokens composes a phrase. A key is defined as an ordered list of phrases. The keys are a critical source to generate general regular expressions. A RE is defined as a sequence of characters that define a search pattern. Currently, the last step, RE generation, is done manually. Table 3.1 explains what the text components are by listing the definition and example for each text component including snippet, token, phrase, key, and RE. The examples in Table 3.1 are not an exhaustive list of keys, phrase or RE we processed.

After discovering all the REs for classifying RT articles, we use the following algorithm to train RE based classifier. The algorithm of training regular RE based classifier involves the following main steps:

Step 1 Initialization: Select N labeled articles to generate initial REs using the RE discovery process explained above. The N articles consist of 3 groups: $1/3$ of N article randomly from each of the three classes. Set the initial REs as current REs $CRE\{r_1, r_2, \dots, r_n\}$.

Step 2 Refinement: Select N new labeled articles to test current REs $CRE\{r_1, r_2, \dots, r_n\}$ and get F-measure value f as an accuracy measure; refine current RE by applying the RE discovery on the misclassified articles including false positive and false negative cases to get the new REs $CRE'\{r_1, r_2, \dots, r_n\}$; then use the new regular expressions $CRE'\{r_1, r_2, \dots, r_n\}$ to get a new F-measure f' .

Step 3 Iteration: Iterate Step 2 by setting CRE to CRE' unless f' stops changing significantly, namely the change rate falls below ε , i.e. $\frac{|f'-f|}{f} \leq \varepsilon$

Step 4 Testing: apply the final regular expressions to test the remaining labeled arti-

cles as a validation. In the experiments reported here we empirically selected N of 10 and ε of 0.01.

Algorithm 1: The algorithm of training RE based classifier

Data: RT articles with labels

Result: Final REs $\{r_1, r_2, \dots, r_n\}$

initialization;

while *the new F-measure stop changing significantly* **do**

 Set new REs to current REs;

 Test current REs with N new labeled articles, get the new F-measure;

 Refine current REs by applying the RE discovery process;

 Return the new REs.

end

3.2.3.2 Naïve Bayes Classifier

Naïve Bayes Classifier is a classical and effective model for text classification [104]. In the article categorization task, we aim to compare the performance of Naïve Bayes Classifier with that of the regular expression based classifier. We use 5-fold cross-validation to partition training data and testing data. In each round, the training data is consisted of 424 randomly selected documents labeled by a domain expert. The rest of the documents are testing data to validate the trained classifier model. After five rounds, each document has four candidate standards labels. The final class for each document is the most voted candidate standards.

3.2.4 Cancer Type Identification

Cancer type identification is straightforward for MEDLINE articles because most of them have already been tagged with MESH terms that indicate cancer types. For the few remaining articles, we use a dictionary-based matching method to identify cancer types in the title and abstract. The look-up dictionary consists of cancer types from the domain exper’s annotation, and terms under Neoplasms [C04] of PubMed

MESH tree structures.

3.3 Experiments

3.3.1 Materials

We selected RT-related clinical articles, also called RT studies in this chapter, in MEDLINE (pubmed.gov) that were published between January 2010 and December 2012 to train and validate models. A total of 668 articles were found using a search strategy described in Supplementary Materials. From the 668 results, we excluded 104 articles due to inadequate information, and 33 articles due to duplication, resulting in a total of 531 articles that were analyzed in our study. All selected full articles were downloaded and extracted into text files for analysis, including both abstracts and full texts, while excluding all figures, tables, and reference lists.

The entire dataset for trends analysis includes additional published articles from January 2013 to August 2015, selected using the same search strategy. A total of 372 articles were found in MEDLINE. After excluding 38 articles due to lack of full texts, and 75 duplicated articles, a total of 259 articles were included in the full dataset amounting to a total of 790 articles.

To develop a gold standard for training and validation, a radiation oncology physician manually reviewed all 531 articles in the training/validation subset and labeled them according to the adverse event grading standards used. This process generated 3 classes, one for each standard grading standards.

After evaluating the two text mining methods using the training/validation dataset, we applied the more accurate method to analyze the entire set of full articles from 2010 to 2015 in terms of the overall usage trends over these years and also in term of usage of the three standards in different cancer types.

3.3.2 Training and Validation of Classifiers

The training of the RE based classifier took 4 iterations to complete. We noticed that F-measure reaches a plateau of $f = 86.7\%$ in the third iteration in Figure 3.1. Thus, we used 30 randomly selected articles in total to learn the REs, and we used the remaining 501 articles to test the resulting classifier.

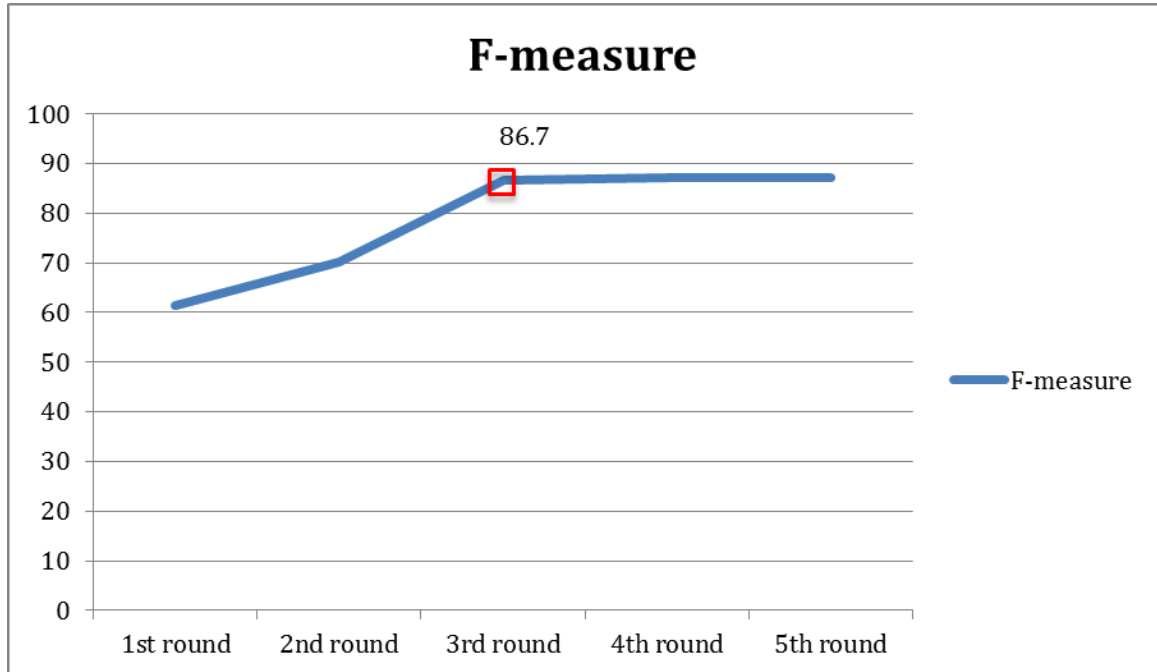


Figure 3.1: During the process of training regular expressions, F-measure increased with iterations and reached plateau in the third iteration.

The validation results show the RE based classifier to be reasonably accurate with a precision of 84.2% and recall of 85.1%. Compared to the RE based classifier, the Naïve Bayes classifier is worse with a precision of 72.1% and recall of 73.8%, but is still comparable to reported text categorization results [105]. Table 3.2 and Table 3.2 separately show the confusion matrix generated by the two classifiers in comparison to the gold results annotated by the radiation oncology expert.

Table 3.2: Confusion matrix of the RE based classifier.

# of Articles	Actual CTCAE	Actual RTOG	Actual LENT-SOMA
Predicted CTCAE	286	44	0
Predicted RTOG	29	268	19
Predicted LENT-SOMA	5	11	26

Table 3.3: Confusion matrix of the Naïve Bayes classifier..

# of Articles	Actual CTCAE	Actual RTOG	Actual LENT-SOMA
Predicted CTCAE	224	48	21
Predicted RTOG	87	261	13
Predicted LENT-SOMA	9	14	11

Based on the categorization results, Figure 3.2 presents the usage trends of adverse event grading standards over the three years in comparison to results from the domain expert. As seen in the figures, the two classifiers show similar trends comparable to those of the domain expert. These results provide an indication that the classifiers have sufficient accuracy for detecting usage trends of adverse event grading standards in clinical articles.

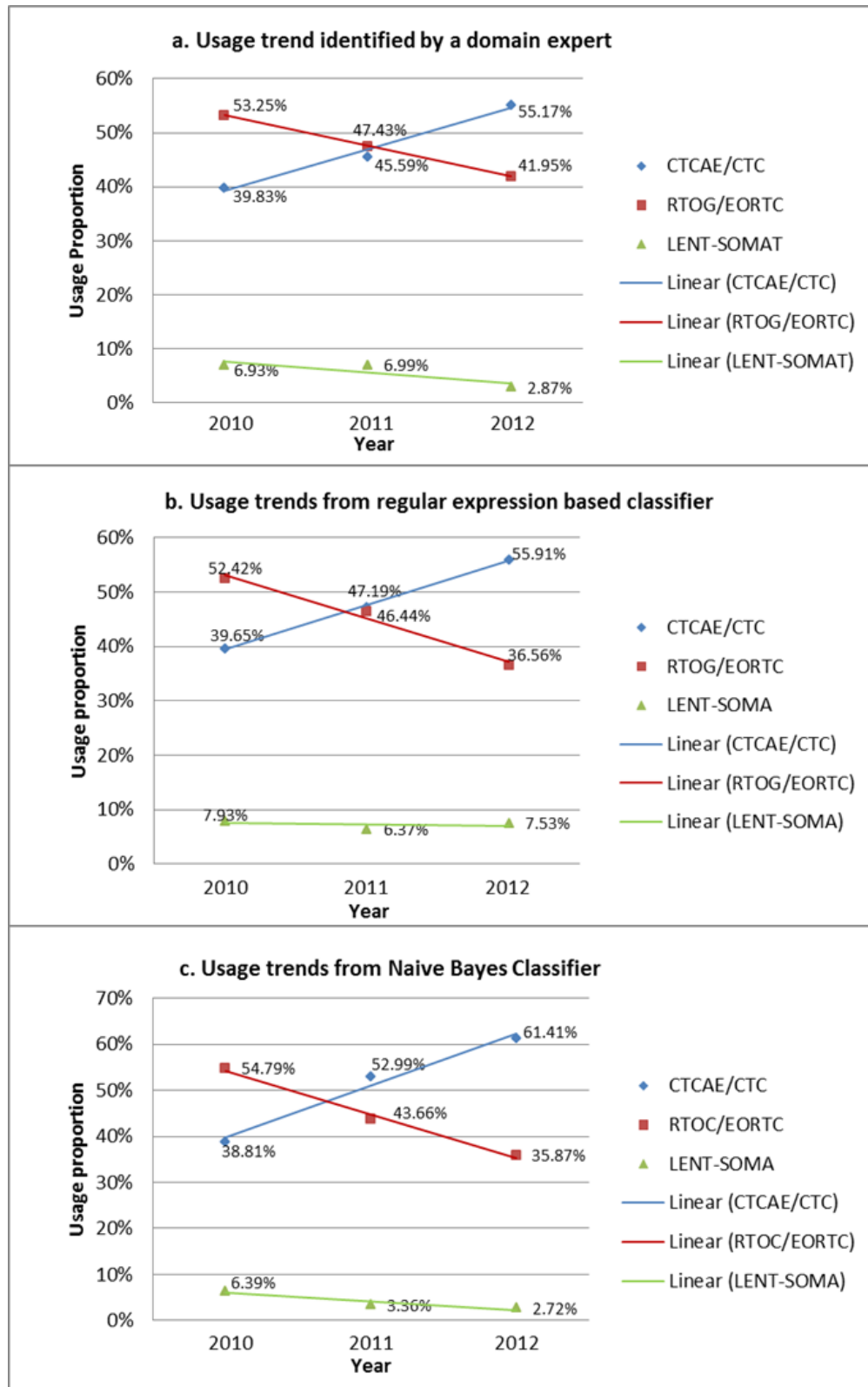


Figure 3.2: Comparing usage trends generated by the two text mining methods with those by domain expert using data from 2010 to 2012. Each line shows the proportion of the RT articles that use a particular grading standard.

3.3.3 Usage Trends of Grading Standards for Radiation-Induced Adverse Events During 2010-2015

Using the more accurate regular expression based classifier, we analyzed the trends of the three standard adverse event grading standards in RT clinical articles since 2010. The overall trends are shown in Figure 3.3. We observe that CTCAE and RTOG continue to be dominant standards in RT articles, each used by almost half of the articles while the LENT-SOMA standards are used by a small percentage of articles. During this period, the usage of RTOG remains relatively stable, that of CTCAE trends slightly up, while that of LENT-SOMA trends slightly down.

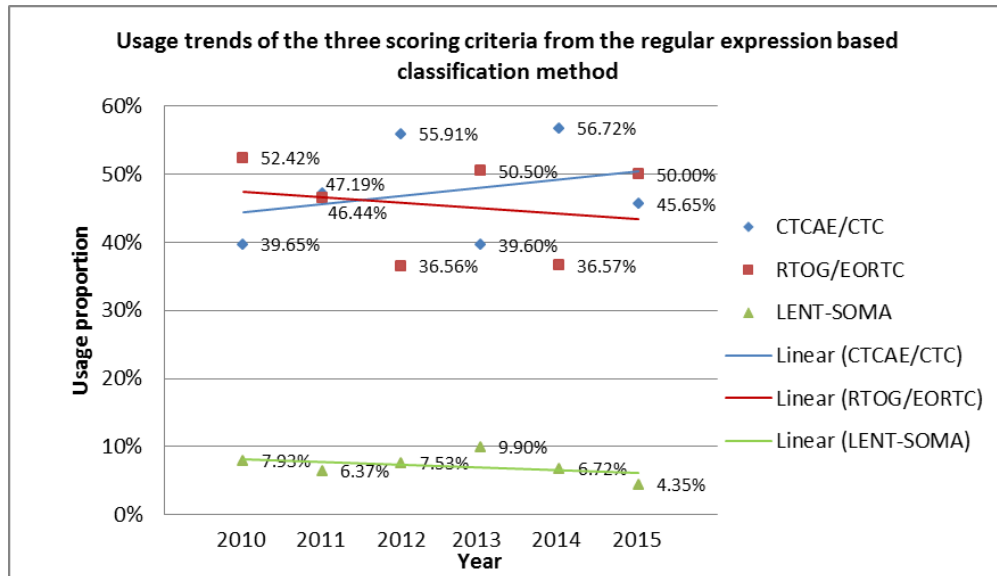


Figure 3.3: The usage trends of the three grading standard for raduaction-induced adverse events in RT articles from 2010 to mid 2015 using the regular expression based classification method. Each line shows the proportion of the articles that use a particular grading standard.

Next, we analyzed the trends by cancer types. Figure 3.4 shows the overall usage of the three standard grading standards in major cancer types over the past five and half years. One interesting finding from this figure is the strong contrast between lung cancer studies that heavily favor CTCAE and the head and neck cancer studies that clearly favor the RTOG standard. We also notice that LENT-SOMA is not only

rarely used, but also used only in select types of cancers, such as the prostate cancer and breast cancer studies. Furthermore, LENT-SOMA is especially not used in lung cancer studies.

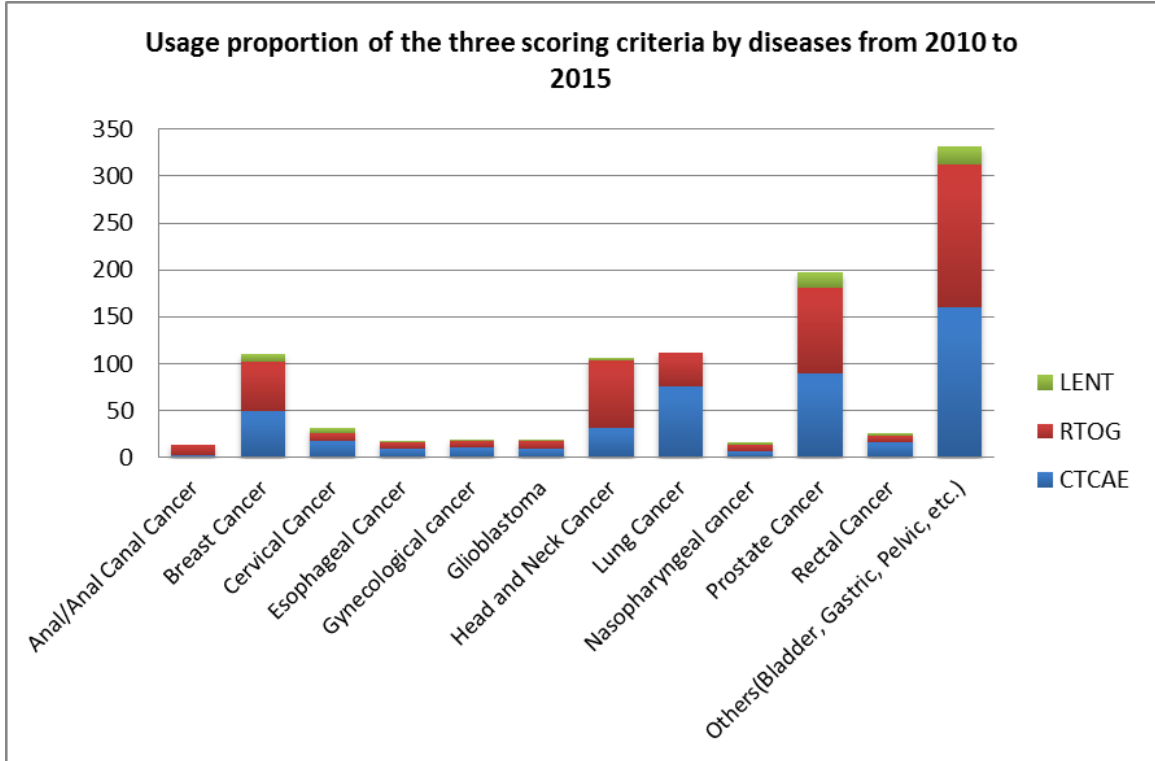


Figure 3.4: The usage proportion of the three grading standard for radiation-induced adverse events grouped by cancer types discussed in RT articles. The last category (Others) includes all other cancer types each with 10 or fewer articles.

3.4 Discussion

We investigated usage trends of the three most commonly used grading standards for radiation-induced adverse events in RT by mining the full text of published literature during 2010-2015. We resorted to mining the full text because the abstract section of clinical articles normally lack details on which grading standards were used in reporting radiation-induced adverse events or normal tissue toxicity after radiation therapy. Mining the full text also supports more detailed analysis of normal tissue toxicities and their adoption of grading standards. With the large and growing number of clinical publications, manual analysis of literature is becoming increasingly

difficult especially when new questions and more comprehensive analyses are needed. The text mining methods provide an important tool for understanding and improving the standards efficiently and continuously monitoring how standards are used for capturing and reporting radiation-induced adverse events in practice.

3.4.1 Error Analysis

There are two potential reasons why the RE based classifier has higher precision than the Naïve Bayes classifier in this study. The first reason is that size of training dataset is relatively small for a Naïve Bayes classifier. The second reason is that the features are relatively simplistic in the current model. However, the method based on RE has its drawbacks too. For example it will have difficulty dealing with articles containing latent information about grading standards.

The accuracy of the classifiers can be further improved. Examples of failure made by the regular expression based classifier include false positive and false negative instances. The majority of failures (57.2%) were traced to the regular expression. The RE only identifies text patterns learned, but ignores the context of why a grading standard is mentioned. In many cases, clinical trials used one grading standards A and mentioned another grading standards B for comparison or reference purpose. In these cases, the RE based classifier might label the clinical trial using grading standards B, when B better matched the learned RE. For example, an article may have a reference of CTCAE like 'XX group applied CTCAE to set questionnaires to grade prostate cancer on 1995' which positively matches regular expressions for CTCAE. However, this clinical study, in fact, used RTOG grading standards to grade toxicity. The next largest source of failure was from the generation of regular expression for RTOG, which contributed 36.5%. Radiation Therapy Oncology Group (RTOG) is an organization, which conducts many clinical trails in the radiation oncology community. The use of RTOG to describe both the clinical trails and the grading standards for adverse events caused confusion in the current classifier. Unrecognized snippet of grading

standards also contributed to classification failures, accounting for 6.3%. This failure is expected since the iteration of learning regular expression stopped on 4th round and our training sample is small. Unrecognized representations of grading standards, such as 'CTCAEs' are used in few clinical trials. We anticipate that expanding training dataset and improving regular expression generation algorithms could overcome these failures.

3.4.2 Limitations

We note this study compares the usage trends of three standards relative to each other. It does not intend to compute the actual utilization rate of individual standards. This study also has the limitation that only English language articles in MEDLINE are included in the analysis. However, given that most RT clinical studies are collected in MEDLINE and reported in English, we believe the findings in this study are representative of the current trends of standards adoption in RT. Furthermore, due to the small number of time points, especially in the validation part, the slightly increasing and decreasing trends that we identified from the data are not statistically significant. The actual trends could be level (slope of zero) for all or some of the three standard grading standards in the past 6 years.

3.4.3 Strategies to Improve CTCAE

From an informatics perspective, it is desirable that the research community adopts one standard for all clinical articles in radiation therapy and we believe that this standard should be CTCAE since it is based on the other two standards and is more up to date. The present study points out a significant challenge in the RT community with continued use of more than two standards in clinical studies. The apparent slow adoption of CTCAE can be attributed to a few factors. First, RTOG is released much earlier and has a longer history of adoption. And it initially contains more comprehensive content than the early versions of CTCAE, such as the late morbidity

effects standards that the early CTCAE did not include. Second, the comprehensive CTCAE v4.0 was released in 2009. But many of the RT clinical articles take many years to complete and therefore, the use of CTCAE may not have been fully reflected in the published articles. Third, we also conjecture that the complexity of the more recent CTCAE versions may have hindered its adoption. While the RTOG standard lacks details in adverse events, its simplicity makes it easier to understand and adopt. This complexity may also explain why LENT-SOMA has a low adoption rate. Finally, there may also be definitions of adverse events in CTCAE that do not properly fit the needs of certain cancer types, such as the head and neck cancer, in radiation treatment.

We propose that the RT community consider three strategies to improve the CTCAE standard. First, in order to make it easier to understand and adopt, we suggest creating subsets of CTCAE that are specific for major cancer types. This is supported by our analysis that shows limited number of RT related adverse event types reported in major cancer types. The subsetting strategy is not new. It has been widely adopted in large terminologies such as the SNOMED CT [106] and UMLS [107]. To illustrate the feasibility of the subsetting strategy, we applied a dictionary based term identification method to extract all adverse events reported in the collected articles. We found a total of 142 different types of adverse events reported in these 790 clinical articles. Figure 3.5 shows the total number of adverse event types reported in all articles of the four major cancer types as well as the average, minimal, and maximal number of adverse event types per article. As we can see, the adverse event types reported in major cancer studies are limited, with total numbers ranging from 30 to 60 within a cancer type and averaging less than 7 different types per article. Table 3.4 shows the top adverse event types with occurrence $> 55\%$ reported for the four major cancer types with 40% or more of respective studies in our six years dataset.

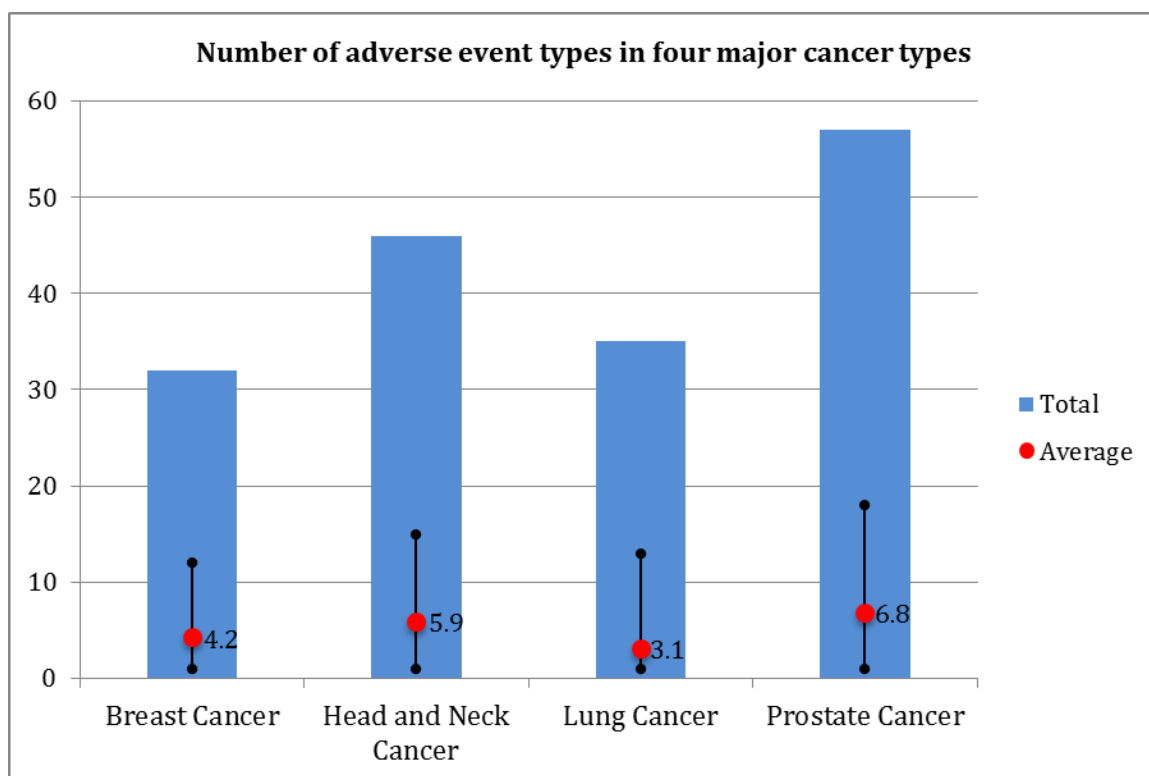


Figure 3.5: Total numbers of adverse event types reported for four major cancer types and average, minimal, and maximal numbers of adverse event types per article. The high-low-dots represent the max-min numbers of adverse event types for each cancer type (in black lines) per article. The total number of adverse event types is the sum of all adverse event types reported in one-type cancer articles (the blue bars).

Second, we believe it is important to reexamine the definitions of adverse events related to certain cancer types that currently favor other grading standards. Finally, we propose that the CTCAE should be represented as a true ontology so that the relationship between adverse events, their affected anatomy, the related synonyms, and severity are explicitly represented. A more formally defined ontology for radiation-induced adverse events will better enable efforts to harmonize, subset, mapping, and improving the standard as well as integrating with other existing clinical standards in the broader clinical research community.

Table 3.4: The top adverse event types reported with occurrence $> 55\%$ for the four major cancer types in RT clinical studies.

Four Major Cancer Types	Top Adverse Events
Breast Cancer	fibrosis
	breast pain
	breast edema
	fat necrosis
	hyperpigmentation
Head and Neck Cancer	dysphagia
	leukopenia
	thrombocytopenia
	hemorrhage
	neutropenia
	xerostomia
	mucositis
Lung Cancer	pneumonitis
	esophagitis
Prostate Cancer	urinary frequency
	urinary obstruction
	hemorrhage
	proctitis
	dysuria
	fatigue

CHAPTER 4: CLINICAL KNOWLEDGE REPRESENTATION AND REASONING FOR RT TREATMENT PLANNING WITH MARKOV LOGIC NETWORKS

Radiation oncologists rely on clinical guidelines and results of clinical research studies to design an effective Radiation Therapy (RT) treatment plan with the goal of maximum damage to cancer cells and minimum effects on normal tissue. As a step toward computerizing the clinical guidelines and clinical trials results in RT planning, this chapter presents an approach and investigates its feasibility of representing the complex and uncertain clinical knowledge in RT using Markov Logic Networks (MLNs). Within this approach, different types of clinical knowledge in RT with associated uncertainty can be extracted from published clinical guidelines and research studies, then be represented into a computerized formal model, and reasoned with evidence for intelligent RT planning. As an example for demonstration we focus on the RT treatment planning scenario for limiting the risk of radiation-induced effects and suggesting dosimetric criteria and prescription dosage. We tested the constructed MLNs by making inferences to predicate the risk of radiation-induced effects given RT dose-volume plan. The initial results show the MLNs prediction of risk is in the range of risk suggested in guidelines.

4.1 Knowledge Engineering with Markov Logic Networks

Formally, a Markov Logic Network (MLN) is defined as a set of pairs (F_i, w_i) , where F_i is a formula in first-order logic and w_i is a real-valued weight. The weighted formulas of MLN define a template for constructing a probabilistic graphical model or Markov network that specifies a distribution over possible worlds. Given different

sets of constants, different grounding Markov networks are produced by the MLN. Each state of grounding Markov network represents a possible world. For example, a possible world described in RT planning studies can be a finite world like 'a 60-years-old male patient with prostate cancer and his whole bladder receiving a dose of 60 Gy in 40 daily fractions, and having RTOG grade 3 bladder toxicity'. Knowledge engineering builds a model of the domain (e.g., radiation oncology) with knowledge representation formalisms and allows one to make inferences or answer certain questions with knowledge reasoning engines (e.g., what is the recommended dose/volume for a sixty-years-old patient with prostate cancer to limit the risk of having RTOG Grade 3 bladder late toxicity below 30%).

Considering the uncertainty and complexity inherent in RT treatment planning, we believe Markov Logic Networks (MLNs) is a suitable representation formalism. Markov Logic Networks generalize both first-order logic and Markov networks by attaching weights to first-order logic formulas. There are two challenges to represent RT knowledge with MLNs. The first challenge is translating narrative statements of RT planning guidelines into first-order logic. Moreover, the translated first-order logic formulas can clarify relations and structures of RT planning knowledge. The second challenge is that RT guidelines and studies are missing grounding data to support weights learning for MLNs and inference with MLNs. Here, we describe an approach to convert narrative clinical guidelines and clinical research studies of RT planning into a computerized representation with MLNs, which allows statistical relational learning and reasoning.

In MLN, formulas can be seen as soft constraints on a set of possible worlds. A possible world is less probable when it violates one formula, but not impossible. The weight associated with formula indicates the constraining degree of the formula: the higher the weight, the greater the difference in log probability between a world that satisfies the formula and one that does not, other things being equal. In other words,

the probability of a possible world is proportional to the exponentiated sum of weights of formulas that satisfied in that world. Collectively, the first step of this method is to learn the structure of MLN from QUANTEC papers; secondly, learn weights for formulas based on statistical data and cases from clinical studies; thirdly, make inference on the risk of having normal tissue effects and the probability of receiving RT dose plan.

4.2 Methods

4.2.1 Materials

The chapter is focusing on representing RT treatment planning knowledge from the set of RT guidelines commonly called the Quantitative Analyses of Normal Tissue Effects in the Clinic (QUANTEC) consisting of 16 organ-based reviews and the 471 referenced clinical research studies by QUANTEC reviews. Important reasoning tasks include predicting the risk rate of having certain radiation-induced adverse events for a specific patient after RT plan, and determining which RT plan is most appropriate to achieve the clinical goals.

4.2.2 Overview of Proposed Framework

Figure 4.1 illustrates the workflow of representing RT planning knowledge with Markov Logic Networks framework. The framework consists of three modules: A) structure learning; B) weight learning; and C) Inference. The structure learning module aims to learn the structure of MLNs, namely first-order logic formulas, from QUANTEC guidelines and RT studies by 1) extracting RT knowledge statements with a Named Entity Recognizer based on CRF method; 2) construct first-order logic formulas by translating the extracted knowledge statements manually. The weight learning module is to determine the weights of the formula in MLNs. In this module, we learn the associated weights for constructed first-order logic formulas in structure learning module from a relational database. The relational database is generated by

simulating RT planning data from the 471 research studies. In the inference module, we make inference on the risk of potential side effects and make the recommendation for RT treatment plan by given query and evidence.

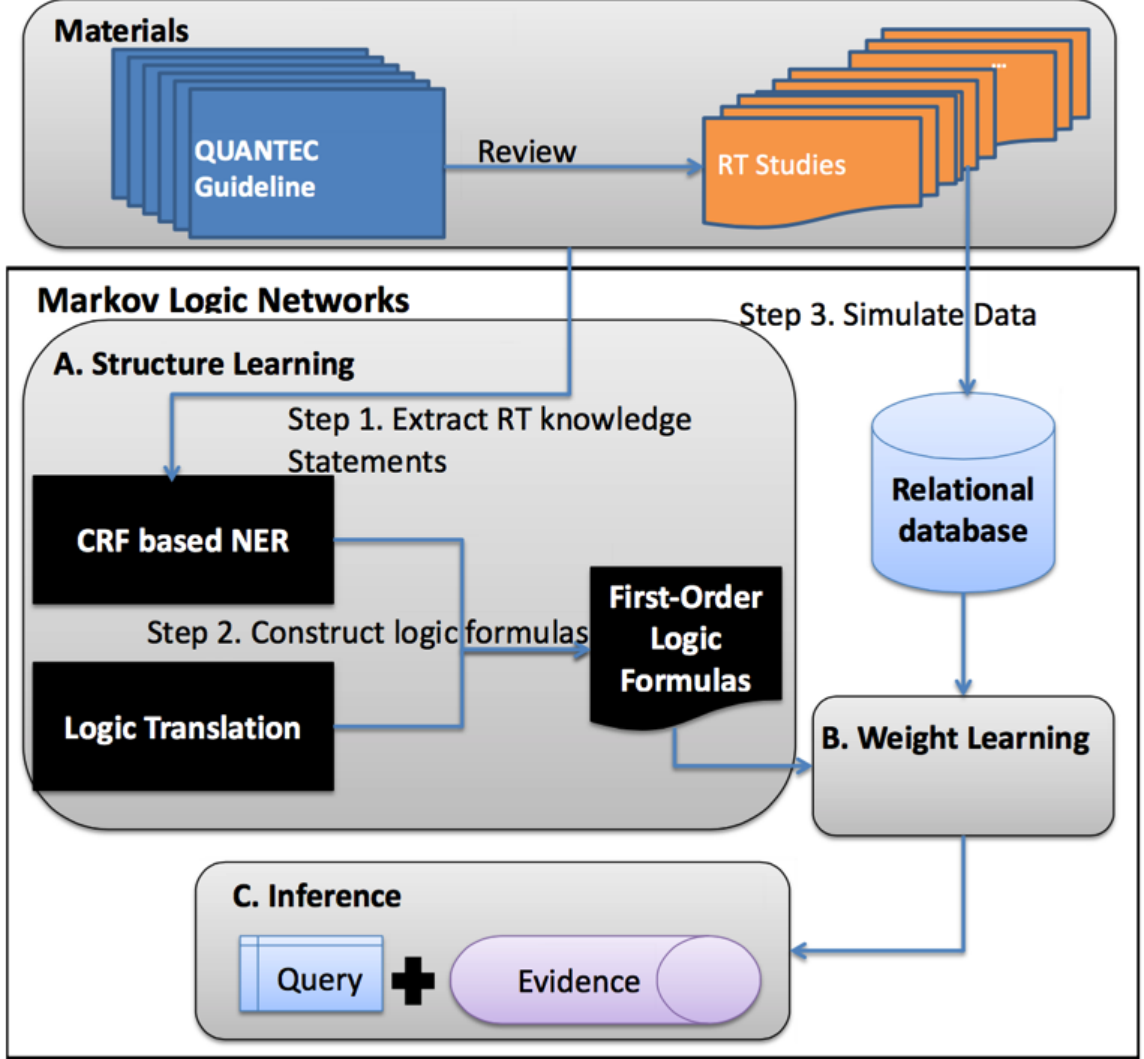


Figure 4.1: The workflow of representing RT planning knowledge with Markov Logic Networks (MLNs) framework. The MLNs framework consists of three modules: A. Structure Learning; B. Weight Learning; and C. Inference.

4.2.3 Structure Learning

Structure learning is currently a manual process. Firstly, we annotate QUANTEC papers manually by extracting statements carrying clinical knowledge for RT treatment planning, which is primarily expressed with the entities and relations that hold

between entities in the domain of RT treatment planning shown in Figure . We extracted the following six classes of clinical knowledge in RT: radiation-induced effects, diagnosis, dose prescription, dosimetric criteria, treatment technique, and target site. We introduced the six classes of clinical knowledge for RT treatment planning in Chapter 1. The statements carrying clinical knowledge for RT treatment planning are simply defined as sentences: 1) containing entities and their relations in the above six classes; 2) occurring in the sections of recommendation, discussion, conclusion, and analysis.

Secondly, we manually translate the extracted knowledge statements into first-order logic formulas. In order to translate the narrative statements in QUANTEC reviews and RT studies, we assume that relations between these entities above could be expressed with first-order logic connectives such as implications, conjunctions, biconditional, and negation. In order to translate the narrative statements in QUANTEC papers, we assume that relations between these entities above could be expressed with first order logic connectives such as implications, conjunctions, biconditional, and negation. During annotation, statements containing potential entities are highlighted, entities and relations within statements are extracted for constructing first-order logic formula. In order to reduce the variation of narrative states in the original papers, we first fit the statements into a set of templates before constructing first-order logic formulas.

We use three types of symbols to construct first-order logic formulas: constants, variables and predicate. Constant symbols represent objects in the domain (e.g., specific patients: Anna, Bob). Variable symbols represent range over objects in the domain. Predicate symbols represent relations among objects in the domain or attributes of objects. Formulas are recursively constructed from predicate symbols applied to a tuple of constants and, variables, or functions using logical connectives and quantifiers. Table 4.1 lists variables, and Table 4.2 lists predicates to define the

structures of MLNs in RT treatment planning.

Table 4.1: The variables for constructing MLNs in RT treatment planning.

Varibales Name	Symbol	Definition
patient	p	The patient receiving RT.
cancer	c	Major cancer type cured in RT.
medical condition	mc	Previous medical intervention on the target site, medical history.
target site	s	Disease site exposing to the prescribed dosage.
RT plan	rt	The treatment plan.
dose prescription	dp	Treatment prescription, ranging from a set of prescribed dose/volume/fraction.
treatment technique	t	Treatment technique modalities used in RT.
dosimetric criteria	dc	Parameters of OAR Constraints
radiation-induced adverse events	ae	Radiation toxicity or side effects associated with RT.
grade	g	The grade of radiation-induced adverse events.
risk	r	The risk of radiation-induced adverse events.

Table 4.2: The predicates for constructing MLNs in RT treatment planning.

Predicates	Definition
$Prostate_Cancer(p)^*$	Whether the patient p having prostate cancer.
$Is_Planned(p, rt)$	Whether the patient p is planned with an RT plan rt .
$Is_Prescribed(rt, dp)$	Whether the RT plan rt is prescribed with dose prescription dp .
$Is_Constrained_Rectum(rt, dc)^*$	Whether the RT plan rt is constrained by dosimetric criteria dc on rectum.
$Risk_of_AE(rt, ae, g, r)$	Whether the radiation-induced adverse event ae at grade g occurs with the risk r after having RT plan rt .
$Use_IMRT(rt)^*$	Whether the RT plan rt uses treatment technique IMRT.
$Smaller_Than_15\%(x)^*$	Whether numerical value x is smaller than 15%.

* Similar predicates can be easily written by replacing 'Prostate_Cancer' with another type of cancers, 'Rectum' with other organ-at-risk, 'IMRT' with other treatment techniques, and '15%' with other numeric value separately.

4.2.4 Weight Learning

Weights associated with formulas in MLN reflect constraining degree of formulas to possible worlds and hold the probabilistic properties of MLN. Although the MLN weights can be assigned by experts in theory, they are usually learned from one or more relational databases containing a set of grounding truth of possible worlds. During weight learning, we make a closed world assumption⁹: the ground atom is assumed to be false if it is absent in the database.

A challenge of representing clinical knowledge in RT with MLNs is that there is

no direct data for weight learning since the resource of clinical knowledge in this dissertation is published studies rather than Electronical Medical Record. We investigated the feasibility of synthesizing a relational database for weight learning by sampling and simulating data from RT research studies that are referenced by QUANTEC papers. The published RT research studies contain statistical information about cohorts, such as patients' age distribution, radiation treatment dosage, and the probabilistic distribution of radiation-induced adverse events over the cohorts. For instance, the QUANTEC paper on bladder [108] referenced a study conducted by Pos et al [109]. In Pos's study, 50 patients with bladder cancer received the total dose of 55 Gy in 20 fractions in 4 weeks on whole bladder. Age of the 50 patients varied from 58 to 93 years with a median of 79 and standard deviation of 7. And the clinical study also provided the distribution of side effect after patients receiving RT, for example, RTOG Grade 1 bladder late toxicity was observed in 15 patients and Grade 2 toxicity was observed in two patients after RT, and so forth. According to this data, we built a relational database of ground atoms by simulating a group of 50 patients $p = (P_1, P_2, \dots, P_{50})$ with 50 different RT treatment plans $rt = (RT_1, RT_2, \dots, RT_{50})$ that fit the patient characteristic data, and side effects data mentioned above. We assume that the age of patient fits normal distribution in the simulated data. Here are examples of ground atoms in the generated relational database for weight learning (only examples not all ground atoms): *Bladder_Cancer*(P_1), *Is_Prescribed*(RT_1 , '*Dose of 55 Gy in 20 fractions in 4 weeks*'), and so on. After generating relational training databases from research studies, we apply weight-learning algorithms provided by Alchemy [62, 63], an open system for MLNs, to learn weights for formulas.

4.2.5 Inference

After learning the structure and weights of MLNs, we can make inference with the MLNs model. The goal of inference in MLNs is to get the most likely probability

distribution of the system. We used Markov chain Monte Carlo (MCMC) inference supported by Alchemy to inference queries with built MLNs and evidence. The idea of MCMC inference is to sample a sequence of states according to their probabilities, and then count the fraction of sampled states where the formula hold. To implement MCMC inference, Alchemy applies three MCMC algorithms: Gibbs sampling, MC-SAT, and simulated tempering. MC-SAT performs orders of magnitude faster than the other two algorithms and applies to any models that can be expressed in MLN. The ideal of MC-SAT is to combine MCMC and SampleSAT.

4.3 A Knowledge Representation Example of Generated MLNs

In structure learning step, we extracted 341 entities related to RT treatment planning in QUANTEC, and translated 61 statements in recommended dose/volume limits section of QUANTEC guidelines into 84 first-order logic formulas. This experiment show preliminary results on MLNs representation of some key knowledge in QUANTEC review on rectum using 29 referenced clinical studies on prostate cancer. We assume that a patient only having one major cancer type for RT treatment, and only receiving one treatment prescription in the RT treatment plan. That is specified by using the syntax '!' following the variables in predicates. Figure 4.2 displays an example of partial MLNs we generated for representing RT knowledge on radiation dose-volume effects of the rectum.

Knowledge Statement:

For patients with prostate cancer, the following dose-volume constraints for conventional fractionation up to 78Gy are provided as a conservative starting point for 3D treatment planning: $V_{50} < 50\%$, $V_{60} < 35\%$, $V_{65} < 25\%$, $V_{70} < 20\%$, and $V_{75} < 15\%$. And the NTCP models predict that following these constraints should limit Grade 2 late rectal toxicity below about 10%.

First-Order Logic Translation:

$\forall p, \exists rt$, IF patient p with prostate cancer, AND p has RT plan rt , AND rt use 3D conventional treatment technique, AND rt has such OAR constraints on rectum: up to 78Gy, $V_{50} < 50\%$, $V_{60} < 35\%$, $V_{65} < 25\%$, $V_{70} < 20\%$, $V_{75} < 15\%$. THEN rt occurs rectal toxicity, AND the risk of the grade 2 rectal toxicity is $\leq 10\%$.

The corresponding MLNs script:

```
// Evidence
ProstateCancer(patient)
Is_Planned(patient, plan)
Use_3D_Conventional(plan)
Is_Constrained_Rectum(patient, plan, criteria)
Is_Prescribed(plan, dose)

// Query
Risk_of_AE(plan, ae, grade, risk)

1.5 ProstateCancer(p) ^ Is_Planned(p, rt) ^ Use_3D_Conventional(rt) ^
Is_Constrained_Rectum(rt, 'up to 78Gy') ^ Is_Constrained_Rectum(rt, 'V50 < 50%') ^
Is_Constrained_Rectum(rt, 'V60 < 35%') ^ Is_Constrained_Rectum(rt, 'V65 < 25%') ^
Is_Constrained_Rectum(rt, 'V70 < 20%') ^ Is_Constrained_Rectum(rt, 'V75 < 15%') ^
Is_Prescribed(rt, dp) => Risk_of_SideEffect(rt, se, g, r) ^ Smaller_Than_10%(r)
```

Figure 4.2: The example of a knowledge statement, its first-order logic translation, and the corresponding MLNs script for representing clinical knowledge on RT treatment plan for curing prostate cancer.

After learning the structure and weights for MLNs, we can apply MLNs to make inference for queries with given evidence. We apply the MLNs to make an inference for a patient named Bob with prostate cancer, and his RT plan followed the dosimetric criteria: $V_{50} < 50\%$, $V_{60} < 35\%$, $V_{65} < 25\%$, $V_{70} < 20\%$, and $V_{75} < 15\%$. V_{50} is a commonly used OAR constraints parameters, which means the OAR volume percentage receiving dose over 50Gy. The query predicate is *Risk_of_AE('Bob', 'late – rectal toxicity', 2, r)*. The predicate indicates the probability of an event occurring

or not in possible worlds. The event is that the risk of a patient named Bob appearing Grade 2 late rectal toxicity is r . The r is a known incidence from RT studies. After inference, Table 4.3 shows the answer to the query *Risk_of_AE('Bob', 'late – rectal toxicity', 2, r)*. The relative probability for each grounding atom indicates the likelihood of the event occurring in the model, instead of the actual probability of the event in the real world. The higher the relative probability is, the more likely the event occurs. According to the inference results, the most likely occurred event is that the risk of Bob having Grade 2 late rectal toxicity is 7% if his RT plan follows the OAR constraints as: $V_{50} < 50\%$, $V_{60} < 35\%$, $V_{65} < 25\%$, $V_{70} < 20\%$, and $V_{75} < 15\%$. Comparing the inference inferencing results with expected risk of toxicity under the recommended dose-volume limits section in QUANTEC paper on the rectum, we found the queried risk of having Grade 2 late rectal toxicity is comparable to the expected risk of late rectal bleeding in QUANTEC review on the rectum. The QUANTEC review on rectum concludes that if a patient with prostate cancer, RT plan should restrict the rectum exposing radiation dosage with $V_{50} < 50\%$, $V_{60} < 35\%$, $V_{65} < 25\%$, $V_{70} < 20\%$, and $V_{75} < 15\%$ to limit the risk of Grade 2 late rectal toxicity below 10%. The risk of a patient having Grade 2 rectal toxicity predicated as 7% from our MLNs inference is consistent with the range $\leq 10\%$ in QUANTEC paper’s recommendation. It suggests the initial validity of the MLNs representation.

Table 4.3: The results of making inference on the query ‘What is the risk of the patient named Bob appearing Grade 2 late rectal toxicity after RT’.

Grounding Atom	Relative Probability
<i>Risk_of_AE('Bob', 'late rectal toxicity', 2, 7%)</i>	0.42
<i>Risk_of_AE('Bob', 'late rectal toxicity', 2, 9%)</i>	0.33
<i>Risk_of_AE('Bob', 'late rectal toxicity', 2, 16%)</i>	0.15
<i>Risk_of_AE('Bob', 'late rectal toxicity', 2, 22%)</i>	0.10

4.4 Discussion

We demonstrated the methods with a small example of the feasibility of using Markov Logic Networks (MLNs) to represent uncertain knowledge of RT planning with complex entities and relationships from narrative RT publications. This representation allows computer inference of queries based on evidence in a probabilistic graphical model. The method of representing clinical knowledge in RT with MLNs provides radiation oncology community a computerized and efficient way to reference a large number of clinical practice guidelines and peer-reviewed clinical research studies during RT treatment planning. And the evidence-driven weight learning process takes patient individual specialization into account. In addition, the paper proposes a general method that can be applied to different domains of health informatics with the Statistical Relational Learning and reasoning model.

However, this method has one limitation in representing RT studies. It is labor-intensive in the understanding the semantics of statements in RT studies to construct first-order logic formulas. To overcome the limitations, we could put more effort on structure (logic) learning by using domain experts' annotation and information retrieval. Ultimately, our hope is that the MLNs based on past clinical studies and guidelines as described in this study can be used as the initial knowledge to learn and accumulate new knowledge from direct clinical study data and thus develop increasingly more sophisticated RT knowledge in the MLN framework.

CHAPTER 5: Quantifying Uncertainty of Hedging Terms in Radiation Oncology Knowledge

Markov Logic Networks (MLNs) use weight to indicate the strength of the formulas constraining the possible worlds represented by MLNs. In short terms, weight in MLNs plays the role to express the uncertainty of clinical knowledge in RT. While we can learn the weights for MNLs formulas with the relational databases, the process of weight learning ignores the uncertain nature of knowledge statements which is caused by human biological systems and biomedical research approaches. The following sentence is an example of uncertainty conveyed in knowledge statement, which is quoted from the QUANTEC paper of lung. The verbal expression 'may' affects the strength of accepting this recommendation in RT guideline to reduce the side effects on lung. It is very common to indicate uncertainty, vagueness and belief to adopt knowledge statements by using verbal expressions likes 'may' in RT publications.

'Limiting the dose to central airways to ≤ 80 Gy **may** reduce the risk of bronchial stricture.'

In this chapter, we present a study to understand the uncertain nature of the hedging terms occurring in RT publications in a quantitative way. Hedging term is defined as a verbal expression denoting probability or uncertainty in RT publications. The study consists of two surveys and the follow-up data analysis. We are interested in the quantification of the uncertainty indicated by hedging terms in RT publications, and their association with contextual information and expert's knowledge background.

5.1 Related Works

Uncertainty usage is common seen phenomena in clinical settings and biomedical texts due to the nature of human language and biological systems. The use of qualitative terms to describe and interpret the uncertainty in medical texts is potentially inaccurate and misunderstanding [110, 111]. In addition, quantification of the uncertainty in medical texts allows computer systems to understand and translate medical texts for further intelligent applications.

Many efforts have been put on probability scoring efforts over decades, which comprehensively conclude that context is crucial to the score [112, 113, 114, 115, 116, 117]. In our study, we aim to investigate how these terms are perceived in the field of radiation therapy and oncology so as to narrow the range of ambiguity and potentially report a probability score of high confidence for this uncertainty (hedging) terms. Since numerical meanings of probability expressions are available in published literatures, instead of scoring context-free hedging terms, we started with scoring hedging terms within a slightly narrow but generic context - radiation therapy to treat cancer resulting in an adverse event. It serves as a standard with which we can compare more context-specific scores of these terms.

Hannauer and his group had a study on the use of uncertainty terms in 100,000 clinical documents to reveal the potential implications of these terms when sharing the documents with patients [118]. They also provided a table of uncertain expressions which we used to select our object hedging terms.

In Kong's study [119], the respondents were given the statement and questions as 'One of the senior physicians in your hospital told you that a particular symptom was in the disease you were discussing. What would be your estimate of the frequency of this symptom in this disease?' One of the 12 probability expressions are filled the blank in the statement. According to Kong' study, they think that not only different contexts but also different formats may affect responses since they are essentially

different measuring devices. They also asked respondents to use four different scales for each response: high probability scale, low probability scale, uniform scale and free choice scale.

In Hobby’s study [120], they used the Visual Analogue Scale to measure the numerical meaning of expressions describing probability in radiologic report. The respondents were asked to indicate the probability of presence of a disease implied by each of the 18 expressions on a visual analogue scale.

In O’Brien’s study [121], the respondents were asked to give a percentage probability rating to a list of 23 words or phrases which could be used in a hypothetical situation to convey to a patient the probability of headache occurring as a side effect from a drug they had prescribed. No drug name or type was specified.

In Timmermans’s study [122], they asked the respondents to assign numerical values with a scale containing three values: point estimate, lower bound, and upper bound. This scale defines both the estimated probability of the hedging term and its probability range. They also asked the respondents several questions with context: 1) whether they would treat the patient or not; 2) how much experience they had had with the specific disease, on a seven-point scale; 3) how much confidence they had in their decisions to treat or not; 4) their interpretations.

5.2 Study Design

5.2.1 Select 14 Hedging Terms as Research Objects

Hedging terms have been studied in isolation from any context or within a context. We selected 14 hedging terms as our research objects from RT publications. The 14 hedging terms are top 14 frequently occurred hedging terms in Radiation Oncology knowledge statements. We did a survey on the hedging terms or verbal expressions carrying uncertainty in RT publications. We counts the frequency for each hedging terms occurred in QUANTEC reviews and their references. The frequency of hedging terms is seen in Figure 5.1, and then we selected the top 14 hedging terms without am-

biguity for our study. The selected hedging terms for this study are: '*common*', '*could be*', '*frequently*', '*high risk*', '*likely*', '*may*', '*might*', '*possible*', '*potential*', '*probably*', '*rarely*', '*risk of*', '*suggest*', and '*usually*'.

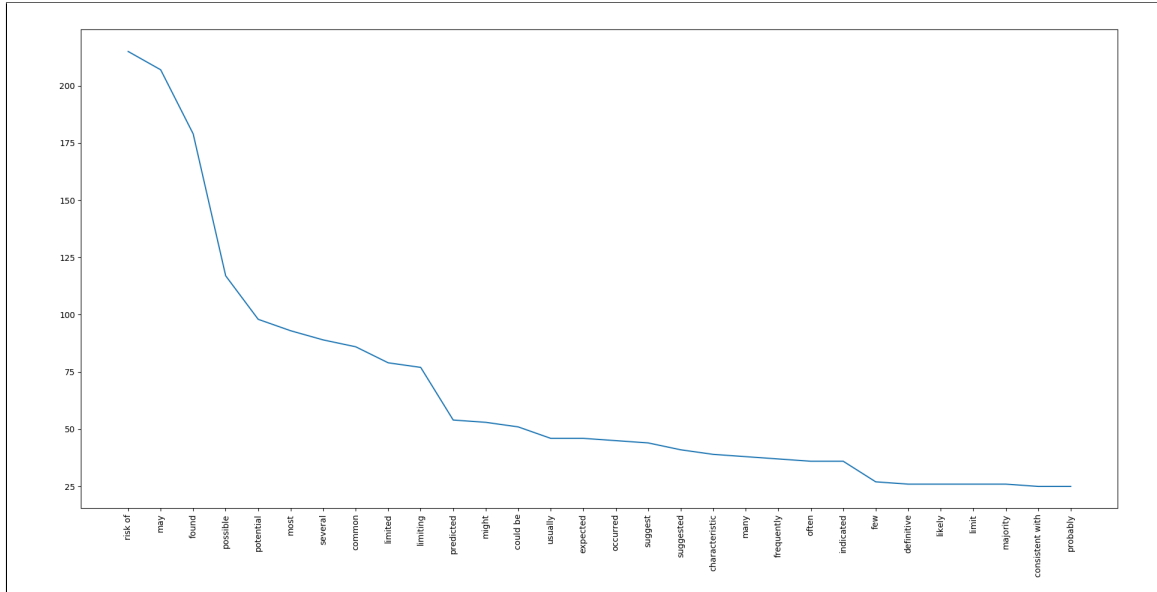


Figure 5.1: The top 30 frequently occurred hedging terms in QUANTEC reviews and RT studies referenced by QUANTEC reviews.

As to the scoring scale for quantifying these 14 hedging terms, we did another survey on the scoring scales for probabilistic degree of verbal expressions in medical documents. We decided to use point estimate, lower bound and upper bound in percentage for scoring the probabilistic belief of the hedging terms.

5.2.2 Experiments: Context-Domain V.S Context-Specific

In the study, we designed two experiments for respondents to score the numerical probability of hedging terms: context-domain experiment and context-specific experiment. We are interested in the effects of contexts put on the quantification of hedging terms. First, in the context-domain situation, context is not complete free, but provide the domain of radiation therapy context for the respondents. In the context-specific situation, contextual information are given, such as organ at risk, target organ, cancer type, and adverse events. In the context-domain experiment, the

respondents were first presented with 14 hedging terms in a random order with only a generic context of radiation therapy domain. The respondents are asked to assign a probability percentage to each hedging term in a hypothetical situation: a patient was treated for cancer with radiation therapy. No connotation of morbidity, sinister adverse event, or parameters of treatment are specified. This is to gain an insight of how individuals of different groups infer probability from only these terms. We did not ask the respondents questions in context-free experiment, because the goal of the study is to investigate the uncertainty of hedging terms in radiation oncology domain, and plenty of work of uncertain expressions have been done in general field.

In context-specific experiment, we provide specific contexts for exact sentences that contain these hedging terms. These sentences are extracted from 470 published clinical studies which are referenced by a Radiation Therapy clinical guideline: QUANTEC. Specific contexts include the organ at risk, target organ, cancer type, and adverse events. This reveals how these hedging terms are scored with specific context. The respondents are asked: 1) to assign a numerical probability for each hedging terms with given specific contexts; 2) to indicate the importance of factors in affecting their choices, such as their experience in the specific context, their confidence on the decision, and the information clarity in their opinion. By comparing the results from the context-domain experiment and context-specific experiment, we can explain variation of results between the two settings, and the factors cause this variation.

5.2.3 Survey Questions

The survey consists of four types of questions. The are (1) the demographic question; (2) the ranking question; (3) the context-domain question; and (4) the context-specific question.

In the demographic question, we ask the respondenst to define their group based on their expertise and knowledge relate to RT, shown in Figure 5.2.

Welcome to the Survey of Quantifying Uncertainty in RT

This study aims to investigate the quantification of uncertain verbal expressions (or hedging terms) in published Radiation Oncology clinical guidelines and clinical studies.

You are asked to give a percentage probability rating to 14 hedging terms with and without context.
Thank you for participating in our survey. Your feedback is important.

Section A. Demographic information

1. Which group most accurately describes you?

☐ Medical Student

☐ Radiation Physicist

☐ Radiation Oncologist

☐ Health Informatician

☐ Lay Person

Figure 5.2: The demographic question asks the respondents to define their group.

In the ranking question, we ask the respondents to rank the 14 hedging terms based on their probabilistic degree from most likely to least likely in Figure 5.3

* 2. Please rank the following 14 hedging terms based on the probabilistic degree they imply. Drag the terms to the box on the right and arrange them in order of most likely to least likely.

common	→
could be	→
frequently	→
high risk	→
likely	→
may	→
might	→
possible	→
potential	→
probably	→
rarely	→
risk of	→
suggest	→
usually	→

Figure 5.3: The ranking question asks the respondents to rank the 14 hedging terms based on their probabilistic degree from most likely to least likely.

In Figure 5.4 and Figure 5.5, the respondents are asked to assign probabilistic scores for each hedging term under two situations separately: context-domain situation and context-specific situation.

3. Please assign three percentage values (including point estimate, lower and upper bounds) to indicate the likelihood of potential adverse events implied by the term **"common"** if a patient was treated for cancer using Radiation Therapy. Scoring scale: 0% for absolutely unlikely – 100% for absolutely certain. *(Please enter number only)*

* Point estimate (%)

* Lower bound (%)

* Upper bound (%)

4. Please assign three percentage values (including point estimate, lower and upper bounds) to indicate the likelihood of potential adverse events implied by the term **"could be"** if a patient was treated for cancer using Radiation Therapy. Scoring scale: 0% for absolutely unlikely – 100% for absolutely certain. *(Please enter number only)*

* Point estimate (%)

* Lower bound (%)

* Upper bound (%)

5. Please assign three percentage values (including point estimate, lower and upper bounds) to indicate the likelihood of potential adverse events implied by the term **"frequently"** if a patient was treated for cancer using Radiation Therapy. Scoring scale: 0% for absolutely unlikely – 100% for absolutely certain. *(Please enter number only)*

* Point estimate (%)

* Lower bound (%)

* Upper bound (%)

Figure 5.4: The context-domain questions ask the respondents to assign each of the 14 hedging terms a probabilistic score under the RT domain situation

"Limiting the dose to the central airways to ≤ 80 Gy **may** reduce the risk of bronchial stricture."

Given context:

- Major cancer: Lung cancer
- Target: the central airways of lung
- Organ at risk: Lung
- Treatment parameters: ≤ 80 Gy
- Adverse event: bronchial stricture

For this sentence, please assign three percentage values (including point estimate, lower and upper bounds) to the term **"may"**. Scoring scale: 0% for absolutely unlikely – 100% for absolutely certain. *(Please enter number only, e.g. if you score the term with 50%, please enter 50 in the box.)*

* Point estimate (%)

* Lower bound (%)

* Upper bound (%)

Figure 5.5: The context-specific question asks the respondents to assign each of the 14 hedging terms a probabilistic score under the RT domain situation and given specific contextual information.

5.3 Results

5.3.1 Demographic Information

We collected responses from 22 participants. Among the twenty-two respondents, there are four medical student, four radiation physicists, three radiation oncologists, seven health informaticians, and four lay persons.

5.3.2 The Ranking of Hedging Terms

The responses of the ranking question are shown as Figure 5.6. Based on the ranking results, we found that only few terms having consistent opinions on their ranking, and most terms having poorly consistent opinions including 'risk of', 'common'.

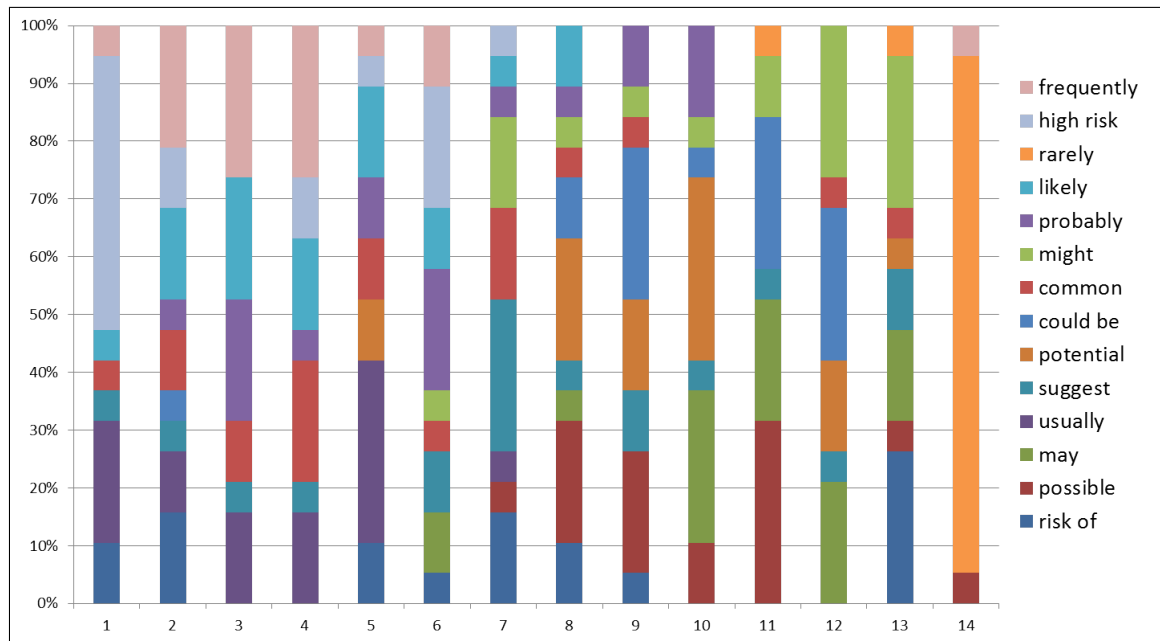


Figure 5.6: The ranking distribution of the 14 hedging terms. The X-axis refers to the ranking orders from top 1 (most likely) to the 14 (least likely). The Y-axis refers to the respondents proportion.

5.3.3 The Probabilistic Scores of Hedging Terms

The Box-Plots in Figure 5.7, Figure 5.8, and Figure 5.9 shows the respondents' scoring for each hedging term on point estimate, upper bound and lower bound respectively. For most of these 14 hedging terms, responses have large variance on

assigning the probabilistic scores.

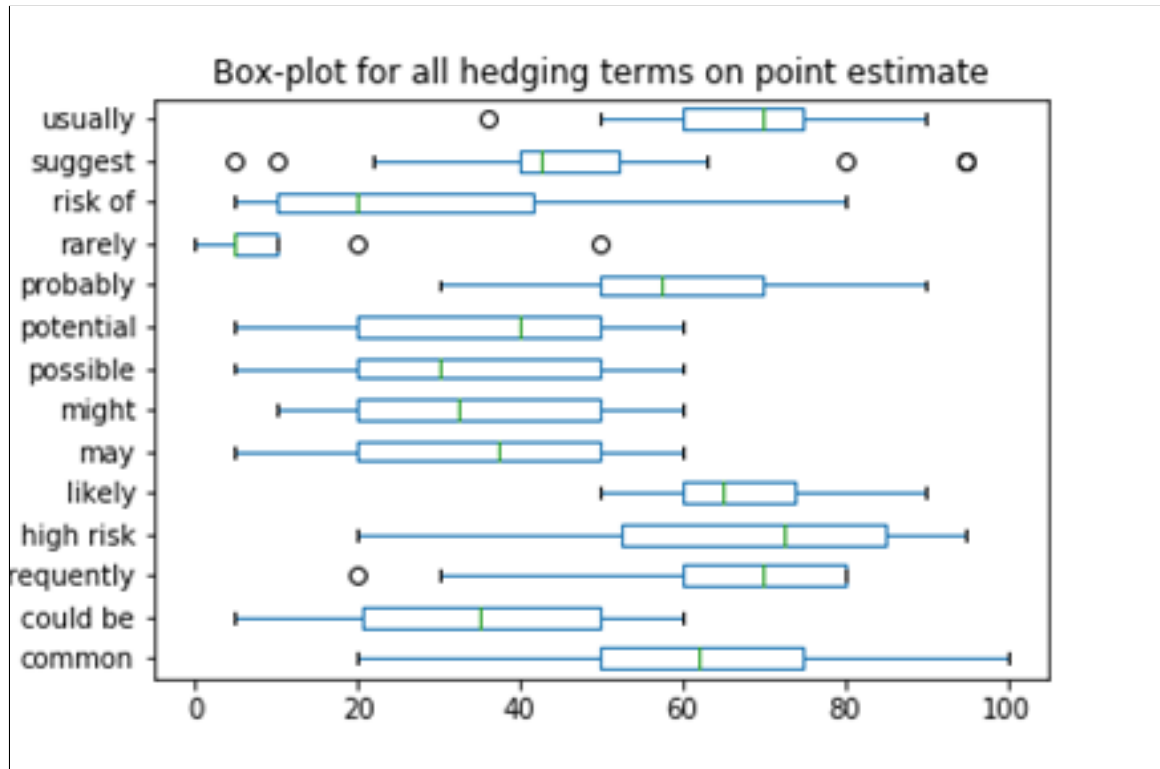


Figure 5.7: The Box Plot of probabilistic scores for the 14 hedging terms on point estimate.

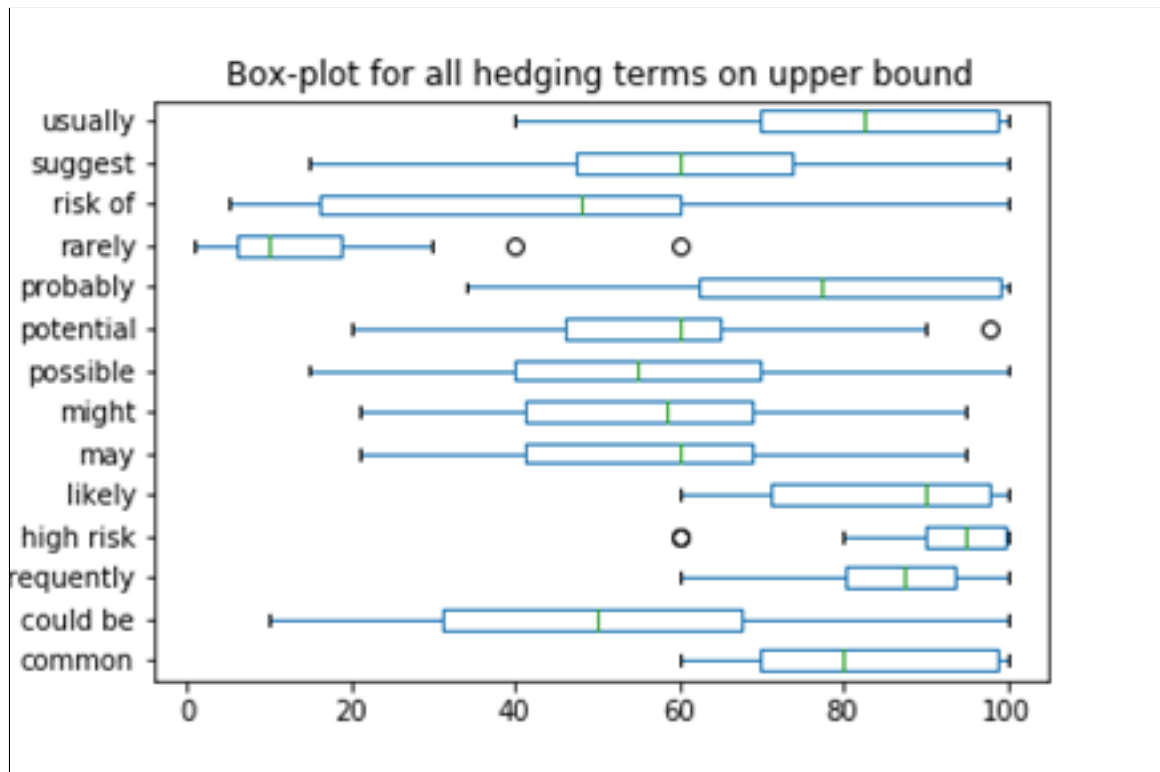


Figure 5.8: The Box Plot of probabilistic scores for the 14 hedging terms on upper bound.

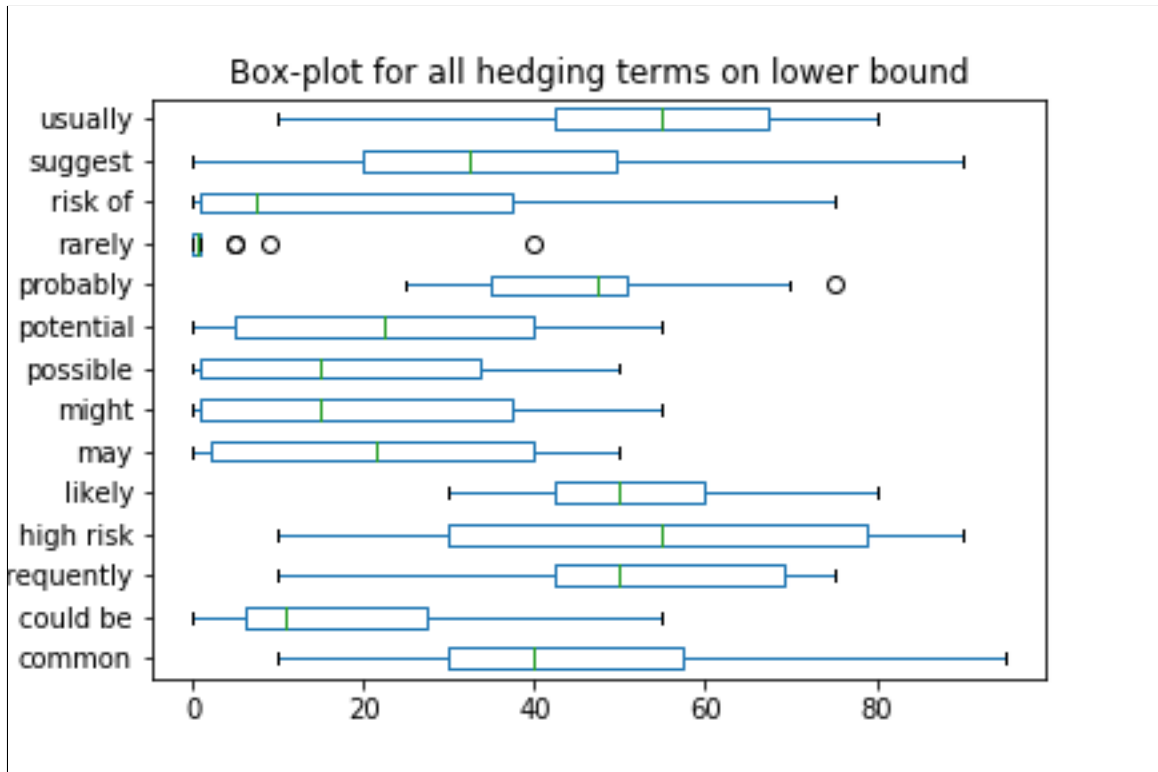



Figure 5.9: The Box Plot of probabilistic scores for the 14 hedging terms on lower bound.

Due to the large variances of probabilistic scores for most hedging terms, we firstly checked their mean values of the probabilistic scores for each hedging terms. Then, we separately explored the probabilistic scores on different respondent groups for each 14 hedging terms. Figure 5.10 lists the mean values of probabilistic score for each hedging term on point estimate. The ranking of their mean values matches the common sense of the usage of these hedging terms. For instance, the hedging term '*probably*' has stronger belief than the term '*possible*'. However, the mean values of probabilistic scores for each hedging terms and their ranking are insufficient to determine the quantitative weights for MLN in RT or other applications.



	Hedging Term	Mean
Most likely	High risk	68.56
	Usually	68.28
	Likely	67.19
	Frequently	65.91
	Common	64.30
	Probably	58.77
	Suggest	47.18
	May	44.24
	Potential	35.78
	Might	33.70
	Could be	33.61
	Possible	33.08
	Risk of	29.26
Least likely	Rarely	8.46

Figure 5.10: The Box Plot of probabilistic scores for the 14 hedging terms.

Thereby, we explored the decisions on assigning probabilistic scores for each hedging terms by different respondent groups. The boxplots displays that different groups have very different opinion on the probabilistic scores for most hedging terms in the following 14 figures including: Figure 5.11, Figure 5.12, Figure 5.13, Figure 5.14, Figure 5.15, Figure 5.16, Figure 5.17, Figure 5.18, Figure 5.19, Figure 5.20, Figure 5.21, Figure 5.22, Figure 5.23, and Figure 5.24.

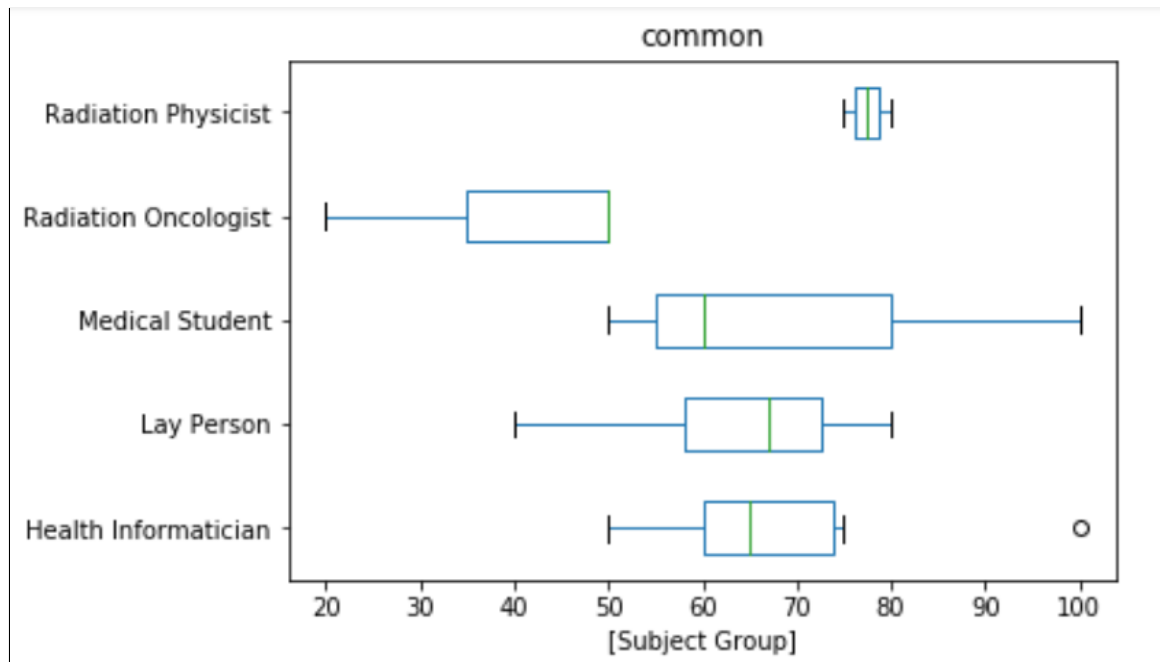


Figure 5.11: The Box Plot of probabilistic scores for the hedging term '*common*'.

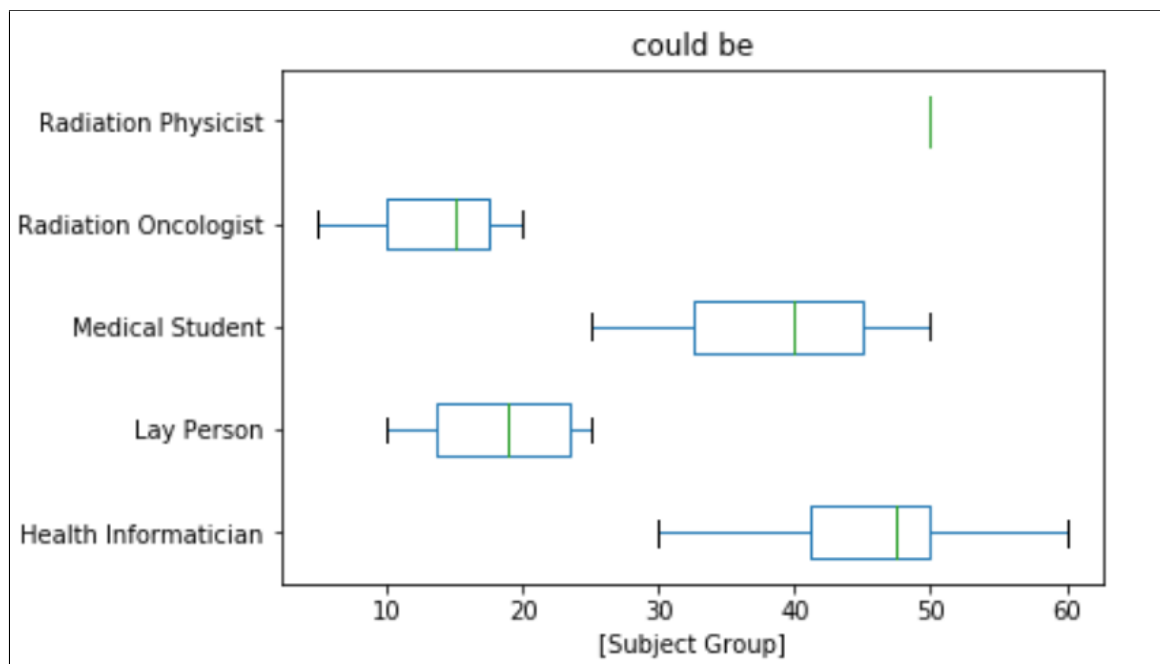


Figure 5.12: The Box Plot of probabilistic scores for the hedging term '*could be*'.

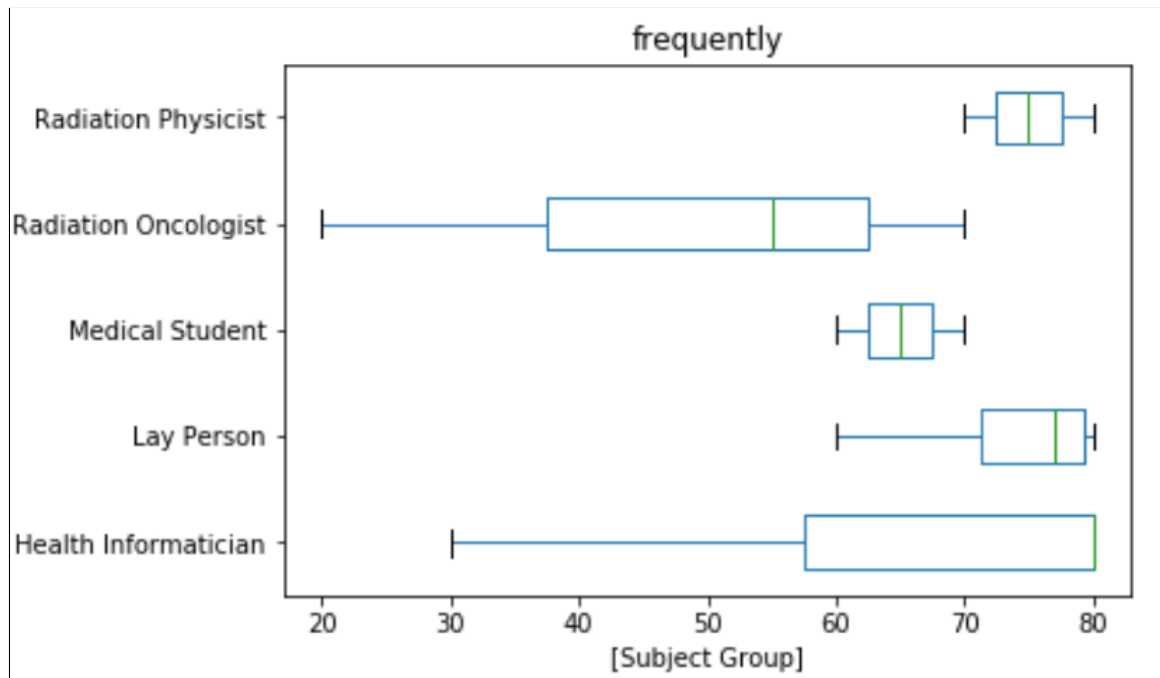


Figure 5.13: The Box Plot of probabilistic scores for the hedging term '*frequently*'.

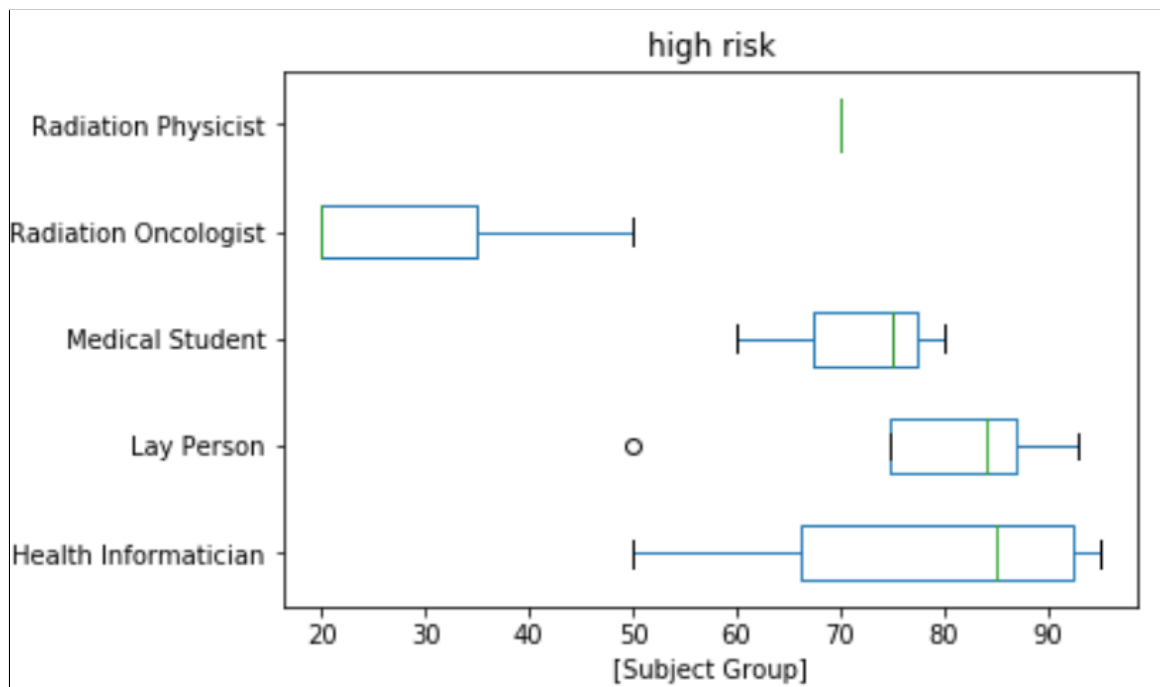


Figure 5.14: The Box Plot of probabilistic scores for the hedging term '*high risk*'.

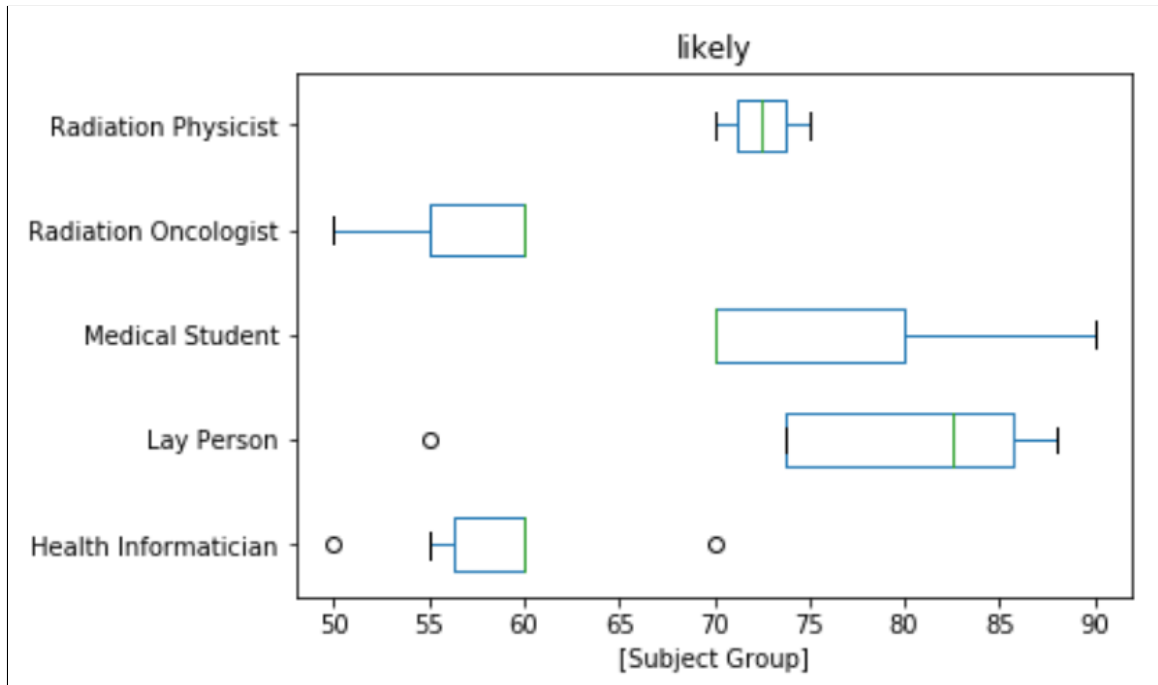


Figure 5.15: The Box Plot of probabilistic scores for the hedging term *'likely'*.

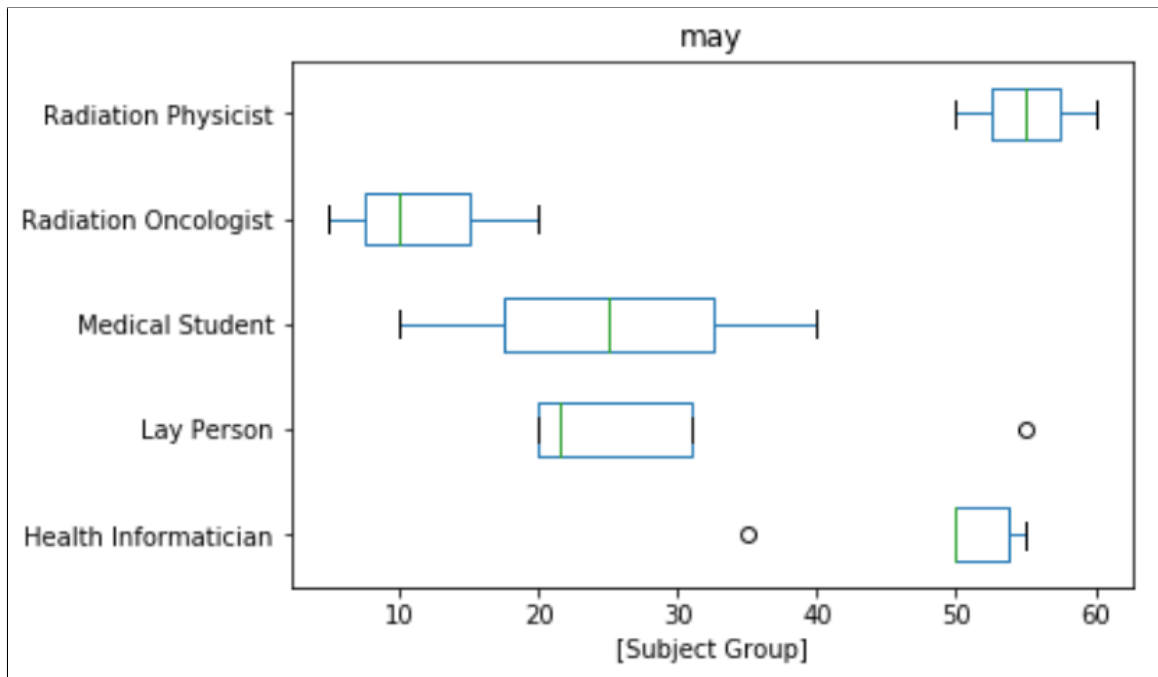


Figure 5.16: The Box Plot of probabilistic scores for the hedging term *'may'*.

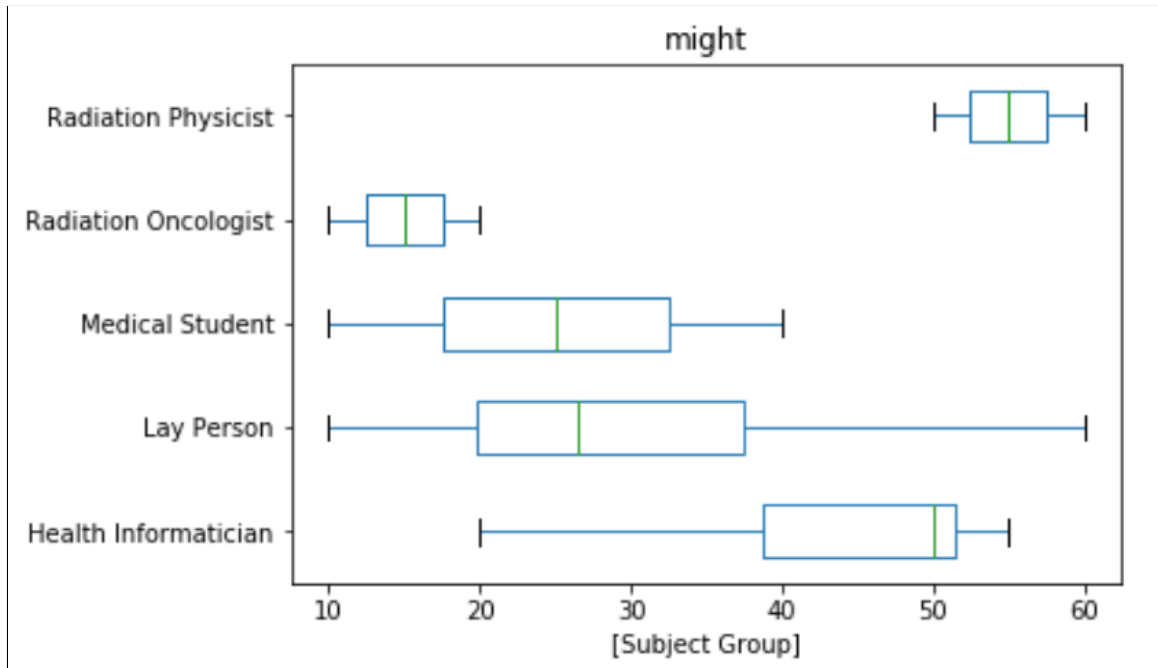


Figure 5.17: The Box Plot of probabilistic scores for the hedging term '*might*'.

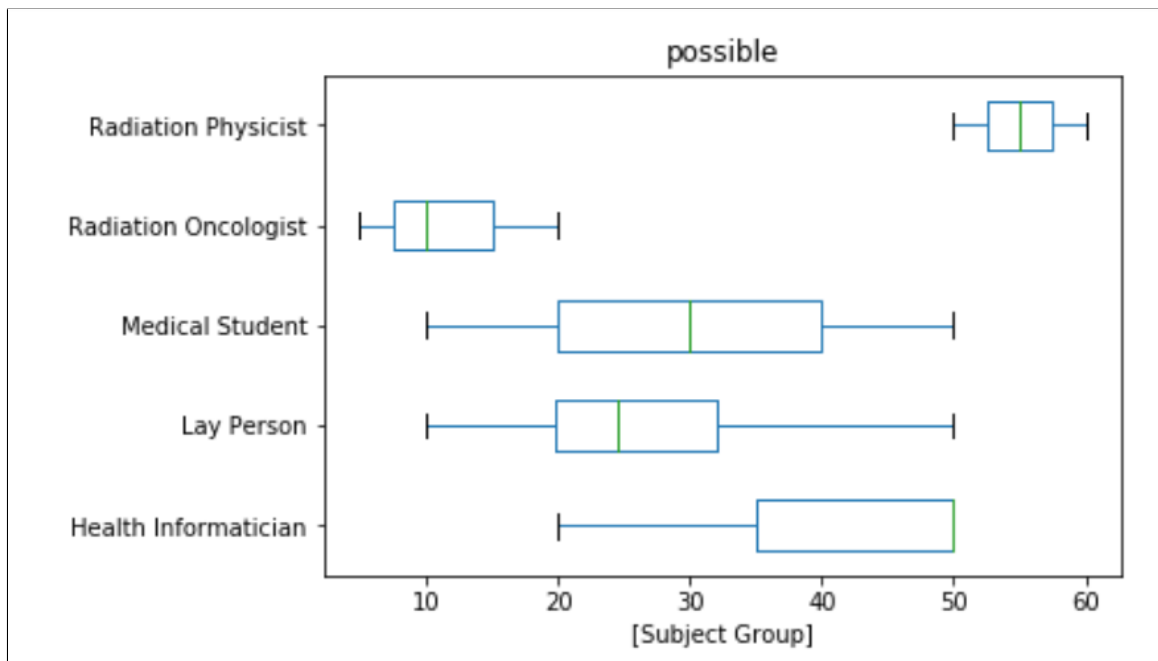


Figure 5.18: The Box Plot of probabilistic scores for the hedging term '*possible*'.

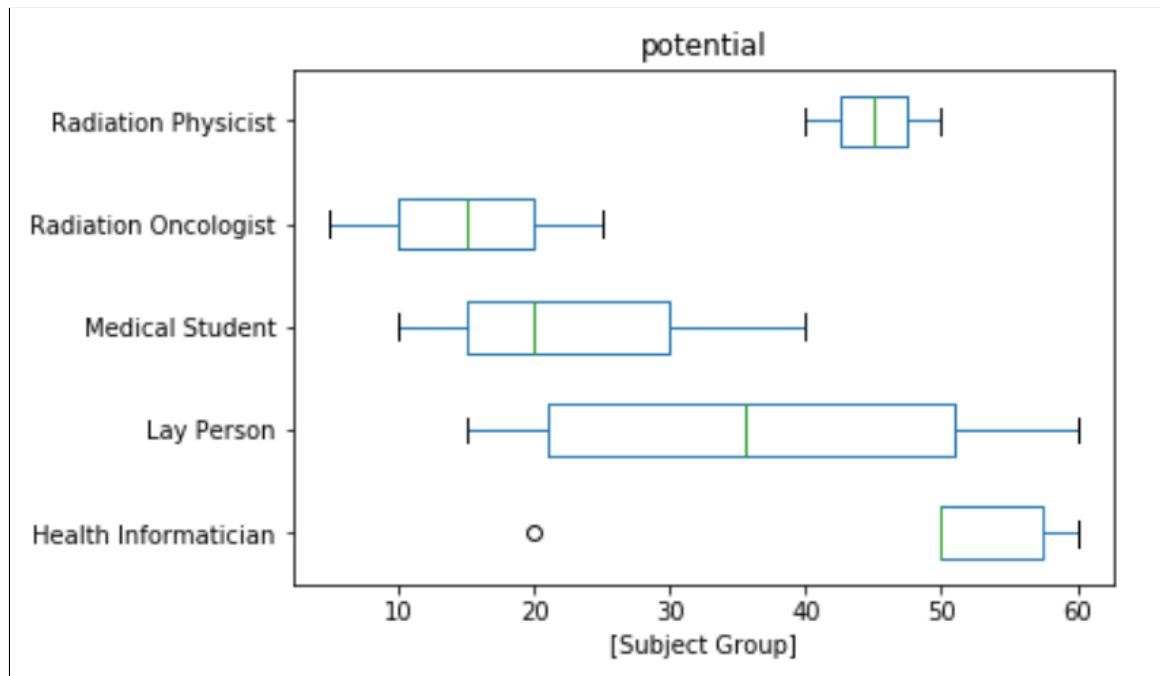


Figure 5.19: The Box Plot of probabilistic scores for the hedging term '*potential*'.

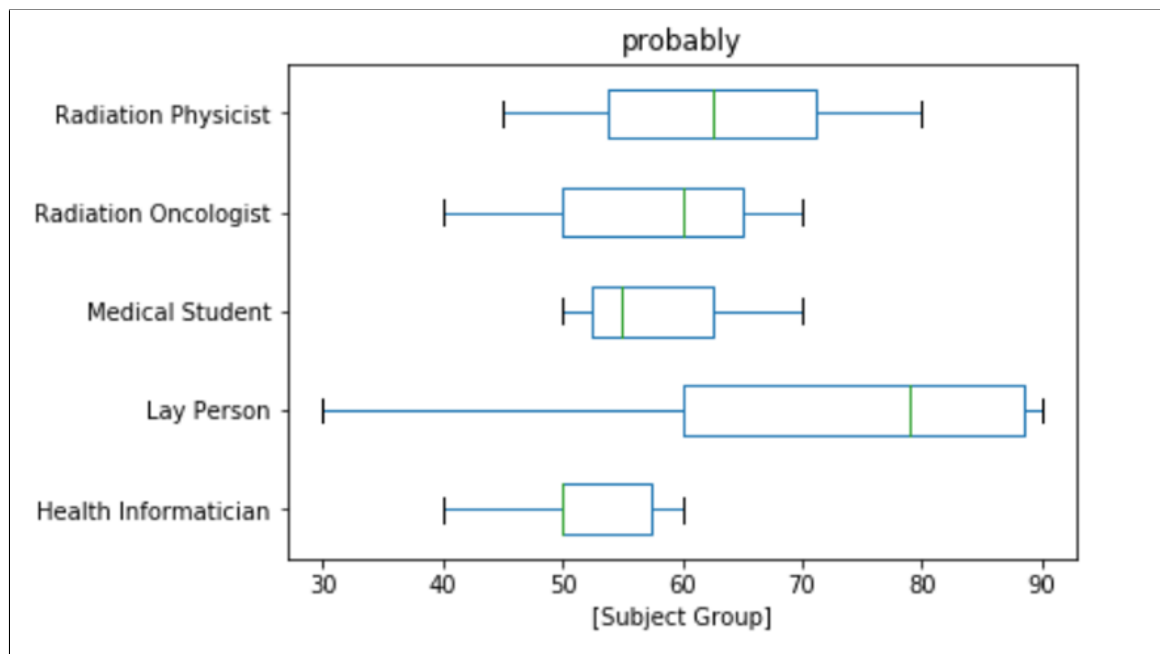


Figure 5.20: The Box Plot of probabilistic scores for the hedging term '*probably*'.

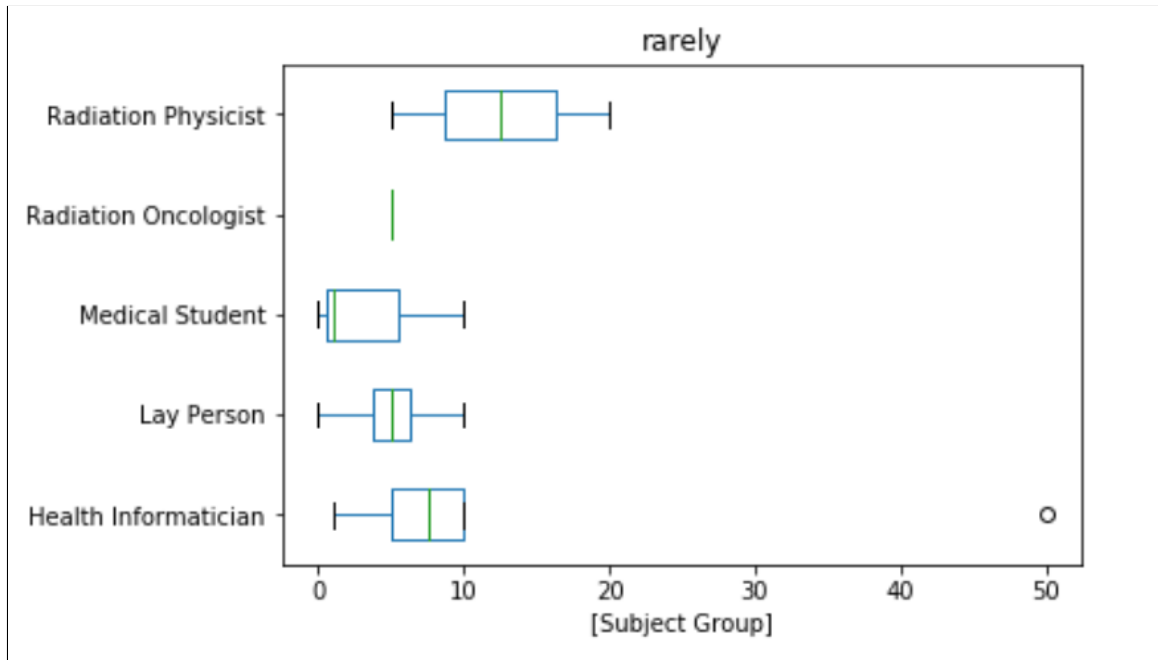


Figure 5.21: The Box Plot of probabilistic scores for the hedging term '*rarely*'.

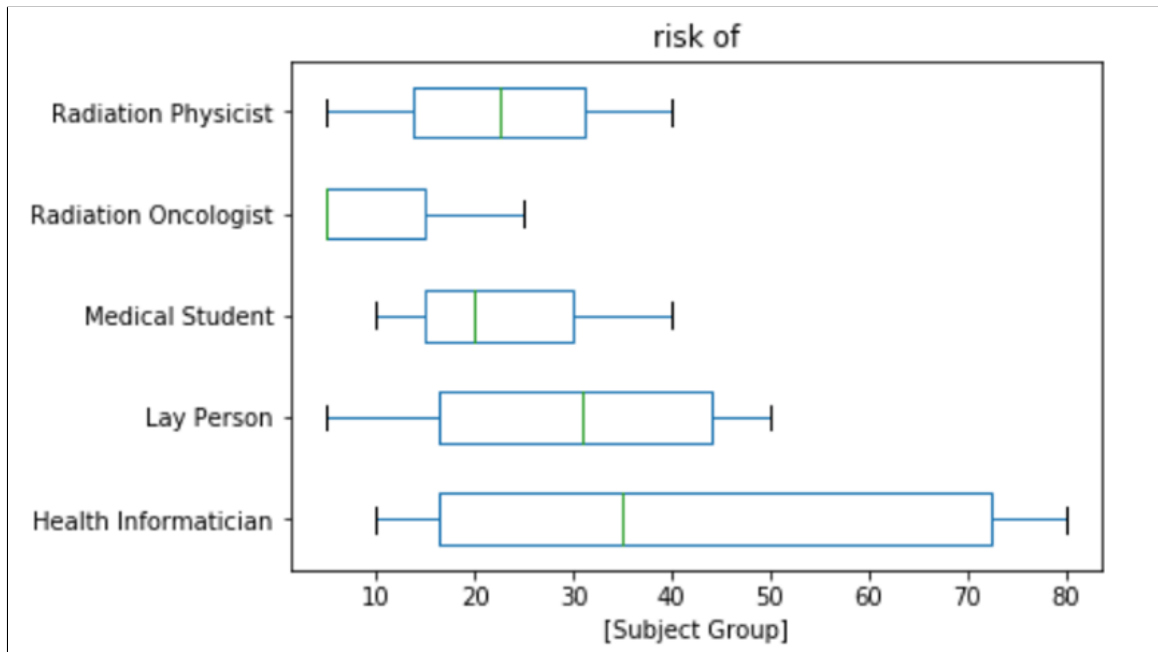


Figure 5.22: The Box Plot of probabilistic scores for the hedging term '*risk of*'.

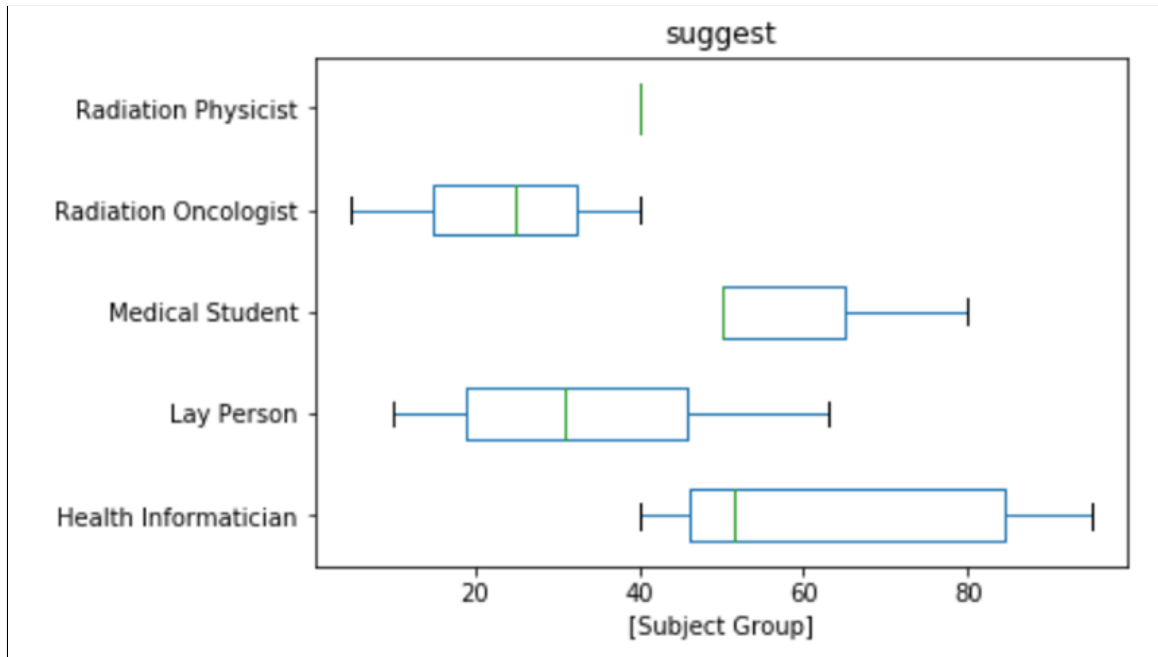


Figure 5.23: The Box Plot of probabilistic scores for the hedging term '*suggest*'.

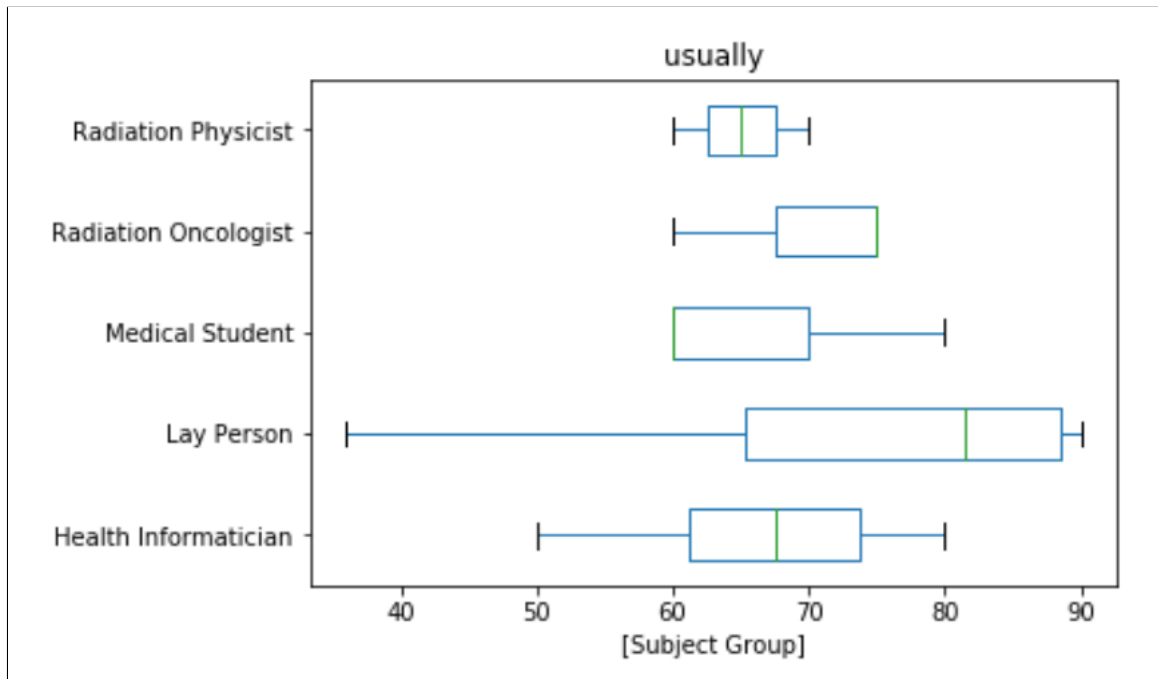


Figure 5.24: The Box Plot of probabilistic scores for the hedging term '*usually*'.

We also made one-way ANOVA analysis on the responses of different groups. Figure 5.25 displays the result of one-way ANOVA on five groups, and Figure 5.26 displays

the result of one-way ANOVA on two groups. The mean values of only 6 hedging terms have significant difference among five respondent groups. The means of probabilistic cores in the 'Radiation Oncologist' group are much lower than other groups, except for term 'usually'. The mean values of only one hedging term ('high risk') have significant difference between lay person and RT expert. There were no significant difference between RT expert and layperson on assigning probabilistic scores to most hedging terms .

	Health Informatician		Radiation Physicist		Lay Person		Medical Student		Radiation Oncologist			
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	f	p
common	69.17	15.92	77.5	15.92	63.5	2.5	70	14.72	40	21.6	1.74411661	0.2003601
could be	45.83	9.32	50	9.32	18.25	0	38.33	6.06	13.33	10.27	10.85654	0.0004309
frequently	66.67	19.72	75	19.72	73.5	5	65	8.02	48.33	4.08	1.07720517	0.4072666
high risk	78.33	17.24	70	17.24	77.75	0	71.67	16.45	30	8.5	4.7663676	0.0137546
likely	59.17	6.07	72.5	6.07	77	2.5	76.67	13.02	56.67	9.43	3.73161043	0.0310631
may	49.17	6.72	55	6.72	29.5	5	25	14.77	11.67	12.25	7.49133769	0.0023353
might	43.67	12.34	55	12.34	30.75	5	25	18.35	15	12.25	3.25038552	0.0469403
possible	41.67	12.13	55	12.13	27.25	5	30	14.45	11.67	16.33	3.56292485	0.0358059
potential	48.33	13.44	45	13.44	36.5	5	23.33	18.23	15	12.47	2.98854355	0.0593531
probably	51.67	6.87	62.5	6.87	69.5	17.5	58.33	24.1	56.67	8.5	0.67041458	0.6239978
rarely	13.5	16.62	12.5	16.62	5	7.5	3.67	3.54	5	4.5	0.58371736	0.6799783
risk of	42.5	29.4	22.5	29.4	29.25	17.5	23.33	17.8	11.67	12.47	0.90662226	0.4884716
suggest	63	22.99	40	22.99	33.75	0	60	19.98	23.33	14.14	2.45679378	0.0977865
usually	66.67	9.86	65	9.86	72.25	5	66.67	21.71	70	9.43	0.1294143	0.9689604

Figure 5.25: The One-Way ANOVA of mean probabilistic scores on five groups, including group 'Lay Person', 'Radiation Physicist', 'Health Informatician', 'Medical Student', and 'Radiation Oncologist'.

	Lay person Mean	Lay person SD	RT Expert Mean	RT Expert SD	f	p
common	67.62	17.29	55	17.29	1.48619676	0.240476
could be	35.62	14.78	28	14.78	0.733358026	0.40444
frequently	68.38	14.67	59	14.67	1.014245364	0.328877
high risk	76.62	15.64	46	15.64	9.502453539	0.007135
likely	68.69	12.94	63	12.94	0.732227504	0.404795
may	37.54	15.53	29	15.53	0.758259714	0.396751
might	35.38	16.48	31	16.48	0.200099773	0.660637
possible	34.54	15.44	29	15.44	0.321231387	0.578734
potential	38.92	17.89	27	17.89	1.495319541	0.2391
probably	58.69	16.61	59	16.61	0.001162174	0.973227
rarely	8.62	12.52	8	12.52	0.009869004	0.9221
risk of	34	24.47	16	24.47	2.125823302	0.164184
suggest	53.31	24.17	30	24.17	3.67332241	0.073328
usually	68.38	14.73	68	14.73	0.002801309	0.958445

Figure 5.26: The One-Way ANOVA of mean probabilistic scores on two groups, including two groups 'Lay Person', and 'Radiation Experts' .

5.4 Conclusion

We conducted a study on quantifying the uncertainty of hedging terms in RT publications. We aim to understand the uncertainty carried by the hedging terms in a quantitative way. We expected to apply the results of quantification as a part of weight learning from expert knowledge for constructing Markov Logic Networks. However, the study concludes that poor consistent opinions among different groups on quantifying probabilistic scores and large variance inside the groups. We should be cautious to use the results as a weight learning method towards MLNs. Due to the small number of respondents participating in the study, we suggest to involve more respondents to obtain reliable results in future.

CHAPTER 6: CONCLUSIONS

In this dissertation, we show that Markov Logic Networks are a promising solution for clinical knowledge representation and reasoning. Towards the practice of EBM in RT treatment planning, we develop an approach to transfer clinical knowledge in RT publications into computerized representation for clinical decision support in RT treatment planning through text mining and knowledge engineering methods. We use text-mining techniques to extract knowledge from RT narrative texts, and apply Markov Logic Network (MLN) to represent and reason clinical knowledge in RT. Markov Logic Network (MLN) is able to represent computerized knowledge of RT with complex entities and relationships from narrative resources. It also allows computer agent make inference of queries based on evidence in a probabilistic graphical model. The method of representing and reasoning RT knowledge with MLN provide radiation oncology community a computerized and efficient way to reference a large number of clinical practice guidelines or peer-reviewed research studies during RT planning. And the evidence-driven weight learning process takes patient individual specialization into account. In addition, the paper proposed a general and semi-automatic method that can be applied into different domains of health informatics with the Statistical Relational Learning and reasoning model.

6.1 Limitations

However, this method has several limits. Currently, we manually translate templates extracted from QUANTEC papers into first-order logics. It is error-prone and labor-intensive in understanding semantics in RT studies to construct first-order logics. To overcome this limit, we could put future effort on logic learning by using

domain experts’ annotation and information retrieval of natural language understanding.

Another drawback of our proposed method is that we ignore the temporal relations when building computerized knowledge representation. And the knowledge extraction from RT text limits the clinical knowledge representation and reasoning. It is difficult to extend the method with rapidly growing RT publications. Therefore in future, a scalable and extendable platform should be implemented to support knowledge extraction from RT text, and representation and reasoning of clinical knowledge in RT. We also need elaborate the evaluation plan, and find the boundary of our proposed model.

6.2 Future Direction of Research Work

In order to overcome the above limitations, we could put efforts on future research work in the following directions. First, inductive logic programming is a promising field of encoding narrative speech into logic formalisms. With inductive logic programming, the structure learning of MLNs with first-order logic formulas can be programmed in a labor-saving and error-less way. Second, there is an interesting problem on how to import temporal information into knowledge representation models. Third, we could conduct a systematical evaluation on the representation of clinical knowledge in RT publications, and a comparison study of different knowledge representation formalisms. In addition, a larger study of quantifying the uncertainty of hedging terms in RT publications is suggested.

REFERENCES

- [1] D. L. Sackett, W. M. Rosenberg, J. M. Gray, R. B. Haynes, and W. S. Richardson, "Evidence based medicine: what it is and what it isn't," 1996.
- [2] S. M. Al-Almaie and N. A. Al-Baghli, "Evidence based medicine: an overview," *Journal of family & community medicine*, vol. 10, no. 2, p. 17, 2003.
- [3] S. E. Straus, P. Glasziou, W. S. Richardson, and R. B. Haynes, *Evidence-Based Medicine E-Book: How to Practice and Teach EBM*. Elsevier Health Sciences, 2018.
- [4] D. E. Malone, "Evidence-based practice in radiology: an introduction to the series," *Radiology*, vol. 242, no. 1, pp. 12–14, 2007.
- [5] M. Staunton, "Evidence-based radiology: steps 1 and 2 - asking answerable questions and searching for evidence," *Radiology*, vol. 242, no. 1, pp. 23–31, 2007.
- [6] B. M. Melnyk, E. Fineout-Overholt, S. B. Stillwell, and K. M. Williamson, "Evidence-based practice: step by step: the seven steps of evidence-based practice," *AJN The American Journal of Nursing*, vol. 110, no. 1, pp. 51–53, 2010.
- [7] I. Sim, P. Gorman, R. A. Greenes, R. B. Haynes, B. Kaplan, H. Lehmann, and P. C. Tang, "Clinical decision support systems for the practice of evidence-based medicine," *Journal of the American Medical Informatics Association*, vol. 8, no. 6, pp. 527–534, 2001.
- [8] A. Georgiou, "Data, information and knowledge: the health informatics model and its role in evidence-based medicine," *Journal of evaluation in clinical practice*, vol. 8, no. 2, pp. 127–130, 2002.
- [9] H. Kilicoglu, D. Demner-Fushman, T. C. Rindflesch, N. L. Wilczynski, and R. B. Haynes, "Towards automatic recognition of scientifically rigorous clinical research evidence," *Journal of the American Medical Informatics Association*, vol. 16, no. 1, pp. 25–31, 2009.
- [10] A. M. Cohen, C. E. Adams, J. M. Davis, C. Yu, P. S. Yu, W. Meng, L. Duggan, M. McDonagh, and N. R. Smalheiser, "Evidence-based medicine, the essential role of systematic reviews, and the need for automated text mining tools," in *Proceedings of the 1st ACM international Health Informatics Symposium*, pp. 376–380, ACM, 2010.
- [11] E. A. Mendonça and J. J. Cimino, "Automated knowledge extraction from medline citations," in *Proceedings of the AMIA Symposium*, p. 575, American Medical Informatics Association, 2000.

- [12] S. Kiritchenko, B. De Bruijn, S. Carini, J. Martin, and I. Sim, “Exact: automatic extraction of clinical trial characteristics from journal publications,” *BMC medical informatics and decision making*, vol. 10, no. 1, p. 56, 2010.
- [13] M. Fiszman, D. Demner-Fushman, H. Kilicoglu, and T. C. Rindflesch, “Automatic summarization of medline citations for evidence-based medical treatment: A topic-oriented evaluation,” *Journal of biomedical informatics*, vol. 42, no. 5, pp. 801–813, 2009.
- [14] D. Li, Z. Wang, L. Wang, S. Sohn, F. Shen, M. H. Murad, and H. Liu, “A text-mining framework for supporting systematic reviews,” *American journal of information management*, vol. 1, no. 1, p. 1, 2016.
- [15] A. Sarker, D. Mollá-Aliod, C. Paris, *et al.*, “Towards automatic grading of evidence,” in *Proceedings of LOUHI 2011 Third International Workshop on Health Document Text Mining and Information Analysis*, pp. 51–58, Citeseer, 2011.
- [16] A. M. Cohen, N. R. Smalheiser, M. S. McDonagh, C. Yu, C. E. Adams, J. M. Davis, and P. S. Yu, “Automated confidence ranked classification of randomized controlled trial articles: an aid to evidence-based medicine,” *Journal of the American Medical Informatics Association*, vol. 22, no. 3, pp. 707–717, 2015.
- [17] D. Wang, M. Peleg, S. W. Tu, E. H. Shortliffe, and R. A. Greenes, “Representation of clinical practice guidelines for computer-based implementations,” *Medinfo*, vol. 10, no. Pt 1, pp. 285–9, 2001.
- [18] M. H. Trivedi, J. Kern, A. Marcee, B. Grannemann, B. Kleiber, T. Bettinger, K. Altshuler, and A. McClelland, “Development and implementation of computerized clinical guidelines: barriers and solutions,” *Methods of information in medicine*, vol. 41, no. 05, pp. 435–442, 2002.
- [19] P. Terenziani, S. Montani, A. Bottrighi, M. Torchio, G. Molino, and G. Correndo, “The glare approach to clinical guidelines: main features,” *Studies in health technology and informatics*, pp. 162–166, 2004.
- [20] A. X. Garg, N. K. Adhikari, H. McDonald, M. P. Rosas-Arellano, P. Devereaux, J. Beyene, J. Sam, and R. B. Haynes, “Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review,” *Jama*, vol. 293, no. 10, pp. 1223–1238, 2005.
- [21] M. Peleg, “Computer-interpretable clinical guidelines: a methodological review,” *Journal of biomedical informatics*, vol. 46, no. 4, pp. 744–763, 2013.
- [22] D. J. Spiegelhalter and R. P. Knill-Jones, “Statistical and knowledge-based approaches to clinical decision-support systems, with an application in gastroenterology,” *Journal of the Royal Statistical Society: Series A (General)*, vol. 147, no. 1, pp. 35–58, 1984.

- [23] G. Kong, D.-L. Xu, and J.-B. Yang, "Clinical decision support systems: a review on knowledge representation and inference under uncertainties," *International Journal of Computational Intelligence Systems*, vol. 1, no. 2, pp. 159–167, 2008.
- [24] L. Ahmadian, M. van Engen-Verheul, F. Bakhshi-Raiez, N. Peek, R. Cornet, and N. F. de Keizer, "The role of standardized data and terminological systems in computerized clinical decision support systems: literature review and survey," *International journal of medical informatics*, vol. 80, no. 2, pp. 81–93, 2011.
- [25] M. A. Musen, B. Middleton, and R. A. Greenes, "Clinical decision-support systems," pp. 643–674, 2014.
- [26] G. Leonardi, A. Bottrighi, G. Galliani, P. Terenziani, A. Messina, and F. Della Corte, "Exceptions handling within glare clinical guideline framework," in *AMIA Annual Symposium Proceedings*, vol. 2012, p. 512, American Medical Informatics Association, 2012.
- [27] T. Minta, *Ontological Representation of Radiation Treatment Data*. PhD thesis, Wake Forest University, 2011.
- [28] K. D. Miller, L. Nogueira, A. B. Mariotto, J. H. Rowland, K. R. Yabroff, C. M. Alfano, A. Jemal, J. L. Kramer, and R. L. Siegel, "Cancer treatment and survivorship statistics, 2019," *CA: a cancer journal for clinicians*, vol. 69, no. 5, pp. 363–385, 2019.
- [29] S. M. Bentzen, "Preventing or reducing late side effects of radiation therapy: radiobiology meets molecular pathology," *Nature Reviews Cancer*, vol. 6, no. 9, p. 702, 2006.
- [30] H. Birgisson, L. Pählman, U. Gunnarsson, and B. Glimelius, "Late adverse effects of radiation therapy for rectal cancer—a systematic overview," *Acta oncologica*, vol. 46, no. 4, pp. 504–516, 2007.
- [31] F. J. Berkey, "Managing the adverse effects of radiation therapy," *Am Fam Physician*, vol. 82, no. 4, pp. 381–8, 2010.
- [32] A. Trotti, A. D. Colevas, A. Setser, V. Rusch, D. Jaques, V. Budach, C. Langer, B. Murphy, R. Cumberlin, C. N. Coleman, *et al.*, "Ctcae v3. 0: development of a comprehensive grading system for the adverse effects of cancer treatment," in *Seminars in radiation oncology*, vol. 13, pp. 176–181, Elsevier, 2003.
- [33] J. D. Cox, J. Stetz, and T. F. Pajak, "Toxicity criteria of the radiation therapy oncology group (rtog) and the european organization for research and treatment of cancer (eortc)," *International Journal of Radiation Oncology Biology Physics*, vol. 31, no. 5, pp. 1341–1346, 1995.
- [34] P. Rubin, L. S. Constine, L. F. Fajardo, T. L. Phillips, T. H. Wasserman, H. Bartelink, J. Denekamp, J.-C. Horiot, F. Mornex, J. Overgaard, *et al.*, "Late

- effects consensus conference: Rtog/eortc,” *Radiotherapy and Oncology*, vol. 35, no. 1, pp. 5–7, 1995.
- [35] S. M. Bentzen, L. S. Constine, J. O. Deasy, A. Eisbruch, A. Jackson, L. B. Marks, R. K. Ten Haken, and E. D. Yorke, “Quantitative analyses of normal tissue effects in the clinic (quantec): an introduction to the scientific issues,” *International Journal of Radiation Oncology* Biology* Physics*, vol. 76, no. 3, pp. S3–S9, 2010.
 - [36] L. B. Marks, R. K. Ten Haken, and M. K. Martel, “Guest editor’s introduction to quantec: a users guide,” *International Journal of Radiation Oncology* Biology* Physics*, vol. 76, no. 3, pp. S1–S2, 2010.
 - [37] J.-E. Bibault, P. Giraud, and A. Burgun, “Big data and machine learning in radiation oncology: state of the art and future prospects,” *Cancer letters*, vol. 382, no. 1, pp. 110–117, 2016.
 - [38] F. Van Harmelen, V. Lifschitz, and B. Porter, *Handbook of knowledge representation*, vol. 1. Elsevier, 2008.
 - [39] U. Grenander and M. I. Miller, “Representations of knowledge in complex systems,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 56, no. 4, pp. 549–581, 1994.
 - [40] H. J. Levesque, “Foundations of a functional approach to knowledge representation,” *Artificial Intelligence*, vol. 23, no. 2, pp. 155–212, 1984.
 - [41] S. Bechhofer, “Owl: Web ontology language,” *Encyclopedia of database systems*, pp. 2008–2009, 2009.
 - [42] F. Bacchus and F. Kabanza, “Using temporal logics to express search control knowledge for planning,” *Artificial intelligence*, vol. 116, no. 1-2, pp. 123–191, 2000.
 - [43] G. Hripcsak, “Writing arden syntax medical logic modules,” *Computers in biology and medicine*, vol. 24, no. 5, pp. 331–363, 1994.
 - [44] G. J. Tso, S. W. Tu, C. Oshiro, S. Martins, M. Ashcraft, K. W. Yuen, D. Wang, A. Robinson, P. A. Heidenreich, and M. K. Goldstein, “Automating guidelines for clinical decision support: knowledge engineering and implementation,” in *AMIA Annual Symposium Proceedings*, vol. 2016, p. 1189, American Medical Informatics Association, 2016.
 - [45] A. Stojadinovic, A. Bilchik, D. Smith, J. S. Eberhardt, E. B. Ward, A. Nissan, E. K. Johnson, M. Protic, G. E. Peoples, I. Avital, *et al.*, “Clinical decision support and individualized prediction of survival in colon cancer: Bayesian belief network model,” *Annals of surgical oncology*, vol. 20, no. 1, pp. 161–174, 2013.

- [46] A. Onisko and R. M. Austin, “Dynamic bayesian network for cervical cancer screening,” in *Foundations of Biomedical Knowledge Representation*, pp. 207–218, Springer, 2015.
- [47] X.-H. Wang, B. Zheng, W. F. Good, J. L. King, and Y.-H. Chang, “Computer-assisted diagnosis of breast cancer using a data-driven bayesian belief network,” *International journal of medical informatics*, vol. 54, no. 2, pp. 115–126, 1999.
- [48] E. Burnside, D. Rubin, and R. Shachter, “A bayesian network for mammography,” in *Proceedings of the AMIA Symposium*, p. 106, American Medical Informatics Association, 2000.
- [49] M. Verduijn, N. Peek, P. M. Rosseel, E. de Jonge, and B. A. de Mol, “Prognostic bayesian networks: I: Rationale, learning procedure, and clinical use,” *Journal of Biomedical Informatics*, vol. 40, no. 6, pp. 609–618, 2007.
- [50] A. Nissan, M. Protic, A. Bilchik, J. Eberhardt, G. E. Peoples, and A. Stojadinovic, “Predictive model of outcome of targeted nodal assessment in colorectal cancer,” *Annals of surgery*, vol. 251, no. 2, pp. 265–274, 2010.
- [51] S. F. Galan, F. Aguado, F. Diez, and J. Mira, “Nasonet, modeling the spread of nasopharyngeal cancer with networks of probabilistic events in discrete time,” *Artificial Intelligence in Medicine*, vol. 25, no. 3, pp. 247–264, 2002.
- [52] A. M. Kalet, *Bayesian networks from ontological formalisms in radiation oncology*. PhD thesis, University of Washington, 2015.
- [53] C. Bettini, O. Brdiczka, K. Henricksen, J. Indulska, D. Nicklas, A. Ranganathan, and D. Riboni, “A survey of context modelling and reasoning techniques,” *Pervasive and Mobile Computing*, vol. 6, no. 2, pp. 161–180, 2010.
- [54] D. Koller, N. Friedman, S. Džeroski, C. Sutton, A. McCallum, A. Pfeffer, P. Abbeel, M.-F. Wong, D. Heckerman, C. Meek, *et al.*, *Introduction to statistical relational learning*. MIT press, 2007.
- [55] K. B. Laskey, “Mebn: A language for first-order bayesian knowledge bases,” *Artificial intelligence*, vol. 172, no. 2-3, pp. 140–178, 2008.
- [56] P. Domingos and M. Richardson, “1 markov logic: A unifying framework for statistical relational learning,” *Statistical Relational Learning*, p. 339, 2007.
- [57] S. Ghosh, N. Shankar, S. Owre, S. David, G. Swan, and P. Lincoln, “Markov logic networks in health informatics,” *In Proceedings of ICML-MLGC*, vol. 2011, 2011.
- [58] M. Niepert, C. Meilicke, and H. Stuckenschmidt, “A probabilistic-logical framework for ontology matching,” in *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

- [59] H. Poon and L. Vanderwende, “Joint inference for knowledge extraction from biomedical literature,” in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 813–821, Association for Computational Linguistics, 2010.
- [60] S. Riedel, H.-W. Chun, T. Takagi, and J. Tsujii, “A markov logic approach to bio-molecular event extraction,” in *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pp. 41–49, Association for Computational Linguistics, 2009.
- [61] S. Natarajan, V. Bangera, T. Khot, J. Picado, A. Wazalwar, V. S. Costa, D. Page, and M. Caldwell, “Markov logic networks for adverse drug event extraction from text,” *Knowledge and information systems*, vol. 51, no. 2, pp. 435–457, 2017.
- [62] S. Richardson, P. Domingos, and M. S. H. Poon, “The alchemy system for statistical relational ai: User manual,” 2007.
- [63] M. Sumner and P. Domingos, “The alchemy tutorial,” 2010.
- [64] F. Niu, C. Ré, A. Doan, and J. Shavlik, “Tuffy: Scaling up statistical inference in markov logic networks using an rdbms,” *Proceedings of the VLDB Endowment*, vol. 4, no. 6, pp. 373–384, 2011.
- [65] L. Snidaro, I. Visentini, and K. Bryan, “Fusing uncertain knowledge and evidence for maritime situational awareness via markov logic networks,” *Information Fusion*, vol. 21, pp. 159–172, 2015.
- [66] A. Kimmig, S. Bach, M. Broecheler, B. Huang, and L. Getoor, “A short introduction to probabilistic soft logic,” in *Proceedings of the NIPS Workshop on Probabilistic Programming: Foundations and Applications*, pp. 1–4, 2012.
- [67] L. D. Raedt and K. Kersting, “Statistical relational learning,” *Encyclopedia of Machine Learning*, pp. 916–924, 2010.
- [68] J. Lee and Y. Wang, “On the semantic relationship between probabilistic soft logic and markov logic,” *arXiv preprint arXiv:1606.08896*, 2016.
- [69] I. Beltagy, K. Erk, and R. Mooney, “Probabilistic soft logic for semantic textual similarity,” in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1210–1219, 2014.
- [70] M. Richardson and P. Domingos, “Markov logic networks,” *Machine learning*, vol. 62, no. 1-2, pp. 107–136, 2006.
- [71] P. Domingos and D. Lowd, “Markov logic: An interface layer for artificial intelligence,” *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 3, no. 1, pp. 1–155, 2009.

- [72] A. Trotti and S. M. Bentzen, “The need for adverse effects reporting standards in oncology clinical trials,” *Journal of Clinical Oncology*, vol. 22, no. 1, pp. 19–22, 2004.
- [73] “Ctcae v4.0 apple app.” <https://itunes.apple.com/us/app/ctcae-v4.0/>.
- [74] T. C. T. E. Program, “Common terminology criteria for adverse events in owl format in ncbo bioportal.” <http://bioportal.bioontology.org/ontologies/CTCAE>.
- [75] T. Gruber, “Ontology,” 2018.
- [76] Y. Ding and S. Foo, “Ontology research and development. part 1-a review of ontology generation,” *Journal of information science*, vol. 28, no. 2, pp. 123–136, 2002.
- [77] N. Guarino, “Some ontological principles for designing upper level lexical resources,” *arXiv preprint cmp-lg/9809002*, 1998.
- [78] M. Uschold and M. Gruninger, “Ontologies: Principles, methods and applications,” *The knowledge engineering review*, vol. 11, no. 2, pp. 93–136, 1996.
- [79] J. R. Reich, “Ontological design patterns: Metadata of molecular biological ontologies, information and knowledge,” in *International Conference on Database and Expert Systems Applications*, pp. 698–709, Springer, 2000.
- [80] C. H. Hwang, “Incompletely and imprecisely speaking: using dynamic ontologies for representing and retrieving information,” in *KRDB*, vol. 21, pp. 14–20, 1999.
- [81] N. F. Noy, D. L. McGuinness, *et al.*, “Ontology development 101: A guide to creating your first ontology,” 2001.
- [82] I. Bedini and B. Nguyen, “Automatic ontology generation: State of the art,” *PRiSM Laboratory Technical Report. University of Versailles*, 2007.
- [83] N. F. Noy, R. W. Fergerson, and M. A. Musen, “The knowledge model of protege-2000: Combining interoperability and flexibility,” in *International Conference on Knowledge Engineering and Knowledge Management*, pp. 17–32, Springer, 2000.
- [84] H. Bohring and S. Auer, “Mapping xml to owl ontologies,” *Marktplatz Internet: Von e-Learning bis e-Payment, 13. Leipziger Informatik-Tage (LIT 2005)*, 2015.
- [85] R. Ghawi and N. Cullot, “Building ontologies from xml data sources,” in *2009 20th International Workshop on Database and Expert Systems Application*, pp. 480–484, IEEE, 2009.
- [86] N. Yahia, S. A. Mokhtar, and A. Ahmed, “Automatic generation of owl ontology from xml data source,” *arXiv preprint arXiv:1206.0570*, 2012.

- [87] L. M. Schriml, C. Arze, S. Nadendla, Y.-W. W. Chang, M. Mazaitis, V. Felix, G. Feng, and W. A. Kibbe, "Disease ontology: a backbone for disease semantic integration," *Nucleic acids research*, vol. 40, no. D1, pp. D940–D946, 2011.
- [88] C. Rosse and J. L. Mejino, "The foundational model of anatomy ontology," in *Anatomy Ontologies for Bioinformatics*, pp. 59–117, Springer, 2008.
- [89] A. Trotti, R. Byhardt, J. Stetz, C. Gwede, B. Corn, K. Fu, L. Gunderson, B. McCormick, M. Morris, T. Rich, *et al.*, "Common toxicity criteria: version 2.0. an improved reference for grading the acute effects of cancer treatment: impact on radiotherapy," *International Journal of Radiation Oncology* Biology* Physics*, vol. 47, no. 1, pp. 13–47, 2000.
- [90] A. Colevas and A. Setser, "The nci common terminology criteria for adverse events (ctcae) v 3.0 is the new standard for oncology clinical trials," *Journal of Clinical Oncology*, vol. 22, no. 14_suppl, pp. 6098–6098, 2004.
- [91] J.-J. Pavy, J. Denekamp, J. Letschert, B. Littbrand, F. Mornex, J. Bernier, D. Gonzales-Gonzales, J. Horiot, M. Bolla, and H. Bartelink, "Late effects toxicity scoring: the soma scale," *International Journal of Radiation Oncology Biology Physics*, vol. 31, no. 5, pp. 1043–1047, 1995.
- [92] J. A. Ajani, S. R. Welch, M. N. Raber, W. S. Fields, and I. H. Krakoff, "Comprehensive criteria for assessing therapy-induced toxicity," *Cancer investigation*, vol. 8, no. 2, pp. 147–159, 1990.
- [93] D. W. Bruner and T. Wasserman, "The impact on quality of life by radiation late effects," *International Journal of Radiation Oncology* Biology* Physics*, vol. 31, no. 5, pp. 1353–1355, 1995.
- [94] S. Dische, M. Warburton, D. Jones, and E. Lartigau, "The recording of morbidity related to radiotherapy," *Radiotherapy and Oncology*, vol. 16, no. 2, pp. 103–108, 1989.
- [95] M. Goleń, K. Skłodowski, A. Wygoda, W. Przeorek, B. Pilecki, M. Sygula, B. Maciejewski, and Z. Kołosza, "A comparison of two scoring systems for late radiation toxicity in patients after radiotherapy for head and neck cancer," *Reports of Practical Oncology & Radiotherapy*, vol. 10, no. 4, pp. 179–192, 2005.
- [96] F. Denis, P. Garaud, E. Bardet, M. Alfonsi, C. Sire, T. Germain, P. Bergerot, B. Rhein, J. Tortochaux, P. Oudinot, *et al.*, "Late toxicity results of the gortec 94-01 randomized trial comparing radiotherapy with concomitant radiochemotherapy for advanced-stage oropharynx carcinoma: comparison of lent/soma, rtog/eortc, and nci-ctc scoring systems," *International Journal of Radiation Oncology* Biology* Physics*, vol. 55, no. 1, pp. 93–98, 2003.
- [97] H. P. van der Laan, A. van den Bergh, C. Schilstra, R. Vlasman, H. Meertens, and J. A. Langendijk, "Grading-system-dependent volume effects for late

- radiation-induced rectal toxicity after curative radiotherapy for prostate cancer,” *International Journal of Radiation Oncology* Biology* Physics*, vol. 70, no. 4, pp. 1138–1145, 2008.
- [98] M. Krauthammer and G. Nenadic, “Term identification in the biomedical literature,” *Journal of biomedical informatics*, vol. 37, no. 6, pp. 512–526, 2004.
 - [99] A. M. Cohen and W. R. Hersh, “A survey of current work in biomedical text mining,” *Briefings in bioinformatics*, vol. 6, no. 1, pp. 57–71, 2005.
 - [100] R. Harpaz, W. DuMouchel, N. H. Shah, D. Madigan, P. Ryan, and C. Friedman, “Novel data-mining methodologies for adverse drug event discovery and analysis,” *Clinical Pharmacology & Therapeutics*, vol. 91, no. 6, pp. 1010–1021, 2012.
 - [101] D. D. A. Bui and Q. Zeng-Treitler, “Learning regular expressions for clinical text classification,” *Journal of the American Medical Informatics Association*, vol. 21, no. 5, pp. 850–857, 2014.
 - [102] F. Sebastiani, “Machine learning in automated text categorization,” *ACM computing surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, 2002.
 - [103] R. J. Mooney and R. Bunescu, “Mining knowledge from text using information extraction,” *ACM SIGKDD explorations newsletter*, vol. 7, no. 1, pp. 3–10, 2005.
 - [104] A. McCallum, K. Nigam, *et al.*, “A comparison of event models for naive bayes text classification,” in *AAAI-98 workshop on learning for text categorization*, vol. 752, pp. 41–48, Citeseer, 1998.
 - [105] Y. Yang, “An evaluation of statistical approaches to text categorization,” *Information retrieval*, vol. 1, no. 1-2, pp. 69–90, 1999.
 - [106] K. Donnelly, “Snomed-ct: The advanced terminology and coding system for ehealth,” *Studies in health technology and informatics*, vol. 121, p. 279, 2006.
 - [107] O. Bodenreider, “The unified medical language system (umls): integrating biomedical terminology,” *Nucleic acids research*, vol. 32, no. suppl_1, pp. D267–D270, 2004.
 - [108] A. N. Viswanathan, E. D. Yorke, L. B. Marks, P. J. Eifel, and W. U. Shipley, “Radiation dose–volume effects of the urinary bladder,” *International Journal of Radiation Oncology* Biology* Physics*, vol. 76, no. 3, pp. S116–S122, 2010.
 - [109] F. J. Pos, G. van Tienhoven, M. C. Hulshof, K. Koedooder, and D. G. González, “Concomitant boost radiotherapy for muscle invasive bladder cancer,” *Radiotherapy and oncology*, vol. 68, no. 1, pp. 75–80, 2003.

- [110] D. L. Mowery, S. Velupillai, and W. W. Chapman, "Medical diagnosis lost in translation: analysis of uncertainty and negation expressions in english and swedish clinical texts," in *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pp. 56–64, Association for Computational Linguistics, 2012.
- [111] M. Foppa, B. S. de Araujo, A. Macari, R. Reichert, and J. R. Goldim, "Limitations in the use of qualitative terms to inform diagnoses," *Archives of internal medicine*, vol. 171, no. 14, pp. 1291–1292, 2011.
- [112] M. Biehl and B. L. Halpern-Felsher, "Adolescents' and adults' understanding of probability expressions," *Journal of Adolescent Health*, vol. 28, no. 1, pp. 30–35, 2001.
- [113] M. M. Christopher and C. S. Hotz, "Cytologic diagnosis: expression of probability by clinical pathologists," *Veterinary clinical pathology*, vol. 33, no. 2, pp. 84–95, 2004.
- [114] M. M. Christopher, C. Hotz, S. Shelly, and P. Pion, "Interpretation by clinicians of probability expressions in cytology reports and effect on clinical decision-making," *Journal of veterinary internal medicine*, vol. 24, no. 3, pp. 496–503, 2010.
- [115] R. E. Mapes, "Verbal and numerical estimates of probability in therapeutic contexts," *Social Science & Medicine. Part A: Medical Psychology & Medical Sociology*, vol. 13, pp. 277–282, 1979.
- [116] M. J. Druzdzel, "Verbal uncertainty expressions: Literature review," *Pittsburgh, PA: Carnegie Mellon University, Department of Engineering and Public Policy*, 1989.
- [117] S. Codish and R. N. Shiffman, "A model of ambiguity and vagueness in clinical practice guideline recommendations," in *AMIA Annual Symposium Proceedings*, vol. 2005, p. 146, American Medical Informatics Association, 2005.
- [118] D. A. Hanauer, Y. Liu, Q. Mei, F. J. Manion, U. J. Balis, and K. Zheng, "Hedging their mets: the use of uncertainty terms in clinical documents and its potential implications when sharing the documents with patients," in *AMIA Annual Symposium Proceedings*, vol. 2012, p. 321, American Medical Informatics Association, 2012.
- [119] A. Kong, G. O. Barnett, F. Mosteller, and C. Youtz, "How medical professionals evaluate expressions of probability," *New England Journal of Medicine*, vol. 315, no. 12, pp. 740–744, 1986.
- [120] J. L. Hobby, B. Tom, C. Todd, P. Bearcroft, and A. K. Dixon, "Communication of doubt and certainty in radiological reports.," *The British Journal of Radiology*, vol. 73, no. 873, pp. 999–1001, 2000.

- [121] B. J. O'Brien, "Words or numbers? the evaluation of probability expressions in general practice.," *JR Coll Gen Pract*, vol. 39, no. 320, pp. 98–100, 1989.
- [122] D. Timmermans, "The roles of experience and domain of expertise in using numerical and verbal probability terms in medical decisions," *Medical Decision Making*, vol. 14, no. 2, pp. 146–156, 1994.