2023

A Learning Health System for Radiation Oncology

Rishabh Kapoor
Virginia Commonwealth University

A Learning Health System for Radiation Oncology

Submitted in partial fulfillment of the requirement for the degree of Doctor of Philosophy at Virginia Commonwealth University

By

Rishabh Kapoor, M.S.

Advisor: Jatinder Palta, PhD

Professor, Department of Radiation Oncology

Virginia Commonwealth University

Richmond, Virginia

July 2023

Acknowledgements

I would like to express my sincere gratitude and appreciation to the following individuals and organizations who have played a significant role in the successful completion of my PhD thesis:

First and foremost, I am deeply indebted to my mentor and PhD advisor, Dr. Jatinder Palta, for his exceptional guidance, unwavering support, and invaluable expertise throughout this journey. His mentorship has not only shaped my professional growth but also ignited my passion for research in the field of radiation oncology. I am truly grateful for his encouragement and for sharing his extensive knowledge, which has been instrumental in my development.

I would also like to extend my heartfelt thanks to all the teachers, graduate assistants, and medical physics residents who have contributed to my education. Their dedication to providing thorough coursework and emphasizing attention to detail has been invaluable in shaping my understanding of the subject matter. Their commitment to fostering a stimulating academic environment has greatly enriched my learning experience.

I am deeply appreciative of the time and effort generously given by my committee members, Drs. William "Ford" Sleeman, Elisabeth Weiss, and Preetam Ghosh. Their insightful feedback, constructive criticism, and scholarly guidance have significantly contributed to the refinement and quality of my research work. I am grateful for their expertise and the valuable perspectives they brought to the table.

I would like to express my gratitude to the Department of Radiation Oncology at Virginia Commonwealth University and the National Radiation Oncology Program office at the Department of Veterans Affairs. Their support and collaboration have been instrumental in providing me with the necessary resources, data, and opportunities to pursue my research. Their commitment to advancing cancer care and promoting innovative approaches in the field has been truly inspiring.

Finally, I would like to acknowledge the unwavering support and understanding of my family throughout this journey. Their encouragement and belief in my abilities have been a constant source of motivation. I am particularly grateful to my late father, Surrender Kapoor, my mother Rekha Kapoor, my sister, and brother-in-law Shagun & Rajiev Grover, and my wife Priyanka Kapoor for their love, understanding, and patience as I balanced the demands of work and school. I would also like to express my heartfelt appreciation to my two daughters, Alana, and Vaani, for their unwavering support and understanding throughout this journey. They have shown incredible patience and resilience as I dedicated long hours to my work and studies, often missing their school activities and weekend outings. Their understanding and love have been a constant source of inspiration for me. I am grateful for their unconditional support, and I am proud to have them as my daughters. Their presence in my life has brought immense joy and meaning, and I am truly grateful for their sacrifices and understanding during this demanding time.

Table of Contents

List of Tables

List of Figures

Statement of Need

Our Learning Health System (LHS) for radiation oncology is a comprehensive platform designed to transform the delivery of radiation therapy and improve patient outcomes. By integrating data from various sources, including clinical data from Electronic Health Records (EHR), dosimetry data from Treatment Planning Systems (TPS), and delivery data from Treatment Management Systems (TMS), and disease-specific clinical templates, the LHS creates a unified knowledgebase by standardizing and structuring data with graphs and ontologies. This integration allows for seamless data interoperability, making patient information findable, accessible, interoperable, and reusable (FAIR) for clinical and research purposes.

Key Features:

- **Comprehensive Patient View:** The LHS provides a real-time and comprehensive view of patient information, enabling radiation oncologists to access vital data in real time, leading to better decision-making and improved patient care. This vast amount of information within the LHS allows AI models to identify patterns and correlations that human analysis may overlook.
- **Standardized Data Representation**: Ontologies play a vital role in the LHS, providing standardized and structured representations of data elements and concepts within the domain. This enables radiation oncologists to query and mine the data using standardized terminologies and ontology-based equivalent concepts. This also ensures that machine learning models built on the LHS are consistent, well-defined, and follow a common vocabulary, making them highly portable and interoperable across different healthcare systems.
- **Promoting Collaboration:** The LHS fosters collaboration among clinicians, researchers, and healthcare institutions, encouraging knowledge sharing and continuous quality improvement. This collaborative approach accelerates advancements in radiation oncology and leads to better patient care.
- **Machine Learning Insights**: Ontology mapping provides "plug and play" functionality for machine learning models, enabling the LHS to analyze vast amounts of data to discover hidden patterns and valuable insights. This new knowledge can then be leveraged to improve radiation therapy techniques and treatment outcomes. Additionally, the LHS utilizes patient similarity techniques, allowing for patient cohort identification and identification of patients with similar attributes for specific research purposes.
- **Integration of New Parameters:** As medical information constantly evolves, the LHS accommodates new parameters and data elements through regular updates and expansion of the ontology. This flexibility ensures that the platform remains relevant and adaptable to the changing landscape of radiation oncology.

The Learning Health System for radiation oncology represents an advancement in how we practice healthcare. By unifying data, standardizing terminologies, and leveraging machine learning capabilities, the LHS empowers radiation oncologists with the tools they need to make informed decisions, collaborate with peers, and continuously improve patient care. With its patient-centered approach and commitment to AI technology, the LHS has the potential to be an impactful solution in radiation oncology.

Abstract

The proposed research aims to address the challenges faced by clinical data science researchers in radiation oncology accessing, integrating, and analyzing heterogeneous data from various sources. The research presents a scalable intelligent infrastructure, called the Health Information Gateway and Exchange (HINGE), which captures and structures data from multiple sources into a knowledge base with semantically interlinked entities. This infrastructure enables researchers to mine novel associations and gather relevant knowledge for personalized clinical outcomes.

The dissertation discusses the design framework and implementation of HINGE, which abstracts structured data from treatment planning systems, treatment management systems, and electronic health records. It utilizes disease-specific smart templates for capturing clinical information in a discrete manner. HINGE performs data extraction, aggregation, and quality and outcome assessment functions automatically, connecting seamlessly with local IT/medical infrastructure.

Furthermore, the research presents a knowledge graph-based approach to map radiotherapy data to an ontology-based data repository using FAIR (Findable, Accessible, Interoperable, Reusable) concepts. This approach ensures that the data is easily discoverable and accessible for clinical decision support systems. The dissertation explores the ETL (Extract, Transform, Load) process, data model frameworks, ontologies, and provides a real-world clinical use case for this data mapping.

To improve the efficiency of retrieving information from large clinical datasets, a search engine based on ontology-based keyword searching and synonym-based term matching tool was developed. The hierarchical nature of ontologies is leveraged to retrieve patient records based on parent and children classes. Additionally, patient similarity analysis is conducted using vector embedding models (Word2Vec, Doc2Vec, GloVe, and FastText) to identify similar patients based on text corpus creation methods. Results from the analysis using these models are presented.

The implementation of a learning health system for predicting radiation pneumonitis following stereotactic body radiotherapy is also discussed. 3D convolutional neural networks (CNNs) are utilized with radiographic and dosimetric datasets to predict the likelihood of radiation pneumonitis. DenseNet-121 and ResNet-50 models are employed for this study, along with integrated gradient techniques to identify salient regions within the input 3D image dataset. The predictive performance of the 3D CNN models is evaluated based on clinical outcomes.

Overall, the proposed Learning Health System provides a comprehensive solution for capturing, integrating, and analyzing heterogeneous data in a knowledge base. It offers researchers the ability to extract valuable insights and associations from diverse sources, ultimately leading to improved clinical outcomes. This work can serve as a model for implementing LHS in other medical specialties, advancing personalized and data-driven medicine.

1. Introduction

1.1 What is a Learning Health System?

Over the past 30 years, there has been an increasing interest in establishing Learning Organizations to tackle the complex challenges faced by society in business, social, and economic domains [1]. In the field of healthcare, the National Academy of Medicine has defined the concept of a Learning Health System (LHS), which aligns science, incentives, culture, and informatics to foster continuous innovation and integrate new knowledge discovery into evidence-based medical practice [2]. Learning Health Systems (LHS) are comprehensive frameworks that integrate research and healthcare delivery to continuously generate knowledge and improve patient outcomes. The reliance on randomized controlled clinical trials, which only capture a small percentage of patient samples (<5%) [3] in controlled environments, is now inadequate and may become irrelevant in the future due to their time-consuming nature, excessive costs, and limited generalizability. The Agency for Healthcare Research and Quality has been promoting the development of LHS as a vital strategy for healthcare organizations to achieve transformative improvements in healthcare quality and value. Recognizing the importance of continuous learning and improving patient care and addressing population health challenges, large-scale healthcare systems are now prioritizing the establishment of infrastructure that facilitates the collection of data from diverse sources such as electronic health records, treatment delivery records, imaging records, patient-generated data, and administrative and claims data. Analyzing this aggregated data can generate new insights and knowledge that can be effectively utilized to enhance patient care and improve outcomes within an LHS. These systems are anticipated to facilitate the synthesis of evidence in scenarios where traditional clinical research structures would be impractical, while also expediting its application in clinical settings. In addition, LHSs offer the advantage of minimal additional costs for data collection compared to resource-intensive conventional randomized controlled trials (RCTs) once a suitable electronic LHS infrastructure is in place [4]. This substantially reduces the obstacles to utilizing clinical data for sequential testing of iterative practice changes, thereby enabling the use of continuous improvement approaches to rapidly optimize treatments.

The challenge in designing the radiation oncology LHS infrastructure is aggregation of data that are both structured and unstructured from disparate data sources. It is extremely difficult to clean, parse, and collate RO data intelligibly, thus making many research and operational tasks that deal with the optimization of quality care, research-based analysis of radiation treatment, and diagnosis-based research and development of computer-aided diagnostic tools at the infrastructural level quite difficult. Several barriers such as limited interoperability amongst different vendor systems, narrative format based clinical data storage in electronic health records (EHRs), reluctance amongst healthcare providers to use form-based data entry due to substantial increase in the number of computer mouse clicks or structuring data entry based on the form requirements, quality of data found in clinical databases, and non-availability of an infrastructure to combine the various silos of databases. For example, radiation treatment planning systems, treatment management systems, EHRs and patient reported outcome systems are some of the major reasons why LHSs have not evolved in radiation oncology. While we are on the cusp of an artificial intelligence (AI) revolution in biomedicine with the fast-growing development of advanced machine learning methods that can analyze complex datasets, there is an urgent need for a scalable intelligent infrastructure that can support these methods. These infrastructures must provide an integration of data

from multiple clinical data sources and semantically interlink clinical concepts for seamless utilization in machine learning models and cohort identification for continuous quality improvement.

1.2 Our Approach of defining the LHS framework for Radiation Oncology

Manual abstraction, collation, curation, and subsequent analysis of healthcare data for quality and outcome assessment of patient treatments are onerous, expensive, and impractical. Advances in computer storage, computing power, and the ability to electronically mine data from disparate sources (e.g., demographics, genetics, imaging, treatment, clinical decisions, and outcomes) have the power to enable big data research in medicine. Developing an ecosystem to make routine clinical data "Artificial Intelligence (AI) ready" will allow us to generate new knowledge from large scale analysis of historical patients' treatments and outcomes to improve care for the future patients. A fundamental barrier in reaching this vision is the lack of strategic development of IT infrastructures that facilitate data aggregation out of clinical systems via federated or centralized infrastructures for data access. Another barrier in reaching our goals is to organize and standardize our knowledge and information in a way that is computer understandable with standardized vocabularies, established taxonomies and interoperability standards. The aim of this research is to build essential components of a comprehensive radiotherapy Learning Health System (LHS). We designed and built an IT infrastructure; Health Information Gateway and Exchange (HINGE) software platform that collates information from all RT data sources, extract/transform/load these data into an internationally standardized semantic interoperable data model such as National Cancer Institute (NCI) Thesaurus and Operational Ontology for Oncology (O3). In order to improve the efficiency of retrieving information from large clinical datasets, we have also developed a novel ontology-based keywork searching and synonym & hyponym-based term matching method. The architecture of proposed LHS infrastructure is presented in Figure 1. This framework is designed to collect clinical data from electronic medical record systems using the HINGE platform. Additionally, delivery data from RO-treatment management systems is obtained through FHIR-based interfaces, and RO-treatment planning systems are accessed via DICOM data export. The abstracted data are consolidated into a common relational database, where data mapping based on ontology and standard taxonomy definitions takes place. Subsequently, the mapped data are transformed into RDF (Resource Description Framework) triple format and uploaded into an RDF-based graph database. Our approach ensures semantic interoperability of the datasets and establishes a framework for universally applicable methods such as data mining, keyword search, semantic search, and ontology-based query expansion. By adopting open semantic ontology-based formats, our system facilitates the availability and interoperability of radiation oncology datasets, thus supporting the execution of large-scale scientific studies.

The core objective of our system is to demonstrate the effective integration of semantic-based data and knowledge from multiple sources using the ontology developed through domain expertise. In this work, we merged concepts from the Radiation Oncology Ontology, NCI Thesaurus, ICD-10, and Units Ontology to construct the ontology. By coordinating the development and integration of tools, our overarching goal is to minimize human error and variability while enabling automated capture of contextual metadata. This assists with tasks such as cohort generation, encumbrance monitoring, data quality assessment, and ontological mapping.

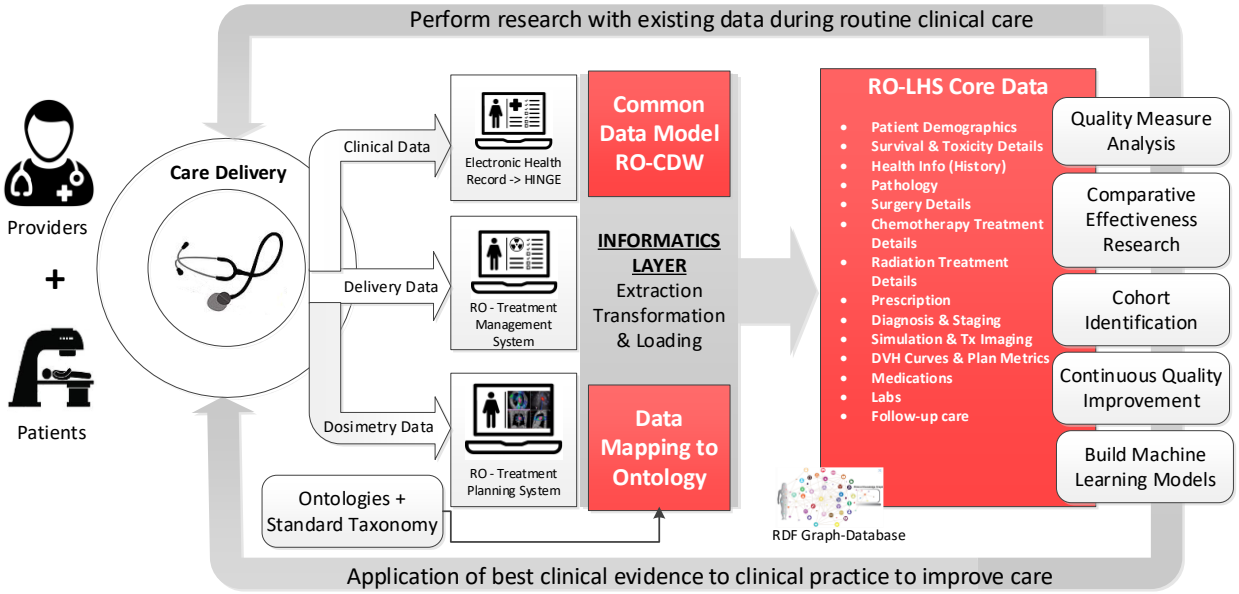


Figure 1: Overall architecture of our Radiation Oncology Learning Health System infrastructure. Here we have the data captured at care delivery from the three data sources and the informatics layer to extract, transform and load this data based on standard taxonomy and ontologies into the RO-LHS core data repository. This repository is the RDF (Resource Description Framework) graph database that stores the data with established definitions and relationships based on the standard terminology and ontology. The data listed in the RO-LHS is made available for subsequent applications such as quality measure analysis, cohort identification, continuous quality improvement and building machine learning models that can be applied back to the care delivery to improve care thus completing the loop for an effective learning health system.

An overview of learning health systems (LHS) and their relevance in radiation oncology, review of existing literature on LHS infrastructure and identification of gaps in the literature and research opportunities is provided in section 2. Section 3 poses four research based specific aims of this work which are addressed in sections 4-7. In section 8, we summarize the main contributions of this work and future research directions are discussed in section 9.

In summary this dissertation is focused on four main topics as follows:

1. Develop a framework for passive data abstraction, extracting clinical, dosimetry, treatment delivery, and outcome data from various sources in radiation oncology. This framework serves as a vital component of the Learning Health System (LHS), enabling comprehensive data collection.
2. Design an Extract, Transform, and Load (ETL) framework to standardize and transform clinical, dosimetry, treatment delivery, and outcome datasets into internationally recognized and semantically interoperable data models and ontologies. This ensures consistent and compatible representation of data for improved collaboration and analysis.
3. Establish a framework for leveraging ontology-based data definitions to create universally applicable tools for keyword search, semantic search, and patient similarity search. These tools enhance data retrieval and analysis by utilizing common data definitions based on ontologies.

4. Application of a framework established to investigate the use of 3D Deep Convolution Neural Networks for radiation pneumonitis prediction following Stereotactic Body Radiotherapy (SBRT).

References:

1. Senge PM. The Fifth Discipline: The Art and Practice of the Learning Organization. New York: Doubleday/Currency; 2006.
2. Olsen L, Aisner D, McGinnis JM. The Learning Healthcare System: Workshop Summary. Washington, DC: Institute of Medicine Roundtable on Evidence-Based Medicine, National Academies Press/National Academy of Sciences; 2007.
3. Nationally representative estimates of the participation of cancer patients in clinical research studies according to the commission on cancer.; Joseph M. Unger and Mark Fleury; Journal of Clinical Oncology 2021 39:28_suppl, 74-74
4. Budrionis A, Bellika JG. The learning healthcare system: where are we now? A systematic review. J Biomed Inform. 2016; 64:87-92.

2. Background

As per the estimations of the World Health Organization, around 10 million individuals succumbed to cancer in the year 2020 [1]. For over 50% of cancer patients, a type of radiation therapy is recommended [2]. With a growing array of treatment choices (for example, beam shapes, radiation modalities, and energy ranges) and the presence of diverse patient groups, it is becoming increasingly challenging to determine the most suitable treatment for each patient. Typically, the assessment of the superiority of one treatment option over another is conducted through clinical trials under the same conditions. These trials evaluate two treatment options for variations in effectiveness and outcome. Most healthcare systems require the highest level of evidence (i.e., type I) through a randomized controlled trial (RCT) before reimbursing a treatment.

Due to advancements in our understanding of cancer tumor phenotypes at a molecular level and the emergence of better medical imaging and diagnostic testing technologies, we have an unprecedented amount of information available to us. However, the challenge now is how to integrate this vast amount of information to determine the best individualized treatment plan for each patient based on their specific needs. Even though patients may have the same tumor stage and risk factors, their response to systemic treatment and radiotherapy can vary widely, as can the toxicity of these treatments on the surrounding healthy organs.

Although higher biologically effective dose levels are generally required to achieve local tumor control rates, it is still common practice in most radiotherapy departments to prescribe the same dose to all patients. However, the emergence of advanced technologies means that we can achieve higher local control rates with hypo-fractionated treatment regimens. It is important to note that achieving this balance between toxicity and tumor control is dependent on the individual treatment plans generated for each patient. Thus, the challenge lies in how to integrate all of the available information to select the most appropriate treatment plan for each patient, ensuring that they receive the optimal dose of treatment to control their tumor while minimizing the potential for toxicity. Some of the factors that underline the need for individualization of treatments are as follows:

- Tumors and patients are less homogeneous than previously assumed, meaning the same treatment can have different toxicities, survival outcomes in patients who have the same type of tumor. For example, different molecular and gene signatures of breast cancer types have very different outcomes. Therefore, it is crucial to consider individual patient characteristics and tumor features when designing a treatment plan [3, 4].

- There has been an increase in the number of treatment options. For example, early-stage prostate cancer can now be treated with conventional external beam RT fractionation, radical prostatectomy, stereotactic radiotherapy, LDR or HDR brachytherapy, high-intensity focused ultrasound, hormone therapy, combination therapies and so on. Similarly, targeted therapies have been rapidly growing in numbers, making it impossible to perform classical randomized clinical trials to compare all new treatment options with the "gold standard" due to the current speed of newer innovations.

- Translating the results of clinical trials to the general patient population and environment is not straightforward. This is because of the higher quality of care in clinical trials and the known selection bias, as trials reach no more than 3% of cancer patients. As a result, it is crucial to evaluate the quality of evidence and match the treated patient characteristics to evidence from the literature [5].

- It has become increasingly difficult for clinicians to find the right evidence for matching the treated patient characteristics to evidence from the literature and evaluate the quality of that evidence. There has been a rapid increase in papers published and a small minority of trial reports are being analyzed in up-to-date systematic reviews which lead to interpretations of the selection data biases and reported results. Therefore, the individualization of treatments has become essential in order to provide the best possible care to each patient based on their specific characteristics and needs [6].

Given the increasing number of treatment options and less homogeneous patient groups, it has become more urgent than ever to make treatment decisions based on robust knowledge and evidence-based medicine. In 2007, the US Institute of Medicine urged healthcare leaders to transform their practices into learning healthcare systems (LHSs), where research and care are integrated and healthcare activities are continuously studied, learned from, and improved [7]. Precision medicine offers promising developments as therapies can be more accurately tailored to specific patient characteristics if data from routine care is systematically used to generate clinical evidence [8].

2.1 The 5 Vs of Big Data

As we endeavor to generate data-driven insights, it is widely recognized that the confidence in our conclusions is highly dependent on five Vs of underlying data: volume, variety, velocity, value, and veracity. Despite being in an era when electronic health record systems are an integral part of patient care continuum, the gaps in standardization, infrastructure, and technical skills are barriers to systematic aggregation and analysis of data at scale. The challenges associated with the five Vs that impact the effective functioning and utilization of an LHS infrastructure is listed below:

- Volume: The term "volume" refers to the magnitude or scale of data, encompassing its quantity ranging from terabytes to zettabytes. The sheer volume of data generated in healthcare poses a challenge for LHS implementation. The increasing amount of data from electronic health records (EHRs), medical imaging, wearables, and genomics can overwhelm the infrastructure's storage and processing capabilities. Managing and analyzing large volumes of data in real-time requires scalable and robust infrastructure.

- Variety: "Variety" indicates the diverse types or formats of data, including structured, unstructured, or semi-structured data. Healthcare data comes in diverse formats, including structured, unstructured, and semi-structured data. LHS infrastructure needs to handle this variety of data sources and integrate them seamlessly. Standardizing and normalizing data from different sources with varying formats and coding systems can be challenging. Ensuring interoperability and compatibility across systems and applications is crucial to make use of the full range of available data. In the field of Radiation Oncology, acquiring structured data from dosimetry datasets is relatively straightforward. However, the primary challenge lies in gathering structured data from clinical data and assessments.

- Velocity: "Velocity" pertains to the speed or rate at which data is generated, processed, and analyzed, highlighting the flow and real-time nature of data. Healthcare data is generated and updated at a rapid pace, requiring real-time processing and analysis. LHS infrastructure should be able to handle the velocity of incoming data streams and enable timely data capture, analysis, and decision-making. It

involves establishing efficient data pipelines, data streaming, and real-time analytics capabilities to keep up with the continuous flow of data.

- Veracity: "Veracity" signifies the quality and reliability of data, encompassing aspects such as accuracy, relevance, predictive value, and the meaningfulness of the information contained within the data. The veracity of healthcare data refers to its accuracy, reliability, and consistency. Ensuring data quality is crucial for making reliable inferences and informed decisions. LHS infrastructure needs to address challenges related to data quality, including data integrity, completeness, and data entry errors. It involves implementing data validation processes, data cleansing techniques, and quality assurance mechanisms.

- Value: "Value" reflects the usefulness and significance of data, specifically in terms of its ability to inform decision-making and contribute to the development of strategies for various purposes. Extracting value from healthcare data is a key goal of LHS implementation. However, realizing the value of data requires overcoming various challenges. LHS infrastructure should enable efficient data analysis and mining techniques to derive meaningful insights and knowledge from the data.

2.2 Role of Knowledge Engineering in LHS

As the digitization and availability of medical data increases due to recent developments and investments in healthcare information technology, such as electronic medical records (EMR), the field of research known as Knowledge Engineering has emerged. This field encompasses data mining, machine learning, and Big Data storage techniques to craft knowledge from large amounts of data. These are difficult tasks, if not impossible, to achieve through human interpretation alone. However, by using historical and daily clinical data, new statistical models can be trained to predict treatment outcomes, such as survival, toxicity, and quality of life [9, 10]. These models can be created by extracting, fitting, and modeling data to find patterns and form algorithms that can predict a patient's likelihood of experiencing severe toxicity from a suggested treatment. The creation of such models requires large amounts of data, and proper validation of these models requires even more datasets from independent and preferably external sources and these datasets should be collected in a data lake [11]. Therefore, it is essential to adequately manage research or trial data as a standard component of EMR used by providers in routine clinical practice.

To effectively reuse clinical data from routine care, it can be classified into three categories: (1) structured and coded data elements that are required to be coded according to standardized terminologies such as diagnosis codes (ICD) and billing codes (CPT), (2) structured but uncoded data elements, such as treatment dose delivery summary and quality of life data, and (3) unstructured free text used for flexible documentation such as clinical notes, findings, and assessments. By utilizing these different types of clinical data, it is possible to generate a more complete and accurate understanding of patient characteristics and outcomes, which can inform and improve future treatments.

2.3 Overview of existing LHS implementations in Radiation Oncology

The American Society of Clinical Oncology (ASCO) has embraced the goal to derive real-world evidence from daily practice and has subsequently launched its own LHS: CancerLinQ [12]. In radiation oncology, the EUROCAT [13], NROR [14], M-ROAR [15], MROQC [16] have attempted to build LHSs to collect and assess practice patterns, perform outcome analyses, and evaluate dosimetry related information. MD Anderson Cancer Center has implemented a system-wide electronic data capture system that records

patients' treatment information [17]. Johns Hopkins Medical Center has launched the Oncospace program that captures RT data containing anatomy, dose distributions and outcomes data in an analytical database [18]. The Mayo Clinic's Department of Radiation Oncology in Florida has linked its radiation oncology information system with Mayo Clinic's internal claims data warehouse along with Mayo's tumor registry which allows for large-scale studies [19]. Whitaker et al. at Mayo Clinic at Rochester developed a patient-reported outcome (PRO) collection and management system to implement a large-scale aggregation of patients' treatment data. The basic idea with all the above-mentioned platforms is the reuse of historical data from routine clinical practice for decisions concerning new patients or to test new hypotheses. This has several obvious advantages, such as the large number of readily available patients and less selection bias compared to clinical trials. A majority of LHS used in radiation oncology are deployed with very specific objectives. An ideal LHS infrastructure should be able to collate comprehensive radiotherapy episodic data that include DICOM-RT data from Treatment Planning System (TPS), treatment data from Treatment Management Systems (TMS), and clinical data from EHR in a single minable data repository where relationships amongst data elements are preserved. The clinical data from the EHR is the most challenging aspect of an ideal LHS infrastructure. It is not surprising that most of the abovementioned systems only deal with data generated within the TPS & TMS platforms. Hence these systems are dealing with limiting datasets that include mostly dosimetry and treatment planning data elements.

For any of these abovementioned systems to be successful, real-world data of acceptable quality and diversity is necessary. This is only possible if data are shared across multiple institutions. Such data sharing is hampered by the lack of standardization of nomenclature used to record the data, differences in data recording patterns where most clinical assessments are recorded in free text based clinical notes, non-existence of clinical tools that aid in capturing discreate data within the clinical workflow and most importantly concerns related to privacy and information security.

2.4 Identification of Gaps in the Literature and Research Opportunities

While there is a growing body of literature on LHS infrastructure, several gaps and research opportunities exist:

- Interoperability Challenges: Further research is needed to address the ongoing challenges related to interoperability and data integration across diverse information systems in radiation oncology. Standardization efforts and technological solutions for seamless data exchange need to be explored.
- Data Quality and Completeness: Studies focusing on improving the quality and completeness of data within LHS are necessary. Enhancing data capture mechanisms, implementing data validation processes, and ensuring comprehensive data collection can contribute to the reliability and usefulness of LHS in radiation oncology. The process of cleaning, parsing, and collating the data into an intelligible format is exceptionally difficult, thus posing obstacles to research and operational tasks
- Advanced Analytics and AI: Exploration of advanced analytics techniques, including fast and convenient patient cohort identification and retrieval based on patient similarity metrics, artificial intelligence (AI) and machine learning, within LHS can offer valuable insights. Developing predictive models for treatment outcomes, automated decision support systems, and risk stratification algorithms can aid radiation oncologists in making informed decisions.

- Implementation Strategies and Barriers: There are practical challenges and barriers associated with implementing LHS in radiation oncology settings including utilizing common patient identifiers to link data from multiple source systems, utilizing common data definitions and semantics, identifying effective implementation strategies, understanding stakeholder perspectives, and evaluating the impact of LHS implementation on workflow and organizational culture.
- Integration of Patient-Centered Outcomes with the treatment tracking process: Further investigation is needed on the impact of LHS on patient-centered outcomes and the engagement of patients in the learning process. Research should focus on assessing patient satisfaction, treatment adherence, shared decision-making, and the integration of patient-reported outcomes within LHS.
- Long-Term Follow-up and Survivorship: Research opportunities exist in the development and integration of LHS components that address long-term follow-up, survivorship care, and monitoring of early and late treatment effects in radiation oncology. Evaluating the effectiveness of LHS in enhancing long-term outcomes and quality of life for cancer survivors is important.

By addressing these gaps and research opportunities, the literature on LHS infrastructure in radiation oncology can be further expanded, thus providing valuable insights, and advancing the field.

2.5 Dealing with Privacy and HIPPA related issues

Privacy and HIPAA (Health Insurance Portability and Accountability Act) issues are critical considerations when implementing a learning health system infrastructure. As patient data plays a central role in such systems, maintaining the privacy and confidentiality of this sensitive information is of utmost importance. HIPAA regulations are designed to protect individuals' health information and ensure its secure handling, storage, and sharing. Therefore, any learning health system infrastructure must comply with HIPAA guidelines to safeguard patient privacy.

One of the key challenges is balancing the need for data sharing and respecting the privacy rights of patients. The infrastructure should employ robust security measures to prevent unauthorized access or breaches of patient data. This includes implementing encryption protocols, access controls, and secure storage practices. Additionally, stringent user authentication mechanisms should be in place to ensure that only authorized personnel can access and utilize patient data.

Another important consideration is the de-identification of patient data. To protect privacy, personally identifiable information (PII) must be removed or anonymized from the data before it is used for research or analysis. De-identification techniques, such as removing direct identifiers or applying data masking, can help mitigate privacy risks.

Furthermore, data governance policies and procedures should be established to ensure compliance with HIPAA regulations. This involves defining clear guidelines for data handling, access, sharing, and retention. Regular audits and monitoring of system activities can help identify any potential privacy breaches and ensure adherence to privacy requirements.

Overall, the design and development of a LHS infrastructure must prioritize privacy and HIPAA compliance. By incorporating robust security measures, de-identification techniques, data governance policies, and user education, healthcare organizations can establish a framework that upholds patient privacy while enabling the benefits of a data-driven learning health system.

2.6 Attributes of an Ideal Learning Health System Infrastructure

An ideal learning health system (LHS) infrastructure for radiation oncology should have the following attributes:

- Data Integration: The LHS infrastructure should enable seamless integration of data from various sources, including electronic health records (EHRs), treatment planning systems, imaging systems, patient-generated data, and administrative and claims data. This comprehensive data integration allows for a holistic view of patient information and facilitates comprehensive analysis.
- Interoperability & Interconnectivity: The LHS infrastructure should ensure interoperability between different healthcare information systems to enable the secure exchange and sharing of data across multiple platforms. This interoperability enables seamless data flow and collaboration between different stakeholders within the radiation oncology ecosystem.
- Data Quality and Standardization: The infrastructure should include mechanisms to ensure data quality and standardization. This involves implementing data validation processes, adhering to standardized data formats, use of standard ontologies and coding systems, and ensuring data accuracy and completeness. High-quality and standardized data are essential for reliable analysis and generating meaningful insights.
- Advanced Analytics and Decision Support: The LHS infrastructure should incorporate advanced analytics capabilities, including artificial intelligence (AI) and machine learning, to analyze large-scale health data and generate actionable insights. This includes developing predictive models, risk stratification algorithms, and decision support tools to aid radiation oncologists in making evidence-based treatment decisions.
- Security and Privacy: Robust security measures should be implemented to protect patient data and ensure privacy compliance. This involves adopting encryption protocols, access controls, and data anonymization techniques to safeguard patient information. Strict adherence to data protection regulations is essential to maintain trust and confidentiality within the LHS infrastructure.
- Real-time Data Capture: The infrastructure should support real-time data capture (access to data within a few minutes when data is generated) to enable timely analysis and decision-making. This includes leveraging technologies such as automatic data extraction, natural language processing, and interoperable data interfaces to capture and update patient data efficiently.
- Stakeholder Engagement: The LHS infrastructure should actively involve all relevant stakeholders, including radiation oncologists, medical physicists, therapists, and researchers. Collaboration and engagement among stakeholders foster a culture of continuous learning, improvement, and shared decision-making within the radiation oncology community.
- Long-term Follow-up and Surveillance: The infrastructure should support long-term follow-up and surveillance of patients to monitor treatment outcomes, late effects, and disease recurrence. This involves integrating mechanisms for tracking and evaluating patient outcomes over an extended period, ensuring continuity of care and proactive management of long-term effects.
- Scalability and Flexibility: The LHS infrastructure should be designed to be scalable and adaptable to accommodate evolving technological advancements and changing healthcare needs. It should have the capability to handle large volumes of data, accommodate future data sources, and incorporate new analytical methodologies as they emerge.

By embodying these attributes, an ideal LHS infrastructure for radiation oncology can effectively support continuous learning, research collaboration, evidence-based decision-making, and improved patient outcomes.

2.7 Efforts for Data Standardization and utilizing Ontologies in Radiation Oncology

With the purpose of standardizing the nomenclature and data elements, ASTRO convened a task force to decide on the minimum data elements (MDE) in radiation oncology which should be standardized and interfaced to the general EHR for continuity of care. [20] In a parallel effort, ASTRO convened an overlapping working group to develop multidisciplinary consensus recommendations on a synoptic radiation treatment summary for continuity of medical care [21]. These two efforts set a precedent in radiation oncology domain of multi-stakeholder participation in standards formulation and dissemination, albeit with limited data elements and scope. The CodeX and minimum Common Oncology Data Elements (mCode) projects provide a platform for integrating domain specific standardizations. Run jointly by Mitre Corporation and HL7, CodeX (Common Oncology Data Elements eXtensions) is a multi-stakeholder, multi-disciplinary HL7 FHIR Accelerator, focuses on developing interoperable data exchange using HL7-FHIR. The minimum Common Oncology Data mCODE provides a subset of high priority cancer care elements around which extensive collaboration with vendors, SNOMED and HL7 is organized for integration into commercial systems. The mCode and CodeX projects provide a platform for integrating domain specific standardizations. Our work will support the mCode and CodeX effort for Radiation Therapy Treatment Data for cancer (RTTD) pilot and use the work products from this effort in designing our infrastructure solution.

Another important aspect to consider before sharing dataset is the use of well-defined common ontologies (e.g., NCI Thesaurus, ROO, SNOMED, OOO) to code the data to enable semantic interoperability. These common ontology terms will potentially serve as a common interface to the data at each institutional site, enabling a common approach to information retrieval thereby creating a common semantic definition-based dataset. This helps ensure that the data is semantically interoperable, meaning that it can be understood and used by different systems in a consistent and standardized way. Some of the most commonly referenced terminologies and ontologies are as follows:

- Foundational Model of Anatomy (FMA): A domain ontology that represents an explicit declarative knowledge about human anatomy. (https://bioportal.bioontology.org/ontologies/FMA)
- NCI Thesaurus: A vocabulary for clinical care, translational and basic research, and public information and administrative activities. (https://bioportal.bioontology.org/ontologies/NCIT)
- Common Terminology Criteria for Adverse Events (CTCAE): The standard classification and severity grading scale for adverse events in cancer therapy clinical trials and other oncology settings (https://bioportal.bioontology.org/ontologies/CTCAE)
- Radiomics Ontology: Broad coverage of not only radiomics features, but also every entity (e.g., software properties, filter properties, features extraction parameters) involved in radiomics computation. (https://bioportal.bioontology.org/ontologies/RO)
- Semantic DICOM Ontology: An ontology for DICOM (https://bioportal.bioontology.org/ontologies/SEDI)
- Radiation Oncology Ontology: The Radiation Oncology Ontology aims to cover the radiation oncology domain with a strong focus on re-using existing ontologies. (Radiation Oncology Ontology - Summary | NCBO BioPortal (bioontology.org)

The foundation of any ontology is a taxonomic backbone which provides a tree or graph-like structure for the classification of entities. An example of the SNOMED ontology is shown in Figure 2. The fundamental advantage to transforming clinical and dosimetry data into standard ontologies is that it enables the transfer, reuse, and sharing of the patient data and seamless integration with other data sources.

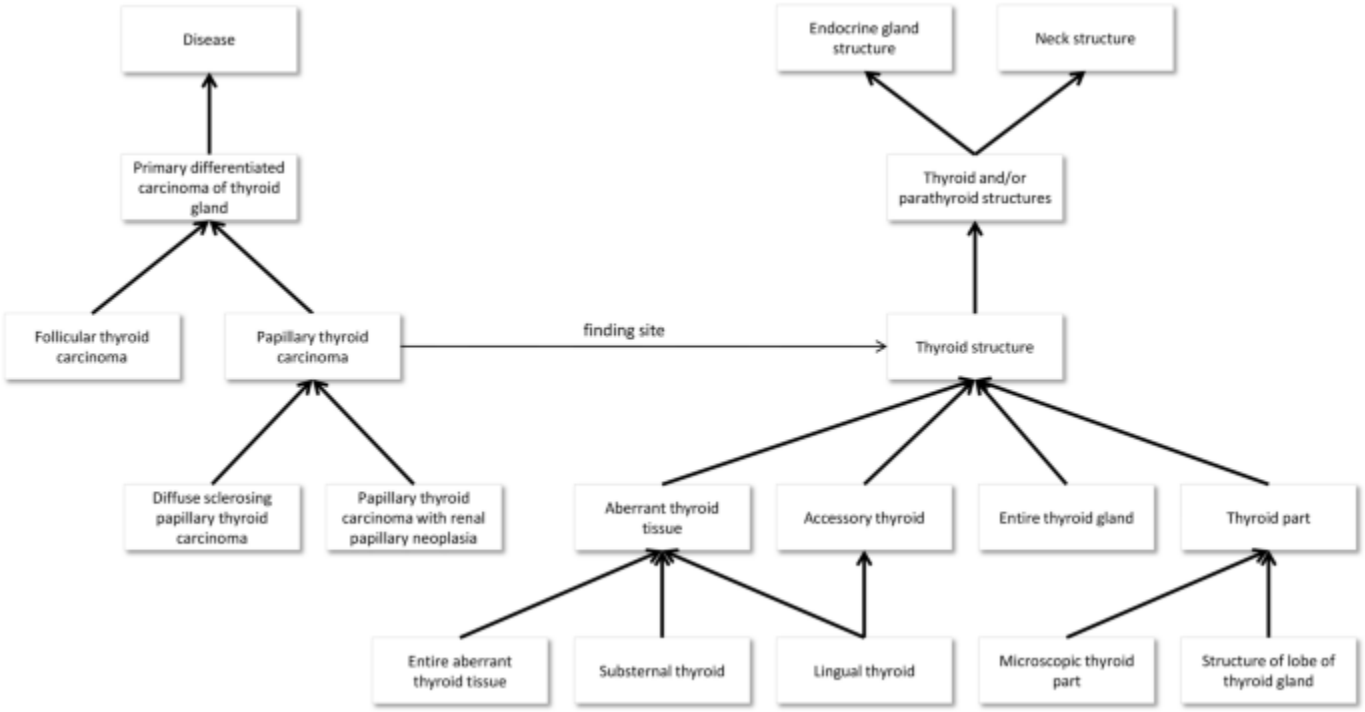


Figure 2: The SNOMED CT Ontology

Figure 2: The SNOMED CT ontology is presented in the form of a directed acyclic graph [22]. Within this graph, the nodes symbolize the concepts within the ontology, such as "papillary thyroid cancer." The relationships denoting the "is-a" connections between a more specific concept and with more general connections are represented by prominent arrows. Additionally, the graph illustrates the finding site attribute, which establishes a link between a disease or condition and an anatomical structure.

The main advantages of using an ontology-based graph database as opposed to traditional relational databases is that the traditional relational databases are designed to cater to a particular application and its software requirements, and data stored is not conducive for clinical research. These databases are not suited to gather data from multiple data sources when the structure of data, schema, data types are unknown. On the other hand, ontology-based graph databases are schema free and designed to store large amount of data with defined interrelationships and the definitions based on universally defined concepts that enable any clinical researcher to query the data without understanding the inherent data structure and schema used to store data in the database. The ontology structure makes querying the data more intuitive for researchers and clinicians because it matches the domain knowledge logical structure. Based on our literature review we found that the use of ontologies and mapping radiation oncology clinical, dosimetry and treatment data in ontology-based definitions have mostly been discussed as conceptual ideas and strategies in the literature [23,24]. There are no actual implementations of this work reported in the public domain.

2.8 Importance of the purpose for data collection

Data collected with a specific purpose, such as optimizing quality care, research-based analysis of radiation treatment, and diagnosis-based research and development of computer-aided diagnostic tools, can help identify patterns and trends that can inform clinical decision-making and improve patient outcomes. Furthermore, having a specific purpose for data collection allows for the creation of customized reports and analytics that can provide valuable insights into the health system's operations, including identifying areas for improvement and facilitating quality improvement initiatives. Ultimately, having a clear purpose for data collection is critical for the success of any radiation oncology LHS infrastructure, as it enables the organization to better understand its patients, improve clinical outcomes, and enhance operational efficiencies.

For example, data collected for the purpose of optimizing quality care can be used to analyze adherence to treatment protocols, track the quality of care rendered to patients, and help in adequate resource allocation. By examining these data, radiation oncologists, hospital administration can identify clinical and process related factors that contribute to successful treatments and develop evidence-based guidelines to improve the quality and effectiveness of care provided. Additionally, the data can help identify variations in practice and outcomes, allowing for the identification of best practices and areas for improvement. Moreover, having a specific purpose for data collection allows for the creation of customized reports and analytics that provide valuable insights into the operations of the radiation oncology health system. By analyzing the collected data, organizations can generate reports and analytics tailored to their specific needs, enabling them to monitor performance, track key performance metrics, identify areas for improvement, and drive quality improvement initiatives. This targeted approach to data analyses and reporting enhances operational efficiencies, streamlines workflows, and facilitates informed decision-making at both the individual patient level and at the broader system level.

The Veterans Health Administration (VHA) has a requirement to gather comprehensive discrete data from all radiotherapy systems that include TPS, RT-EMR, and EHR to monitor the quality of radiotherapy delivered to Veterans, determine practice variations, and identification of the care gaps in the VHA. The work in this dissertation is to leverage these data requirements for the VHA system and build a software system to evaluate quality of care in the VHA system and create a learning health system infrastructure that incorporates real-time data that enables continuous improvement of care delivery. This should lead to improved patient outcomes, increased efficiency, and reduced healthcare costs. The project will serve as a model for the implementation of learning health systems in other medical specialties, contributing to the advancement of personalized and data-driven medicine.

References:

1. Ferlay J, Colombet M, Soerjomataram I, Parkin DM, Piñeros M, Znaor A, Bray F. Cancer statistics for the year 2020: An overview. Int J Cancer. 2021 Apr 5. doi: 10.1002/ijc.33588. Epub ahead of print. PMID: 33818764.
2. Delaney G, Jacob S, Featherstone C, Barton M. The role of radiotherapy in cancer treatment. Cancer 2005;104(6):1129–1137. doi:10.1002/cncr.21324

3. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012 Oct 4;490(7418):61-70. doi: 10.1038/nature11412. Epub 2012 Sep 23. PMID: 23000897; PMCID: PMC3465532
4. Starmans MH, Lieuwes NG, Span PN, Haider S, Dubois L, Nguyen F, van Laarhoven HW, Sweep FC, Wouters BG, Boutros PC, Lambin P. Independent and functional validation of a multi-tumour-type proliferation signature. Br J Cancer. 2012 Jul 24;107(3):508-15. doi: 10.1038/bjc.2012.269. Epub 2012 Jun 21. PMID: 22722312; PMCID: PMC3405210.
5. Movsas B, Moughan J, Owen J, Coia LR, Zelefsky MJ, Hanks G, Wilson JF. Who enrolls onto clinical oncology trials? A radiation Patterns Of Care Study analysis. Int J Radiat Oncol Biol Phys. 2007 Jul 15;68(4):1145-50. doi: 10.1016/j.ijrobp.2007.01.051. Epub 2007 Apr 9. PMID: 17418963.
6. Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? PLoS Med. 2010 Sep 21;7(9):e1000326. doi: 10.1371/journal.pmed.1000326. PMID: 20877712; PMCID: PMC2943439.
7. Olsen L, Aisner D, McGinnis JM, Institute of Medicine (US). Roundtable on Evidence-Based Medicine . *The learning healthcare system: workshop summary*. Washington, DC: The National Academies Press; 2007:XIII.
8. Visvanathan K, Levit LA, Raghavan D, et al. Untapped potential of observational research to inform clinical decision making: American Society of Clinical Oncology research statement. *J Clin Oncol*. 2017;35:1845-1854.
9. El Naqa I, Bradley J, Blanco AI, Lindsay PE, Vicic M, Hope A, et al. Multivariable modeling of radiotherapy outcomes, including dose–volume and clinical factors. Int J Radiat Oncol • Biol • Phys 2006;64(4):1275–1286. doi:10.1016/j.ijrobp.2005.11.022
10. Carvalho S, Leijenaar RTH, Velazquez ER, Oberije C, Parmar C, van Elmpt W, et al. Prognostic value of metabolic metrics extracted from baseline positron emission tomography images in non-small cell lung cancer. Acta Oncol 2013;52(7):1398–1404. doi:10.3109/0284186X.2013.812795
11. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and ElaborationThe TRIPOD Statement: Explanation and Elaboration. Ann Intern Med 2015;162(1):W1–W73. doi:10.7326/M14-0698
12. Sledge JG, Miller RS, Hauser R. CancerLinQ and the future of cancer care. In American Society of Clinical Oncology educational book. American Society of Clinical Oncology. Annual Meeting (pp. 430-434)
13. Lambin P, Zindler J, Vanneste BGL, et al. Decision support systems for personalized and participative radiation oncology. Advanced Drug Delivery Reviews. 2017;109:131-153.
14. Palta JR, Efstathiou JA, Bekelman JE, Mutic S, Bogardus CR, McNutt TR, Gabriel PE, Lawton CA, Zietman AL, Rose CM. Developing a national radiation oncology registry: From acorns to oaks. Pract Radiat Oncol. 2012 Jan-Mar;2(1):10-7. doi: 10.1016/j.prro.2011.06.002. Epub 2011 Jul 14. PMID: 24674031.
15. Mayo CS, Kessler ML, Eisbruch A, et al. The big data effort in radiation oncology: Data mining or data farming? Adv Radiat Oncol. 2016;1(4):260-271. doi:10.1016/j.adro.2016.10.001
16. Moran JM, Feng M, Benedetti LA, et al. Development of a model web-based system to support a statewide quality consortium in radiation oncology. Pract Radiat Oncol. 2017;7(3):e205-e213. doi:10.1016/j.prro.2016.10.002

17. Pasalic D, Reddy JP, Edwards T, Pan HY, Smith BD. Implementing an Electronic Data Capture System to Improve Clinical Workflow in a Large Academic Radiation Oncology Practice. JCO Clin Cancer Informatics. 2018;(2):1-12. doi:10.1200/CCI.18.00034
18. McNutt TR, Evans K, Wu B, et al. Oncospace: All Patients on Trial for Analysis of Outcomes, Toxicities, and IMRT Plan Quality. Int J Radiat Oncol • Biol • Phys. 2010;78(3):S486. doi:10.1016/j.ijrobp.2010.07.1139
19. Waddle MR, Kaleem T, Niazi SK, et al. Cost of Acute and Follow-Up Care in Patients with Pre-Existing Psychiatric Diagnoses Undergoing Radiation Therapy. Int J Radiat Oncol. 2017;99(5):1321. doi:10.1016/j.ijrobp.2017.09.023
20. Hayman JA, Dekker A, Feng M, Keole SR, McNutt TR, Machtay M, Martin NE, Mayo CS, Pawlicki T, Smith BD, Kudner R, Dawes S, Yu JB. Minimum Data Elements for Radiation Oncology: An American Society for Radiation Oncology Consensus Paper. Pract Radiat Oncol. 2019 Nov;9(6):395-401. doi: 10.1016/j.prro.2019.07.017. Epub 2019 Aug 21. PMID: 31445187.
21. Christodouleas JP, Anderson N, Gabriel P, Greene R, Hahn C, Kessler S, Mayo CS, McNutt T, Shulman LN, Smith BD, West J, Williamson T. A Multidisciplinary Consensus Recommendation on a Synoptic Radiation Treatment Summary: A Commission on Cancer Workgroup Report. Pract Radiat Oncol. 2020 Nov-Dec;10(6):389-401. doi: 10.1016/j.prro.2020.01.002. Epub 2020 Jan 24. PMID: 31988040.
22. Filice R, Kahn Charles. Biomedical Ontologies to Guide AI Development in Radiology, Journal of Digital Imaging 34 (1) Nov 2021, 10.1007/s10278-021-00527-1
23. Lambin P, Roelofs E, Reymen B, Velazquez ER, Buijsen J, Zegers CM, Carvalho S, Leijenaar RT, Nalbantov G, Oberije C, Scott Marshall M, Hoebers F, Troost EG, van Stiphout RG, van Elmpt W, van der Weijden T, Boersma L, Valentini V, Dekker A. 'Rapid Learning health care in oncology' - an approach towards decision support systems enabling customised radiotherapy'. Radiother Oncol. 2013 Oct;109(1):159-64. doi: 10.1016/j.radonc.2013.07.007. Epub 2013 Aug 28. PMID: 23993399.
24. Price G, Mackay R, Aznar M, McWilliam A, Johnson-Hart C, van Herk M, Faivre-Finn C. Learning healthcare systems and rapid learning in radiation oncology: Where are we and where are we going? Radiother Oncol. 2021 Nov;164:183-195. doi: 10.1016/j.radonc.2021.09.030. Epub 2021 Oct 4. PMID: 34619237.

3. Specific Aims of Our Research

Our overall aim is to coordinate the development and integration of tools that will minimize human error and variability as well as enable automated capture of contextual metadata to assist with cohort generation, monitoring of encumbrances, data quality assessment and ontological mapping. Initially, we focused on two of the most common cancer diagnoses in the US, lung cancer (>235,000 cases per year) and prostate cancer (>248,000 cases per year) by bootstrapping the process and accelerating data availability through datasets from the Veterans Health Administration (VHA) radiation oncology quality surveillance program [1]. The specific aims of our proposed research and development effort are as follows:

Specific Aim 1 (SA1): Design and develop a comprehensive learning health system for radiation oncology that is based on passive abstraction and collation of clinical, dosimetry, treatment delivery, and outcome data. ***Rationale***: A software platform that is designed to passively abstract and collect data from multiple data sources in radiation oncology is an essential component of Radiation Oncology LHS. It serves as a centralized platform for data collection and analysis. Such a system can help improve the quality and efficiency of radiation oncology practices by leveraging data-driven insights. The use of passive data abstraction and collation ensures the completeness and accuracy of the data, while reducing the burden on healthcare providers. The integration of multiple sources of data provides a comprehensive understanding of the patient journey and treatment outcomes, allowing for continuous quality improvement and optimized patient care in radiation oncology. ***In Chapter 4, We introduce the design and implementation framework of an integrated data abstraction, aggregation, and storage, curation, and analytics software: The Health Information Gateway and Exchange (HINGE), which collates data for cancer patients receiving radiotherapy.***

Specific Aim 2 (SA2): Extract, Transform and Load (ETL) clinical, dosimetry and treatment datasets into internationally standardized semantic interoperable data models such as NCI Thesaurus & Radiation Oncology Ontology. ***Rationale*****: The** use of internationally recognized data models ensures that data are consistent and can be easily shared and compared across different institutions and countries. Such data model framework helps advance the understanding and advancement of radiation oncology by providing a comprehensive and standardized data resource for researchers and clinicians. The ETL process allows for efficient and reliable transfer of data, reducing the risk of errors and improving the quality of the data. By leveraging semantic interoperability, the extracted, transformed, and loaded data is easily Findable, Accessible, Interoperable, and Reusable (FAIR), allowing for better and more informed decision making in radiation oncology. ***In Chapter 5, We discuss the ETL process, data model frameworks, ontologies, and present a real-world clinical use case with clinical and dosimetry records mapped with this data pipeline using FAIR concepts.***

Specific Aim 3 (SA3): Build a framework for utilizing the common ontology-based data definitions to create universally applicable data mining, keyword search, semantic search, and query expansion methods that make use of standard concepts defined with terminologies and ontologies. ***Rationale*****: Ontology**-based data leads to improved efficiency and accuracy in retrieving relevant information from large clinical data sets. Standard concepts defined with terminologies and ontologies provide a shared understanding of the data being analyzed. Additionally, such a framework allows for the development of more advanced and sophisticated methods for data analyses, as the use of ontologies provides a more

structured and semantically rich representation of the data. A search engine that utilizes ontology-based keyword searching, synonym-based term matching that leverages the hierarchical nature of ontologies to retrieve patient records based on parent and children classes, connects to the BioPortal database for relevant clinical attributes retrieval is highly desirable. To identify similar patients, a method involving text corpus creation and vector embedding models are employed, using cosine similarity and distance metrics is needed. ***In Chapter 6, We discuss the design framework of our data mining, keyword search tool and present the results of four different vector embedding models for patient similarity analysis.***

Specific Aim 4 (SA4): Application of a framework established to investigate the use of 3D Deep Convolution Neural Networks for radiation pneumonitis prediction following Stereotactic Body Radiotherapy (SBRT). ***Rationale*****:** It is essential to test the AI readiness of the LHS in radiation oncology. We propose to determine the feasibility and effectiveness of using 3D Deep Convolution Neural Networks (3D-CNN) for predicting radiation pneumonitis following Stereotactic Body Radiotherapy (SBRT). Radiation pneumonitis is a common side effect of SBRT, and early and accurate prediction of its onset is crucial for timely intervention and improved patient outcomes. The use of 3D Deep Convolution Neural Networks, which are a type of artificial intelligence and machine learning algorithm, has the potential to provide a more accurate and efficient method for predicting radiation pneumonitis compared to traditional methods. The application of the established framework is to allow for a systematic investigation of the use of these networks and demonstrate the effectiveness of the 3D CNN networks for improving patient care in radiation oncology; thus, closing the loop for an effective learning health system. **In Chapter 7, We discuss the design, methodology and predictive results of the two popular 3D-CNN models with input from radiographic and dosimetric datasets of primary lung tumors and surrounding lung volumes to predict the likelihood of radiation pneumonitis (RP)**

3.1 Overall Impact of proposed Research:

Despite the availability of many important clinical and imaging databases such as, The Cancer Imaging Archive (TCIA), The Cancer Genome Atlas (TCGA), NIH data commons, clinical data science researchers still face severe technical challenges in accessing, interpreting, integrating, analyzing, and utilizing the semantic meaning of heterogeneous data and knowledge from these disparately collected and isolated data sources [2, 3]. These tasks pose huge challenges for most clinical data science researchers. Even if data are available and accessible, it still presents a formidable task of cleaning such data for LHS because of inconsistent data formats, syntaxes, notations, and schemas in data sources. These limitations severely hamper the consumption of data and inherent knowledge stored in these data sources. Furthermore, this requires the researcher to learn multiple software systems, configurations, and access requirements which leads to significant increase in time and complexity for scientific research.

Robust learning health system in radiation oncology requires comprehensive clinical and dosimetry data. Furthermore, advanced machine learning models and AI require high fidelity and high veracity data to improve the model performance. Scalable intelligent infrastructure that can provide data from multiple data sources and can support these models are not yet prevalent [4, 5]. **Our proposed LHS framework provides an integrated approach for capturing data from multiple sources and then structure the data in a knowledge base with semantically interlinked entities for seamless consumption in machine learning methods. The use of such an infrastructure solution will allow researchers to mine novel associations from multiple, heterogeneous, and multiple domain sources simultaneously and gather relevant knowledge to provide feedback to the clinical providers for obtaining better clinical outcomes**

for patients on a personalized basis. Additionally, the project has the potential to serve as a model for the implementation of learning health systems in other medical specialties, contributing to the advancement of personalized and data-driven medicine.

References:

1. Hagan M, Kapoor R, Michalski J, Sandler H, Movsas B, Chetty I, Lally B, Rengan R, Robinson C, Rimner A, Simone C, Timmerman R, Zelefsky M, DeMarco J, Hamstra D, Lawton C, Potters L, Valicenti R, Mutic S, Bosch W, Abraham C, Caruthers D, Brame R, Palta JR, Sleeman W, Nalluri J. VA-Radiation Oncology Quality Surveillance Program. Int J Radiat Oncol Biol Phys. 2020 Mar 1;106(3):639-647. doi: 10.1016/j.ijrobp.2019.08.064. Epub 2020 Jan 23. PMID: 31983560.
2. McNutt TR, Bowers M, Cheng Z, Han P, Hui X, Moore J, et al. Practical data collection and extraction for big data applications in radiotherapy. Med Phys. 2018 Oct;45(10):e863–9.
3. Mayo CS, Phillips M, McNutt TR, Palta J, Dekker A, Miller RC, et al. Treatment data and technical process challenges for practical big data efforts in radiation oncology. Med Phys. 2018 Oct;45(10):e793–810.
4. Jochems A, Deist TM, van Soest J, Eble M, Bulens P, Coucke P, Dries W, Lambin P, Dekker A. Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital - A real life proof of concept. Radiother Oncol. 2016 Dec;121(3):459-467. doi: 10.1016/j.radonc.2016.10.002. Epub 2016 Oct 28. PMID: 28029405.
5. Zerka F, Barakat S, Walsh S, Bogowicz M, Leijenaar RTH, Jochems A, Miraglio B, Townend D, Lambin P. Systematic Review of Privacy-Preserving Distributed Machine Learning From Federated Databases in Health Care. JCO Clin Cancer Inform. 2020 Mar;4:184-200. doi: 10.1200/CCI.19.00047. PMID: 32134684; PMCID: PMC7113079.

4. Automated data abstraction for quality surveillance and outcome assessment in radiation oncology

4.1 Introduction

Advanced technologies in health care are bringing a sharper focus on clinical outcome assessment and the assessment of healthcare quality. Manual abstraction, collation, and subsequent analysis of healthcare quality from patient treatment and outcome data are onerous, expensive, and impractical. Advances in computer storage, computing power, and the ability to electronically mine data from disparate sources (e.g., demographics, genetics, imaging, treatment, clinical decisions, and outcomes) have enabled big data research in medicine. The evolution of several initiatives in the realm of interconnectivity of healthcare data sources and the availability of advanced computing frameworks have opened doors for answering a broad array of questions related to quality, safety, and outcomes of patients' clinical care efficiently, objectively, and in a cost-effective manner.

In the radiation oncology domain, large amounts of data are captured routinely across several clinical systems over the course of a patient's treatment as shown in Figure 3.

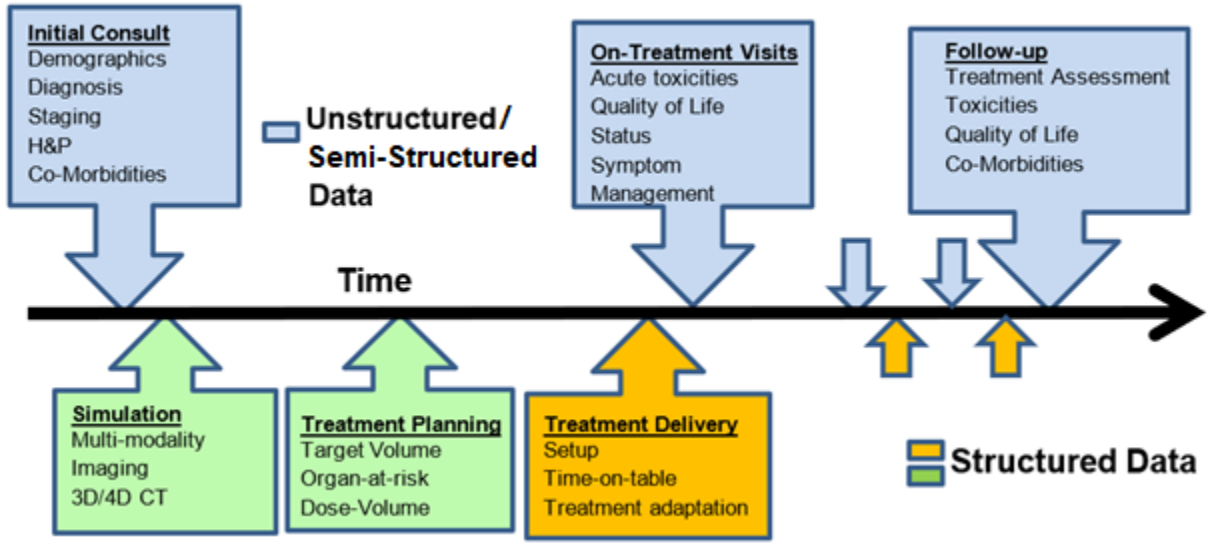


Figure 3: The sequential radiation treatment workflow

Figure 3: The sequential radiation treatment workflow: initial patient consultation, simulation, treatment planning, treatment delivery, on-treatment evaluation, and follow up. The clinical data are in unstructured and/or semi-structured data formats, whereas simulation treatment planning and treatment delivery data are inherently in a structured format.

The electronic health record (EHR) is used to document clinical data that typically includes; demographic information, medical history, medications, laboratory test results, and radiology reports. The physician assessments are often stored in unstructured free text from which key data elements are difficult to abstract for any subsequent data mining efforts. For each patient receiving radiotherapy treatment, the

clinical documentation in EHR typically includes; (1) a detailed initial consultation note, (2) a simulation note describing the treatment simulation procedure, (3) a treatment planning note documenting the prescription and proposed treatment plan, (4) a weekly On Treatment Visit (OTV) note from the staff physician documenting a review of the patient's treatment progress and any acute side effects, (5) a treatment summary or survivorship care plan for the patient and referring physician at the completion of therapy, and (6) routine follow-up notes tracking disease outcomes and any late toxicities. These clinical notes are usually dictated on a telephone, transcribed, and imported into the EHR as preliminary documents. These free-text formatted notes are then reviewed, edited, and finalized. There is a wealth of information in clinical notes for big data applications, but the challenge is to capture and abstract these data in discrete format as part of the regular clinical workflow. However, the treatment planning data including the radiotherapy plan, images, dose, structure set, and dose-volume information from the treatment planning system (TPS) are in structured formats (DICOM-RT). Additionally, the treatment management system (TMS) that contains information regarding the radiation treatment delivery, fractions, visits, etc., is also structured.

The challenge in radiation oncology is to aggregate data, which are both structured and unstructured from disparate data sources. It is extremely difficult to clean, parse and collate the data intelligibly, thus making many research and operational tasks that deal with the optimization of quality care, research-based analysis of radiation treatment, diagnosis-based research, and development of computer-aided diagnostic tools at the infrastructural level quite difficult. Additionally, the lack of interconnectivity and interoperability of RT software systems have made the process of data sharing/transfer cumbersome and challenging. Unfortunately, valuable clinical and radiation treatment data remain trapped behind proprietary software systems. There are Natural Language Processing (NLP) methods that can be employed to extract structured data from clinical narratives dictated in the EHR. These methods utilize text mining symbolic methods approaches such as named entity recognition (NER) based on dictionary lookup and information extracting (IE) relying on pattern matching. Each of these methods provide far from ideal results in gathering accurate structured information from the clinical notes since these methods have to deal with the idiosyncrasies of clinical sub-language due to the use of non-standard ontologies and data dictionaries as well a high degree of spelling and grammatical errors [1]. The accuracy of these approaches can potentially improve when there is a comprehensive cancer ontology used to enable semantic representation of textual information found in clinical narratives. The utilization of these not perfect methods for extracting structure data from the EHR can adversely affect the outcome assessment and predictive analytics modules specifically considering the sensitivity of the data elements to both tasks. Therefore, the structured template-based approach alleviates these concerns and makes structured data capture more credible for clinical use in production quality assurance, outcomes, and big-data analytics platforms.

For the Veteran Health Administration's National Radiation Oncology Program (NROP) office, we developed an integrated enterprise-wide data curation, storage, and analytics portal, called HINGE (Health Information Gateway and Exchange). HINGE is a web-based electronic structured data capture system which has electronic data sharing interfaces with the EHR, TMS and TPS with a specific goal to collect accurate and comprehensive data and to determine clinical practice variations, outcomes, and gaps in treatment quality, and to compare the effectiveness of various treatment modalities and ultimately enable big data analytics in radiation oncology. It is an automatic data aggregator that collates data from different radiotherapy clinical systems/IT applications. It processes radiotherapy treatment

data for quality assessment, predictive analytics and other enterprise-driven clinical informatics solutions with a single online data portal and provides benchmark data and quality improvement tools for individual providers. Additionally, HINGE's design and infrastructure caters to the imminent need for a research-based practice environment and is cognizant of the role of advanced modern computational strategies involving big-data predictive analytics and clinical informatics. Because we realized that achieving these objectives for the whole cancer domain would be extremely challenging, we restricted our scope to two disease sites (prostate and lung cancer). The promise does not come without challenges and hence there were significant technical and workflow related challenges with the actual extraction and aggregation of data from disparate radiotherapy information sources.

4.2 Impetus for automated radiotherapy data abstraction

The Veterans health Administration (VHA), which is the largest integrated health care system in the United States, provides care at 1243 health care facilities, including 170 VA medical centers and 1063 outpatient sites of care of varying complexity. It serves more than 9 million enrolled veterans each year. Forty of the large VA medical centers offer onsite radiation oncology services with oversight from the National Radiation Oncology Program (NROP) office. In 2016, the NROP office embarked on a pilot project to monitor the quality of radiotherapy delivered, determine practice variations, and identification of the care gaps in the VHA. The pilot effort addressed intermediate risk and high-risk prostate cancers (CaP), stage IIIA/B non-small cell lung cancers (NSCLC) and limited stage small cell lung cancers (SCLC). These disease site presentations were selected for the pilot because RT is pivotal in the treatment of these cancers, which together represent more than 60% of patients receiving RT in the VA.

For over 50 years, radiation oncologists have conducted clinical trials to explore new treatment techniques, schedules, and modalities for specific tumor types. These trials have led to improved outcomes, reduced toxicity, and the development of care standards that benefit the broader radiation oncology community. The American College of Radiology's Quality Research in Radiation Oncology (QRRO) program recognized that these care standards based on trial results were reflected in the practice patterns of many radiation oncology practices. Using data from these practices, QRRO investigators developed clinical performance measures that could be quantified and compared to national averages, enabling robust evaluations of practice quality. Building on this success, the VA quality surveillance program, VA-ROQS, integrated these measures and demonstrated that reliable data could be obtained from a national practice base. The VHA NROP office collaborated with the American Society for Radiation Oncology (ASTRO) to establish clinical quality measures (CQM) by which individual patient care would be assessed and compared with the national VHA practice. ASTRO assembled disease site panels comprised of nationally recognized experts who were asked to identify CQM for each phase of patient management by the radiation oncologist as well as dose/volume metrics for the evaluation of quality of radiation treatment plans. The genesis of CQM was the seminal body of work done by the American College of Radiology's Quality Research in Radiation Oncology program [2, 3]. ASTRO panels defined CQM in three categories: currently expected performance measures, those anticipated for the near future (aspirational CQM) or CQM for surveillance only. Methods were developed for manual data abstraction, analytic methods for DICOM-RT data, data curation and the data scoring system. Web-based user interfaces were also developed to report patient scores to their VHA radiation oncologists and aggregate data for benchmarking. Data elements for 1660 patients from the 40 VA radiation oncology practices were abstracted from the electronic medical records, treatment management and planning systems as part of the pilot. The pilot demonstrated that clinical measures provide a tangible means to quantify and improve

quality of care [4]. It also proved that manual data abstraction is time consuming, onerous, and very expensive. It clearly established the need for IT infrastructures for automatic data abstraction and curation [5].

There have been several IT initiatives by research groups in radiation oncology to develop integrated data analysis platforms for either outcome studies and/or decision support systems [6]. There are many large databases such as Surveillance, Epidemiology and End Results (SEER) program established by the National Cancer Institute (NCI) in 1973 and Center of Medicare and Medicaid Services (CMS) that collect data from large number of cancer patients treated over time [7]. The data in these databases include demographics, cancer incidence, clinical and survival factors, but fail to include detailed clinical and treatment information. Some of the data analysis from the SEER database suggests that the database lacks information about the radiation dose, technique, and radiotherapy receipt [8].

The University of Michigan has spearheaded the development of two robust data aggregation systems known as M-ROAR [9] and MROQC [10]. These innovative systems have been designed to effectively collect and evaluate various practice patterns, enabling the analysis of outcomes and the assessment of dosimetry-related information. MD Anderson has taken a significant stride by implementing a comprehensive electronic data capture system that operates across its entire network. This system serves the vital purpose of recording and storing crucial treatment information pertaining to patients [11]. Meanwhile, Johns Hopkins University has launched the Oncospace program, an impressive initiative aimed at capturing and analyzing radiotherapy data. This program goes beyond simple data collection, as it acquires and integrates essential details such as anatomy, dose distributions, and outcomes into an analytical database [12]. The Mayo Clinic's Department of Radiation Oncology in Florida has taken a forward-thinking approach by establishing a linkage between its radiation oncology information system and Mayo Clinic's internal claims data warehouse. Furthermore, by connecting with Mayo's tumor registry, this integration facilitates large-scale studies in the field of radiation oncology [13]. At Mayo Clinic in Rochester, Whitaker et al. have devised a patient-reported outcome (PRO) collection and management system. This sophisticated system allows for the aggregation of treatment data from a substantial number of patients, enabling researchers to gain valuable insights on a larger scale. While many existing platforms have specific objectives, the HINGE software stands out for its comprehensive approach to radiotherapy data collation. In addition to capturing DICOM-RT data from TPS (Treatment Planning System) and treatment data from TMS (Treatment Management System), the HINGE software also integrates clinical data from EHR (Electronic Health Records). This comprehensive integration empowers researchers and practitioners with a holistic view of patients' radiotherapy episodes. The overarching goal of HINGE is to meet the following objectives:

- to allow healthcare institutes to assess their practices/treatment outcomes and make improvements at a systemic level
- to better equip and assist the physician with complimentary/supplementary information to aid their clinical decision-making process
- to create systems which would allow for the research and development of tools that relate to machine learning, artificial intelligence, and big data analytics
- to allow for ease of data interoperability, data access and exchange for third-party applications/programs

- to foresee the future trends in the healthcare industry and subsequently design data platforms in alignment with the upcoming technologies

4.3 Overview of the HINGE platform

The crucial data elements required to assess the quality of radiotherapy planning and delivery and to build decision support systems are distributed across disparate clinical systems and are recorded along each sequential step of the radiation treatment (from initial Consult to follow-up). Figure 4 shows a brief overview of the HINGE architecture. HINGE is a real-time data analytics portal connecting the EHR, TPS and TMS. The HINGE local application is hosted on a central cloud server which is accessible to each local facility via standard web-browsers and the HINGE Central Server is also hosted at a HIPPA compliant secure cloud server. Radiation Oncologists enter the information via smart disease-specific templates (user interface) which are discretized and stored in the database. HINGE Local also communicates with EHR, TPS and TMS and imports/exports relevant patient data. The data are anonymized and sent to HINGE Central Server for data analytics and display of results on an interactive Web-based dashboard for quality managers, physicians, and hospital administration.

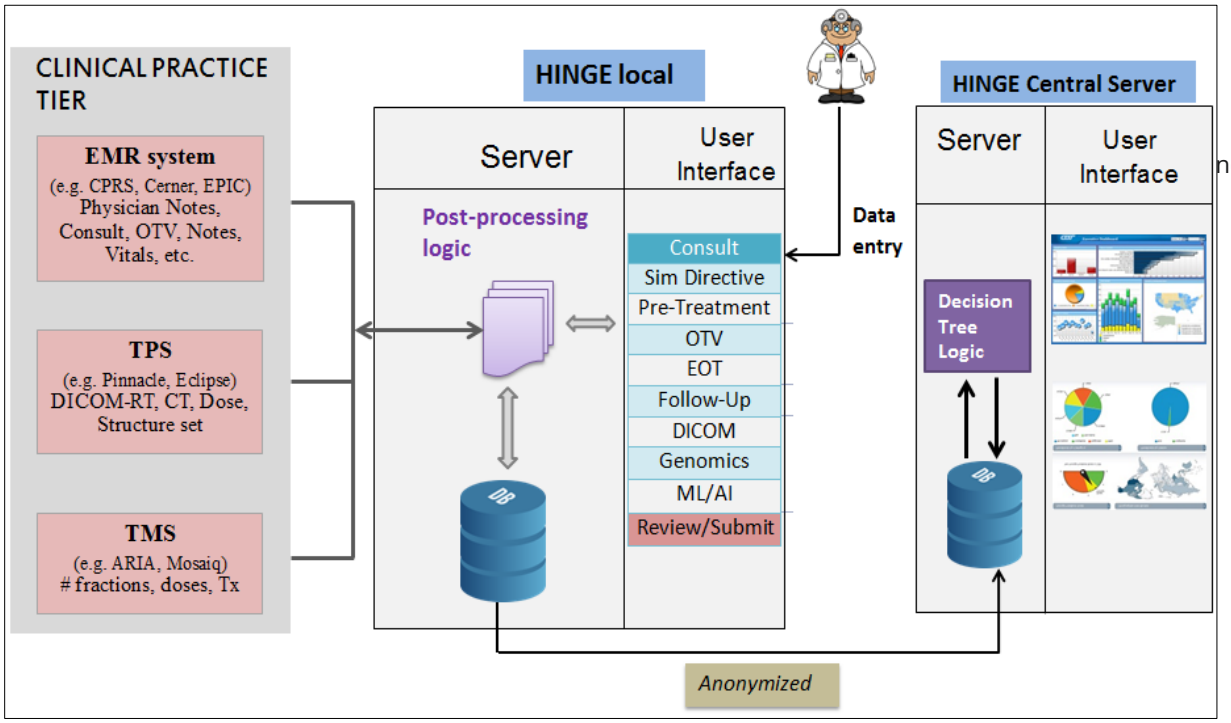


Figure 4: Overview of the architecture of Health Information Gateway and Exchange (HINGE) software platform. The clinical workflow templates (Consult, Sim Directive, etc.) in the HINGE local are automatically populated with data that are available in clinical practice systems that include electronic health record (EHR), treatment planning system (TPS), and treatment management system (TMS). The complete radiotherapy data are sent to the HINGE central server, where it is evaluated for data integrity, curated, and prepared for visualization by end users in a web-based graphical user interface (GUI).

4.3.1 Data standardization

HINGE software is designed to facilitate the process of extract, transform and load (ETL) of data throughout. Disparate systems are used to collect clinical, treatment and process data, however the lack

of uniformity in data syntax and semantics makes it extremely difficult for data aggregation and analysis. Unfortunately, at the present time there is a paucity of ontologies with radiation oncology specific terms. HINGE deploys 'smart' disease-specific templates meant for data entry/viewing as part of its user interface for radiation oncologists. These templates facilitate the physicians and the clinical staff to enter the relevant clinical information in a discrete manner. Figure 4 shows the overview of the components of the HINGE application. The goal of HINGE is to provide a method to collect comprehensive radiotherapy episodic data including DICOM-RT from the TPS, daily treatment data from the TMS and clinical data from the EHR. All these data are collected in discrete and structured data formats. This approach will allow healthcare institutes to assess their practices for system-wide improvements, provide supplementary information to aid physicians in their decision making and help foresee changes in the healthcare field. HINGE also improves the ease of data interoperability and data access between third party applications by using industry standard protocols and services. The templates have embedded critical data elements (data farming) that are required for quality assurance/assessment (QA) analyses. These critical data elements are used to score the disease site specific clinical quality measures that are listed in the paper from Hagan et. al [4]. Most commonly, electronic case report form templates are utilized routinely to collect structured data in randomized controlled trials, but these templates are limited to trial-specific data elements and those are entered in addition to routine clinical documentation. These templates are utilized as part of the routine clinical workflow and documentation and are the starting point for the physicians to record their assessments. The templates mimic the radiotherapy workflow from – consultation, simulation, treatment, end-of-treatment, to follow-up care. The templates are interfaced with the EHR, allowing data such as allergies, drug list, lab values and vitals to be automatically populated into the template from the EHR database. Thus, the templates facilitate the entry of data in a structured discrete format, along with simultaneously allowing free-text data entry sections for recording additional observations. However, much of the data such as TNM staging, Performance status, treatment intent, status, previous cancer encounters with RT, chemotherapy or surgery as the treatment modality, prescription, toxicity grades, simulation, treatment planning directive, survivorship data elements, etc. are entered in discrete format. We used all predefined radiotherapy data nomenclature (AJCC [17], CTCAE [15, 16], AAPM TG 263 [14]) and defined additional ones where no standard data definitions existed. Examples of these additional nomenclatures include discrete data fields used to describe past medical history, molecular testing status, plan of care including treatment options discussed with the patient, recommended therapy, and patient selected treatment to name a few. Automatic calculation of assessment scores and graphical indication of treatment progress are rendered in these templates. At each encounter, after the data entry has concluded, HINGE prepares the data into a textual narrative note format by utilizing user specified template boilerplate narratives and embedding these discrete data elements. The full note narratives are then exported to the EHR for medical records via an interface for the purpose of maintaining clinical documentation and continuity of care since the patient might be subsequently seen at other clinical services within the hospital. These templates have been developed with specific UI based design considerations from the physician end users. These templates are specifically designed to save physicians' time/effort and enhance their ease of access by incorporating technical UI/UX features like – least amount of page scrolling, reducing the number of mouse clicks, data entry in a lateral motion within the HINGE application, positioning high-utility patient details on the top of the page, business logic for auto calculation of certain data elements such as NCCN risk groups based on staging, Gleason, and PSA values etc. In addition, auto population of subsequent note templates (e.g., end of treatment template) with discrete data from previous templates (consult, treatment planning directive

template) also saves physician time that they can spend with the patients rather than just dictating notes in the EHR. he templates also perform extensive data entry validation, data-completeness check at the entry level and provide helpful error messages, suggestions, highlighting of critical elements, etc. These templates are disease site specific and relevant data entry fields appear based on the diagnosis and treatment site codes. The templates also prepopulate the data fields from TPS and TMS so that the physicians do not have to make redundant entries.

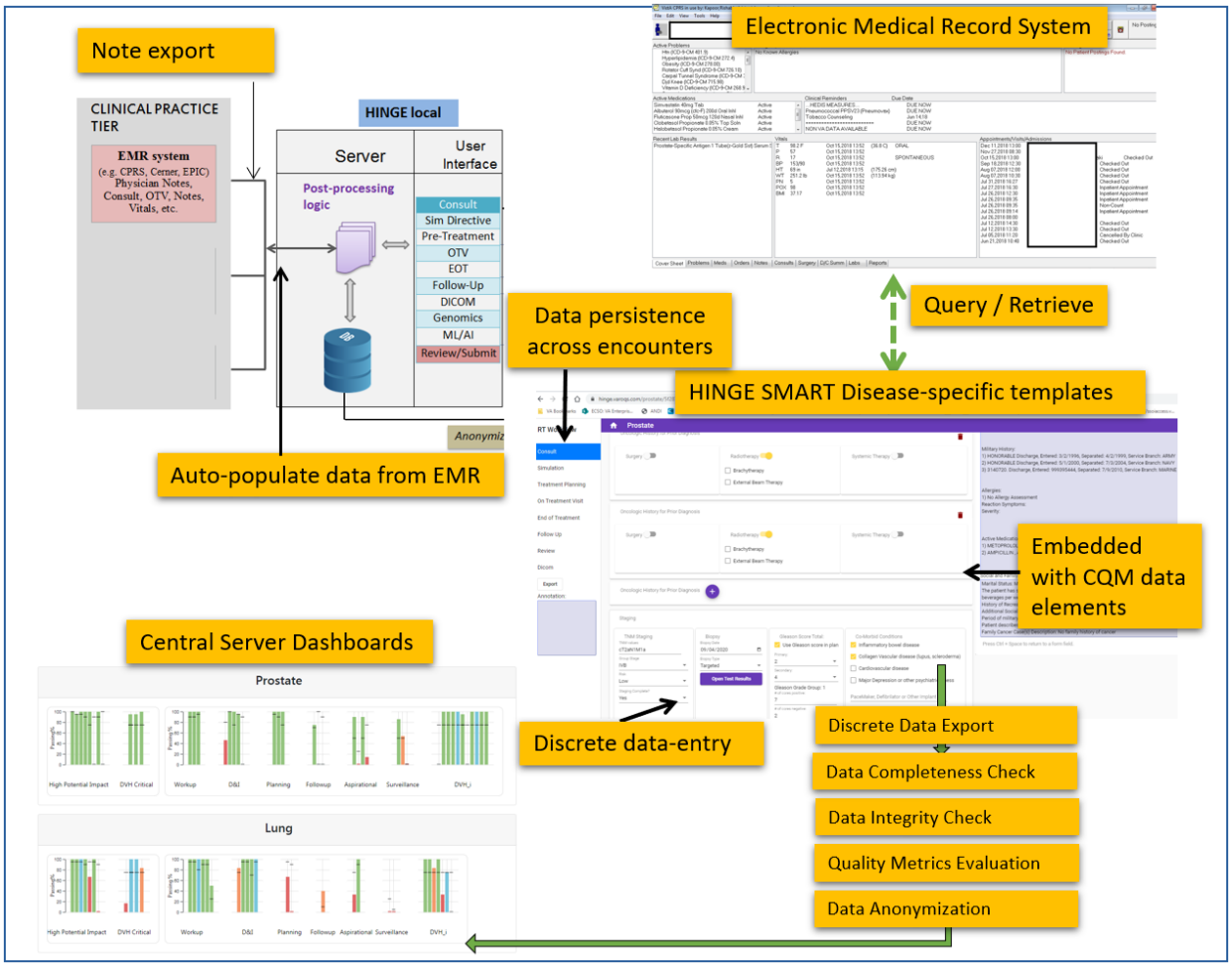


Figure 5: Overview of the components of the HINGE application. Discrete clinical data abstracted via query/retrieve from the Electronic Medical Record (EHR) and populated in the HINGE SMART Disease-specific templates UI. Discrete and free-text data is transcribed by the providers in the Disease specific templates. SMART templates have business logic to auto calculate scores, perform auto-population of subsequent templates with discrete data, report any missing value or value outside a defined range and abstract the data elements for clinical quality measure (CQM) analysis. A free text narrative note is generated from all these discrete data elements and interfaced to the EHR as part of the clinical documentation. All the data from these SMART templates are checked for completeness, integrity and anonymized before exporting it to the Central Server Dashboard where data visualization tools (charts, graphs with flagging of outliers etc.) are deployed to analyze the CQMs, clinical and dosimetry data for a cohort of patients.

4.4.2. Integration with Radiotherapy Data sources

4.4.2.1 EHR-HINGE integration:

HINGE is designed to communicate with VHA's EHR, i.e., VISTA. HINGE has employed an external interface which is able to communicate (query/retrieve) with VISTA and fetch required patient details such as demographics, vitals, labs, medications, surgery, pathology, encounter, allergies, hazardous material exposure, problems list and survival information etc. The list of data types retrieved from the EHR is shown in the Figure 6. Most of the information exists in discrete format when it is retrieved from the EHR. The interface is also able to retrieve information such as health history, surgery and radiology reports that only exist as clinical free-text notes from the EHR. All this information is pre-populated in the clinical note templates in HINGE for the physician or care team member convenience. The physician or the care team member tasked to complete and sign the template in the HINGE software are required to enter their clinical assessments in discrete format in the templates. There is section where free-text narrative text is entered in the template to describe the discrete data or medical rationale behind the entered assessments. Additionally, after the note is completed by the physician, the discrete data is converted into a textual note and exported to VISTA via this interface. Specifically, this interface-based design allows HINGE to be oblivious to the underlying EHR system (VISTA, Cerner, EPIC, etc.). It helps in its portability and allows it to be functional even if the EHR system changes by isolating the business logic of the integration strictly within the interface.

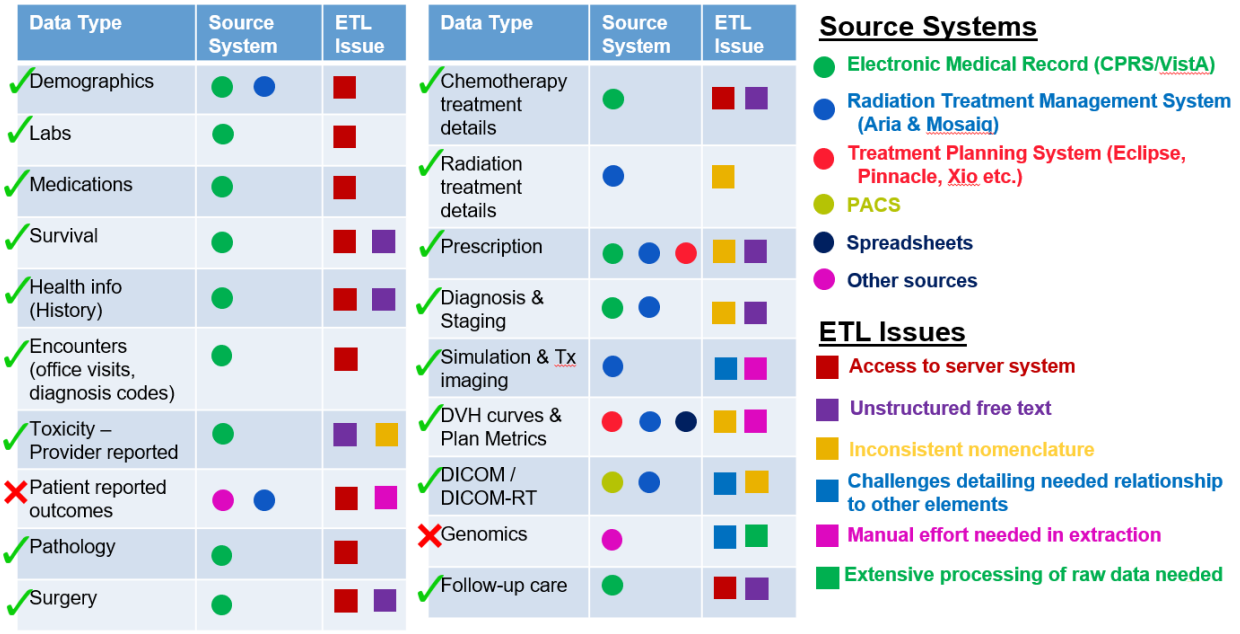


Figure 6: List of the data types utilized in Radiation Oncology domain. *List of the data types utilized in Radiation Oncology domain, source system where the data resides, extract/transfer/load (ETL) issues. Access to server system, unstructured free-text and inconsistent nomenclature are amongst the major ETL issues across the various source systems. HINGE application gathers data types (green tick) from all the mentioned source systems except patient reported outcomes and genomic data.*

4.4.2.2 TPS- HINGE integration:

HINGE is able to import DICOM-RT data from any Treatment Planning System (TPS) that conforms to the Integrating the Healthcare Enterprise – Radiation Oncology (IHE-RO) [18] defined profiles. The VA system utilizes all the TPS products sold in the marketplace and hence it is imperative that we conform to one standard interoperability solution provided by IHE-RO to pull data for an enterprise-wide application such as HINGE. We have deployed a free, open source and light weight DICOM server known as Orthanc [29] to collect DICOM-RT datasets from any commercial treatment planning system. Orthanc is a simple, yet powerful standalone DICOM server designed to support research, and query/retrieve functionality of DICOM datasets. One of the major challenges with examining patients' DICOM-RT data is the lack of standardized target and organ at risk (OAR) nomenclatures, prescription formatting, and ambiguity regarding dose-volume histogram metrics, etc. across several disease-sites. This impedes any research into examining dosimetric effects of practice patterns longitudinally. To resolve this issue, an initiative to introduce the standardizing nomenclature for radiotherapy was implemented under TG-263 [14]. HINGE deploys this naming convention within its system, requiring treatment planners to match the deemed organs at risk (OARs) to their TG-263 names. HINGE automatically suggests the equivalent TG-263 names for the listed OARs for the planner (Figure 6). In addition to simple text mapping, Machine Learning can be used to automate the process of relabeling physician specified structure set names to the TG-263 defined names. Structure name mapping scripts are written in python programing language to support this process. Success in this approach has been shown using target and organ at risk text labels [19], geometric information [20,21,22] and radiomics features [23], all found in the DICOM structure set, dose, and reference imaging (CT) datasets. All these methods have shown reasonably good accuracy over many different structure types and the HINGE platform has the capability of deploying such methods as it has all of the treatment planning DICOM files as well as access to cloud-based Machine Learning frameworks including AWS Elastic Map Reduce and Deep Learning Containers. The software calculates and displays DVH from the uploaded dataset to the dosimetrist for final verification and selection of the key target and OAR sites. Based on the DVH dose constraint-based quality measures, the appropriate pass/fail/acceptable variation status is stored in the database before the complete dataset is uploaded to the central server dashboard.

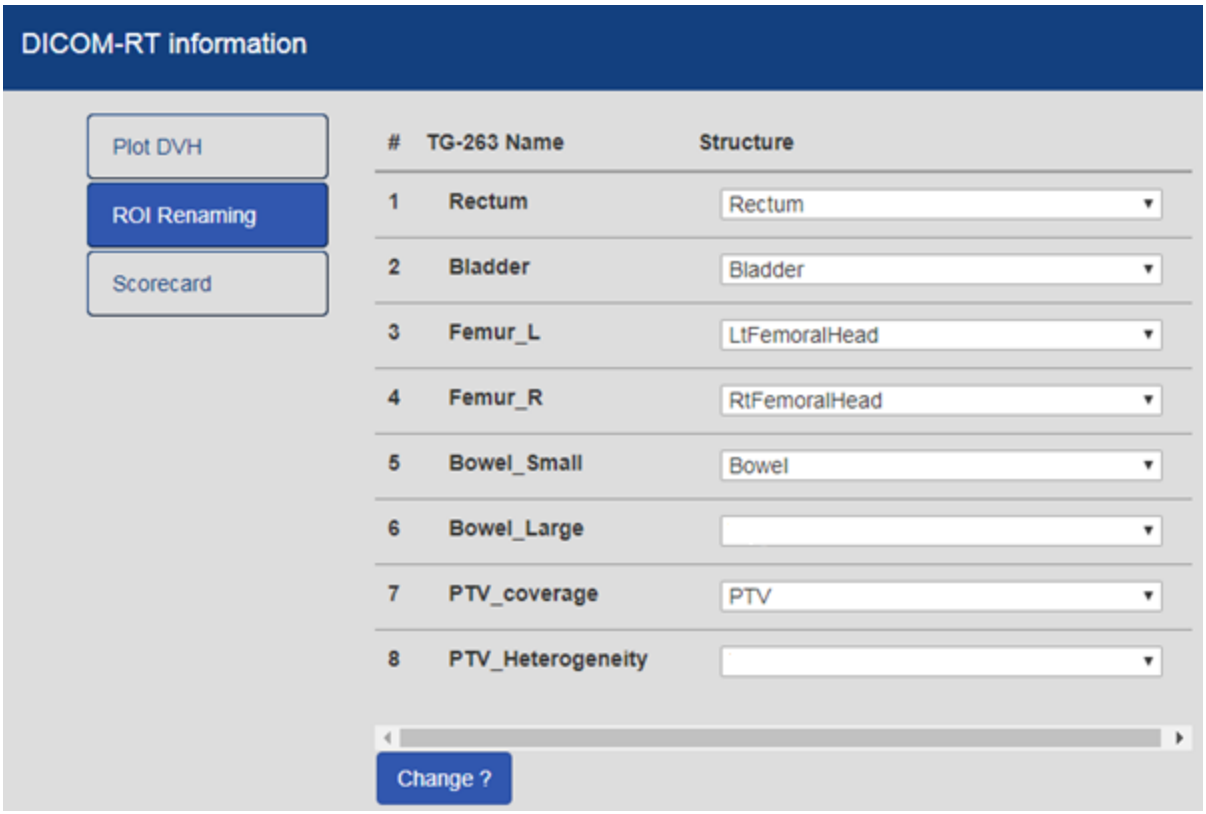


Figure 7: Screen capture of the user interface for selecting the appropriate structures for target and OAR renaming in the HINGE application.

4.4.2.3 TMS- HINGE integration:

Existing treatment management systems are primarily designed to optimize the clinical workflow and therefore lack utilities required to facilitate Big Data applications. For HINGE to assume the role as the one-stop-shop for managing radiotherapy data, it must have access to the data present in treatment management systems such as Varian Aria/ Elekta Mosaiq products. To achieve this goal, we have created a Docker container, named HINGE-Broker, which runs on the same network as TMS's underlying database. Within HINGE-Broker, a Python script provides access to the TMS database using the SQLAlchemy toolkit and a Node application exposes a web API for HINGE to make remote queries. The TMS database contains discrete elements such as patient demographics, prescribed prescriptions, delivered dose and a listing of all clinical notes.

However, TMS software upgrades may result in changes to the underlying database schema and potentially different versions of TMS deployed across a healthcare system will also require the support of multiple methods of data access. In addition, some versions of TMS require very complicated, non-intuitive SQL queries for retrieving current dose information which potentially makes this approach very sensitive to schema changes. Although the TMS software provides methods for directly entering many discrete data elements, these tools are often underutilized, resulting in little information that could be used for further studies. The data extracted from TMS is used to populate the simulation, under-

treatment, end of treatment summary templates in HINGE and are made available for physicians to view/edit. This allows for HINGE to access treatment delivery data in a discretized manner.

Radiotherapy TMS does support the storage and display of Word documents within the application. We have partially addressed the lack of discrete treatment data by creating Word templates for the under-treatment visit and end of treatment summary notes with tagged fields. Using a Word macro, these discrete template fields can be exacted as JSON which can then be stored in a MongoDB database for analysis.

4.4.3. Quality Assurance/Analyses of Radiotherapy Data

With data standardization and clinical integration described in the preceding sections, HINGE software is able to aggregate the critical data elements from the entire clinical workflow to score all clinical quality measures (CQM) for each patient. Once a patient's treatment is deemed complete for export, the anonymized data are uploaded to the HINGE's central server. The business logic for deriving the CQMs based on the patient data resides on the central server (see Figure 4). After the data is received, the CQMs are calculated and are available for viewing on the visual dashboards on the central server via a Web portal. Many of the CQMs are not straightforward and require extensive decision-tree logic to construe a 'pass' or a 'fail.' HINGE has deployed such decision-tree logic (Figure 8) in its system to calculate the CQM scores automatically for each CQM. The HINGE system utilizes the web-based dashboard software that provides real-time access to aggregate Clinical Quality Measure (CQM) scores for a cohort of patients. The calculation of these scores is automated and occurs in real-time, allowing for immediate feedback. The dashboard software is designed to be user-friendly and provides detailed information on the CQM scores to both the quality manager and individual physicians. By leveraging the decision tree logic implemented on HINGE's central server, the software calculates and displays the CQM scores to the respective stakeholders. This real-time feedback is generated within minutes of data submission to the central server, ensuring that physicians can promptly assess their performance and compare it against their peers across the VA enterprise. The dashboard software serves as a valuable tool for physicians, enabling them to benchmark their treatment practices and outcomes against their colleagues. The scores and evaluation of clinical and treatment quality measure results are presented in an easily understandable format, empowering physicians to gain insights into their performance and

make informed decisions to improve patient care.

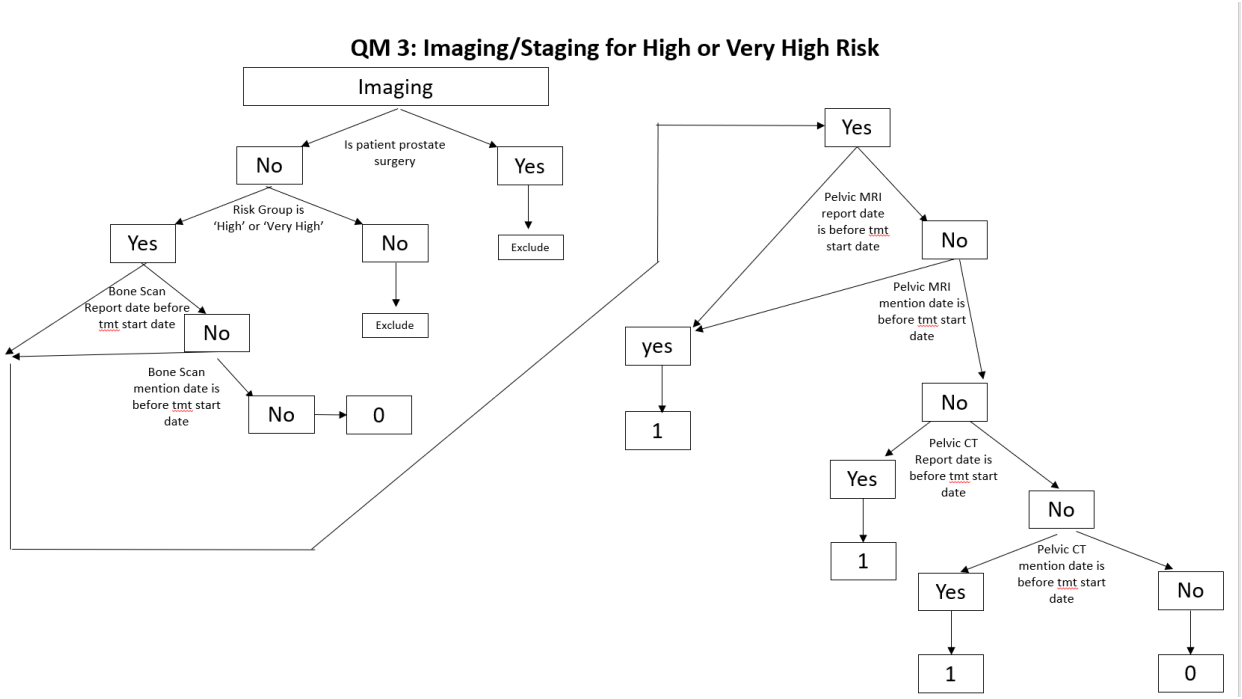


Figure 8: Example of a decision tree logic for a clinical quality measure. Data from the HINGE Smart templates, TPS and TMS modules are utilized with these decision trees to generate pass/fail [1/0] for each of the disease site specific clinical quality measures.

4.5 Dashboard Analytics

HINGE visual dashboards available through the central server display visual plots, charts, and graphs each detailing the performance of every VA practice for each CQM (Figure 9). The dashboard provides a vantage position for every physician and the quality managers to assess the performance of each CQM relating to its current standing, comparison with expert-defined thresholds and their peers' performance nationwide. This allows the physician to understand the quality of RT care they deliver and ways to improve it. The dashboard also provides the quality managers with insightful information to investigate performance or systemic issues, marshal resources and design effective health policy solutions to improve RT care.

In addition to viewing the information through the dashboard, the central server functions as the data warehouse since all the patient data from local VA facilities is exported to this data warehouse. Thus, it is poised for large enterprise analytics that involve data mining, outcomes research, comparative effectiveness, machine learning and other large database interrogation queries.

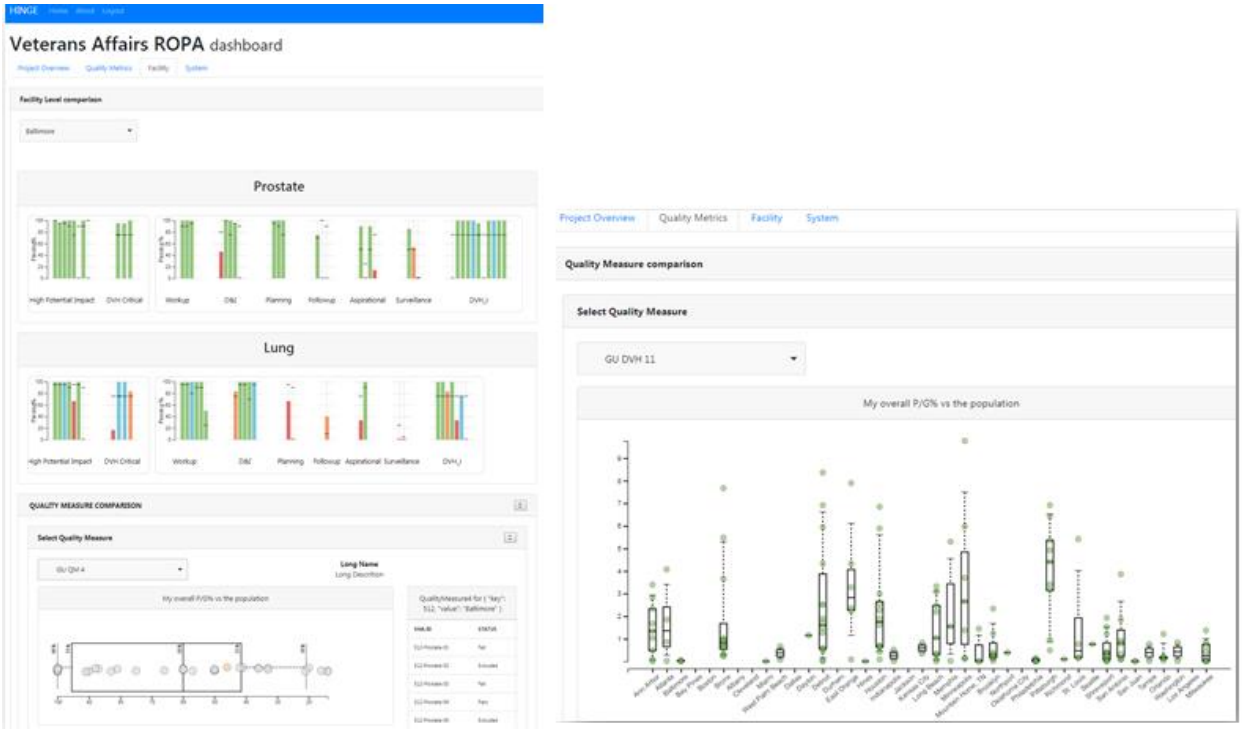


Figure 9: Screen capture of the HINGE dashboard application showing data from 40 VA practices.

4.6 Data Anonymization

By way of architectural design, HINGE is split into HINGE local and HINGE central server (Figure 3). HINGE local is the application facing the local VHA facility connected to the local environment and resources such as EHR, TPS and TMS. The HINGE local application is hosted on the cloud computing platform hosted by VA's Enterprise cloud (VA-EC) – Amazon Web Services (AWS) and the application is running multiple instances with the database siloed into partitions for each local VA center. HINGE central server is hosted on a VA-EC as well where data is captured from each of the HINGE-local instances. After a radiation treatment is completed and when the end of treatment note is generated from the HINGE software, the physician/clinical staff are prompted by the software via notifications to review the patient data and push the data to the central server. After the clinical staff at a local facility initiates the export of data, the data is anonymized and presented for review and approval. In this review, the treatment data collected from all clinical and treatment management notes, TPS and TMS (including DICOM-RT) without the protected health information (PHI) are presented to the physician for final approval. The de-identified data are compliant as per HIPAA policies. After approval, the data are exported from HINGE local to the central server via a secure encrypted channel.

4.7 Data Security

Data security is paramount and a crucial component of any healthcare organization and infrastructure. The numbers of cyber-attacks on the healthcare industry are constantly growing for the purposes of medical identity theft and Medicare fraud. HIPPA regulations [24] sets specific guidelines for maintaining the privacy and security of any information system deployed in the healthcare domain. The HIPPA security rules outlines the administrative, physical, and technical security measures that an organization must take to ensure confidentiality, integrity, and availability of healthcare datasets [25, 26]. HINGE is utilizing administrative safeguards where documented policy and procedures are established to create a uniform

process that clinical users follow to maintain patient privacy and information security in the software. HINGE also employs technical safeguards where no PHI/PII is shared within or with other interfaced application without appropriate network (SSL) and software encryption. The VHA has very stringent physical safeguards in place where the data centers housing the HINGE application has locks and security system to protect from PHI data breaches associated with break-ins. Keeping the healthcare data confidential, available, and maintaining integrity have direct relationships with HIPAA compliance.

Confidentiality is the act of ensuring that patient's health data are kept completely undisclosed to unauthorized entities. HINGE is integrated with the VA's single sign on (SSO) and 2 factor authentication (2FA, token key and password) system where enterprise-wide access control measures are undertaken by the VA's central IT office. Having the HINGE software run on the cloud environment leads to an increase in the risk of data compromises, as the data becomes accessible to an augmented number of sub-systems. In the HINGE software design architecture, we have made the application tools self-contained thereby mitigating the risk that comes with connecting with third party vendor tools. The interfaces with the EHR's also are unidirectional with the intention to pull the data with software encryption built in.

Integrity is important to make sure that the healthcare data captured by HINGE is accurate and consistent and not modified in any way. Treatment decisions based on erroneous data can have serious and adverse consequences on patients' health. HINGE utilizes checksum or a hash, before using the data and if integrity check fails, the application reports an error in an audit trail and terminates the transaction without processing the data.

For the HINGE application to be successful and serve its purpose, the information must be available at all times in spite of service disruptions due to hardware failure, system upgrades, power outages, and denial of service attacks. The deployment of the application is on two separate AWS availability zones with load balancers and multiple redundant copies of the MongoDB backend database to ensure high availability.

4.8 Testing and deployment of the platform

The validation process of the HINGE platform involved the use of a meticulously collected dataset comprising 1660 patient clinical and dosimetry records from 40 VA Radiation Oncology clinics. This dataset was specifically tailored for the pilot study described in section 4.2 and included patients with prostate cancer, non-small cell lung cancer (NSCLC), and small cell lung cancer (SCLC). Each patient's record consisted of approximately 80-100 clinical data elements, which were entered into the disease site specific HINGE templates. With the auto-population feature from the patient's electronic medical records, the physician assessments in the HINGE software require minimal data entry. Based on our preliminary analysis, completing the template, and generating a textual narrative note for initial consultations or follow-up care would take less than 10 minutes. This streamlined process saves time for physicians, allowing them to focus more on patient care and decision-making rather than manual data entry. By leveraging the comprehensive patient data already available in the electronic medical records, the HINGE software reduces the burden on physicians while maintaining the accuracy and integrity of the information. The software intelligently populates relevant sections of the template, ensuring that the physician has access to a comprehensive overview of the patient's clinical history and treatment details.

To analyze and score the data, the HINGE platform leveraged a decision tree logic implemented in a dashboard software. The data from the templates were exported to this dashboard software, which is accessible at https://varoqs.com. The software provided a visual representation of the data through plots,

graphs, and charts, offering insights into the performance of each VA practice for every Clinical Quality Measure (CQM) under consideration. As part of the validation process, the manually entered data in the HINGE template underwent a meticulous manual check to validate the accuracy and functionality of the decision tree logic. The data was analyzed to score 22 prostate clinical quality measures (CQM), 12 prostate dosimetry measures examining coverage of the planning target volume and doses to bowel, femurs, bladder, and rectum. 26 NSCLC CQMs and 15 dosimetry measures, 26 SCLC CQMs and 21 dosimetry measures were also scored with this dataset. There was a total of 35,303 clinical and 12,565 dose constraint based DVH data elements used for scoring these quality measures. Any instances of failure for these measures were closely analyzed using visualization tools that depicted the decision tree and highlighted the specific nodes where the failure occurred (shown in figure 10). This level of detailed analysis allowed us to identify and address any discrepancies or issues in the decision tree logic.

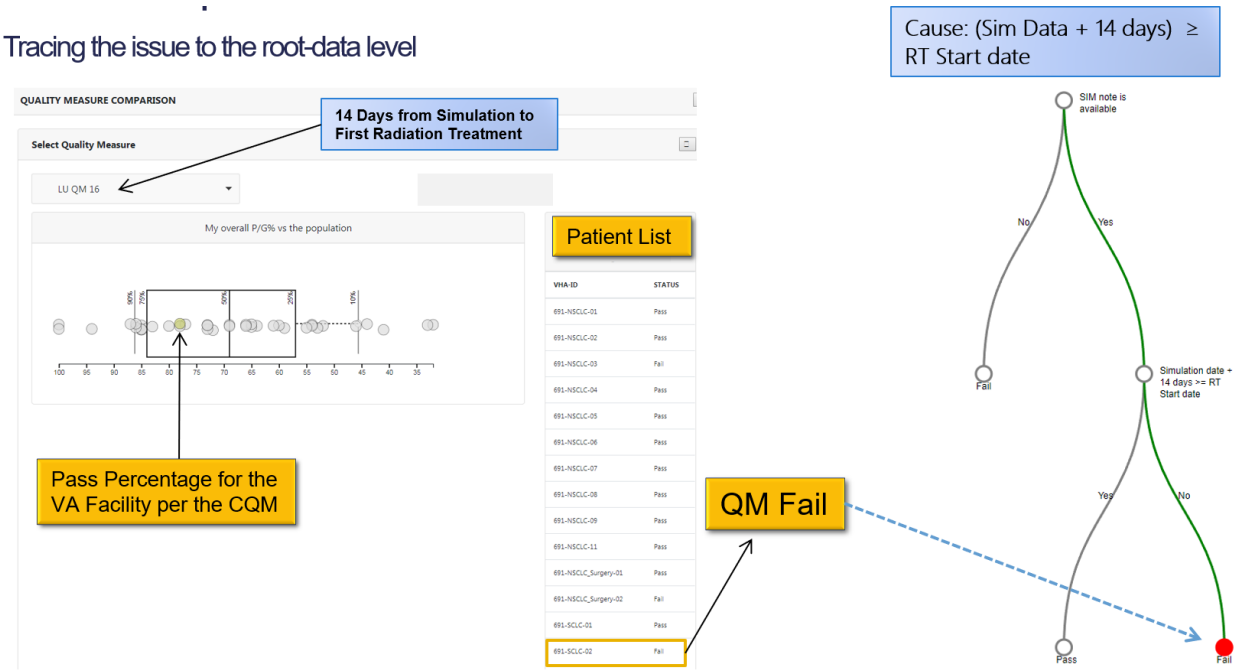


Figure 10: Tracing the CQM failures in the HINGE Dashboard Portal

The platform has also been undergoing further refinement through workshops conducted with VA Subject Matter Expert (SME) physicians. These physicians received training on using the disease site specific HINGE templates and were tasked with entering sample cases to provide valuable feedback. These workshops have proven to be instrumental in improving the quality of the templates and the discrete data generated by the platform. The interactive sessions allowed the physicians to provide insights into the usability and functionality of the templates, leading to iterative improvements and enhancements. This collaborative approach ensured that the HINGE platform was aligned with the specific needs and workflows of the VA healthcare system. While we have made significant progress in establishing the data collection framework through the utilization of the HINGE software, it is currently not in active operation within the VA system. It is because the software is undergoing evaluation for compliance with the

cybersecurity framework established by the VA, based on the risk management NIST framework. This evaluation ensures that the platform meets the necessary security standards to safeguard patient data and protect against potential cybersecurity threats. The rigorous assessment process includes evaluating the software's infrastructure, data protection measures, access controls, encryption protocols, data backups & disaster recovery planning, incident response planning, and vulnerability management. Once the evaluation is complete and the software is deemed compliant, it will be deployed within the VA system, providing healthcare professionals with a reliable and secure tool for data entry, analysis, and decision support in the field of radiation oncology.

4.9 HINGE Information Technology and Deployment Architecture

The HINGE web applications are deployed on the VA cloud platform hosted by Amazon Web Services (AWS) using ECS containers and Fargate services. The backend database where all the template data is stored is Mongodb and the front-end user interface is designed using the Angular web application framework. The architecture relies on EC2 instances to run MongoDB, which serves as the repository for web form data. Additionally, AWS S3 is utilized for storing a large collection of medical imaging and treatment planning objects in the DICOM format. The cloud-based architecture for HINGE, depicted in Figure 11, incorporates essential data backup solutions. It includes two availability zones and a load balancer system to ensure efficient network traffic routing to available servers, enabling timely fulfillment of web requests. This architecture is implemented across three separate environments: developmental, pre-production, and production. These environments play crucial roles in hosting the application during development and testing phases before its deployment on the production environment. This multi-environmental setup allows for thorough evaluation and refinement of the software. To ensure security, continuous cybersecurity measures are performed, including vulnerability analysis and the implementation of remediation measures.

The overall cost for hosting these three environments and conducting ongoing cybersecurity evaluations, along with applying necessary remediation measures and routine software updates and patching, amounts to approximately less than half a million dollars (FY 2022 estimate). This investment covers the infrastructure, maintenance, and security measures required to support the HINGE platform and ensure its reliable operation.

The HINGE information technology and deployment architecture provide a robust and scalable foundation for the platform's functionalities. It leverages the capabilities of AWS services, such as ECS, Fargate, MongoDB, and S3, to securely store and process data while ensuring high availability and performance. By adhering to industry best practices and employing a multi-environment approach, the architecture promotes the development and deployment of a reliable and secure application for radiation oncology practitioners and researchers.

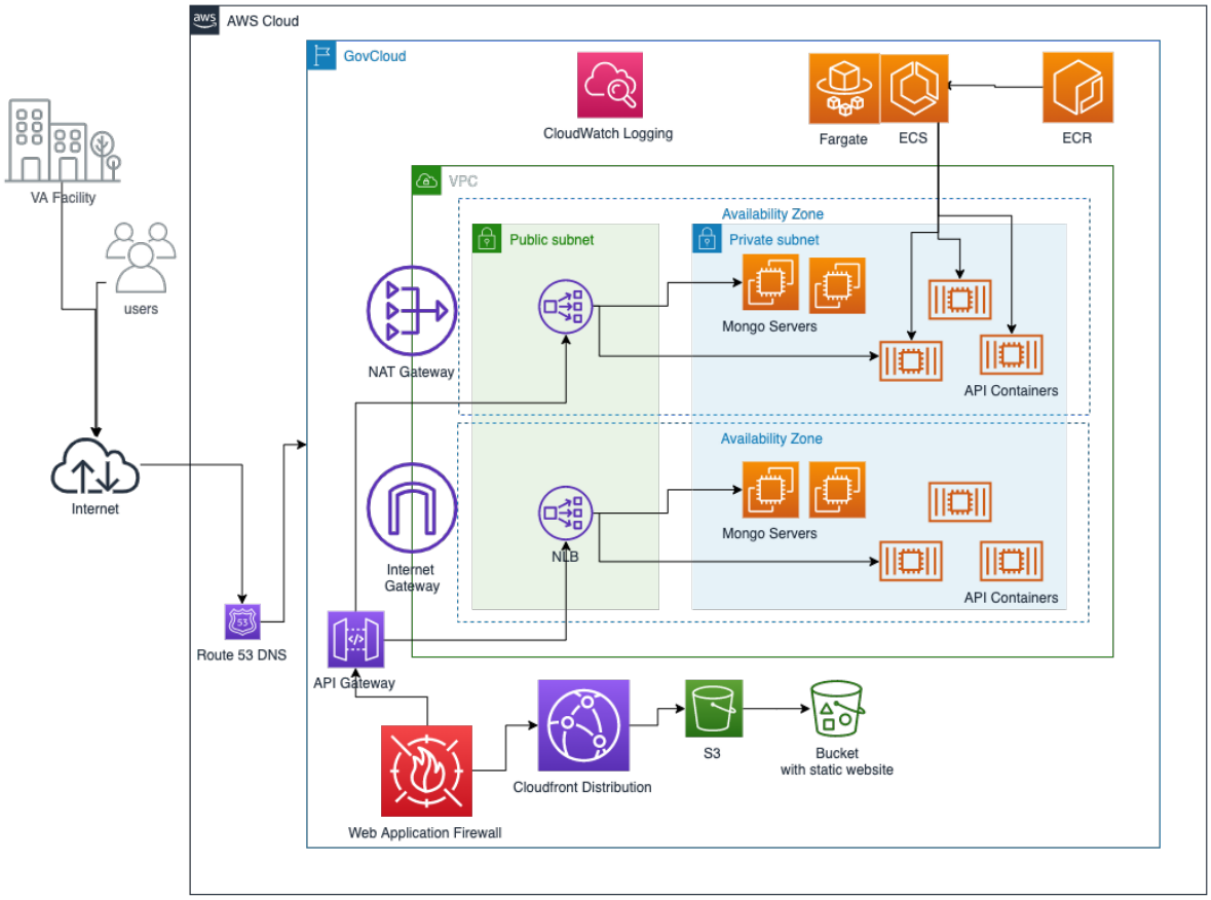


Figure 11: High-level cloud architecture for the HINGE platform

4.10 Discussion

Collation of comprehensive population based clinical information, radiation treatment planning, delivery, and health outcome information is essential for any robust radiation oncology quality surveillance and outcome assessment program. The HINGE software platform allows passive real time assessment of a radiotherapy quality of care. Building the HINGE software presented several challenges, which were addressed through careful consideration and implementation. The challenges and their corresponding solutions are as follows:

- Consensus among VA subject matter expert physicians: Obtaining agreement and alignment among the VA subject matter expert physicians regarding the specific data elements to be included in disease site-specific templates was a major challenge. Physicians may have different preferences, opinions, or practices, making it difficult to establish a unified approach. For example, when developing a template for prostate cancer, one physician may emphasize certain clinical indicators or treatment outcomes, while another physician may prioritize different aspects. Fortunately, the VA quality measures that referenced evidence-based guidelines were the guiding force behind the achieving consensus regarding which data elements should be included in the templates. Achieving consensus required extensive communication, collaboration, and negotiation among the physicians to identify

and prioritize the essential data elements for each template. Throughout the process, we played the consensus building role by actively listening to the physician SMEs, clarifying points of confusion, and ensuring that discussions remained focused and productive.

- Variation in clinical workflows: Radiation oncology workflows can vary across different disease sites, treatment modalities, and across different Radiation Oncology clinics. Designing templates that accommodate these variations while capturing the necessary data elements posed a challenge. The templates designed in HINGE, with the collaboration and input from VA Subject Matter Expert (SME) physicians, were aimed at addressing the majority of the clinical data documentation requirements and recommendations set forth by accrediting bodies such as the American College of Radiology (ACR), ASTRO's APEx bodies, and the Joint Commission in addition to the billing-based documents requirements.
- Integration with EHR: Clinical and operations data in radiation oncology are collected from various systems, leading to a lack of uniformity in data syntax and semantics. HINGE has built an integration with the VHA's EHR (VISTA) to retrieve patient details, such as demographics, vitals, labs, medications, and allergies. An external electronic interface was built with HL7 Fast Healthcare Interoperability Resource (FHIR) specifications to communicate with the EHR system, ensuring data field pre-populate the HINGE templates and the clinical free-text notes generated in HINGE are available in the EHR for billing and continuity of care purposes.
- Integration with Treatment Planning System (TPS): HINGE required the ability to import DICOM-RT data from any TPS conforming to the Integrating the Healthcare Enterprise - Radiation Oncology (IHE-RO) profiles. Standardizing target and organ at risk (OAR) nomenclatures, prescription formatting, and dose-volume histogram metrics across disease sites was a challenge. HINGE adopted the TG-263 naming convention for target structures and suggested equivalent TG-263 names for OARs in RT-Structure Sets from the TPS to facilitate consistent data mapping. Machine learning techniques were also explored to automate the relabeling process.
- Integration with Treatment Management System (TMS): Existing TMS lacked utilities necessary for export of treatment summary information that include the delivered dose details, treatment modality, technique dates of treatment to track the progress of treatment. HINGE developed a Docker container, HINGE-Broker, to access the TMS database and retrieve discrete elements such as patient demographics, prescribed prescriptions, delivered dose, and clinical notes using an FHIR interface. Changes in TMS software and different versions deployed across healthcare systems were addressed by supporting multiple methods of data access and using Word templates with tagged fields to extract discrete treatment data.
- Lack of radiation oncology-specific ontologies: Currently, there is a paucity of ontologies with radiation oncology-specific terms, making it difficult to establish standardized data definitions. To address this issue, HINGE deployed disease-specific templates for data entry and viewing, enabling radiation oncologists to enter relevant clinical information in a discrete and structured manner. HINGE also mapped the structured data to standard ontologies, ensuring interoperability and facilitating data aggregation and analysis (reported in section 5).
- Template usability and user experience: The design of the templates needed to prioritize usability and provide a smooth user experience for radiation oncologists. Physicians should find it intuitive and efficient to enter the required data elements within the templates with fields getting auto calculated (e.g., NCCN risk score based on staging, Gleason score and PSA values), auto populated from a previously entered note template and having the clinical narrative note built from the discrete data

entry and pushed in the EHR. Balancing the need for comprehensive data capture with ease of use presented a challenge, requiring iterative feedback loops and usability testing to refine the templates based on physician input.

With the help of HINGE, quality managers will be able to grade every treatment against established rubric of nationwide norms of outcome, toxicity, and treatment delivery. Some additional benefits include reducing the burden of basic data collection for quality analysis and creating a single-point capture of source data, which protects data integrity by eliminating manual transcription of data from multiple sources, provide better data traceability and provenance, while reducing the need for data queries, data cleaning, and source data verification — processes that within themselves hold the potential for errors. Furthermore, data collected in the HINGE can be used to create decision support models in the clinical systems that enable clinicians to improve the quality and safety of care rendered to the patients. With the increasing emphasis on delivery of value-based health care, the HINGE system can not only quantify value and quality of care but also aggregate outcome data using standard templates / data elements. An alternative approach to collecting data for quality surveillance and outcome research is to leverage natural language processing (NLP) for the extraction of discrete data from unstructured clinical documentation. However, there are several challenges with this approach. The free-text clinical notes have many different taxonomies, vocabularies, terms, or abbreviations that are often used by clinicians since there are currently no standards that are universally adopted in radiation oncology domain. The lack of standardization of information presented in the free text notes makes traditional NLP solutions difficult to implement. Syed et al. [27] recently reported on an integrated Machine Learning/NLP model using the fast Text algorithm [28] for standardizing the organs-at-risk names in the DICOM RT structure set files with the TG-263 specified standard names. The results for prostate and lung datasets reported high F1 scores on OAR names but low scores on tumor/target names due to a wide variability of non-standard names utilized for targets. In many cases, even when presented a consistent vocabulary/taxonomy, it is challenging for the NLP algorithm to decern information since much of the clinical meaning in free-text blobs is context based and it requires specific decision tree logics with multiple expression values to extract a single data element.

Another key feature of the next release of software will be the integration of Patient Reported Outcomes (PRO). We plan to deploy an infrastructure with public patient facing Web-based tools to capture longitudinal PRO data within HINGE to facilitate earlier interventions, rapid symptom management, and track patient reported quality of life assessments. A similar public facing Web-based tool will be deployed to collect radiotherapy treatment data from community radiation oncology providers that are currently treating over 60% of veteran cancer patients. Such a strategy will allow us to aggregate radiotherapy data for over 45000 veterans treated annually in the community and at 41 VHA sites. This has the potential for big data outcome research in radiation oncology and high-quality continuity of care. Finally, the development of future versions of HINGE software will be coordinated with the medical and surgical oncology programs to ensure harmonization of clinical workflow templates amongst all cancer care specialties.

For big data and smart healthcare techniques to succeed in medicine, it is imperative that all stakeholders – physicians, physicists, nurses, clerks, and commercial vendors work together on how and what data needs to be collected. The funding agencies such as National Institute of Health and National Cancer Institute should direct their resources to support the work around integrating the clinical practice with automating and streamlining clinical workflow around structured data collection methodologies and

define clinically meaningful measures of care rendered to our patients. The process used to create the HINGE database and model can be replicated for all domains of medicine where each domain is responsible to define their own workflow templates, clinical measures and data analysis tools that can be used as feedback to the practice for quality improvement.

References:

1. Spasić I, Livsey J, Keane JA, Nenadić G. Text mining of cancer-related information: review of current status and future directions. Int J Med Inform. 2014 Sep;83(9):605-23. doi: 10.1016/j.ijmedinf.2014.06.009. Epub 2014 Jun 24. PMID: 25008281.

2. The Clinical Research Center; a vital part of the ACR mission, Fleishon HB, Wald C, Korn R, Rosenthal S, Fredericks N. J Am Coll Radiol. 2011 Jun;8(6):422-7. PMID:21636057

3. Shifting the focus to practice quality improvement in radiation oncology. Crozier C, Erickson-Wittmann B, Movsas B, Owen J, Khalid N, Wilson JF. J Healthc Qual. 2011 Sep;33(5):49-57.PMID:23845133

4. Hagan M, Kapoor R, Michalski J, Palta J, et al. VA-Radiation Oncology Quality Surveillance Program. Int J Radiat Oncol Biol Phys. 2020; 106(3):639-647. doi:10.1016/j.ijrobp.2019.08.064

5. Caruthers D, Brame S, Palta JR, et al. Development and Implementation of Quality Measures for the Survey Based Performance Assessment of Radiation Therapy in the VA. Int J Radiat Oncol. 2017;99(2):E391-E392. doi:10.1016/j.ijrobp.2017.06.1539

6. Matuszak MM, Fuller CD, Yock TI, et al. Performance/outcomes data and physician process challenges for practical big data efforts in radiation oncology. Med Phys. 2018;45(10):e811-e819. doi:10.1002/mp.13136

7. Seer.cancer.gov. (2017). Surveillance, Epidemiology, and End Results Program. [online] Available at: http://seer.cancer.gov/ [Accessed 13 Nov. 2017].

8. Jagsi, R., Abrahamse, P., Hawley, S., Graff, J., Hamilton, A. and Katz, S. (2011). Underascertainment of radiotherapy receipt in Surveillance, Epidemiology, and End Results registry data. Cancer, 118(2), pp.333-341.

9. Mayo CS, Kessler ML, Eisbruch A, et al. The big data effort in radiation oncology: Data mining or data farming? Adv Radiat Oncol. 2016;1(4):260-271. doi:10.1016/j.adro.2016.10.001

10. Moran JM, Feng M, Benedetti LA, et al. Development of a model web-based system to support a statewide quality consortium in radiation oncology. Pract Radiat Oncol. 2017;7(3):e205-e213. doi:10.1016/j.prro.2016.10.002

11. Pasalic D, Reddy JP, Edwards T, Pan HY, Smith BD. Implementing an Electronic Data Capture System to Improve Clinical Workflow in a Large Academic Radiation Oncology Practice. JCO Clin Cancer Informatics. 2018;(2):1-12. doi:10.1200/CCI.18.00034

12. McNutt TR, Evans K, Wu B, et al. Oncospace: All Patients on Trial for Analysis of Outcomes, Toxicities, and IMRT Plan Quality. Int J Radiat Oncol • Biol • Phys. 2010;78(3):S486. doi:10.1016/j.ijrobp.2010.07.1139

13. Waddle MR, Kaleem T, Niazi SK, et al. Cost of Acute and Follow-Up Care in Patients with Pre-Existing Psychiatric Diagnoses Undergoing Radiation Therapy. Int J Radiat Oncol. 2017;99(5):1321. doi:10.1016/j.ijrobp.2017.09.023

14. Mayo C, Moran JM, Xiao Y, et al. AAPM Task Group 263: Tackling Standardization of Nomenclature for Radiation Therapy. Int J Radiat Oncol. 2015;93(3):E383-E384. doi:10.1016/j.ijrobp.2015.07.1525

15. Ctep.cancer.gov. (2017). Common Terminology Criteria for Adverse Events (CTCAE). [online] Available at: https://ctep.cancer.gov/protocoldevelopment/electronic_applications/ctc.htm [Accessed 15 Nov. 2017].

16. Basch, E., Pugh, S., Dueck, A., et.al (2017). Feasibility of Patient Reporting of Symptomatic Adverse Events via the Patient-Reported Outcomes Version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE) in a Chemoradiotherapy Cooperative Group Multicenter Clinical Trial. International Journal of Radiation Oncology*Biology*Physics, 98(2), pp.409-418.

17. AJCC. (2017). American Joint Committee on Cancer. [online] Available at: https://cancerstaging.org/Pages/default.aspx [Accessed 15 Nov. 2017].

18. Rengan R, Kapoor R, Palta J. et.al. Addressing connectivity issues: The Integrating the Healthcare Enterprise-Radiation Oncology (IHE-RO) initiative; Practical Radiation Oncology; VOLUME 1, ISSUE 4, P226-231, OCTOBER 01, 2011

19. Syed, Khajamoinuddin, et al. "Integrated natural language processing and machine learning models for standardizing radiotherapy structure names." Healthcare. Vol. 8. No. 2. Multidisciplinary Digital Publishing Institute, 2020.

20. Rhee, D., et al. "TG263-Net: A deep learning model for organs-at-risk nomenclature standardization." MEDICAL PHYSICS. Vol. 46. No. 6. 111 RIVER ST, HOBOKEN 07030-5774, NJ USA: WILEY, 2019.

21. Yang, Qiming, et al. "A novel deep learning framework for standardizing the label of OARs in CT." Workshop on Artificial Intelligence in Radiation Therapy. Springer, Cham, 2019.

22. Sleeman IV, William C., et al. "A Machine Learning method for relabeling arbitrary DICOM structure sets to TG-263 defined labels." Journal of Biomedical Informatics 109 (2020): 103527.

23. Sleeman IV, C., et al. “Relabeling Non-Standard to Standard Structure Names Using Geometric and Radiomic Information”. MEDICAL PHYSICS. Vol. 47. No. 6. 111 RIVER ST, HOBOKEN 07030-5774, NJ USA: WILEY, 2020.

24. Accountability Act. Health insurance portability and accountability act of 1996. Public Law. 1996;104:p. 191.

25. Metri P., Sarote G. Privacy issues and challenges in cloud computing. International Journal of Advanced Engineering and Technology. 2011;5(1):5–6.

26. Andress J. The Basics of Information Security: Understanding the Fundamentals of InfoSec in Theory and Practice. Boston, MA, USA: Syngress; 2014.

27. Syed K, Sleeman IV W, Ivey K, Hagan M, Palta J, Kapoor R, Ghosh P. Integrated natural language processing and machine learning models for standardizing radiotherapy structure names. InHealthcare 2020 Jun (Vol. 8, No. 2, p. 120). Multidisciplinary Digital Publishing Institute.

28. 20. Wu S, Manber U. Fast text searching: allowing errors. Communications of the ACM. 1992 Oct 1;35(10):83-91

29. Orthanc DICOM Server - Available at: https://www.orthanc-server.com/

5. Extract, Transform and Load (ETL) Clinical, Dosimetry and Treatment datasets into Internationally Standardized Semantic Interoperable Data Models

The adoption of electronic health records (EHRs) in patient's clinical managements is rapidly increasing in healthcare but the use of data from EHR in clinical research is lagging. The utilization of patient-specific clinical data available in EHR has the potential to accelerate learning and bring value in several key topics of research including comparative effectiveness research, cohort identification for clinical trial matching and quality measure analysis. [1, 2]. However, there is an inherent lack of interest in the use of data from the EHR for research purposes since the EHR was never designed for research. The modern EHR technology has been optimized for capturing health details for clinical record keeping, patient management, scheduling, ordering, and capturing data from external sources such as laboratories, diagnostic imaging, and capturing encounter information for billing purposes [3]. Many data elements collected in routine clinical care, which are critical for oncologic care, are not collected as structured data elements nor with the same defined rigor as those in clinical trials [4, 5]. In this chapter, we set out to contribute to the advancement of the science of Learning Health Systems (LHS) by presenting a detailed description of the technical characteristics and infrastructure that were employed to design an LHS specifically with a knowledge graph approach.

5.1 What are ontologies and why are they important to us?

In the past few years, there has been a focus on obtaining high-quality data regarding patients, their treatments, and the outcomes in healthcare. Various advancements and growing interest in personalized medicine [6], genetic profiling [7][8], and machine learning [9][10] highlight the significance of having substantial amounts of reliable data. These areas of focus rely heavily on extensive, high-quality data to drive advancements and improve healthcare practices. Despite the open availability of many important databases and knowledge bases, biomedical researchers still face severe logistical and technical difficulties when integrating, analyzing, and visualizing heterogeneous data and knowledge from these diverse and isolated sources. These tasks pose a steep learning curve for most biomedical researchers. Researchers need to be aware of the sources where the data and knowledge relevant to their research exist. Depending on the availability and the accessibility, biomedical researchers need exhaustive computational resources and extensive programming skills to query and explore the data and knowledge sources. The heterogeneity across these sources, in terms of formats, syntaxes, notations and schemas, severely stymies the systematic consumption of data and knowledge stored in these sources. The biomedical researcher ends up learning multiple systems, configurations, and access requirements, significantly increasing the complexity and time of scientific research.

Given these considerations, it is crucial to establish a system that effectively organizes and standardizes our knowledge, enabling seamless sharing and adaptation to new information. Ontologies have emerged as a promising approach, especially in light of the extensive digitization of information [11]. The term "ontology" originates from the Greek word denoting the study of existence or the essence of things. While it has long been employed in philosophical contexts, its application in the realm of information science has yielded a more contemporary and comprehensive definition. In essence, modern ontologies serve as computer-interpretable descriptions of human knowledge pertaining to specific domains or areas of the

world. Ontologies serve as representations of universal concepts, defined classes, and relationships. In essence, with the help of these universal concepts, we can capture the fundamental characteristics of entities in reality, establishing a shared essence or "natural kind." By delving into the meaning of underlying data, ontologies provide robust semantic frameworks for specific knowledge domains. To achieve this, ontologies employ well-crafted definitions for the entities they encompass. These definitions can be both human-readable and machine-readable [12]. Human-readable definitions take the form of descriptive text that explains the denotation of a given term, while machine-readable definitions consist of formal, logical axioms.

Over time, collaborative efforts have emerged to develop tools and repositories for ontologies, offering invaluable resources to the biomedical community. One such resource is BioPortal, which is supported by the National Center for Biomedical Ontology. BioPortal serves as an open repository for biomedical ontologies, enabling access to ontologies developed in various formats such as OWL, RDF-Triple, RDF-XML, and OBO [13]. Through a user-friendly web interface, individuals can explore ontologies, add notes, provide reviews, and examine mappings between different ontologies. BioPortal currently hosts over 800 ontologies, encompassing diverse resources such as the Proteomics Standards Initiative, the OBO library, and the Semantic Type Ontology of the Unified Medical Language System (UMLS) [14]. With over 100,000 terms, the NCI Thesaurus (NCIT) includes wide coverage of cancer terms as well as mapping with external terminologies. NCIT is a product of NCI Enterprise Vocabulary Services (EVS) and its vocabularies consists of public information on cancer, definitions, synonyms, and other information on almost ten thousand cancers and related diseases, seventeen thousand single agents and related substances, as well as other topics that associated with cancer. This includes comprehensive details such as definitions, synonyms, and other pertinent information that aids in understanding and describing specific cancer types. By encompassing a wide range of cancer classifications, the NCIT caters to diverse research needs and facilitates effective collaboration across different domains within the field of oncology.

5.2 Standardization Challenges and Considerations in Radiation Oncology Data Sharing

The treatment of cancer requires a multidisciplinary approach which typically includes providers who represent disparate clinical disciplines. The care coordination between all the provider teams is critically important and is done via sharing clinical notes in the EHR. Furthermore, in order to improve the quality of care, to evaluate cancer therapy outcomes, and perform research in an automated fashion, abstracting data from multiple sources is required. Due to the lack of standardization in oncology the ability to perform the care coordination becomes difficult and affects patient safety. The use of a standard ontology, taxonomy and data dictionary is the need of the hour if we would like to realize the true potential of Big Data analytics in the oncology domain.

One of the key areas of work by experts in the field has been the standardization of nomenclature labels applied to cancer targets, normal tissues, and treatment planning regions in the patient's body. This standardization is an important precursor to any pooling of the data for analysis, treatment plan evaluation, population-based studies, and clinical trials. A task group (TG-263) was formed at American Association of Physicists in Medicine (AAPM) (medical physicists' professional organization) to develop consensus recommendations on nomenclature used in radiation oncology. The task group reviewed existing ontologies from the Foundation Model of Anatomy (FMA) in radiation oncology domain. This is an open source and well-maintained ontology by the University of Washington, Seattle. FMA provides a detailed list of structure labels that phenotypically represent the human body. Each of the structure labels is associated with a unique identifier code called FMAID. Many of the structure labels recommended by

this group came from the FMA ontology. There were many structure labels that were only specific to radiation oncology domain (e.g., SpinalCord_PRV) that were not included in FMA but the concepts labels from FMA were used to define these specific labels.

The Systematized Nomenclature of Medicine-Clinical Terms SNOMED-CT is a standard terminology that has developed and defined clinical concepts that are used to improve the consistent recording and utilization of information in the electronic health records. SNOMED-CT terminologies and clinical concepts are built to enable clinical information to be recorded in a consistent manner with the purpose to facilitate evidence-based healthcare and create a link between clinical records and clinical guidelines, enhancing the quality of care rendered to patients. It also enables the use of clinical decision support systems that can potentially check the clinical record for discrete minable information and provide real-time clinical advice. In addition, it supports the sharing of appropriate discrete clinical information with other clinical providers involved in the care of the patients who might be utilizing other electronic health record systems, thus allowing standardizing flow and processing of this information for all providers. FMA and SNOMED-CT concept labels are linked to unique numerical codes but there were many structure labels that are utilized in radiation oncology and were absent from SNOMED-CT and FMA defined ontologies. It was concluded that both FMA and SNOMED-CT are great ontologies, but they did not meet the requirements for anatomical, target structure labels that were required for radiation oncology. Both these ontologies were weakly associated with the task groups recommendations of the anatomical and target structure labels. SNOMED-CT codes are added for meeting the requirements for the structure labels but the vendor platforms where these labels are recorded are yet to implement the export of these SNOMED based codes based on the associated target or anatomical labels.

DICOM (Digital Imaging and Communication in Medicine) [16] is an international standard that is utilized for storing and communicating medical imaging and related data. Its predecessor, the ACR-NEMA (American College of Radiology–National Electrical Manufacturers Association) standard, was published in 1982, followed by a second version, ACR-NEMA 2.0, in 1988, neither of which addressed computer networking issues. The current version of this standard was introduced in 1992. The use of this standard has enabled the integration of medical imaging and radiation oncology treatment planning and delivery devices from multiple manufacturers. This standard is widely adopted by radiology, cardiology, ophthalmology, pathology, dentistry, oncology, and hospital infrastructure support vendors. Since radiation oncology is very imaging intensive, it was the first domain to be introduced in DICOM after radiology and there were five radiation oncology specific information object definitions (IODs) that were introduced in 1997. These information objects include RT Structure Set, RT Plan, RT Dose, RT Image, and RT Treatment Record, which is further divided into RT Beams Treatment Record, RT Brachy Treatment Record, and RT Treatment Summary Record. These DICOM objects are by far the only readily exportable and sharable records in Radiation Oncology domain so far.

Another important aspect that is impeding data sharing is the inability for clinicians to electronically share radiotherapy treatment summary information from radiation oncology information systems to electronic health record systems (EHRs) that are used by healthcare systems for care coordination amongst multiple clinical disciplines. There is high variation in documentation of radiation therapy–specific data and sharing between information systems is often done manually rather than automatically, leading to a potential breakdown of efficiency and accuracy. For the past two years, IHE-RO Technical Committee has been working to develop a FHIR (Fast Health Interoperability Resources) based interoperability profile called Exchange of Radiotherapy Summary (XRTS) to seamlessly bridge this critical communication gap and make

the minimal treatment summary information readily available to reuse and share across systems [17]. For this effort, IHE-RO has partnered with CodeX, an HL7 FHIR Accelerator, and is contributing to the Radiation Therapy Treatment Data (RTTD) use case (Radiation Therapy Treatment Data for Cancer - CodeX - Confluence (hl7.org)). The RTTD project team consists of AAPM, ASTRO with its minimum data elements initiative (https://www.astro.org/Patient-Care-and-Research/Clinical-Practice-Statements/Minimum-Data-Element), clinical subject matter expert physicists, physicians, Radiation Oncology and EHR vendors to build FHIR-based data communication protocols. HL7 FHIR is a next generation standards framework created by HL7 which combines the best features of HL7's v2, v3 and CDA product lines while leveraging the latest web standards and applying a tight focus on implementation. HL7 FHIR is a highly versatile standard that finds extensive application in diverse scenarios, such as facilitating EHR-based data sharing, enabling seamless communication between servers within large healthcare institutions, and supporting the exchange of clinical context-based information. The HINGE development and its integration of electronic health records (EHRs) have yielded significant benefits, particularly in the development of the IHE-RO and CodeX interoperability profiles. These profiles have played a crucial role in extracting valuable treatment summary information from RO-Treatment Management Systems and seamlessly integrating it with EHRs using FHIR specifications. As the chair of the IHE-RO workgroup for the past four years, I have been leading the charge with standardization efforts aimed at creating a vendor-independent solution. This solution has been successfully implemented using the HINGE platform. The process involved defining the data elements with their respective SNOMED codes, establishing interface specifications, and configuring the clinical workflow in the Treatment management software. Rigorous testing was conducted using clinical cases of increasing complexity to ensure the efficacy of these interfaces. Figure 12 illustrates the comprehensive list of data elements defined in the treatment summary data exchange, encompassing details at the RT course level (Radiotherapy Course summary), individual treatment phase level (Radiotherapy Treated Phase), and their corresponding plan summary data elements. These interface specifications have been effectively implemented in both the Varian Aria software and the HINGE software. Successful testing was carried out at the IHE-RO 2023 XRTS workshop, where discrete data was seamlessly gathered and auto-populated into our On-treatment visit and End of treatment summary templates within the HINGE platform.

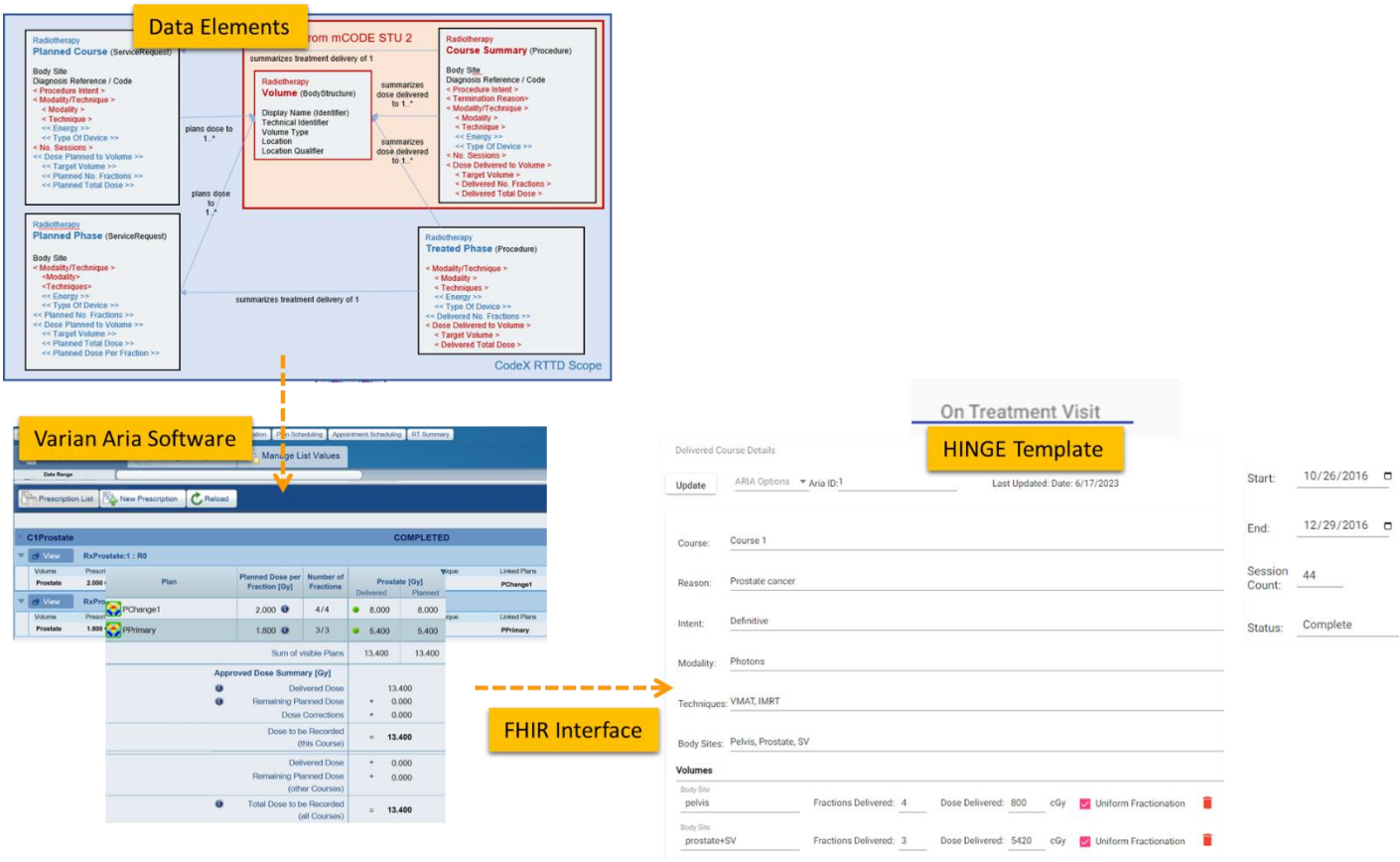


Figure 12: Consensus treatment summary data elements defined with CodeX and IHE-RO effort is implemented in the Varian Aria software and interfaced to the HINGE software via the FHIR interfaces. These treatment summary elements are then auto-populated in the On-treatment visit note templates in the HINGE software and saves time for the physicians by avoiding manual transcription of this data in HINGE.

Another effort is underway to address the lack of data standardization in electronic health records (EHRs), Radiation Oncology Information Systems (ROIS), treatment planning systems (TPSs), and other cancer care and outcomes databases. This effort aims to create a standardized ontology for clinical data, social determinants of health (SDOH), and other radiation oncology concepts. The American Association of Physicists in Medicine's Big Data Science Committee (BDSC) engaged stakeholders to optimize the integration of diverse perspectives and develop the Operational Ontology for Oncology (O3) [18]. O3 includes key elements, attributes, value sets, and relationships that are of clinical significance and likely to be available in EHRs. Recommendations are provided for different stakeholders, such as device manufacturers, clinical care centers, researchers, and professional societies, on how to best use and develop O3. Implementing these recommendations will facilitate the aggregation of information, creating large and representative datasets that adhere to FAIR principles (findable, accessible, interoperable, and reusable). This effort emphasizes the utilization of comprehensive datasets and advanced analytic techniques, including artificial intelligence (AI), to transform patient management and improve outcomes by leveraging the increased access to information derived from larger datasets.

5.3 Our Approach with the Extract, Transform and Load Pipeline

We created a data pipeline from HINGE to export discrete clinical data in JSON based format. These data are then fed to the Extract, Transform and Load (ETL) processor. An overview of the data pipeline is shown in figure 13. ETL is a three-step process where the data is first extracted, transformed (cleaned, formatted), and loaded into an output RO-Clinical Data Warehouse (RO-CDW) repository. The RO-CDW relational database structure comprises of 15 data tables with primary and foreign keys that allow for interrelationships to be established amongst various data tables storing the specific logical information. Since HINGE templates do not function as case report forms and they are formatted based on an operational data structure, data cleaning process is performed with some basic data preprocessing, including cleaning, and checking for redundancy in the dataset, ignoring null values, making sure each data element has its supporting data elements populated in the dataset. As there are several types of datasets, each dataset requires a different type of cleaning. Therefore, multiple scripts for data cleaning have been prepared. The following outlines some of the checks that have been performed using the cleaning scripts.

- Data type validation: We verified whether the column values were in the correct data types (e.g., integer, string, float). For instance, the "Performance Status Value" column in a patient record should be an integer value.
- Cross-field consistency check: Some fields require other column values to validate their content. For example, the "Radiotherapy Treatment Start Date" should not be earlier than the "Date of Diagnosis." We conducted a cross-field validation check to ensure that such conditions were met.
- Mandatory element check: Certain columns in the input data file cannot be empty, such as "Patient ID Number" and "RT Course ID" in the dataset. We performed a mandatory field check to ensure that these fields were properly filled.
- Range validation: This check ensures that the values fall within an acceptable range. For example, the "Marital Status" column should contain values between 1 to 9.
- Format check: We verified the format of data values to ensure that they were consistent with the expected year-month-day (YYYYMMDD) format.

The main purpose of this step is to ensure that the dataset is of high quality and fidelity when loaded in RO-CDW. In the data loading process, we have written SQL and .Net-based scripts to transform the data into RO-CDW compatible schema and load them into Microsoft's SQL Server 2016 database. When the data are populated, unique identifiers are assigned to each data table entry and interrelationships are maintained within the tables so that the investigators can use query tools to query and retrieve the data, identify patient cohorts, and analyze the data.

We have deployed a free, open source and light weight DICOM server known as Orthanc [19] to collect DICOM-RT datasets from any commercial treatment planning system. Orthanc is a simple, yet powerful standalone DICOM server designed to support research, and query/retrieve functionality of DICOM datasets. Orthanc provides a RESTful API that makes it possible to program using any computer language where DICOM tags stored in the datasets can be downloaded in a JSON format. We used the python plug-in to connect with the Orthanc database to extract the relevant tag data from the DICOM-RT files. Orthanc was able to seamlessly connect with the Varian Eclipse planning system with the DICOM DIMSE C-STORE protocol [20]. Since the TPS conforms to the specifications listed under the Integrating the Healthcare Enterprise – Radiation Oncology (IHE-RO) profile, the DICOM-RT datasets contained all the relevant tags

that were required to extract data. One of the major challenges with examining patients' DICOM-RT data is the lack of standardized organs at risk (OAR) and target names, and ambiguity regarding dose-volume histogram metrics, and multiple prescriptions mentioned across several treatment techniques. With the goal of overcoming these challenges, the AAPM TG 263 initiative has published their recommendations on OAR and target nomenclature. The ETL user interface deploys this standardized nomenclature and requires the importer of the data to match the deemed OARs with their corresponding standard OAR and target names. In addition, this program also suggests a matching name based on an automated process of relabeling using our published techniques (OAR labels [21], radiomics features [22], and geometric information [23]). We find that these automated approaches provide an acceptable accuracy over the standard prostate and lung structure types. In order to gather the dose volume histogram data from the DICOM-RT dose and structure set files, we have deployed a DICOM-RT dosimetry parser software. If the DICOM-RT dose file exported by the treatment planning system (TPS) contains DVH information, we utilize it. However, if the file lacks this information, we employ our dosimetry parser software to calculate the DVH values from the dose and structure set volume information.

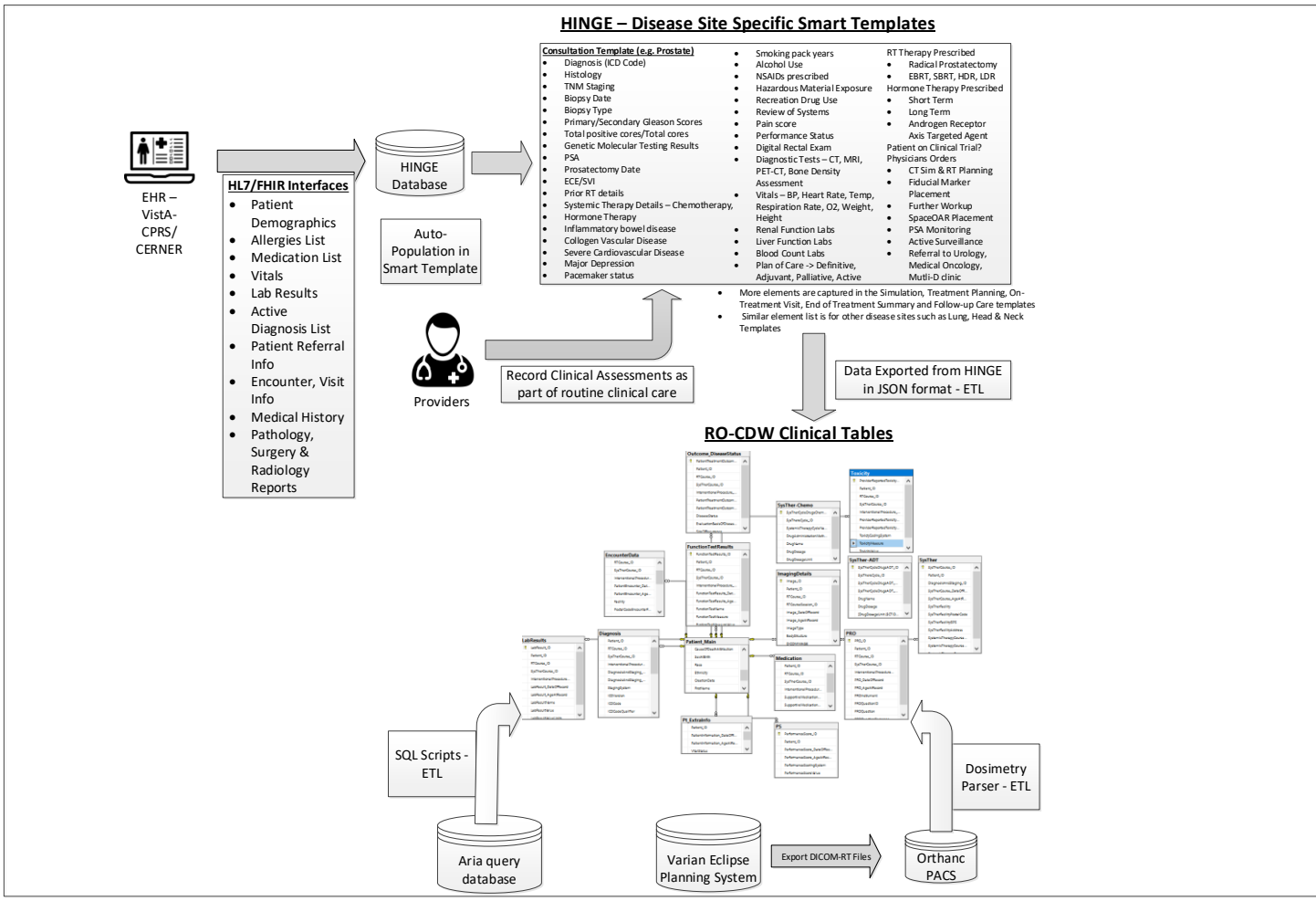


Figure 13: Overview of the data pipeline to gather clinical data into the RO-Clinical Data

As part of this pipeline, we have built HL7/FHIR interfaces between the EHR system and HINGE database to gather pertinent information from the patient's chart. This data is stored in the HINGE database and

used to auto-populate disease sites specific smart templates that depict the clinical workflow from initial consultation to follow-up care. The providers record their clinical assessments in these templates as part of their routine clinical care. Once the templates are finalized and signed by the providers in HINGE, the data is exported in JSON format and using an ETL process, we can load the data in our RO-Clinical Data Warehouse relational SQL database. Additionally, we use SQL stored procedures to extract, transform and load data from the Varian Aria data tables and extraction of dosimetry DVH curves to our RO-CDW.

5.4 Mapping data to standardized terminology, data dictionary, ontologies, and use of Semantic Web technologies

For data to be interoperable, sharable outside the single hospital environment and usable for the various requirements of an LHS, the use of standardized terminology and data dictionary is a key requirement. Specifically, clinical data should be transformed following FAIR (Findable, Accessible, Interoperable, and Reusable) principles [24]. An ontology describes a domain of classes and is defined as a conceptual model of knowledge representation. The use of Ontologies and Semantic Web technologies play a key role in transforming healthcare data with the FAIR principles. The use of ontologies enables the sharing of information between disparate systems within the multiple clinical domains. An ontology acts as a layer above the standardized data dictionary and terminology where explicit relationships, i.e., predicates, are established between unique entities. Ontologies provide formal definitions of the clinical concepts used in the data sources and renders the implicit meaning of the relationships among the different vocabulary and terminologies of the data sources explicitly. For example, it can be determined if two classes and data items found in different clinical databases are equivalent or if one is a subset of another. Semantic level information extraction and query are possible only with the use of ontology-based concepts of data mapping.

A rapid way to look for new information on the internet is to use a search engine such as Google. These search engines return a list of suggested web pages devoid of context and semantics and require human interpretation to find useful information. Semantic Web is a core technology that is utilized in order to organize and search for specific contextual information on the web. Semantic Web, which is also known as Web 3.0 is an extension of the current World Wide Web (WWW) via a set of W3C data standards [25] with a goal to make internet data machine readable instead of human readable. For automatic processing of information by computers, the Semantic Web extensions enable data (text, meta data on images, videos, etc.) to be represented with well-defined data structures and terminologies. To enable the encoding of semantics with the data, web technologies such as Resource Description Framework (RDF), Web Ontology Language (OWL) and SPARQL Protocol and RDF Query Language are used. RDF (Resource Description Framework) is a standard for sharing data on the web.

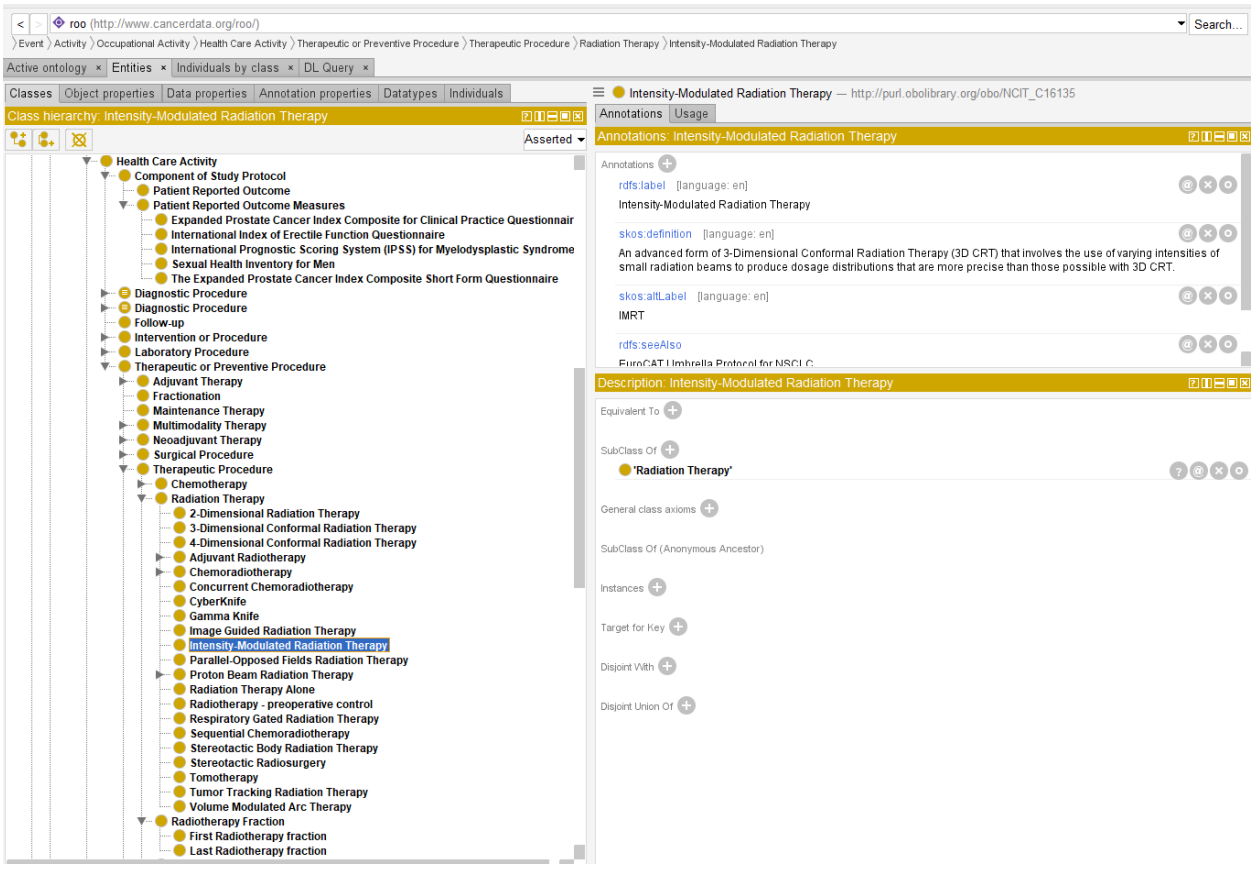


Figure 14: Screen capture of the ontology editor tool Protégé for inspecting and adding the key classes, properties, and relationships to the Radiation Oncology Ontology (ROO) based on classes defined in the NCI Thesaurus and SNOMED ontologies that align with our RO-CDW data elements.

We utilized an existing ontology known as Radiation Oncology Ontology (ROO) [26] available on the NCBO Bioportal website [27]**.** The main role of ROO is to define a broad coverage of main concepts used in the radiation oncology domain. The ROO currently consists of 1,183 classes with 211 predicates that are used to establish relationships between these classes. Upon inspection of this ontology, we noticed that the collection of classes and properties were missing some critical clinical elements such as smoking history, CTCAE v5 toxicity scores, diagnostic procedures such as Gleason scores, PSA levels, patient reported outcome measures, KPS performance status scales and radiation treatment modality. We utilized the ontology editor tool Protégé [28] for adding these key classes and properties in the updated ontology file (Figure 14). We reused entries from other published ontologies such as the National Cancer Institute Thesaurus (NCIT) [29], International Classification of Disease, version 10 (ICD-10) [30], Dbpedia [31] ontologies. We carefully studied each data element definition in ROO and NCIT to make sure that the selected codes are adequately aligning with the data elements captured in our RO-CDW database. We added 216 classes (categories defined in Table 1) with 19 predicate elements to the ROO. With over 100,000 terms, the NCI Thesaurus (NCIT) includes wide coverage of cancer terms as well as mapping with external terminologies. NCIT is a product of NCI Enterprise Vocabulary Services (EVS) and its vocabularies consists of public information on cancer, definitions, synonyms, and other information on almost ten thousand cancers and related diseases, seventeen thousand single agents and related substances, as well as other topics that associated with cancer.

Categories	Number of classes
Race, Ethnicity	5
Tobacco Use	4
Blood Pressure + Vitals	3
Laboratory tests (e.g., Creatinine, GFR, etc.)	20
Prostate specific diagnostic tests (e.g., Gleason score, PSA, etc.)	10
Patient reported outcome	8
CTCAE v5	152
Therapeutic Procedures (e.g., Immunotherapy, Targeted Therapy, etc.)	6
Radiation Treatment Modality (e.g., photon, electron, proton, etc.)	7
Units (cGy)	1

Table 1: Additional classes added to the Radiation Oncology Ontology (ROO) and used for mapping with our dataset

In order to leverage and validate the ontology we had defined; we undertook a meticulous mapping process that involved integrating our data stored in a clinical data warehouse relational database with the concepts and relationships outlined in the ontology. This mapping process linked each component (column headers, values) of the SQL relational database to its corresponding clinical concept (classes, relationships, and properties) in the ontology. To perform the mapping, the SQL database tables are analyzed and matched with the relevant concepts and properties in the ontology. This can be achieved by identifying the appropriate classes and relationships that best represent the data elements from the SQL relational database. For example, if the SQL relational table provides information about a patient's smoking history, the mapping process would identify the corresponding class or property in the ontology that represents smoking history. A correspondence between the table columns in the relational database and ontology entities was established. An example of this mapping is shown in Figure 15A. We used the D2RQ mapping language to map the relational database schema to RDF ontology-based vocabulary. The D2RQ mapping is a direct mapping method where all the data and individual columns from the relational database is directly mapped to the ontology-based structure and translation, or logic is employed for derived data fields where direct mapping is not feasible. An example screenshot of the schema mapping file used by the D2RQ platform is shown in Figure 15B This mapping language is executed by the D2RQ platform that connects to SQL database, reads the schema, perform the mapping, and generates the output file in turtle syntax. Each SQL table column name is mapped to its corresponding class using the d2rq:ClassMap command. These classes are also mapped to existing ontology-based concept codes such as NCIT:C48720 for T1 staging. In order to define the relationships between two classes, d2rq:refersToClassMap command is used. The properties of the different classes are defined using the d2rq:PropertyBridge command. To ensure machine readability and facilitate interoperability with other RDF databases, we assigned Unique Resource Identifiers (URIs) to each entity within the ontology. These URIs serve as unique identifiers and enable seamless linking and integration of data across different

databases and systems. Table 2 provide a listing of key data elements that are used to map between our Clinical Data Warehouse relational database and ontology-based graph database include the standard NCIT and Radiation Oncology Ontology (ROO) codes used for the mapping.

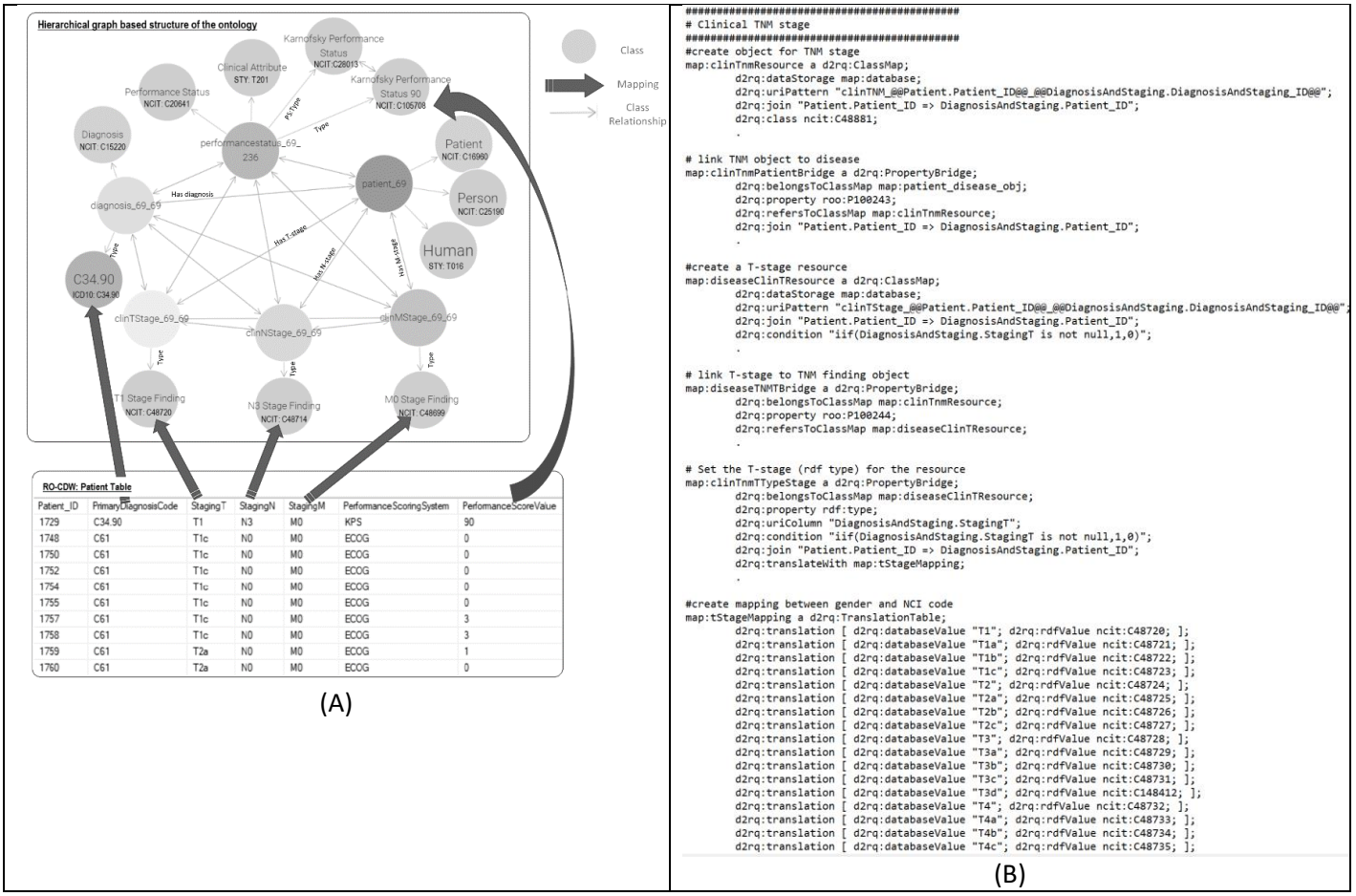


Figure 15 (A): Overview of the data mapping between the relational RO-CDW database and the hierarchical graph-based structure based on the defined ontology

The top rectangle displays an example of the various classes of the ontology and their relationships including the NCI Thesaurus, ICD-10 codes. The bottom rectangle shows the relational database table and the solid arrows between the top and bottom rectangles display the data mapping.

(B): An example screenshot of the mapping file used by the D2RQ platform to perform the mapping between the SQL table data elements and values to ontology formatted codes. The first block (TNM stage) defines the mapping for each TNM value entry in the SQL database. The "ClassMap" property in the D2RQ script defines the mapping between the column name in the SQL relational database and the corresponding class in the ontology. The next blocks in the example define the "PropertyBridge" which is used in the D2RQ script to define the relationship between the different classes. In the example shown above, the "PropertyBridge" between the TNM class to the T Stage classes (e.g., T1 or T1a, etc.).

Table 2		
Category	**Attribute**	**Codes/Datatypes**
Patient Details	Patient ID	NCIT: C16960
	Race	NCIT: C17049
	Ethnicity	NCIT: C16564
	Date of Birth	NCIT: C68615
	Date of Death	NCIT: C70810
	Sex at Birth	Male: NCIT: C16576 Female: NCIT: C20197
	Cause of Death	NCIT: C99531
Other Patient Details	Vital Status	NCIT: C25717 Alive: NCIT: C37987 Deceased: NCIT: C28554
	Tobacco Use History	NCIT: C181760 Smoker: NCIT: C67147 Former Smoker: C67148
	Smoking Pack Years	NCIT: 127063
	Patient Height	NCIT: C25347
	Patient Weight	NCIT: 25208
	Blood Pressure	NCIT: C54706
	Heart Rate	NCIT: C49677
	Temperature	NCIT: C25206
Diagnosis and Staging	Staging System	
	Diagnosis	NCIT: C15220
	ICD Version	ICD:10
	ICD Code	ICD 10 codes e.g., C61
	Histology	Adenocarcinoma: NCIT: C2852, Ductal Carcinoma: NCIT: C36858, etc.
	Clinical TNM Staging	NCIT: C48881
	Pathological TNM Staging	NCIT: C48739
	Staging-T	T1: NCIT: C48720, T2, etc.
	Staging-N	N0: NCIT: C48705, N1, etc.
	Staging-M	Mx: NCIT: C48704, M0, etc.
	Biopsy obtained via imaging	NCIT: C17369
Prostate Specific Elements	Had Prostatectomy	NCIT: 15307
	Prostatectomy Margin Status	NCIT: 123560
	Primary Gleason Score	NCIT: C48603
	Secondary Gleason Score	NCIT: 48604

	Tertiary Gleason Score	NCIT: 48605
	Total Number of Prostate Tissue Cores	NCIT: 148277
	Number of Positive Cores	NCIT: 148278
	Prostate Specific Antigen Level	NCIT: 124827
Patient Reported Outcome	Patient Reported Outcome	NCIT: 95401
	PRO Instruments	EPIC-26: NCIT: C127367, AUA IPSS: NCIT: C84350 IIEF: NCIT: C103521 EPIC-CP: NCIT: C127368 SHIM: NCIT: C138113
	PRO Question Response	Integer
Performance Score	Scoring System	KPS: NCIT: C28013 ECOG: NCIT: C105721 ZUBROD: NCIT: C25400
	Performance Score Value	ECOG 1: NCIT: C105723, KPS 10: NCIT: C105718, etc.
Toxicity Reporting	Coding System	CTCAE v5: NCIT: C49704 RTOG: NCIT: C19778
	Toxicity Measure	Erectile dysfunction: NCIT: C55615, Fatigue: NCIT: C146753, etc.
	Toxicity Grade	Erectile dysfunction Grade 1: NCIT: C55616, Fatigue Grade 1: NCIT: C55292, etc.
Treatment Procedures	Therapy Included in the Treatment Procedure	Radiation Therapy: NCIT: C15313, Systemic Therapy: NCIT: C15698, Surgical Procedure: NCIT: C15329, Hormone Therapy: NCIT: C15445
	Agents used - Hormone Therapy	String
	Drugs Used - Chemotherapy	String
RT Treatment Course	Radiation Treatment Modality	Photon: NCIT: C88112, Electron: NCIT: C40428, Proton: NCIT: C17024, etc.
	Radiation Treatment Technique	IMRT: NCIT: C16135, SBRT: NCIT: C118286, 3D CRT: NCIT: C116035, etc.
	Target Volume	PTV: NCIT: C82606, CTV: NCIT: C112912, GTV: NCIT: C112913, etc.
	Prescribed Radiation Dose	ROO: C100013 - Float
	Radiation Dose Units	cGy: NCIT: C64693, Gy: NCIT: C18063
	Number of prescribed fractions	NCIT: C15654 - Float
	Organs at Risk - Structure	Bladder: NCIT: C12414, Rectum: NCIT: C12390, Heart: NCIT: 12727, etc.
	Delivered Radiation Dose	ROO: C100013 - Float
	Number of delivered fractions	NCIT: C15654 - Float
	Start date of RT Course	Date
	End date of RT Course	Date
Dose Volume Histogram	DVH Constraint	NCIT: C112816 - String
	DVH Value	Float

	DVH Value Units	Gy: NCIT: C18063 cGy: NCIT: C64693 %: UO: 0000187
NCIT: National Cancer Institute Thesaurus, ROO: Radiation Oncology Ontology, UO: Units Ontology, ICD-10: International Classification of Diseases, Version 10		

Table 2: Key data elements that are used to map between our Clinical Data Warehouse relational database and ontology-based graph database. This table shows some examples of the codes used for the purpose of this mapping.

5.5 Importing Data in Knowledge based Graph-based database

The output file from the D2RQ mapping step is in Terse RDF Triple Language (turtle) syntax. This syntax is used for representing data in the semantic triples, which comprise a subject, predicate, and object. Each item in the triple is expressed as a Web URI. In order to search data from such formatted datasets, the dataset is imported in Knowledge graph databases. RDF database, also called as Triplestore, is a type of graph database that stores RDF triples. The knowledge on the subject is represented in these triple formats consisting of subject, predicate, and object. RDF knowledge graph can also be defined as labeled multi-diagraphs which consists of a set of nodes which could be URIs or literals containing raw data, and the edges between these nodes represent the predicates [32]. The language used to reach data is called SPARQL — Query Language for RDF. It contains ontologies that are schema models of the database. Although SPARQL adopts various structures of SQL query language, SPARQL uses navigational-based approaches on the RDG graphs to query the data which is quite different than the table join based storage and retrieval methods adopted in relational databases. In our work, we utilized the Ontotext GraphDB software [33] as our RDF store and SPARQL endpoint.

5.6 Validating the Pipeline with Real-World Datasets

With the aim to test out the data pipeline and infrastructure, we used our clinical database that has 1660 patient clinical and dosimetry records. These records are from patients treated with radiotherapy for prostate, non-small cell lung cancer and small cell lung cancer disease. There are 35,303 clinical and 12,565 dose constraints based DVH data elements that are stored in our RO-CDW database for these patients. All these data elements were mapped to the ontology using the D2RQ mapping language, resulting in 504,180 RDF tuples. In addition to the raw data, these tuples also defined the interrelationships amongst various defined classes in the dataset. An example of the output RDF tuple file is shown in Figure 16 displaying the patient record relationship with diagnosis, TNM staging etc. All the entities and predicates in the output RDF file have a URI, which is resolvable as a link for the computer program or human to gather more data on the entities or class. For example, the RDF viewer would be able to resolve the address http://purl.obolibrary.org/obo/NCIT_48720 to gather details on the T-stage such as concept definitions, synonym, relationship with other concepts and classes etc. We were able to achieve a mapping completeness of 94.19% between the records in our clinical database and RDF tuples. During the validation process, we identified several ambiguities or inconsistencies in the data housed in the relational database, such as indication of use of ECOG instrument for performance status evaluation but missing values for ECOG performance status score, record of T stage but nodal and metastatic stage missing and delivered number fractions missing with the prescribed dose information. To maintain data integrity and accuracy, the D2RQ mapping script was designed to drop these values due to missing or incomplete data or ambiguous information. Additionally, the validation process thoroughly examined the interrelationships among the defined classes in the dataset. We verified that the relationships and associations between entities in the RDF tuples accurately reflected the relationships present in the

original clinical data. Any discrepancies or inconsistencies found during this analysis were identified and addressed to ensure the fidelity of the mapped data. To evaluate the accuracy of the mapping process, we conducted manual spot checks on a subset of the RDF tuples. This involved randomly selecting samples of RDF tuples and comparing the mapped values to the original data sources. Through these spot checks, we ensured that the mapping process accurately represented and preserved the information from the clinical and dosimetry data during the transformation into RDF tuples. Overall, the validation process provided assurance that the pipeline effectively transformed the clinical and dosimetry data stored in the RO-CDW database into RDF tuples while preserving the integrity, accuracy, and relationships of the original data.

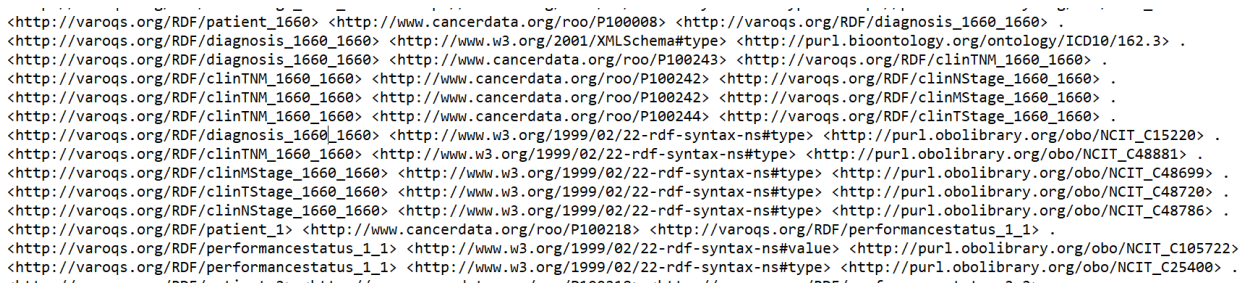


Figure 16: Example of the output RDF tuple file

5.7 Visualization of data in ontology based graphical format

Visualizations on ontologies play a key role for users to understand the structure of the data and work with the dataset and its applications. This has an appealing potential when it comes to exploring or verifying complex and large collections of data such as ontologies. We utilized the Allegrograph Gruff toolkit [34] that enables users to create visual knowledge graphs that display data relationships in a neat graphical user interface. The Gruff toolkit uses simple SPARQL queries to gather the data for rendering the graph with nodes and edges. These visualizations are useful because they increase the users' understanding of data by instantly illustrating relevant relationships amongst class and concepts, hidden patterns, and data's significance to outcomes. An example of the graph-based visualization for a prostate and non-small cell lung cancer patient is shown in Figures 17 and 18. Here all the nodes stand for concepts and classes and the edges represent relationships between these concepts. All the nodes in the graph have unique resource identifiers (URI) that are resolvable as a Web link for the computer program or human to gather more data on the entities or classes. The color of the nodes in the graph visualization are based on the node type and there are inherent properties of each node that include the unique system code (NCIT code or ICD code etc.), synonyms terms, definitions, value type (string, integer, floating point number etc.). The edges connecting the nodes are defined as properties and stored as predicates in the ontology data file. The use of these predicates enables the computer program to effectively find the queried nodes and their interrelationships. Each of these properties are defined with URIs that are available for gathering more detailed information on the relationship definitions. The left panel in figure 17 and 18 shows various property types or relationship types that connect the nodes in the graph. Using SPARQL language and Gruff visualization tools, users can query the data without having any prior knowledge of the relational database structure or schema, since these SPARQL queries are based on universal publish classes defined in the NCI's Thesaurus, Units Ontology, ICD-10 ontologies.

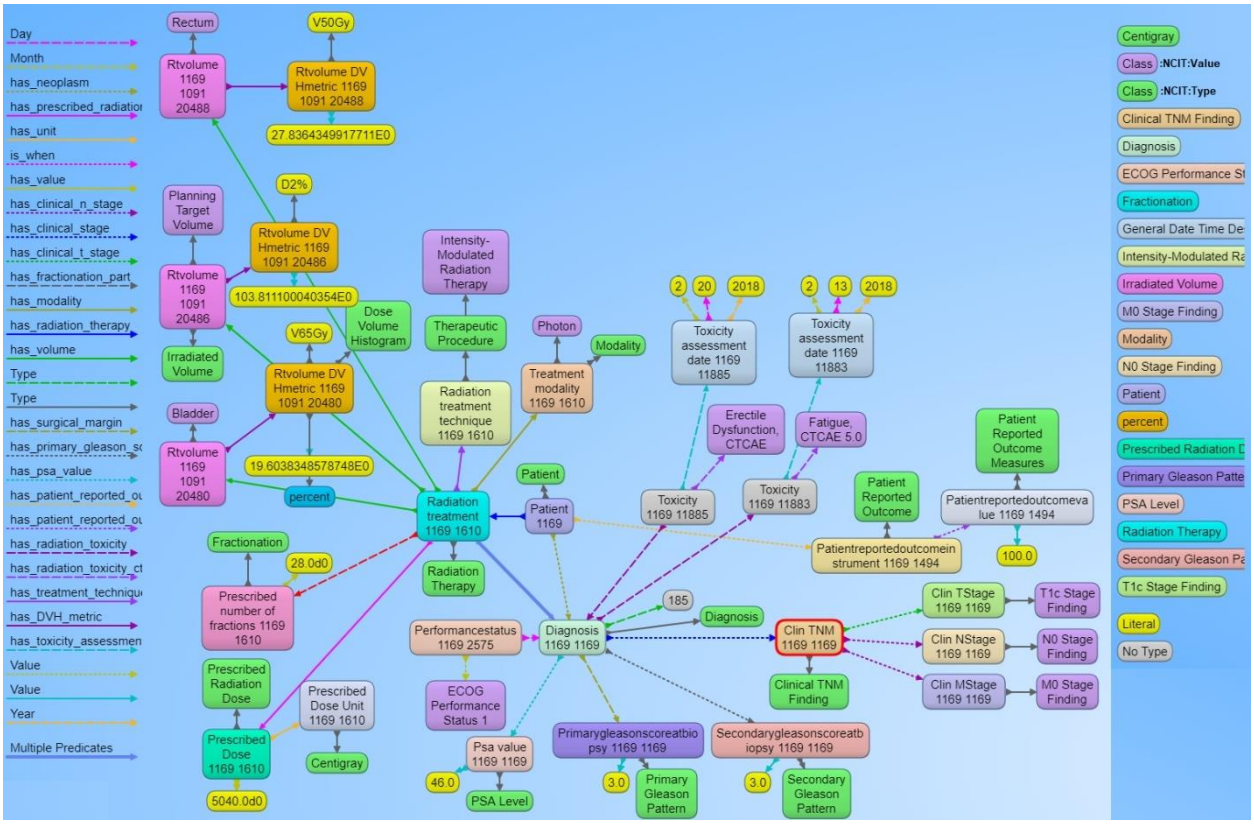


Figure 17: Example of the graph structure of a prostate cancer patient record based on the ontology

Each node in the graph are entities that represent objects or concepts and have a unique identifier and can have properties and relationships to other nodes in the graph. These nodes are connected by directed edges representing relationships between the information, such as the relationship between the diagnosis node and the radiation treatment node. Similarly, there are edges from the diagnosis node to the toxicity node and further to the specific CTCAE toxicity class, indicating that the patient was evaluated for adverse effects after receiving radiation therapy. The different types of edge relationships from the ontology that are used in this example are listed on the left panel of the figure. The right panel shows different types of nodes that are used in the example.

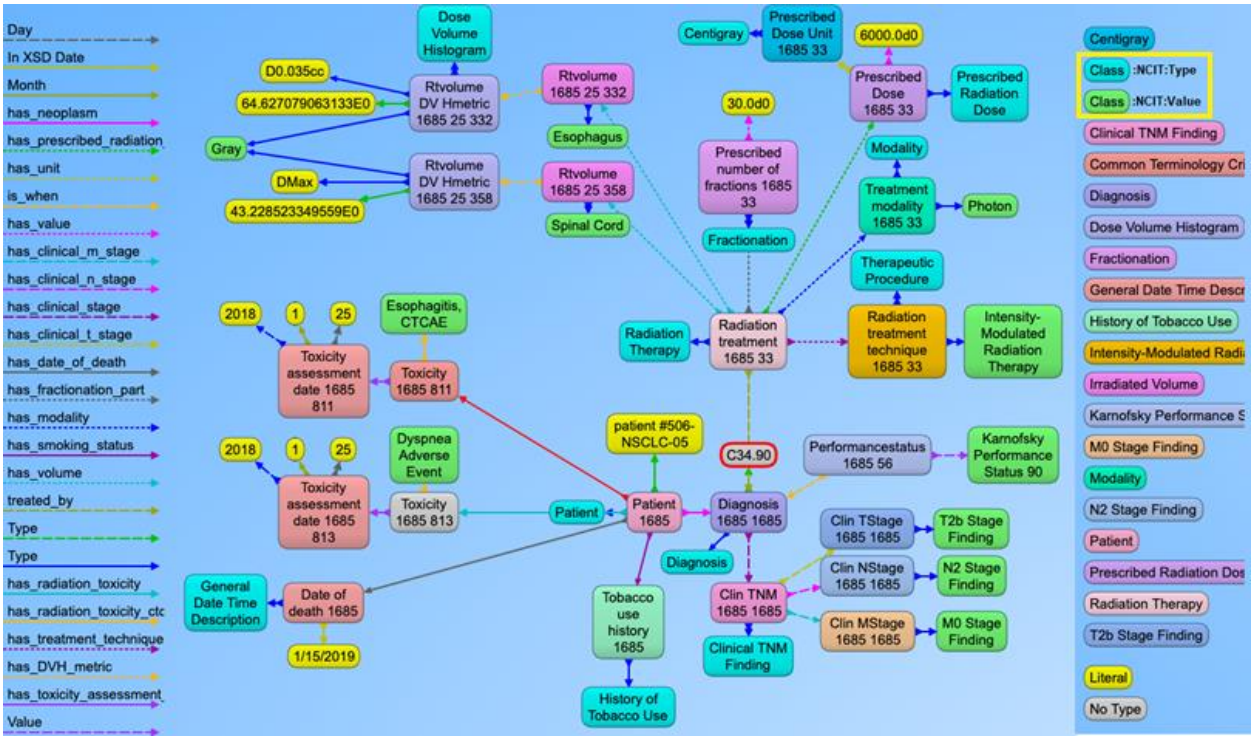


Figure 18: Example of the graph structure of a non-small cell lung cancer (NSCLC) patient based on the ontology

This has a similar structure to the previous prostate cancer example with NSCLC content. The nodes in green and aqua blue color (highlighted in the right panel) indicate the use of NCIT classes to represent the use of standard terminology to define the context for each node present in the graph. For simpler visualization, the NCIT codes and URIs are not displayed with this example.

Finally, these SPARQL queries can be used with commonly available programming languages like python and R via representational state transfer (REST) application programming interfaces (APIs). We also verified that data from the SPARQL queries and the SQL queries from the CDW database to verify accuracy of the mapping. Our analysis found no difference in the resultant data from the two query techniques. The main advantage of using the SPARQL method is that the data can be queried without any prior knowledge of the original data structure based on the universal concepts defined in the ontology. Also, the data from multiple sources can be seamlessly integrated in the RDF graph database without the use of complex data matching techniques and schema modifications that is currently required with relational databases. This is only possible if all the data stored in the RDF graph database refers to published codes from the commonly used ontologies.

5.8 Comparison between our Knowledge Graph-based Solution and Traditional Relational Database-based solution

This section presents the benchmark tests conducted on a Microsoft SQL 2016 Relational Database-based Solution (SQL-DB or RO-CDW) and a Knowledge Graph (KG-DB) Solution from Ontotext GraphDB software to evaluate their querying performance. The tests included two query types designed to compare the performance of these two data modeling technologies. It is important to note that the implementation of

the technologies may influence the results, but the tools used for the experiments are commonly employed by data model designers, making the results indicative of typical implementation performance.

- The first query involved retrieving data from two tables in the relational database to obtain a list of patients diagnosed with prostate cancer. This query aimed to provide a simple functionality that could be replicated in both the relational database and the knowledge base (KB) graph database. (Figure 19a shows this query for relational SQL database and 19b shows this query for Knowledge base graph database)
- The second query was more complex, involving data retrieval from multiple tables and filtering based on diagnosis, TNM staging, prescribed dose, number of fractions and Erectile Dysfunction Grade 1 toxicity. This query aimed to test the performance of the two technologies in a scenario that required more extensive data processing and filtering. (Figure 19c shows this query for relational SQL database and 19d shows this query for Knowledge base graph database)

```
SELECT Patient.[Patient_ID]  ,[PatientMRN]
  FROM Patient
    inner join DiagnosisAndStaging on Patient.Patient_ID = DiagnosisAndStaging.Patient_ID
  WHERE DiagnosisAndStaging.ICDCode='C61'
```

(a)

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX roo: <http://www.cancerdata.org/roo/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX icd: <http://purl.bioontology.org/ontology/ICD10/>
select * where {
        ?patient roo:P100008 ?diagnosis .
        ?diagnosis xsd:type icd:C61.
    ?patient rdfs:value ?patient_MRN
}
```

(b)

```
SELECT Patient.[Patient_ID]
    ,[PatientMRN], NFractionsDelivered, StagingT, StagingN, StagingM, TargetVolume
    ,[TargetVolumeDose]
    ,[DoseUnit] , [ToxicityMeasure]
    ,[ToxicityValue], ProviderReportedToxicity_DateOfRecord
FROM Patient
    inner join DiagnosisAndStaging on Patient.Patient_ID =DiagnosisAndStaging.Patient_ID
    inner join RTCourse On RTCourse.Patient_ID = Patient.Patient_ID
    Inner join RTTreatedPlan on RTTreatedPlan.RTCourse_ID = RTCourse.RTCourse_ID
    Inner join RTPhase on RTPhase.RTCourse_ID = RTCourse.RTCourse_ID
    inner join RTPhaseTargetDose on RTPhaseTargetDose.RTPhase_ID = RTPhase.RTPhase_ID
    Inner join ProviderReportedToxicity on ProviderReportedToxicity.Patient_ID = RTCourse.Patient_ID
WHERE DiagnosisAndStaging.ICDCode='C61' and ToxicityMeasure = 'Erectile dysfunction'
```

(c)

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX roo: <http://www.cancerdata.org/roo/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX icd: <http://purl.bioontology.org/ontology/ICD10/>
PREFIX varoqs:<http://www.varoqs.org/varoqs/>
PREFIX ncit:<http://purl.obolibrary.org/obo/NCIT_>
select * where {
        ?patient roo:P100008 ?diagnosis .
        ?diagnosis xsd:type icd:C61.
    ?patient varoqs:V100013 ?toxicity.
    ?toxicity varoqs:V100014 ncit:C55615.
    ?patient rdfs:value ?patient_MRN.
    ?diagnosis roo:P100243 ?TNM.
    ?TNM roo:P100241 ?Mstage.
    ?TNM roo:P100242 ?Nstage.
    ?TNM roo:P100244 ?Tstage.
    ?patient roo:P100301 ?RT_course.
    ?RT_course roo:P100023 ?RT_Rx_Dose.
    ?RT_Rx_Dose roo:P100042 ?RT_Rx_Dose_value.
    ?RT_course roo:P100269 ?No_fractions.
      ?No_fractions roo:P100042 ?No_fractions_val.
}
```

(d)

Figure 19: (a) Simple query for relational SQL database, (b) Simple query for Knowledge Based Graph database. (c) Complex query with multiple inner join statements for relational SQL database (d) Complex query with Knowledge based Graph database

To analyze the performance of the two data modeling technologies, benchmarking tests were conducted using the Microbench v0.8 Python framework. This framework allows programmers to assess the performance of Python routines. The benchmarking tests focused on measuring query response time, scalability, and concurrency handling. The Microbench framework executed the routines successively within a preconfigured period, known as an iteration, and the framework was configured to run multiple iterations. Some iterations were allocated as warm-up iterations. The framework provided the number of times each routine was executed during the specified period for each iteration. After completing all benchmarking tests, the average values for each benchmark across all iterations were obtained. A lower response time indicated better performance for the tested routine. To ensure consistent and controlled testing conditions, benchmarking tests were deployed in Docker containers. Two Docker containers were

created and deployed, with one container dedicated to each database technology. This approach helped minimize disturbances from different operating system processes. To evaluate the scalability metric, the dataset was replicated five times (5X) in both the relational database (SQL-DB) and the knowledge graph-based (KG-DB) solution. This allowed for the assessment of changes in response time as the dataset size increased. Figure 20 provides the results of the benchmarking tests. For a single request, the knowledge graph (KG-DB) has a response time of 75ms (milliseconds) for the simple query which is similar to the response time from relational SQL-DB, but the response time (202ms) is significantly higher for complex SQL-DB query. As the number of concurrent requests increases, the SQL-DB has higher response times than KG-DB. To assess scalability, the dataset was replicated five times (5X Dataset). The SQL-DB exhibits an increase in response time with the larger dataset with concurrent queries, while the KG-DB generally maintains a lower response time. Overall, based on the provided results, it can be observed that the KG-DB performs relatively well in terms of query response times and scalability compared to the SQL-DB. However, it is important to note that these results are specific to the dataset and testing conditions used in your experiment.

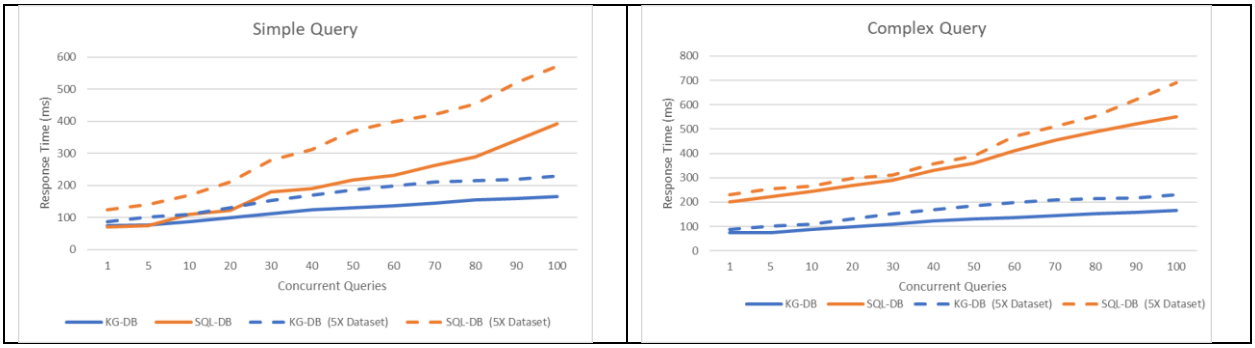


Figure 20: Results showing the query execution times for single and concurrent queries and with increasing the dataset size for relational SQL-DB and KG-DB

Updating the LHS infrastructure and the KG database to integrate new parameters from the clinical workflow depends on several factors. The ease of integration is influenced by the compatibility of the new parameters with the existing data structure and ontology. If the new parameters align well with the current structure and ontology, the process is smoother. However, substantial changes to the data structure or ontology may introduce complexity and require additional effort. In cases where updates or additions to the ontology are needed, modifications to the schema and the creation of new classes and relationships may be necessary. Fortunately, our LHS design is built with scalability in mind. It allows for the seamless linking of new data from future patient encounters and other clinical domains, such as medical oncology and surgery, without significant changes to the data pipeline and IT resources. This flexibility ensures that the LHS infrastructure can accommodate evolving data needs and supports the integration of diverse clinical information. Table 3 shows some additional comparison metrics between our knowledge graph solution and the relational database solution.

Comparison Metrics	Knowledge Graph-based Solution	Relational Database-based Solution
Scalability and Performance	• Minimal increase in response time as dataset size increases • Highly scalable with linking new data from future patient encounters and data from other clinical domains. • Is able to handle complex queries due to optimized knowledge graph traversal methods.	• Increase in response time as dataset size increases • Performance may degrade with large datasets or complex queries due to table joins and indexing limitations
Data Integration and Interlinking	• Efficient integration of data from multiple sources and linking through semantic relationships in the knowledge graph	• Limited ability to integrate and establish relationships between data from different tables in the database
Data Discovery and Accessibility	• Enhanced data discoverability and accessibility due to ontology-based indexing and semantic querying	• Relatively limited data discoverability and accessibility through traditional SQL queries
Semantic Enrichment	• Relationships among data fields are established and used for searching for the patient cohort. • Allows searching for synonyms, hyponym terms that are not present in the dataset and gather patients that have similar attributes.	• Relationships among data fields need to be manually established. • Each synonym and hyponym term needs to be manually annotated in the dataset. • Limited querying flexibility primarily based on structured SQL queries
Data Analysis and Visualization	• Enables advanced data analytics, visualization, and identification of trends and patterns in patient outcomes through graph-based analysis	• Limited data analysis capabilities and visualization options compared to graph-based analytics
Data Reusability and Interoperability	• Supports data reusability and interoperability by adhering to FAIR principles (Findable, Accessible, Interoperable, and Reusable)	• Relational databases offer limited data reusability and interoperability without additional integration efforts

Table 3: Comparison between knowledge graph-based ontology-specific search solution and the traditional relational database-based solution from the various oncology data sources.

5.9 Discussion

Robust learning health system in radiation oncology requires comprehensive clinical and dosimetry data. Furthermore, advanced machine learning models and AI require high fidelity and high veracity data to improve the model performance. Scalable intelligent infrastructure that can provide the data from multiple data sources and can support these models are not yet prevalent [35, 36]. Our infrastructure solution provides an integrated approach to capture data from multiple sources and then structure the data in a knowledge base with semantically interlinked entities for seamless consumption in machine learning methods. The use our infrastructure solution will allow researchers to mine novel associations from multiple, heterogeneous, and multiple domain sources simultaneously and gather relevant knowledge to provide feedback to the clinical providers for obtaining better clinical outcomes for patients on a personalized basis, which will enhance the quality of clinical research.

We have shown the process to transform clinical traditional database schemas into a knowledge graph-based database with the use of ontologies. The main advantages of using an ontology-based graph database as opposed to SQL based relational databases is that the relational databases are designed to cater to a particular application and its software requirements, and data stored is not conducive for clinical

research. These databases are not suited to gather data from multiple data sources when the structure of data, schema, data types are unknown. On the other hand, ontology-based Knowledge graph databases are schema free and designed to store large amount of data with defined interrelationships and the definitions based on universally defined concepts that enable any clinical researcher to query the data without understanding the inherent data structure and schema used to store data in the database. The ontology structure makes querying the data more intuitive for researchers and clinicians because it matches the domain knowledge logical structure [37]. Each data node in the graph has a unique URI that is useful to transform the data using the FAIR concepts. The FAIR guidelines ensure that the data and knowledge is findable, by assigning a globally unique and persistent identifier to each data field. To make the data accessible, these data can readily be shared with almost no pre or post processing requirements. Interoperability can be achieved by using standard ontologies to represent the data and once the data is shared and merged with data from other domains, it can be reused for multiple applications for the benefit of patient care. Ontologies provide a shared understanding of data elements, enabling consistent interpretation of information across multiple institutions. This consistency in data meaning is crucial for federated queries, as it ensures that queries can be formulated using standardized terms and concepts that are understood uniformly by all participating institutions. These approaches enable federated queries where each hospital maintains its local knowledge graph that represents its specific radiation oncology data but can securely collaborate and gain insights from a collective pool of knowledge without sharing individual patient data. Federated queries involve formulating standardized queries that can be executed across multiple local knowledge graphs simultaneously. These queries leverage the common ontology-based definitions and consistent representation of data structures to retrieve relevant information from each hospital's knowledge graph. By adhering to common ontology terms and relationships, federated queries can effectively integrate data from multiple hospitals, facilitating cross-institutional analysis and knowledge sharing. Traditional methods with artificial intelligence and machine learning techniques do not address the issues of data sharing, and interpretability amongst multiple systems and institutions. With this approach, hospitals can leverage the collective intelligence within the federated knowledge graph to gain insights, identify patterns, and conduct research without compromising patient privacy and data security.

Additionally, ontologies can be used to enhance data analysis by allowing for more precise querying and reasoning over the data. For example, an ontology-based query might retrieve all patients who received a certain type of radiation treatment, while an ontology-based reasoning system might infer that a certain treatment plan parameter or dose constraint is contraindicated for a certain type of cancer [38]. Ontologies allow for query expansion and mapping capabilities, which are essential for federated queries. When a query is executed across multiple databases, the use of ontologies enables automatic expansion of the query to include synonymous terms or related concepts from different institutions. The integration of ontologies not only facilitates effective data analysis but also supports more informed decision-making in the field of clinical research.

References:

1. Lambin P, Roelofs E, Reymen B, Velazquez ER, Buijsen J, Zegers CM, Carvalho S, Leijenaar RT, Nalbantov G, Oberije C, Scott Marshall M, Hoebers F, Troost EG, van Stiphout RG, van Elmpt W, van der Weijden T, Boersma L, Valentini V, Dekker A. 'Rapid Learning health care in oncology' -

an approach towards decision support systems enabling customised radiotherapy'. Radiother Oncol. 2013 Oct;109(1):159-64. doi: 10.1016/j.radonc.2013.07.007. Epub 2013 Aug 28. PMID: 23993399.
2. Price G, Mackay R, Aznar M, McWilliam A, Johnson-Hart C, van Herk M, Faivre-Finn C. Learning healthcare systems and rapid learning in radiation oncology: Where are we and where are we going? Radiother Oncol. 2021 Nov;164:183-195. doi: 10.1016/j.radonc.2021.09.030. Epub 2021 Oct 4. PMID: 34619237.
3. Nordo AH, Eisenstein EL, Hawley J, et al. A comparative effectiveness study of eSource used for data capture for a clinical research registry. Int J Med Inform. 2017;103:89-94.
4. Coleman N, Halas G, Peeler W, et al. From patient care to research: a validation study examining the factors contributing to data quality in a primary care electronic medical record database. BMC Fam Pract. 2015;16:11. doi: 10.1186/s12875-015-0223-z.
5. Spasić I, Livsey J, Keane JA, Nenadić G. Text mining of cancer-related information: review of current status and future directions. Int J Med Inform. 2014 Sep;83(9):605-23. doi: 10.1016/j.ijmedinf.2014.06.009. Epub 2014 Jun 24. PMID: 25008281.
6. https://ghr.nlm.nih.gov/primer/precisionmedicine/initiative
7. S.J. Schrodi, S. Mukherjee, Y. Shan, G. Tromp, J.J. Sninsky, A.P. Callear, et al. Genetic-based prediction of disease traits: prediction is very difficult, especially about the future; Front Genet, 2 (5) (2014 Jun), p. 162
8. L. Jostins, J.C. Barrett; Genetic risk prediction in complex disease; Hum Mol Genet, 20 (R2) (2011), pp. R182-R188
9. K. Kourou, T.P. Exarchos, K.P. Exarchos, M.V. Karamouzis, D.I. Fotiadis; Machine learning applications in cancer prognosis and prediction; Comput Struct Biotechnol J, 1 (13) (2015), pp. 8-17
10. S. Yousefi, F. Amrollahi, M. Amgad, C. Dong, J.E. Lewis, C. Song, et al.; Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models; Sci Rep, 7 (1) (2017), p. 11707
11. R. Arp, B. Smith, A.D. Spear; Building ontologies with basic formal ontology; Mit Press (2015)
12. T.R. Gruber; A translation approach to portable ontology specifications; Knowledge Acquisition, 5 (2) (1993), pp. 199-220
13. N.F. Noy, et al.; BioPortal: ontologies and integrated data resources at the click of a mouse; Nucleic Acids Res, 37 (Web Server issue) (2009), pp. W170-W173
14. P.L. Whetzel, et al.; BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications; Nucleic Acids Res, 39 (Web Server issue) (2011), pp. W541-W545
15. Rosse C, Mejino Jr JL. A Reference Ontology for Biomedical Informatics: The Foundational Model of Anatomy. Journal of Biomedical Informatics. 2003 Dec 1;36(6):478-500.
16. Mildenberger P, Eichelberg M, Martin E. Introduction to the DICOM standard. European radiology. 2002 Apr;12(4):920-7.
17. Vorisek CN, Lehne M, Klopfenstein SAI, Mayer PJ, Bartschke A, Haese T, Thun S. Fast Healthcare Interoperability Resources (FHIR) for Interoperability in Health Research: Systematic Review. JMIR Med Inform. 2022 Jul 19;10(7):e35724. doi: 10.2196/35724. PMID: 35852842; PMCID: PMC9346559.
18. Mayo CS, Feng MU, Brock KK, Kudner R, Balter P, Buchsbaum JC, Caissie A, Covington E, Daugherty EC, Dekker AL, Fuller CD, Hallstrom AL, Hong DS, Hong JC, Kamran SC, Katsoulakis E,

Kildea J, Krauze AV, Kruse JJ, McNutt T, Mierzwa M, Moreno A, Palta JR, Popple R, Purdie TG, Richardson S, Sharp GC, Shiraishi S, Tarbox L, Venkatesan AM, Witztum A, Woods KE, Yao J, Farahani K, Aneja S, Gabriel PE, Hadjiiski L, Ruan D, Siewerdsen JH, Bratt S, Casagni M, Chen S, Christodouleas J, DiDonato A, Hayman J, Kapoor R, Kravitz S, Sebastian S, Von Siebenthal M, Xiao Y. Operational Ontology for Oncology (O3) - A Professional Society Based, Multi-Stakeholder, Consensus Driven Informatics Standard Supporting Clinical and Research use of "Real -World" Data from Patients Treated for Cancer: Operational Ontology for Radiation Oncology. Int J Radiat Oncol Biol Phys. 2023 May 25:S0360-3016(23)00525-4. doi: 10.1016/j.ijrobp.2023.05.033. Epub ahead of print. PMID: 37244628.

19. Orthanc DICOM Server - Available at: https://www.orthanc-server.com/
20. DICOM DIMSE- Available at https://dicom.nema.org/dicom/2013/output/chtml/part07/sect_7.5.html
21. Syed K, Sleeman W, Ivey K, et al. Integrated natural language processing and machine learning models for standardizing radiotherapy structure names. Healthcare. 2020;8:120
22. Sleeman C. Relabeling Non-Standard to Standard Structure Names Using Geometric and Radiomic Information. Med Phys. 2020:47. No. 6. 111 RIVER ST, HOBOKEN 07030-5774, NJ USA: WILEY.
23. Sleeman IV WC, Nalluri J, Syed K, et al. A Machine learning method for relabeling arbitrary DICOM structure sets to TG-263 defined labels. J Biomed Inform. 2020;109:103527
24. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data. 2016;3:160018.
25. Semantic Web at W3C: Available at: https://www.w3.org/standards/semanticweb
26. Traverso A, van Soest J, Wee L, Dekker A. The radiation oncology ontology (ROO): Publishing linked data in radiation oncology using semantic web and ontology techniques. Med Phys. 2018 Oct;45(10):e854-e862. doi: 10.1002/mp.12879. Epub 2018 Aug 24. PMID: 30144092.
27. Radiation Oncology Ontology - Summary | NCBO BioPortal - Available at: https://www.bioontology.org
28. Noy NF, Crubezy M, Fergerson RW, et al. Protégé-2000: an open-source ontology-development and knowledge-acquisition environment. AMIA Annu. Symp. Proc. AMIA Symp. 953; 2003.
29. National Cancer Institute Thesaurus - Summary | NCBO BioPortal - Available at: https://www.bioontology.org
30. International Classification of Diseases, Version 10 - Summary | NCBO BioPortal - Available at: https://www.bioontology.org
31. DBpedia ontology – Available at: https://www.dbpedia.org
32. Urbani, J and Jacobs, C., Adaptive Low-level storage of very large knowledge graphs, arXiv, 2020, 2001/09-78v1. http://arxiv.org/abs/2001.09078
33. Ontotext GraphDB – Available at: https://www.ontotext.com
34. Gruff - AllegroGraph software – Available at: https://allegrograph.com/
35. Jochems A, Deist TM, van Soest J, Eble M, Bulens P, Coucke P, Dries W, Lambin P, Dekker A. Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital - A real life proof of concept. Radiother Oncol. 2016 Dec;121(3):459-467. doi: 10.1016/j.radonc.2016.10.002. Epub 2016 Oct 28. PMID: 28029405.
36. Zerka F, Barakat S, Walsh S, Bogowicz M, Leijenaar RTH, Jochems A, Miraglio B, Townend D, Lambin P. Systematic Review of Privacy-Preserving Distributed Machine Learning From

Federated Databases in Health Care. JCO Clin Cancer Inform. 2020 Mar;4:184-200. doi: 10.1200/CCI.19.00047. PMID: 32134684; PMCID: PMC7113079.
37. Min H, Manion FJ, Goralczyk E, Wong YN, Ross E, Beck JR. Integration of prostate cancer clinical data using an ontology. J Biomed Inform. 2009 Dec;42(6):1035-45. doi: 10.1016/j.jbi.2009.05.007. Epub 2009 Jun 2. PMID: 19497389; PMCID: PMC2784120.
38. Yan, J., Wang, C., Cheng, W. et al. A retrospective of knowledge graphs. Front. Comput. Sci. 12, 55–74 (2018). https://doi.org/10.1007/s11704-016-5228-9

6. Design Framework for Ontology-based Keyword Search and Patient Similarity Techniques

6.1 Statement of Problem

The sharing of medical information among various stakeholders, including hospitals, clinicians, and pharmaceutical companies, faces a significant challenge due to the proliferation of medical terms. Even within a single hospital, different clinicians often employ distinct terminology when referring to the same diagnosis, while symptoms are inconsistently recorded in varying levels of detail in patient records. For instance, one clinician may document a patient's diagnosis as "Pineoblastoma," while another might use the synonymous term "PNET of Pineal Gland." Instead of the more specific term "Pineoblastoma," a generic term like "Brain Neoplasm" might be recorded in the patient's record, with the former term considered a subtype of the latter (a hyponym). Although coding systems such as SNOMED or ICD 10 and ontologies like the National Cancer Institute (NCI) Thesaurus encode terms and their relationships, they do not prevent clinicians from using different hyponyms, hypernyms, or synonyms within an electronic medical record (EMR).

In order to search for relevant records, clinicians or researchers or data abstractors currently follow a laborious process. First, they need to access the NCI thesaurus. Then, they have to identify all the synonymous terms of "brain tumor" listed in the thesaurus, which amounts to seven terms in this case. Finally, data abstractor must utilize these terms in a query to the relational database to retrieve the desired records, such as “Brain Neoplasm,” “Stomatitis.” Although this approach resembles the one used in PubMed for retrieving medical articles, it is clearly inefficient as it necessitates significant manual effort. The inefficiency becomes even more apparent when data abstractor considers the hyponyms of "brain tumor" to retrieve patient records that refer to specific subtypes of brain tumors, like "Pineoblastoma" or "Thalamic Neoplasm." With 233 such terms in the NCI thesaurus, it becomes practically impossible for data abstractor to manually extract all this information from the NCI and conduct a comprehensive search for the appropriate records. A certain level of automation is imperative in this process.

This section focuses on two key aspects: ontology-based keyword search and patient similarity techniques. The ontology-based keyword search aims to provide an efficient method for searching healthcare data by utilizing the rich semantic relationships captured in ontologies. By incorporating synonym-based term matching and leveraging the hierarchical structure of ontologies, this approach enables precise and context-aware searches. It allows users to retrieve relevant information based on clinical terms while considering both parent and children classes, thus ensuring comprehensive and accurate results.

In addition to keyword search, the chapter also explores patient similarity techniques. Understanding patient similarity is crucial for various healthcare applications, such as cohort identification, personalized medicine, and decision support systems. By employing advanced embedding models and distance metrics, the chapter presents a framework to measure the similarity between patients based on their clinical attributes. These techniques facilitate the identification of patients with similar profiles and support data-driven decision-making in healthcare.

To validate the effectiveness of the proposed framework, comprehensive evaluations are conducted. The chapter discusses various evaluation metrics and methodologies to assess the performance of the

ontology-based keyword search tool and patient similarity techniques. These evaluations aim to demonstrate the accuracy, efficiency, and usability of the developed solutions in real-world healthcare scenarios.

Overall, the chapter provides a design framework that combines ontology-based keyword search and patient similarity techniques, offering a comprehensive approach to enhancing search capabilities and uncovering hidden insights within healthcare data. By leveraging the power of ontologies and semantic models, this framework contributes to the advancement of learning health systems and enables more efficient and effective healthcare information retrieval and analysis.

6.2 Literature Review

To access the required documents from a document corpus, various methods and strategies are employed. The following are some of the approaches associated with identifying word similarity. Haolin Wang et al. developed a two-stage query expansion technique using latent semantic relationships, but it was not suitable for large datasets [1]. Youcef Djenouri et al. used data mining and Bees Swarm Optimization for document retrieval, but it had slow execution and poor similarity scores [2]. Fei Li et al. proposed a similarity measure based on WordNet and Wikipedia, but it had limitations in accuracy due to weight-based statistical approaches [3]. Oscar Araque et al. developed a semantic similarity measure for text terms, but it had poor feature extraction and overall performance [4]. NH Mahadzir et al. explored semantic similarity measures for disambiguating Malay and English terms, but there were issues with terms missing from the source (WordNet) [5]. Jingxiang Zhang et al. analyzed food safety incidents using semantic templates, but their method had limitations in large-scale data similarity analysis [6]. This approach is not suitable for finding similarities between words in large data due to high execution time. Despite the mentioned review techniques, several challenges such as inaccurate similarity scores, long execution times, and inadequate feature extraction persist in most information retrieval models. Additionally, none of these techniques have been utilized with healthcare ontology-based patient graphs. Consequently, an Ontology-Based Semantic Retrieval of patient graph records using the word embedding models has been developed to address these limitations.

6.3 Ontology Keyword Based Searching Tool Architecture

To provide an effective method to search the graph database, we built an ontology-based keyword search engine that utilizes the synonym-based term matching methods. Another advantage of using ontology-based term searching is realized by using the class parent-children relationships. Ontologies are hierarchical in nature with the terms in the hierarchy often forming a directed acyclic graph (DAG). For example, if we are searching for patients in our database with clinical stage T1, the matching patient list will only comprise patients that have T1 stage NCI Thesaurus code (NCIT: C48720) in the graph database. These matching patients will not return any patients with T1a, T1b, T1c sub-categories that are children of the parent T1 staging class. We built this search engine where we can search on any clinical term and its matching patient records based on both parent and children classes are abstracted. The method that is used in this search engine is as follows. When the user wants to use the ontology to query the graph based medical records, the only input necessary is the clinical query terms (q-terms) and an indication of whether the synonyms should also be considered while retrieving the patient records. The user has the option to specify the multiple levels of child class search and parent classes to be included in the search parameters. The software will then connect to the Bioportal database via REST API and perform the search to gather the matching classes for the q-terms and the options specified in the program [7, 8].

Using the list of matching classes, a SPARQL based query is generated and executed with our patient graph database and matching patient list and the q-term based clinical attributes are returned to the user. In order to find patients that have not the same but similar attributes based on the search parameters; we have designed a patient similarity search method. First, a text corpus is created by performing breadth-first search walks on each patient's individual knowledge graph. This involves traversing the graph starting from the patient node and exploring its neighboring nodes in a breadth-first manner. As the traversal progresses, relevant attributes and information associated with each node are collected and added to the corpus. This allows for the extraction of meaningful textual data from the knowledge graph. Once the text corpus is constructed, it serves as a representation of the patient's attributes and their relationships within the knowledge graph. This corpus contains information from matched patients who share similar attributes or characteristics. It captures relevant details such as diagnoses, treatment information, clinical outcomes, and other pertinent data points. By leveraging this corpus, similarity analysis techniques can be applied to identify patients who exhibit similar patterns or profiles.

The extracted text corpus undergoes several preprocessing stages, including padding, stemming, tokenization, case folding, and stop word removal, to prepare it for further analysis. In the padding stage, the text is transformed into a set of word arrays, allowing for more effective preprocessing in subsequent steps. Stemming is applied to reduce words to their base or root form, ensuring consistent representation for word embedding techniques. Tokenization breaks the text into tokens, which can be words, symbols, or specific units with distinct meanings. Stop word removal eliminates insignificant words that provide less information to the model. Case folding involves converting all words to lowercase to ensure uniformity in representation. This process treats lowercase and uppercase forms of words as equivalent. Furthermore, removing unwanted characters helps structure the text. In sentiment analysis or emotion recognition tasks, words like 'I', 'we', 'us', 'a', 'an', 'the', etc., which are considered stop words, are removed as they contribute less to the overall meaning.

We utilized four vector embedding models, namely Word2Vec [9], Doc2Vec [10], GloVe [11], and FastText [12], to train and generate vector embeddings. The description of these models is provided in section 6.3. The text corpus used for training is obtained from the Bioportal website, which encompasses NCIT, ICD, and SNOMED codes, as well as class definition text, synonyms, and hyponyms terms. We trained these models with the training dataset for 100 epochs on CPU hardware. These trained models are subsequently utilized to generate embeddings for the individual patient text corpus obtained earlier. The Cosine similarity, Euclidean distance, Manhattan distance and Minkowski distance metrics are employed to measure the distance between the matched patients and all patient feature vectors. Figure 21 shows the design architecture of the software system. The main purpose of this search engine is to provide the users with a simple interface to search the patient records.

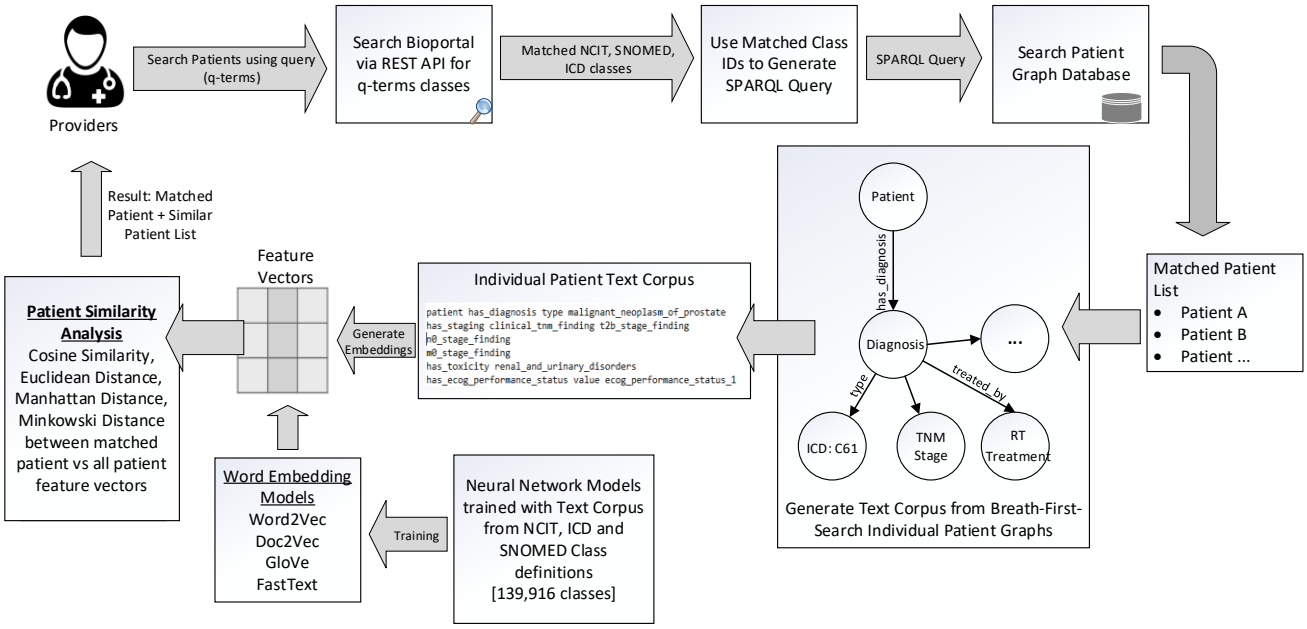


Figure 21: Design architecture for the Ontology based keyword search system.

When the user wants to query the patient graph database to retrieve matching records, the only input necessary is the medical terms (q-terms) and an indication to include any synonym, parent, or children terminology classes in the search. The software queries the Bioportal API and retrieves all the matching NCIT, SNOMED, ICD-10 classes to the q-terms. A SPARQL query is generated and executed on the graph database SPARQL endpoint and the results indicating the matching patient records and their corresponding data fields are displayed to the user. Our architecture includes the generation of text corpus from breath first search of individual patient graphs and using word embedding models to generate feature vectors to identify similar patient cohorts.

6.4 Description of Word Embedding Models: Word2Vec, Doc2Vec, GloVe, and FastText

In natural language processing (NLP) and text analysis, Word2Vec, Doc2Vec, GloVe, and FastText are popular models. For creating embeddings for words or documents, each model uses a different approach, capturing semantic relationships between words and documents. Here is a brief description of each model and its differences:

Word2Vec: Word2Vec is one of the most widely used embedding models that represents words as dense vectors in a continuous vector space. It employs two primary architectures: CBOW and Skip-gram. CBOW predicts target words using context words, while Skip-gram predicts target words based on context words. Through training on substantial text data, Word2Vec effectively captures semantic relationships between words.

Doc2Vec extends Word2Vec to capture embeddings at the document level. It represents documents, such as paragraphs or entire documents, as continuous vectors in a similar way to how Word2Vec represents individual words. This model architecture is also known as Paragraph Vector, learns document representations by incorporating word embeddings and a unique document ID during the training process. This enables the model to capture semantic similarities between different documents.

GloVe: GloVe (Global Vectors for Word Representation) is another popular model for generating word embeddings. This model uses the global matrix factorization and local context window methods to generate the embeddings. GloVe constructs a co-occurrence matrix based on word-to-word co-occurrence statistics from a large corpus and factorizes this matrix to obtain word vectors. It considers the global statistical information of word co-occurrences, resulting in embeddings that capture both syntactic and semantic relationships between words.

FastText: FastText is a model developed by Facebook Research that extends the idea of Word2Vec by incorporating information about subwords. Instead of treating each word as a single entity, FastText model represents words as bags of character n-grams (subword units). By considering subwords, FastText can handle out-of-vocabulary words and capture morphological information. This model enables better representations for rare words, inflections, and compound words. FastText also supports efficient training and retrieval, making it useful for large-scale applications.

In summary, Word2Vec focuses on word-level embeddings, Doc2Vec extends it to capture document-level embeddings, GloVe emphasizes global word co-occurrence statistics, and FastText incorporates subword information for enhanced representations. The choice of model depends on the specific task, data characteristics, and requirements of the application at hand.

6.5 Evaluation Metrics for Measuring Patient Similarity

- **Cosine Similarity**:

Cosine similarity measures the cosine of the angle between two vectors. It calculates the similarity between vectors irrespective of their magnitudes. The cosine similarity between vectors A and B is computed using the dot product of the vectors divided by the product of their magnitudes:

Cosine Similarity = (A.B) / (||A|| * ||B||) (1)

- **Euclidean Distance**:

Euclidean distance is a popular metric to measure the straight-line distance between two points in Euclidean space. In the context of vector spaces, it calculates the distance between two vectors in terms of their coordinates. The Euclidean distance between vectors A and B with n dimensions is calculated as:

Euclidean Distance = $\sqrt{(A[1]-B[1])^2+(A[2]-B[2])^2+\ldots+(A[n]-B[n])^2}$ (2)

- **Manhattan Distance**:

Manhattan distance, also known as city block distance or L1 distance, measures the sum of the absolute differences between the coordinates of two vectors. It represents the distance traveled along the grid-like paths in a city block. The Manhattan distance between vectors A and B with n dimensions is calculated as:

Manhattan Distance = |A[1] - B[1]| + |A[2] - B[2]| + ... + |A[n] - B[n]| (3)

Manhattan distance is commonly used in clustering algorithms, such as k-means, and in applications where the direction of differences is less important than the magnitude.

- **Minkowski Distance**:

Minkowski distance is a generalization of both Euclidean and Manhattan distances. It measures the distance between two vectors in terms of their coordinates, with a parameter p determining the degree of the distance metric. The Minkowski distance between vectors A and B with n dimensions is calculated as:

$$\text{Minkowski distance} = (|A[1] - B[1]|^p + |A[2] - B[2]|^p + ... + |A[n] - B[n]|^p)^{(1/p)} \quad (4)$$

When p = 1, it is equivalent to the Manhattan distance, and when p = 2, it is equivalent to the Euclidean distance. Minkowski distance considers the magnitude and direction of differences between the data points in all dimensions. It is more flexible than Manhattan distance and can handle different types of data distributions.

These metrics provide different ways to quantify the similarity or dissimilarity between vectors, each with its own characteristics and use cases.

6.6 Results

For effective searching of discrete data from the RDF (Resource Description Framework) graph database, we built an ontology-based keyword searching Web tool. The public website for this tool is https://hinge-ontology-search.anvil.app. Here we are able to search the database based on keywords (q-terms). The tool is connected to the BioPortal via REST API [8] and finds the matching classes or concepts and renders the results including the class name, NCIT code and definitions. We specifically used the NCI Thesaurus ontology for our query which is 112MB in size and contains approximately 64000 terms. The search tool can find the classes based on synonym term queries where it matches the q-terms with the listed synonym terms in the classes (figure 22A). The tool has features to search the child and parent classes on the matching q-term classes. Screenshot of the web tool with the child class search is shown in Figure 22B. The user can also specify the level of search which indicates if the returned classes should include classes of children of children. In the example in Figure 22B, we are showing the q-term used for searching "fatigue" while including the child classes up to one level and the return classes included the fatigue based CTCAE class and the grade 1, 2, 3 fatigue classes. Once all the classes used for searching are found by the tool, it searches the RDF graph database for matching patient cases with these classes. The matching patient list including the found class in the patient's graph is displayed to the user. This tool is convenient for the end users to abstract cohorts of patients that have particular classes or concepts in their records without the user learning and implementing the complex SPARQL query language. Based on our evaluation, we found that the average time taken to obtain results is less than five seconds per q-term if there are less than 5 child classes in the query. The maximum time taken is 11 seconds for a q-term that has 16 child classes. Table 4 provides the validation summary of a complex query with eight search q-terms. The major reasons for the tool missing a few patients were the miscoding of some staging attributes (for e.g., patients had a T1Xa, T2e stage). The automated data pull by the tool was not able to identify these edge cases and hence some patients were misrepresented by the search tool. Overall, we were able to achieve good accuracy (0.995) and F1 score (0.994) for such a complicated data query.

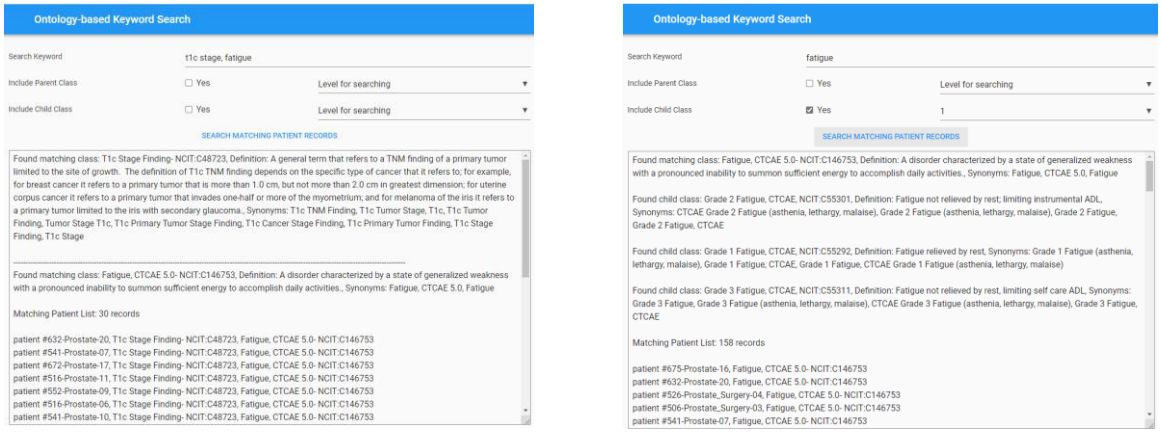


Figure 22: Screenshot of the Ontology-based keyword search portal. A) Search performed using two q-terms returns results with definitions of the matching classes from the Bioportal and the corresponding patient records from the RDF graph database. B) Search performed to include child class up to 1 level on the matching q-term class. Returned results display the matching class, child classes with Fatigue CTCAE grades and matching patient records from the RDF graph database.

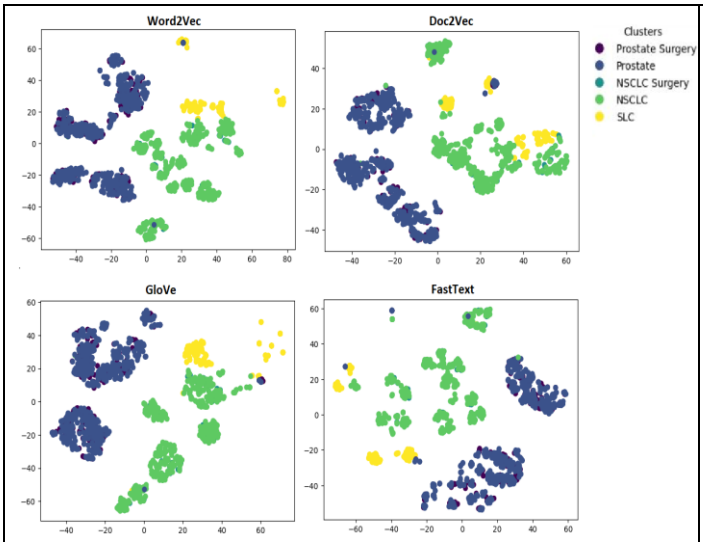
	Predicted Positive	**Predicted Negative**
Actual Positive	498 (TP)	2 (FN)
Actual Negative	3 (FP)	497 (TN)

Validation Measure	**Result**
Accuracy = $\frac{(TP+TN)}{(TP+FP+FN+TN)}$	0.995
Precision = $\frac{TP}{(TP+FP)}$	0.994
Recall= $\frac{TP}{(TP+FN)}$	0.996
F1 Score= $\frac{2*(Recall*Precision)}{(Recall+Precision)}$	0.994

Table 4: Validation of keyword search tool results with eight q terms (PSA value, Primary Gleason Score, T1 stage, Nodal Status, Fractionation, ECOG performance Status, DVH[Rectum], CTCAE Fatigue) with manually curated patient list.

For evaluating the patient similarity-based word embedding models, we evaluated the quality of the feature embedding based vectors produced by using the technique called t-Distributed Stochastic Neighbor Embedding (t-SNE) and cluster analysis with a predetermined number of clusters set to five based on the diagnosis groups for our patient cohort. Clustering methods identify similar groups of data in a data set collection. This method can reveal the local and global features encoded by the feature vectors and thus can be used to visualize clusters within the data. It is important to have prior knowledge of the data set, as this algorithm takes the number of clusters as input. It partitions the “n” data points into “k” clusters in which each data point belongs to the cluster with the nearest mean. We applied t-SNE to all 1660 patient feature-based vectors produced via the four word embedding models. The t-SNE plot is shown in Fig. 23A, the disease data points can be grouped into five clusters with varying degrees of

separability and overlap. The analysis of patient similarity using different embedding models revealed interesting patterns. The Word2Vec model showed the highest mean cosine similarity of 0.902, indicating a relatively higher level of similarity among patient embeddings. In contrast, the Doc2Vec model exhibited a lower mean cosine similarity of 0.637 (Fig. 23B). The GloVe model demonstrated a moderate mean cosine similarity of 0.801, while the FastText model achieved a similar level of 0.855. Regarding distance metrics, the GloVe model displayed lower mean Euclidean and Manhattan distances, suggesting that patient embeddings derived from this model were more compact and closer in proximity. Conversely, the Doc2Vec, Word2Vec and FastText models yielded higher mean distances, indicating greater variation and dispersion among the patient embeddings. These findings provide valuable insights into the performance of different embedding models for capturing patient similarity, facilitating improved understanding and decision-making in the clinical domain.



(A)

Metrics	Word2Vec	Doc2Vec	GloVe	FastText
Cosine Similarity Mean [best score≈ 1]	**0.902**	0.637	0.801	0.855
Euclidean Distance Mean [best score=low distance value]	7.665	9.785	**0.999**	13.054
Manhattan Distance Mean [best score=low distance value]	136.72	174.67	**17.373**	232.88
Minkowski Distance Mean [best score=low distance value]	3.179	4.055	**0.428**	5.416

(B)

Figure 23: (A) Annotation embeddings produced by Word2Vec, Doc2Vec, GloVe and FastText, a 2D-image of the embeddings projected down to 3 dimensions using T-SNE technique. (B) Results of the evaluation metrics used to measure patient similarity. (A) Each point indicates one patient and color of a point indicates the cohort of the patient based on the diagnosis-based cluster. A good visualization result is that the points of the same color are near each other. (B). Word2Vec model had the best cosine similarity, and the GloVe model had the best Euclidean, Manhattan and Minkowski distance suggesting that patient embeddings derived from this model were more compact and closer in proximity.

6.7 Case Study

We conducted a patient similarity analysis using the Word2Vec embedding model and cosine similarity scores to identify similar patients based on clinical attributes. The results of the analysis are presented below.

We randomly selected a patient from the cohort as the target patient and compared their word embeddings with other patients to find the most similar ones. We utilized the cosine similarity score for gathering the top 5 similar patients. The patient attributes for the target patient and the top 5 similar patients found in our cohort using the Word2Vec model are as follows:

	Patient Text Corpus	**Cosine Similarity Score**
Target Patient	patient *has_diagnosis* malignant_neoplasm_of_prostate *has_clinical_stage clinical_tnm_finding has_t_stage* t2a_stage_finding *has_n_stage* n0_stage_finding *has_m_stage* m0_stage_finding *has_histology* adenocarcnoma *has_performance_state* ecog_performance_status_0 *has_primary_gleason_score* gleason_score_3 *has_secondary_gleason_score* gleason_score_4 *has_radiation_treatment type* intensity_modulated_radiation_therapy *has_radiation_toxicity* erectile_dysfunction *has_radiation_toxicity_ctcae_grade* erectile_dysfunction_grade_1 *has_radiation_toxicity* urinary_frequency *has_radiation_toxicity_ctcae_grade* urinary_urgency_grade_1 *has_radiation_toxicity* nocturia	
Similar Patient 1	patient *has_diagnosis* malignant_neoplasm_of_prostate *has_clinical_stage clinical_tnm_finding has_t_stage* t2a_stage_finding *has_histology* adenocarcnoma *has_performance_state* karnofsky_performance_status_100 *has_primary_gleason_score* gleason_score_3 *has_secondary_gleason_score* gleason_score_4 *has_radiation_treatment type* intensity_modulated_radiation_therapy *has_radiation_toxicity* erectile_dysfunction *has_radiation_toxicity_ctcae_grade* erectile_dysfunction_grade_1 *has_radiation_toxicity* urinary_frequency *has_radiation_toxicity* nocturia	0.92
Similar Patient 2	patient *has_diagnosis* malignant_neoplasm_of_prostate *has_clinical_stage clinical_tnm_finding has_t_stage* t2_stage_finding *has_n_stage* nx_stage_finding *has_histology* adenocarcnoma *has_performance_state* karnofsky_performance_status_90 *has_primary_gleason_score* gleason_score_3 *has_secondary_gleason_score* gleason_score_4 *has_radiation_treatment type* intensity_modulated_radiation_therapy *has_radiation_toxicity* erectile_dysfunction *has_radiation_toxicity_ctcae_grade* erectile_dysfunction_grade_1 *has_radiation_toxicity* urinary_frequency	0.90
Similar Patient 3	patient *has_diagnosis* malignant_neoplasm_of_prostate *has_clinical_stage clinical_tnm_finding has_t_stage* t2_stage_finding *has_n_stage* nx_stage_finding *has_m_stage* mx_stage_finding *has_histology* adenocarcnoma *has_performance_state* karnofsky_performance_status_100 *has_primary_gleason_score* gleason_score_3 *has_secondary_gleason_score* gleason_score_4 *has_radiation_treatment type* intensity_modulated_radiation_therapy *has_radiation_toxicity* rash_desquamation *has_radiation_toxicity* urinary_frequency *has_radiation_toxicity* nocturia	0.84
Similar Patient 4	patient *has_diagnosis* malignant_neoplasm_of_prostate *has_clinical_stage clinical_tnm_finding has_t_stage* t1c_stage_finding *has_n_stage* n0_stage_finding *has_m_stage* mx_stage_finding *has_histology* adenocarcnoma *has_performance_state* karnofsky_performance_status_90 *has_primary_gleason_score* gleason_score_3 *has_secondary_gleason_score* gleason_score_4 *has_radiation_treatment type* intensity_modulated_radiation_therapy *has_radiation_toxicity* dysuria *has_radiation_toxicity* urinary_frequency *has_radiation_toxicity_ctcae_grade* urinary_urgency_grade_1 *has_radiation_toxicity* diarrhea *has_radiation_toxicity* nocturia	0.82
Similar Patient 5 Attributes	patient *has_diagnosis* malignant_neoplasm_of_prostate *has_clinical_stage clinical_tnm_finding has_t_stage* t1c_stage_finding *has_n_stage* n0_stage_finding *has_m_stage* m0_stage_finding *has_histology* adenocarcnoma *has_performance_state* ecog_performance_status_1 *has_primary_gleason_score* gleason_score_3 *has_secondary_gleason_score* gleason_score_4 *has_radiation_treatment type* intensity_modulated_radiation_therapy *has_radiation_toxicity* dysuria *has_radiation_toxicity* urinary_frequency *has_radiation_toxicity_ctcae_grade* urinary_urgency_grade_1 *has_radiation_toxicity* diarrhea *has_radiation_toxicity* nocturia	0.79

Table 5: Randomly selected patient text corpus (Target) and the top 5 similar patients text corpus utilizing cosine similarity scores using the Word2Vec model.

The analysis of these results reveals both shared attributes and attribute variations between the target patient and similar patients, we can consider the following key points:

- Shared attributes: The target patient and similar patients share several clinical attributes such as the diagnosis of malignant neoplasm of the prostate, clinical TNM findings, histology (adenocarcinoma), and primary and secondary Gleason scores. These shared attributes indicate a common disease type and some similar characteristics among the patients.
- Attribute variations: Despite the shared attributes, there are notable variations in certain attributes between the target patient and similar patients. For example:
    - Similar Patient 1: The performance status of the target patient (ECOG performance status 0) differs from the similar patient (Karnofsky performance status 100). ECOG and KPS are two different scales to measure the performance status of the patient and ECOG performance status score of 0 is equal to KPS score of 100. We also find the KPS score of 90 for similar patients 2 & 4.
    - Similar Patient 2: The target patient has radiation toxicity related to erectile dysfunction, while the similar patient does not have this toxicity.
    - Similar Patient 3: The target patient has radiation toxicity related to nocturia, whereas the similar patient does not. This implies variations in the urinary symptoms experienced during treatment.
    - Patients 4 and 5 have similar radiation toxicities (dysuria, urinary frequency, urinary urgency, and diarrhea) but differ from the exact list of toxicities listed for the target patient (erectile_dysfunction, urinary frequency, urinary urgency, nocturia).
- Cosine similarity scores: The cosine similarity scores provide a measure of similarity between the target patient and each similar patient. Higher scores indicate a higher degree of similarity in the clinical attributes. In this case, the top-ranked similar patient, Similar Patient 1, has the highest cosine similarity score of 0.92, indicating a strong similarity with the target patient based on the clinical attributes considered in the analysis.

6.8 Discussion

The ontology-based keyword search program that can then be used to query the RDF graph database by clinicians and researchers based on any keyword/s. The software can match the patient records based on the synonyms and hyponyms of the search keywords and provide a list of patient records with an exact match and patients who have similar attributes in their clinical record. We also analyzed patient similarity using four different embedding models where Word2Vec model achieved highest mean cosine similarity indicating higher level of similarity among patient embedding vectors. This suggests that the Word2Vec model captures semantic relationships better, leading to more comparable patient representations. When examining distance metrics, the GloVe model stood out with lower mean Euclidean and Manhattan distances. This indicates that patient embeddings derived from the GloVe model are more compact and closer in proximity, signifying a more clustered distribution of similar patients. The choice of which model is better for an application depends on the specific requirements and priorities. If the ability to capture semantic relationships and identify patients with similar attributes is crucial, the Word2Vec model may be more suitable. Conversely, if compactness and clustering of similar patients are of primary importance, the GloVe model may be preferred. These findings provide valuable insights into the performance and characteristics of the different models, enabling researchers and practitioners to make informed decisions about which model best suits their specific requirements. Our designed search tool is useful for cohort identification and can potentially be used to identify patients and their inherent data for quality measure analysis, comparative effectiveness research, continuous quality improvement and most importantly to support the use, training, and evaluation of machine learning models directly for streaming clinical data.

It is important to consider the limitations of the analysis. The analysis is solely based on the categorical clinical attributes, and other relevant factors, such as DVH scores that are continuous numerical variables have not been considered for our patient similarity analysis. This is because the word embedding models require the input features included in its dictionary before it can generate the vectors. For numerical variables it is not possible to include all the numerical attributes in the training datasets for the word embedding models. Additionally, the word embedding model and cosine similarity scores have their own limitations and may not capture the full complexity of patient similarity. These results provide a starting point for exploring patient similarity and can guide further analysis and investigation. It would be valuable to validate the findings using additional patient data, evaluate the clinical significance of attribute variations, and assess the impact of patient similarity on treatment outcomes and prognosis.

Moreover, it is important to acknowledge the limitations of the word embedding model and cosine similarity scores employed in the analysis. These techniques may not fully capture the intricacies and nuances of patient similarity. Different aspects of patient data, such as demographics, medical history, and treatment details, may require more sophisticated similarity measurement techniques to account for their multidimensional nature. Exploring alternative or advanced similarity measurement methods could potentially improve the accuracy and effectiveness of the ontology-based search tool. To further strengthen the findings and conclusions derived from the analysis, it is crucial to validate the results using a larger and more diverse patient dataset. Moreover, it is important to be aware of potential biases in the data, such as underrepresentation of certain demographic groups, diagnosis category or treatment type. Understanding and mitigating these biases will contribute to the robustness and fairness of the ontology-based patient search tool.

Furthermore, the current word embedding model and cosine similarity scores used in the analysis may have limitations in capturing the intricacies of patient similarity. Exploring alternative or advanced similarity measurement methods can potentially improve the accuracy and effectiveness of the ontology-based search tool. Novel techniques such as Similarity Network Fusion (SNF), as proposed in [12], offer promising avenues for integrating diverse data types and capturing both shared and complementary information. SNF utilizes patient similarity networks inferred from different data samples and combines them into a single patient similarity network using a nonlinear combination method. Investigating the applicability and benefits of SNF and similar approaches in the context of the ontology-based search tool can enhance its capability to identify coherent patient subtypes and derive clinically relevant insights.

Validating the findings and conclusions derived from the analysis using larger and more diverse patient datasets is another crucial research direction. Validation ensures that the identified patient similarities and search results are consistent across different populations and can be generalized to broader contexts. Future studies should focus on acquiring and analyzing comprehensive datasets that encompass various patient cohorts, treatment modalities, and disease types to validate the effectiveness and robustness of the ontology-based search tool.

In addition, efforts should be made to improve the interpretability and explainability of the ontology-based search tool. Techniques like rule-based reasoning, natural language generation, or case-based reasoning can be employed to generate explanations that are tailored to the user's context and easily comprehensible. Providing insights into the importance and contribution of different features or attributes in the search results can enhance interpretability. By quantifying the relevance or impact of specific attributes in the search algorithm, users can understand why certain patients or treatments were

prioritized. Incorporating uncertainty estimation methods can help users gauge the reliability and confidence of the search results. Uncertainty measures, such as confidence intervals or probabilistic approaches, can provide information about the level of uncertainty associated with the results. By quantifying the uncertainty, users can better interpret and contextualize the search outcomes, making more informed decisions. Providing meaningful explanations for the search results can enhance the tool's usability and trustworthiness for clinicians and researchers. Exploring techniques to generate explanations based on the ontology structure and the specific attributes driving the search results can facilitate better understanding and utilization of the tool in clinical decision-making.

Conducting rigorous evaluation studies and user feedback sessions are crucial for assessing the interpretability and explainability of the ontology-based search tool. User studies can gather insights into how clinicians and researchers perceive and interpret the search results, identify areas of confusion or improvement, and guide the refinement of the tool's interpretability features.

References:

1. Wang H., Zhang Q., Yuan J.; Semantically enhanced medical information retrieval system: a tensor factorization-based approach; IEEE Access, 5 (2017), pp. 7584-7593
2. Djenouri Y., Belhadi A., Belkebir R.; Bees swarm optimization guided by data mining techniques for document information retrieval; Expert Syst. Appl., 94 (2018), pp. 126-136
3. Li F., Liao L., Zhang L., Zhu X., Zhang B., Wang Z.; An efficient approach for measuring semantic similarity combining WordNet and Wikipedia; IEEE Access, 8 (2020), Article 184318-184338
4. Araque O., Zhu G., Iglesias C.A.; A semantic similarity-based perspective of affect lexicons for sentiment analysis; Knowl.-Based Syst., 165 (2019), pp. 346-359
5. Mahadzir N.H., Omar M.F., Nawi M.N.M.; Semantic similarity measures for Malay-English ambiguous words; J. Telecommun. Electron. Comput. Eng. (JTEC), 10 (1–11) (2018), pp. 109-112
6. Zhang J., Chen M., Hu E., Wu L.; Data mining model for food safety incidents based on structural analysis and semantic similarity; J. Ambient Intell. Humaniz. Comput. (2020), pp. 1-15
7. N.F. Noy, et al.; BioPortal: ontologies and integrated data resources at the click of a mouse; Nucleic Acids Res, 37 (Web Server issue) (2009), pp. W170-W173
8. P.L. Whetzel, et al.; BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications; Nucleic Acids Res, 39 (Web Server issue) (2011), pp. W541-W545
9. Mikolov, Tomas; et al. (2013). "Efficient Estimation of Word Representations in Vector Space". arXiv:1301.3781
10. Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32 (ICML'14). JMLR.org, II–1188–II–1196.
11. Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
12. P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching Word Vectors with Subword Information; arXiv preprint arXiv:1607.04606

7. 3D Deep Convolution Neural Network for Radiation Pneumonitis Prediction Following Stereotactic Body Radiotherapy

This chapter focuses on the implementation of a learning health system framework for predicting radiation pneumonitis, a potential side effect following stereotactic body radiotherapy (SBRT) treatment. SBRT is a highly precise and effective technique for delivering radiation therapy, but it can increase the risk of pneumonitis, which is a severe inflammatory response in the lungs. The primary objective of this chapter is to present a novel approach using a 3D deep convolutional neural network (CNN) for predicting the occurrence of radiation pneumonitis in patients undergoing SBRT. By leveraging the power of artificial intelligence and deep learning techniques, this approach aims to provide accurate and reliable predictions that can aid in clinical decision-making and improve patient outcomes.

The chapter begins by providing an overview of radiation pneumonitis and its impact on patients' quality of life and treatment outcomes. It highlights the need for effective prediction models and provides a review of the published models for these predictions.

Next, the chapter delves into the principles and architecture of deep convolutional neural networks, emphasizing their ability to learn complex spatial features from volumetric medical imaging data. It discusses the advantages of using 3D CNNs model architecture over traditional 2D approaches in capturing three-dimensional information and extracting meaningful patterns from medical images.

Subsequently, the chapter outlines the methodology employed for training and validating the 3D CNN model. It covers aspects such as data acquisition, preprocessing, network architecture, training strategies, and performance evaluation metrics. Special attention is given to addressing challenges specific to radiation pneumonitis prediction, including limited data availability and class imbalance.

Furthermore, the chapter discusses the integration of the developed predictive model within a learning health system framework. It highlights the importance of data collection, aggregation, and analysis to continuously improve the model's performance and facilitate knowledge sharing among healthcare providers.

Finally, the chapter concludes with a discussion to identify salient regions within the input 3D image dataset via an integrated gradient technique that provided important details of the tumor surrounding volume in the patient RP stratification. Overall, this chapter provides valuable insights into the implementation of a learning health system framework for radiation pneumonitis prediction following SBRT. It highlights the potential of deep learning techniques to enhance patient care, improve treatment outcomes, and contribute to the advancement of precision medicine in radiation oncology.

7.1 Introduction

Stereotactic body radiotherapy (SBRT) is the standard of care for medically inoperable patients with early-stage NSCLC resulting in excellent local control and typically low treatment-related morbidity [1, 2]. Among the most common complications observed with SBRT are radiation pneumonitis (RP) and pulmonary fibrosis. Due to the smaller treatment volumes, the incidence of RP is less than in locally advanced lung cancers and is generally observed in ≤10% of patients after up to 6 months from treatment completion [3]. Despite its lower incidence, RP is a serious side effect with potentially lethal outcomes in this population with typically severely compromised lung function. Radiographic signs of RP are observed

with CT images that indicate ground-glass opacities and patchy or confluent consolidations in the lung tissue [4]. These imaging characteristics are seen within 3-6 months post-treatment [5]. The diagnosis of RP is often subjective and is typically based on clinical evaluation and radiological findings. While several risk factors associated with RP have been identified, such as dose- and volume-dependent factors [3] or interstitial lung disease [6], the prediction of the individual RP risk is difficult and complex.

Various studies have been performed to assess lung density changes on CT as a metric of parenchymal lung changes after conventional radiotherapy (RT) and SBRT. Dose-dependent increases in regional lung density with conventional RT [7] and SBRT [8, 9] were identified, but no quantitative correlation has been established for predicting RP. Furthermore, for several pneumonitis patients, no major lung density changes have been observed [10]. Studies have also shown the average increase in lung density is related to the percent reduction in pulmonary function tests indicating functional lung changes [11]. Lung density changes, location of the tumor (upper vs lower lobe), total lung volume, and radiation field design are some of the attributes that contribute to radiation-induced toxicities [12]. There is a need for multiple higher-order pattern recognition metrics and techniques that can capture and model the intricacies of these toxicity patterns.

Deep learning approaches have shown great promise in the medical imaging domain with image-recognition tasks where intricate biological interactions are extracted more effectively without defining these features manually. As opposed to the subjective visual assessment of images by trained physicians and extracting engineered features such as radiomics, these deep learning methods automatically identify and quantitatively evaluate complex patterns in the dataset and select the most robust features. Deep learning methods have performed better than their traditional statistical counterparts in many imaging tasks such as multimodality image registration [13], automatic contouring [14], and survival analysis [15]. Convolutional neural networks (CNNs) are a class of deep learning methods that combine different sizes of imaging filters with a network of neurons through a series of interconnected linear and non-linear layers. As part of the training, the CNN image filters learn high- and low-level imaging features, eventually making predictions on the desired outputs. Li et al. first applied a 3D CNN model for the evaluation of treatment response in locally advanced esophageal squamous cell carcinoma [16]. Ibragimov et al. applied CNNs to 3D dose distributions of the rectum surface for toxicity predictions [17]. With a few exceptions, most of these studies lack generalization of their models and results due to insufficient data – under 100 patients. Many of these studies have used 2D data for their efforts or alternatively, used 3D datasets with a limited volume in and around the tumor region only. It is important to note that none of these methods have been utilized as a part of the clinical routine yet. Since there are very few published and shared 3D CNN models that have been trained on medical or general images, there are no medical-to-medical transfer learning approaches being applied to solve similar imaging classification problems until now.

In this study, we investigate 3D convolutional neural networks with the input of radiographic and dosimetric characteristics in the lung tumor and surrounding lung volume to predict the likelihood of RP for NSCLC patients treated with SBRT. Reliable prediction of pneumonitis risk may guide individualized treatment approaches and reduce pulmonary toxicity. We designed an analytical setup with a dataset of NSCLC patients imaged before radiotherapy (pre-treatment imaging), and 3 and 6 months after treatment (post-treatment imaging), to discover the prognostic power of CNNs as a binary RP classification problem.

7.2 Methods

7.2.1 Dataset

This study includes a total of 193 primary lung cancer patients treated with SBRT from 2008 to 2020 at our facility. Following approval by the institutional review committee, clinical data, radiographic images, and dose distribution matrices were extracted from the medical record, PACS, and the treatment planning system to create a database for analysis. Only patients who had a follow-up visit with CT scans at 3 and 6 months were included in this study. Table 1 details the clinical and dose prescription characteristics of the patients included in this study. The internal gross tumor volume (iGTV) was delineated on the maximum intensity projection (MIP) images from the 4D CT (Brilliance Big Bore, Philips, Amsterdam, Netherlands) dataset, or on a single-phase image and propagated over all phases of the 4D CT. For patients treated in breath hold, an iGTV was created based on 3 repeat breath-hold CTs. The pixel spacing for these images is 0.98-1.37mm with a slice thickness of 3mm, in-plane matrix size of 512x512 acquired with 120-140 kVP. The planning target volume (PTV) was generated by adding an expansion margin of typically 5 mm. The dose covering 95% of the PTV volume was 102.5 ± 4.2% of the prescription dose. Image-guidance using cone beam CT was applied to align the target daily prior to treatment delivery. RP was clinically evaluated at 3- and 6-month follow-up visits based on clinical symptoms and radiographic findings.

7.2.2 Imaging and Treatment Planning Dataset

The pre-treatment images used for treatment planning comprised of either an end inspiration CT scan for patients treated with breath hold or the 30% phase or average image set of a 4D CT scan. Post-treatment imaging was performed with comfortable inspiration breath hold on diagnostic CT scanners at 3- and 6-month intervals after the completion of the radiation treatment. These CT scans were acquired typically with end inspiration scanning techniques. All thoracic CT scans for follow-up visits were acquired on either GE or Siemens CT scanners and reconstructed with sharp kernels. A total of 579 3D CT images and 193 dose datasets were analyzed during this study.

7.2.3 Image Registration

To account for changes in the anatomy between the CT scans used at treatment planning (pre-treatment) and follow-up, deformable image registration was performed. These radiographic changes were due to post-treatment volume loss, distortion, fibrosis, and tumor regression and had an impact on the overall lung architecture. Because follow-up images were high-quality images and essentially free of artifacts, follow-up CTs could be registered directly to planning CTs (either inspiration BH or average images from 4D CTs). The follow-up CT scans were first rigidly registered to the baseline CT scan used for treatment planning. This step aligned the two 3D datasets slice by slice using a clinically utilized image registration software (MIM Maestro version 7.0). Visual inspection of the automatic rigid registration was performed with manual adjustments to the translation and rotation parameters using a box-shaped mask to align the spine and vertebral structures. In the next step, the deformable registration algorithm from MIM Maestro was utilized. To avoid the algorithm to perform non-physiological and non-realistic deformations, radiographic lung changes and PTV regions of the follow-up CT and planning CT were given a value of -250 HU prior to the registration for the algorithm to avoid overfitting to small anatomical discrepancies. The corresponding 120 cm^3 regions on longitudinal image sets were co-registered to minimize potential differences in the alignment of patient anatomy while preserving radiographic lung changes due to radiotherapy All registrations were individually verified based on the overlay of the two image sets and matching the regions of interest such as lung tissue, PTV, chest wall, airways, trachea, etc. The rigid and deformable registrations were repeatedly performed until a visually acceptable solution was found for

each dataset [Figure 24]. In the final step, the radiotherapy 3D dose distribution (RT Dose) and contours (RT Structure Set) were mapped onto the follow-up CT scans. All datasets including the registered CT, RT Dose, RT Plan (for planning dataset only), and RT Structure Set were exported in DICOM format and stored in a folder.

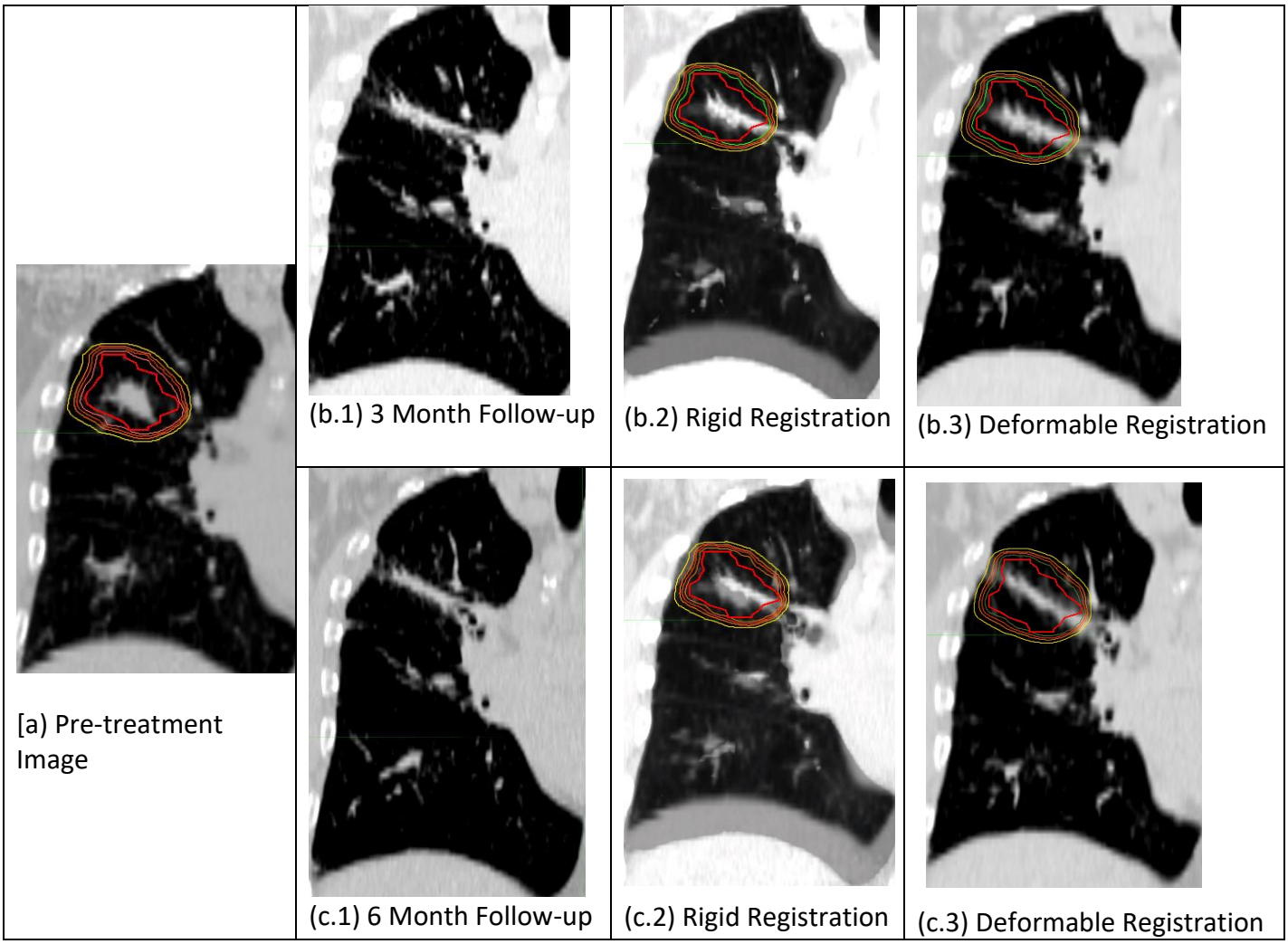


Figure 24: Example of Image Registration. (a) baseline pre-treatment (used for treatment planning) CT scan (coronal section) with the PTV (red) and isodose lines. (b.1) 3-month follow-up CT scan. (b.2) rigid registration with pre-treatment CT scan and 3 month follow-CT scan. (b.3) deformable registration with pre-treatment CT scan and 3-month follow-up CT scan with PTV volume and isodose curves. (c.1) 6-month follow-up CT scan. (c.2) rigid registration with pre-treatment CT scan and 6 month follow-CT scan. (c.3) deformable registration with pre-treatment CT scan and 6-month follow-up CT scan with PTV volume and isodose curves.

7.2.4 Data preprocessing for deep learning

All imaging, RT Dose and RT Structure Set files were imported in a custom-built software coded in python v3.7.13. All datasets were resampled into isotropic voxels of unit dimensions to ensure comparability, where 1-unit voxel is equal to $1mm^3$. These interpolations were carried out using the nearest neighbor interpolation methods for images, dose, and contour datasets. Using the full 3D tumor and lung volume

contours from the RT Structure Sets, both the center of mass and the bounding box of the PTV was computed. Using this center of mass and bounding box, 3D isotropic patches of size 120 x 120 x 120 were extracted from the imaging and dose datasets. These 3D patch extractions were manually verified and shifted in three dimensions to maximize the capture of the lung and PTV volume within the patch size. The 3D dose patches for each dataset were overlaid on the imaging patches to visually verify image-dose registrations. The imaging patches were normalized to a 0-1 range using the upper and lower HU bounds (-1024 to 2048). The dose patches were normalized to a 0-1 range using 0-60Gy dose range and any value greater than 60Gy was normalized to 1. The high-density regions outside the lung volume such as bone tissue were patched out of the input samples since these tend to be non-informative and can potentially confound the deep learning models.

Due to the fact that there is a class imbalance in our dataset where the number of patients (154 patients or 79.8%) who did not indicate radiation pneumonitis toxicity (no RP) greatly outnumber the patients who showed pneumonitis symptoms (26 patients – grade 1 (13%) (RP1), 13 patients – grade >=2 (6.5%) (RP2), data augmentation techniques were applied to randomly oversample [18] the patches with the minority class (RP1 & RP2), yielding the training size of 1182 input samples. This technique created new input samples for training without actually altering the visual characteristics of the images. These augmentations included random flipping of the 3D image and dose patches along the left-right and superior-inferior axes, random translations ±10 voxels in three-dimensional space, and random rotations of 5, 7, and 10 degrees along the longitudinal axes. These class-specific perturbations were applied to the initial test set to make our DL models robust and to reduce bias and generalization errors. These basic data augmentation techniques have been the most popular approach for recent medical imaging research [19]. With the help of these augmentation and oversampling techniques, we created an equal number of images and dose samples for the three classes used in the prediction models. This technique may also help with overfitting as it produces new patches with similar properties to the original data but also fills previously unoccupied feature space. Furthermore, similar image augmentation techniques were utilized in real-time during training for the purpose of avoiding overfitting and making the models generalizable. The total input samples were defined in an 80:20 ratio in training and testing input samples. The testing samples were not exposed to the training process. The training input samples were further split in an 80:20 ratio between the final training and validation set. The split was stratified by the three prediction classes, which ensured that an equal percent of data is taken from each class for training. Figure 25 shows the study design. The models were trained on the training dataset and used to predict the test dataset to evaluate the model performance. Once the model was trained and validated, we tested the model with the testing input samples (unseen by the model during training).

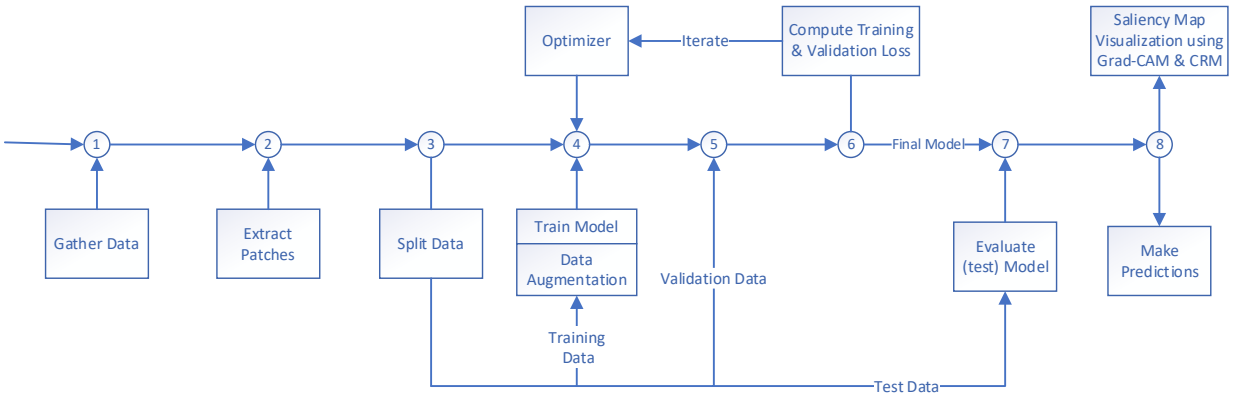


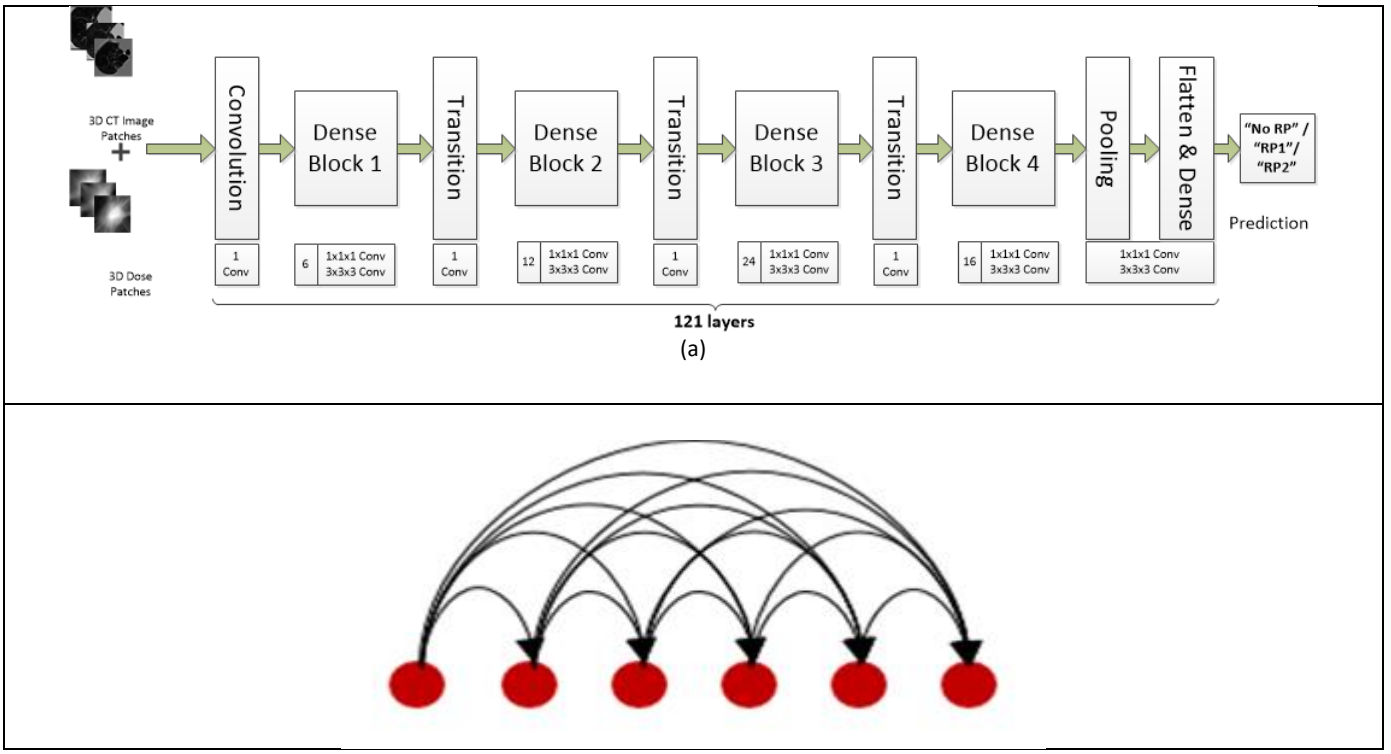
Figure 25: The study design of the proposed model for radiation pneumonitis prediction classification

7.2.5 Deep learning Architecture

Traditionally, deep learning for medical imaging datasets has been confined to 2D convolutional neural networks (CNNs) where predictions are made on a per slice basis in two-dimensional space, and then final predictions are obtained through highest probability, voting, or other methods based on the prediction results from all slices in the dataset. These methods work fine with 2D medical images such as computed radiography (CRs) and x-rays. With three-dimensional images, the algorithm loses the contextual information of the shape, size, and texture of the lesions, tumor, and organs between the slices of images, and hence the prediction performance is low. We used the 3D CNN models for this study which are very much like 2D CNNs except that it uses 3D convolution and max pooling layers.

For our study, we modified the published 2D CNN models, namely DenseNet-121 and ResNet-50 [20, 21]. These models use deeper neural networks which are generally more difficult to train. These models have the residual learning framework to ease the training of networks that are substantially deeper than those used with basic general 2D images such as VGG-16 [22] etc. Using substantially deeper networks has been shown to be more accurate and more efficient to train if they contain shorter connections between the layers that are closer to the input and closer to the output layer [23]. For our models, we modified the 2D convolution kernels of the 2D basic blocks to a 3D convolution kernel and built a 3D basic block. The basic architecture [Figure 26a] and structure [Figure 26b] of the DenseNet-121 & ResNet-50 [Figure 26c] were unchanged. With modification of the 3D basic blocks, we built 3D DenseNet-121 & 3D ResNet-50 models. With traditional deep learning models, there is a tendency for the prediction accuracy to decrease as the depth of the layers increases beyond a certain number. This problem was solved by passing features from the lower layers to the higher layers thus eliminating the vanishing gradient problem. Here all the features learned by the first few layers can be utilized throughout all the subsequent layers thus reducing the number of trainable parameters [Figure 26a]. The ResNet-50 model is built to train deeper CNNs by creating shortcuts (skip connections) between the front and back layers. DenseNet-121 is built with the same concept but it establishes dense connections of all the previous and subsequent layers. It has been reported that DenseNet achieves comparable performance to ResNet-50 with fewer parameters and less computation costs [24].

The input channels for these models were the 120 x 120 x 120 cube image patches and the same sized 3D dose patch as an additional input channel and the output were either three classes (No RP, RP1, RP2) or two classes (No RP, Yes RP) with the purpose to perform binary classification of these input patches. Since we were performing a binary classification task, the output layer of the network is a dense layer whose activation function is sigmoid, and the output values range between 0 and 1. Since CNN models are typically trained with millions of image sets and due to be unavailability of such a dataset for our study, we utilized the publicly available Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI) [25] data set to shallowly train our CNN models. These datasets are being utilized to train CNN models for lung nodule detection and are comprised of 1000 lung CT images. We trained our model for 20 epochs with a binary classification output layer with the purpose of initializing the convolution filter within our model to recognize the presence of the lung lesion in the input datasets. The IDRI dataset was utilized to initialize and pre-train the model weights before utilizing the model for actual training with our CT and dose datasets. Training details are as follows: we used the gradient-based stochastic optimizer called Adam with a learning rate of 0.001 with a decay rate of 0.96 and decay step of $1x10^4$, a batch size of 4, and trained both models for 100 epochs with a total of 11.3M trainable parameters for DenseNet-121 and 46.2M trainable parameters for ResNet-50 models. In order to avoid overfitting, we utilized early stopping techniques where the training monitored the loss function and stopped training with a patience value (the number of epochs to wait before early stop if no progress on the validation set) of 10. Our model was trained multiple times with an early stopping value of 2-20. We observed that a stopping value of more than 10 showed that the model loss was constant, and accuracy showed minor three decimal places improvements indicating overfitting. We also included batch normalization layers in order to improve convergence and generalization in the models. TensorFlow 2.8 was used to train and test CNN. Google's web-based coding note called Collaboratory, or Collab, was used to execute code on Google's cloud-based NVIDIA Tesla P100 GPU servers with 25.46GB available memory size.



(a)

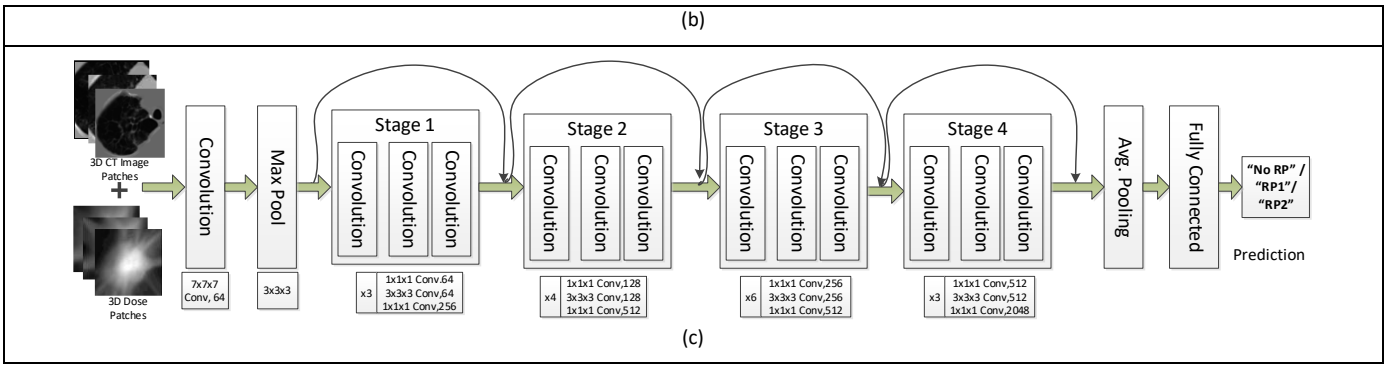


Figure 26: a) DenseNet 121 – 3D Architecture: A deep DenseNet with four dense blocks. The transition layers in between the successive dense blocks are responsible for changing the feature-map sizes via convolution and pooling operations. This architecture has 121 layers with interconnected layers in a feed forward fashion to ensure maximum information flow between layers in the network. (b) The connections between the 121-layer blocks of the DenseNet-121 CNN network where there are direct connections from any layer to all subsequent layers. The connection between different layer blocks increases variation in the input of subsequent layers via feature reuse and improves efficiency. With this architecture the vanishing gradient and loss problems are resolved since each layer has direct access to the gradients from the loss function and the original input signal, leading to an implicit deep supervision. (c) ResNet-50 – 3D Architecture: A residual network of 50 parameter layers where the subtraction of features is learned from the input of that layer by using shortcut connections which are shown as curved arrow.

7.2.6 Visual Explanation of Convolution Neural Network Predictions: Explainable AI – Integrated Gradients

Explainable AI is a research field in machine learning interpretability techniques whose aim is to understand machine learning model predictions and explain them in human understandable terms to build trust with stakeholders. In computer vision, a saliency map is an image that can be in the form of a heat map that shows each input voxel's unique quality with the goal to represent the image and the predictions from the DL models into something that is more meaningful and easier to analyze. These heat maps predominately help explain how a model made its decision on a particular dataset although these explanations are not guaranteed to make sense to human experts and these explanations are also not considered to follow any known rules or decision trees that are traditionally used for image classification algorithms. These heat map techniques do not make the CNN model interpretable where accuracy and human understandable relationships could be derived between the inputs and the output of the CNN models. Our purpose with using these saliency heat map techniques was to qualitatively review the model and gather more insights into model predictions. This is clearly a developmental goal where we are testing if these maps can be usefully created for our 3D datasets and understand the general areas in the image where the network is focused to make the predictions. Explainability and interpretability of the deep learning model is a topic of research [26] and with this work, we are trying to use industry-standard techniques that have been successfully utilized to highlight areas in real-world non-medical photographic images for CNN model predictions. Here we are briefly describing the integrated gradient methods that are used to localize and highlight important regions of interest (ROIs) for a particular category within an input image set with the purpose to explain the CNN model predictions. The feature attribution heat maps help to highlight frequently missed features in the 3D imaging datasets thus complementing human judgment by physicians with accurately grading the dataset.

7.2.7 Integrated Gradients (IG)

To visualize the highlighted map of the most important features in an input image set with respect to predictions made, we utilized the integrated gradient method (IG) [27]. The computation of gradients of the model output with respect to the input features provides an analog for feature importance. IG [28] uses the gradient information of a target class flowing back into the last convolution layer to form visual heat maps from the CNN-based DL models. This method provides pixel-based maps that measure the contribution of each pixel in the input image to a predicted pneumonitis class. The contributions are measured relative to a baseline image, which is intended to provide no information to the model. For this study, we used a black image as the baseline. We verified that our trained algorithm predicted no prediction for this baseline. For each 3D image-set, we generated a path of 100 steps, in which each step was interpolated between the blank baseline image-set and the target image-set. For each output pneumonitis class, we summed model gradients over each pixel and took the absolute value. We then surfaced the heatmap for the pneumonitis class with the highest score.

7.2.8 Evaluation Metrics

The dataset used in this work is highly imbalanced with a smaller number of samples with an RP status than the non-RP status. The metrics used to evaluate the performance of these models need to be agnostic to the data imbalance. Since we are evaluating a multi-class problem, we have used macro-averaged metrics instead micro-averaged ones. The overall performance of a multi-class classifier is commonly obtained by taking an average of individual class performances. The advantage of using these metrics is that micro-average assigns equal weight to each sample (or instance), whereas macro-average assigns equal weight to each type (or class), and with highly imbalanced datasets, reporting micro-averaged performance would be misleading because the class with fewer samples (i.e., rare class) are given less importance than the class with more samples. With macro-averaged metrics, equal importance is given to all classes irrespective of the number of samples in each class.

The expressions for the macro-averaged metrics are as follows.

$$Precison_{macro} = \frac{1}{N}\sum_{c=1}^{N}\frac{TP_c}{TP_c + FP_c} (4)$$

$$Recall_{macro} = \frac{1}{N}\sum_{c=1}^{N}\frac{TP_c}{TP_c + FN_c} (5)$$

$$F_1Score_{macro} = \frac{1}{N}\sum_{c=1}^{N} 2.\frac{Precision_c - Recall_c}{Precision_c + Recall_c} (6)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} (7)$$

TP (True Positive): when the model predicted as positive, and the ground truth is positive (e.g., an 'RP1' sample is identified as 'RP1' by the model)

TN (True Negative): when the model predicts as negative, and the ground truth is negative (e.g., an 'RP1' sample is not identified as a 'No RP' sample by the model)

FP (False Positive) (Type I Error): when the model predicted as positive, but ground truth is negative (e.g., the model predicts the sample to the 'RP1' class, but the ground truth is false)

FN (False Negative) (Type II Error): when the model predicted as negative, but the ground truth is positive (e.g., the model does not predict the sample to the 'RP1' class, but the ground truth is true)

Accuracy is the proportion of correct predictions over the total number of samples. Recall (Sensitivity) is defined by calculating the total predicted positives out of the total number of actual positives. Precision (Positive Predictive Value) is defined by calculating the total number of actual positive samples out of predicted positive samples. F1 score is the harmonic average of the Precision and Recall values and is a widely used evaluation measure for a classification problem.

We also used the confusion matrices to assess the model performance. The correct / incorrect number of predictions are shown by the count values which are further divided into individual classes. A confusion matrix contains information about the model's confusion in predicting between classes and performance of a model. The performance of the models was also assessed via the areas under the receiver operating characteristic curve (AUCs). Finally, the AUCs of the two models (DenseNet-121 & ResNet-50) were compared by the Mann-Whitney U test with the Bonferroni correction. Two-sided $P<0.05$ were considered to indicate statistical significance.

7.3 Results

7.3.1 Clinical characteristics

The patients' clinical and dosimetric characteristics are provided in Table 6. The patients had follow-up visits with the treating radiation oncologist 3 months following the completion of radiation treatment and subsequently every 3 months thereafter. Clinical RP was evaluated and scored by the treating oncologists as part of the patient's assessments according to the common terminology toxicity criteria (CTCAE) version 5 [29]. Our cohort is quite homogeneous (95% have stage I-IIA, there are no major variations in performance status, 93% were treated with BED ≥100 (as generally recommended). We do not expect to find any imbalance in our dataset based on this homogeneity. The only factors that vary and might influence RP risk are PTV size and dose, both of which are already inherently included in the models. Patients with higher stages had either oligometastatic disease with a small primary lung cancer and isolated metastases outside the chest or had lung cancers in separate lobes where one of the tumors was resected. Fractionation schedules were selected based on tumor size and location. P-values are computed from the chi-square test of independence to determine the significant association between the specific variable listed in Table 6 and RP status. Since none of the p-value are less than 0.05, we fail to reject the null hypothesis. This means we do not have sufficient evidence to say that there is an association between the clinical characteristics and RP status.

Clinical Characteristics		P-value
Staging (No of patient)	IA (n=147)	0.58

	IB (n=27) IIA (n=10) IIB (n=2) IIIA (n=3) IV (n=4)	
Gender	Male (n=99) Female (n=94)	0.22
Median Age (range)	69.2 ± 10.4 years	0.54
Karnofsky score (%)	80 ± 10	0.90
PTV Volume	36.05 ± 29.6 cc	0.17
Prescription Dose (Gy) / fractions (No of patients)	48 Gy / 4 (n=142) 50 Gy / 5 (n= 26) 40 Gy / 5 (n= 12) 60 Gy / 5 (n=6) 60 Gy / 8 (n=5) 45 Gy / 5 (n=2)	0.11
RP Status (Number of patients)	No RP [RP0] (n=154) RP Grade 1 [RP1] (n=26) RP Grade>=2 [RP2] (n=13)	

Table 6: Patient Characteristics and Treatment Regimen for training and validating the 3D CNN models

7.3.2 Prediction performance of the 3D DenseNet-121 model

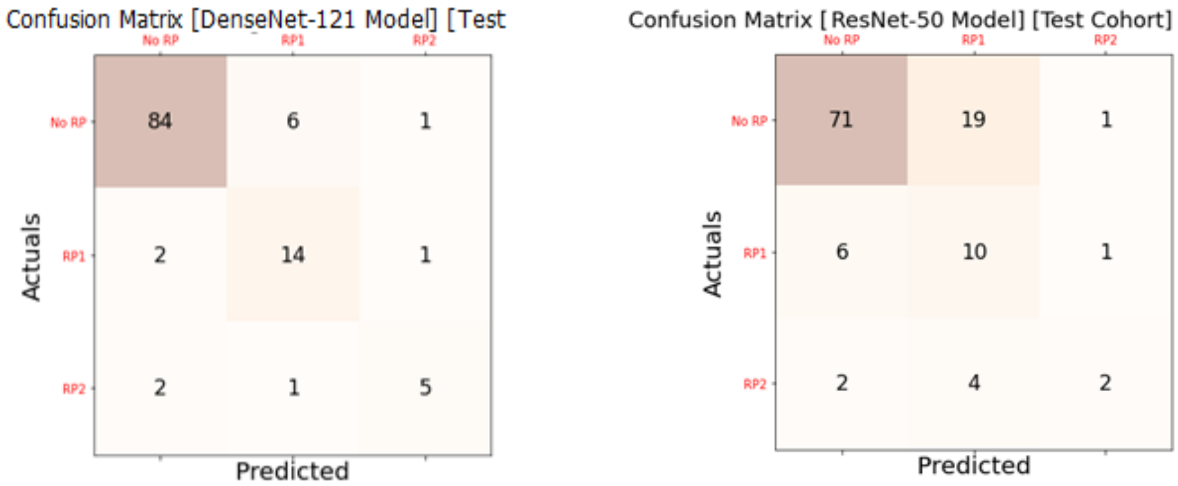
In assessing the ability of CNN models to quantify radiographic traits and characteristics of the lung and tumor region, we performed an analysis based on the test dataset that was never exposed or seen by the model during training. In this section, we present the results and prediction evaluation of our models. Here we made two copies each for the 3D DenseNet-121 and ResNet-50 models, the first two models were trained to predict three classes (No RP [RP0}, RP Grade 1 [RP1], RP Grade >=2 [RP2]) and the next two models were trained to predict two classes (No RP, Yes RP (including all grades)). All models were trained with 80% of pre-treatment and follow-up CT datasets with the 3D dose patches and tested using the unseen remaining 20% of the dataset. All models were independently trained, validated, and tested. Figure 27 shows the confusion matrix and ROC and True Positive Rate vs False Positive Rate (precision vs recall) curves for the 3D DenseNet-121 and ResNet-50 models for the three-class prediction. For three class predictions, the DenseNet-121 model had an AUC_{macro}=0.91 and F1 $score_{macro}$=0.81, whereas the ResNet-50 model had an AUC_{macro}=0.72 and F1 $score_{macro}$=0.54 [Table 7]. Mann Whitney U test performed for pair-wise comparison of AUCs among the two model types had a p-value of 0.017 indicating that the three-class DenseNet-121 showed significantly better performance than the ResNet-50 model**.** Figure 27 shows the confusion matrix, ROC, and True Positive Rate vs False Positive Rate (precision vs recall) curves for the 3D DenseNet-121 & ResNet-50 models. The test dataset contains 116 data samples, 91 of which are 'RP0' samples, 17 were 'RP1' samples and 8 were 'RP2' samples. From the confusion matrix shown in figure 27, the DenseNet-121 model can accurately identify 84 'RP0', 14 'RP1', and 8 'RP2' samples for three class predictions. In total, this model could accurately identify a total of 103 data samples (88%). The ResNet-50 model could accurately identify a total of 83 data samples (72%). For two class predictions, the DenseNet-121 model had an AUC_{macro}=0.84 and F1 $score_{macro}$=0.77, whereas the ResNet-50 model had an AUC_{macro}=0.71 and F1 $score_{macro}$=0.68 [Table 7]. The Mann Whitney U test performed for pair-wise comparison of AUCs among the two model types had a p-value of 0.527. Figure 28 shows the confusion

matrix, ROC, and True Positive Rate vs False Positive Rate (precision vs recall) curves for the 3D DenseNet-121 & ResNet-50 model. The test dataset contains 116 data samples, 91 of which are 'No RP' samples and 25 of which are 'Yes RP' samples. From the confusion matrix shown in figure 28, the DenseNet-121 model can accurately identify 80 'No RP' and 17 'Yes RP1' samples. In total, this model could accurately identify a total of 97 data samples (83%). The ResNet-50 model could accurately identify a total of 89 data samples (77%).Since these models showed higher accuracy than prediction models that utilize dose volume histogram and dose function histogram [30] based prediction models (AUC = 0.73) using support vector machine techniques, logistic regression classifiers [31] (AUC values 0.64-0.75), we are confident in the DenseNet-121 model's ability to accurately assess whether a patient would have RP or not, even when predicting new, unseen patients.

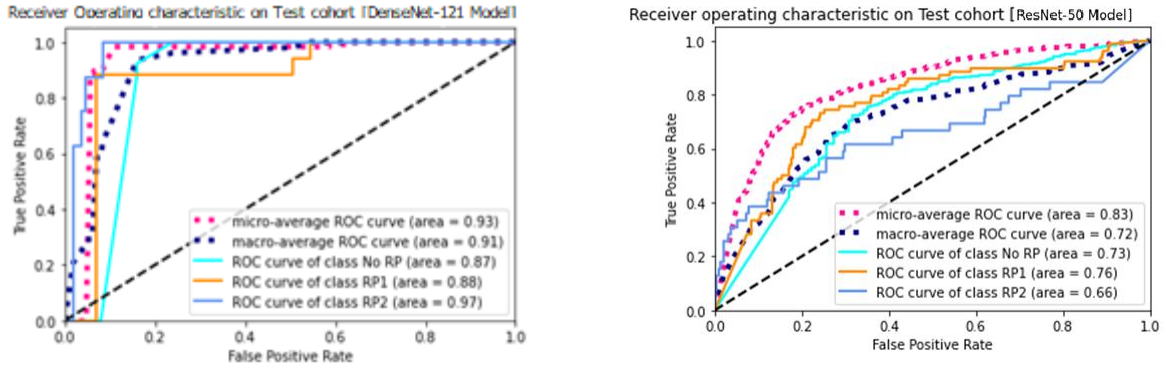
Training Dataset	**Prediction classes**	**Testing Dataset**	**Model**	**Precision** macro	**Recall** macro	**F1 score** macro	**Accuracy**	**AUC** macro	**p-value**
Pre-Treatment and Follow-up Dataset [80% split]	Three class prediction	**Test Cohort [20% split]**	**DenseNet-121**	**0.84**	**0.78**	**0.81**	**0.89**	**0.91**	**0.017**
			ResNet-50	**0.57**	**0.59**	**0.54**	**0.72**	**0.72**	
	Two class prediction	**Test Cohort [20% split]**	**DenseNet-121**	**0.76**	**0.78**	**0.77**	**0.84**	**0.84**	**0.527**
			ResNet-50	**0.67**	**0.69**	**0.68**	**0.77**	**0.71**	

Table 7: Macro-averaged Precision, Recall, F1 score and overall accuracy for the four models based on the test cohort (not seen or trained on the model) and training cohorts.

Three Class Prediction Models



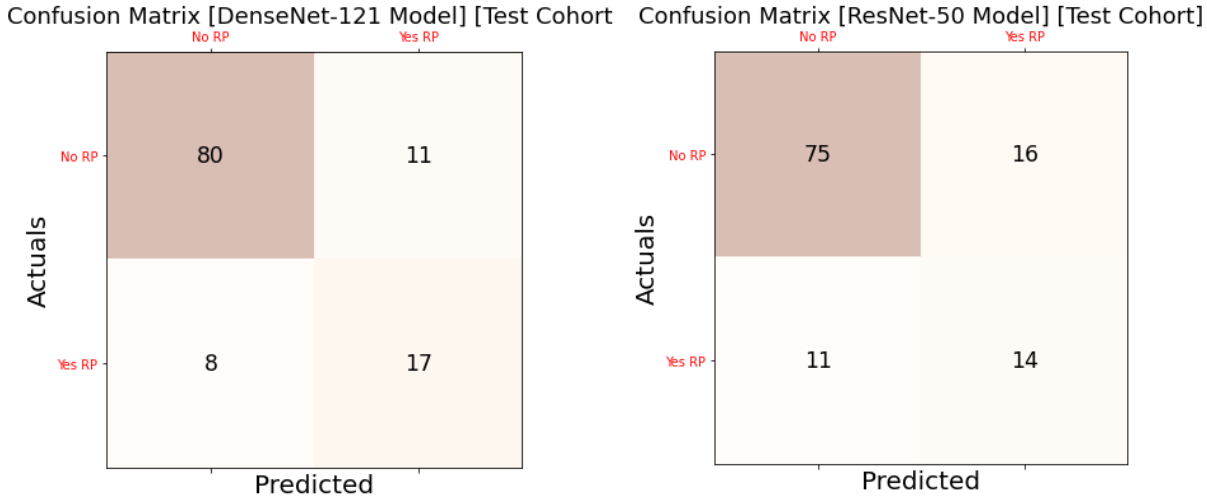
(a) (b)



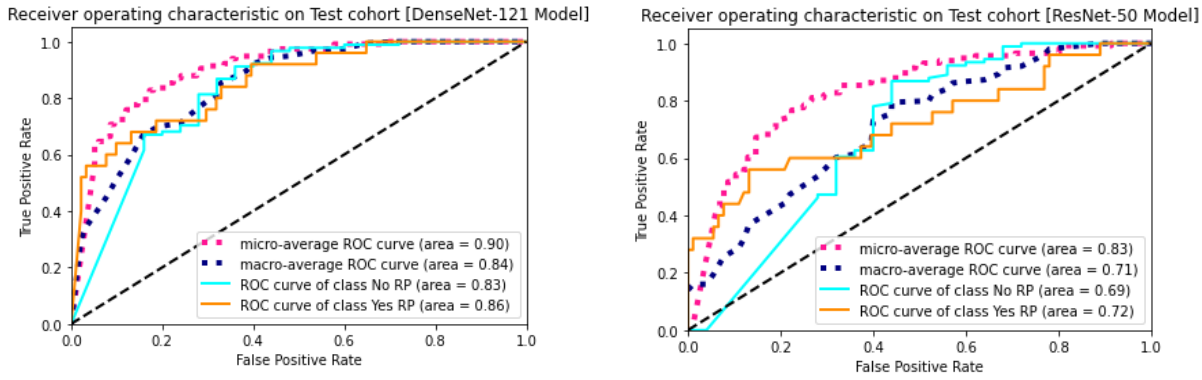
(c) (d)

Figure 27: Evaluation of the 3D Dense-121 vs ResNet-50 model trained with 3D image + 3D dose patches from the pre-treatment and follow-up datasets. (a) Confusion Matrix for three class prediction with the 20% sample set [Test set] that was not seen or trained on the DenseNet-121 model. (b) Confusion Matrix for the ResNet-50. Darker color cells demonstrate more accurate predictions, and the diagonal shows the labels predicted correctly. (c) Prognostic power (ROC) and True Positive Rate vs False Positive Rate curves derived from the test set for Dense-121 model. (d) Same chart for the ResNet-50 model.

Two Class Prediction Models



(a) (b)

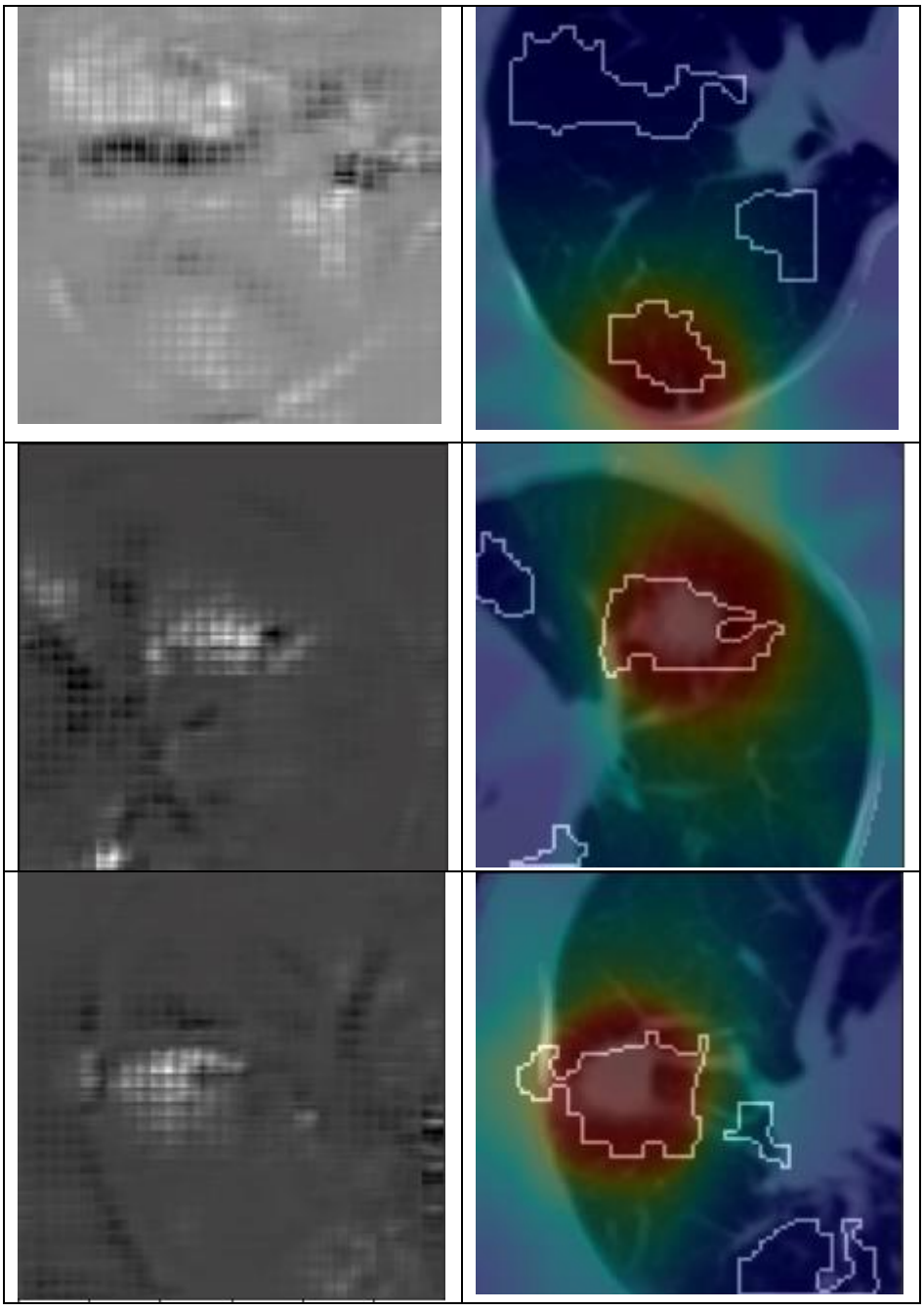


(c) (b)

Figure 28: Evaluation of the 3D Dense-121 vs ResNet-50 model trained with 3D image + 3D dose patches for the two-class prediction.

7.3.3 Localization Evaluation

We also analyzed the 3D DenseNet-121 model built to predict the three classes by rendering the integrated gradient (IG) heat maps that provide information on the importance of each region and voxel relative to the final prediction class. The IG map can help understand how the model scores the input data with learned features (e.g., dose distribution pattern, imaging features) from different regions of the input. Generally, regions of the IG map with bright regions correspond to regions whose learned features are more important for RP prediction. We observed that the network highlighted the regions of the tumor and its interface with the parenchyma or pleura regions of the lung which shows that these regions are critical for the model in discriminating toxicity from non-toxicities. Based on our qualitative analysis of thirty RP1 + RP2 patient cases, these heat maps agreed with the tumor and dose regions and their surrounding voxels and also included some voxels outside the tumor or dose regions, as shown in Figure. 29. Relatively higher density regions that included the tumor and the surrounding regions had the most contributions to the predictions. In some patients, the heat maps showed a rather focused area, whereas in other patient cases, these heat maps were rather widespread covering large parts of the lung. None of these models aim at providing physiological modeling using the tissue type characteristics of the tumor and lung structure and are using the information present in the image textures and patterns to pick out unique traits and attributes that are used in the eventual classification. As such, while showing the high dose and tumor area as important for RP prediction is reasonable from the clinical perspective, identification of other lung areas as relevant for RP through the IG process cannot be reasonably verified based on the current understanding of RP development. Examples of the IG heat maps are shown in Figure 29 for five patient cases who had grade 2 RP.



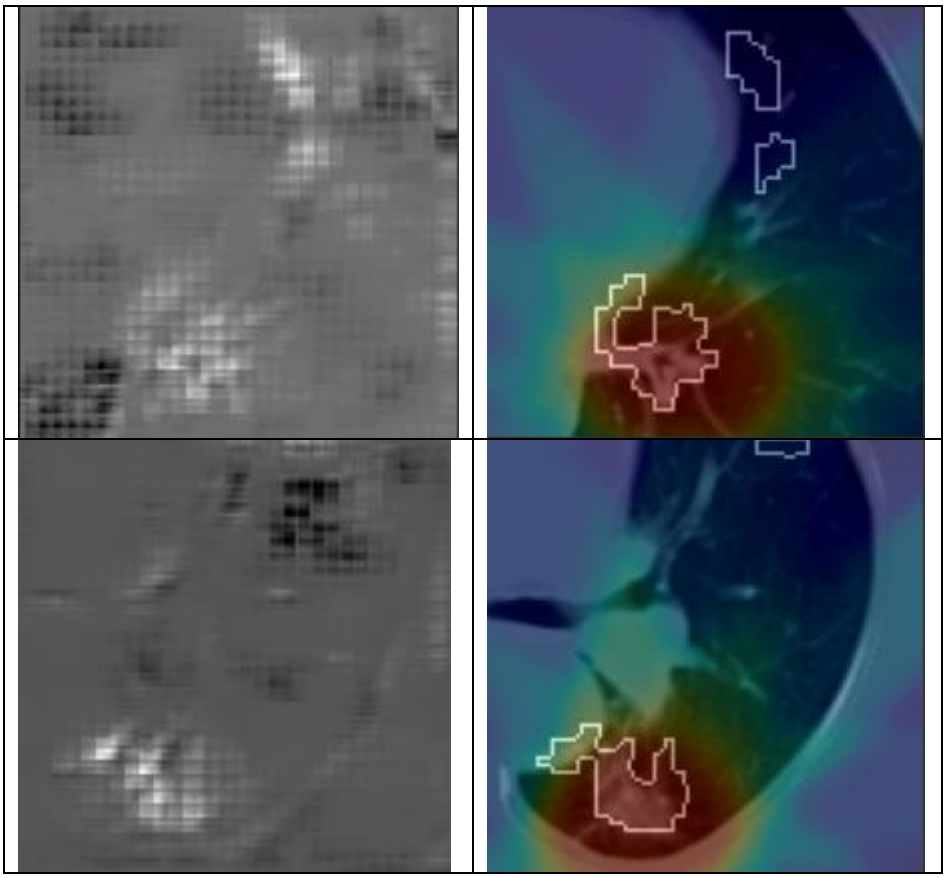


Figure 29: Visual display of the most important areas of the input 3D dataset that have the most contributions to maximize the outputs of the final prediction layer used to predict a RP case. The rows represent five different patient samples (axial slice) that encompass the PTV volume. The first column displays the Integrated Gradient heat maps. Bright (white) regions represent positive gradients, and dark (black) regions show negative gradients. The second column represents the CT patch annotated dose maps (displayed as heat map) and contours of voxel regions that are in top 50% of IG maps.

7.4 Discussion

This is one of the first studies to predict post-SBRT pneumonitis using a comparatively large CT dataset. We demonstrated the ability of CNN models to significantly stratify patients into none vs grade 1 vs grade 2 or higher clinical RP groups. We chose imaging and dose datasets for patients with early-stage lung cancer since these datasets are in general much cleaner (no pneumonia, atelectasis, or other pathologic changes) than images of patients with larger tumors. Also, radiation dose distributions are more localized and conformal, and post-RT changes are easily identifiable. Despite its low incidence, RP is a serious side effect with potentially lethal outcomes in this population with typically severely compromised lung function. Often patients are already on supplemental oxygen or on oral steroids before treatment. RP grade 2 or higher, even if not lethal, can lead to further impairment of a patient's lung function, requiring initiation of oxygen supplement, reduced mobility, loss of work, need to go on disability, or limited ability to do activities of daily living.

This study applies a common deep learning approach used in non-medical applications such as predicting image features from millions of 2D images to the medical imaging domain. Very few deep learning studies to date have explored stratification based on outcome parameters, with most studies exploring tasks such

as image segmentation [32] or malignancy detection [33]. One of the many advantages of deep learning approaches is the automation of feature extraction whereas the traditional radiomics approaches have relied on complex manual feature extraction and selection techniques using shallower neural networks such as support vector machines and random forest algorithms [34]. One of the major drawbacks of this approach is the dependence on the manually extracted engineered features where certain fine and minuscule imaging patterns may be neglected or missed and hence be unavailable for the machine learning models to be utilized for prediction purposes. The deep learning inputs are comprised of 3D voxel cubes that allow the network to consider not only the tumor volume but also the surrounding regions that, according to our study, appear to have high predictive power. More recently, deep learning approaches have become popular and are used as the de-facto standard for machine learning on medical images [35, 36].

Since our data is highly imbalanced (No RP vs Yes RP), accuracy may not be the right measure to evaluate the performance of the models since they are based on how many samples, both positively and negatively, were correctly classified. Higher accuracy scores can be obtained by correctly classifying all the samples from the majority (No RP) class. Though class balancing approaches and data augmentation techniques have been utilized on the training dataset, the AUC and F1 scores on the test dataset were calculated by taking the harmonic mean between the precision and recall values. This provides a good measure to evaluate the model performance with imbalanced test datasets. In this study, we built 3D DenseNet and ResNet-50 models of three (No RP, RP1, RP2) and two (No RP, Yes RP) classes as outputs. The Dense-Net-121 models performed better than the ResNet-50 models with statistically significant results for the 3-class prediction model. This could be due to the DenseNet-121 model architecture where dense connections are established between all previous and subsequent layers and features learned in the top layers (e.g., ground glass opacity within the lung or consolidation changes in the lung parenchyma, etc.) with coarse convolution layers would also contribute to the eventual decision making for this classification. The three-class prediction AUC_{macro} (0.91) was better than the two-class prediction (0.84) for the DenseNet-121 model. With ML approaches, the assumption is that effects of radiotherapy on -in our case - lung tissue are complex and not deterministically defined. Useful models depend on the quality of input data to produce reasonable predictions. One thought is therefore that models learn better from well-characterized feature classes that show obvious differences (by differentiating between RP1 and RP2), as opposed to creating a mixed feature class of RP1 and RP2 together. There could also be other factors that are inherently related to RP0, 1, and 2 that we are not aware of which nonetheless can lead to a clearer differentiation between RP0, 1, and 2 as opposed to RP0 versus RP1 and 2 combined. On the other hand, we cannot exclude that the number of samples from one of the two minority classes (RP1 or RP2) when concatenated might be randomly oversampled more than the other class. This would create more samples of one of the RP1 or RP2 classes in the combined Yes RP class used for the two class prediction models. This would be the reason the two class models with mixed feature class of RP1 and RP2 would have a lower AUC than the three-class model since the model would learn features for the class that has more input samples and not learn features for both classes in a balanced manner.

Our results using the images and dose patches as model inputs are consistent with published literature on deep learning-based models for radiotherapy toxicity predictions. Ibragimov et al. [17] reported that the AUC for a CNN-based prediction model for hepatobiliary toxicities with liver SBRT cases was 0.79. They also reported that combining the CNNs with the 3D dose patched increased the AUC to 0.85. Zhen et al. [35] also reported an AUC of 0.89 for the CNN models trained using transfer learning from a pre-trained

VGG16 model to predict rectum toxicities in cervical cancer radiotherapy. Su et al. [36] investigated the use of an artificial neural network model with three fully connected feed-forward networks using CT images of 142 patients treated with three-dimensional conformal radiotherapy and achieved an AUC of 0.85. Others have used Support Vector Machines (SVM) [37] (AUC = 0.72) or logistic regression [38, 39] (AUC=0.68) for these prediction models.

We also attempted to identify salient regions within the input 3D image dataset via an integrated gradient technique that provided important details of the tumor surrounding volume in the patient RP stratification. Based on our visual evaluation of IG maps with thirty patients (fifteen each for No RP and Yes RP) samples cases, these techniques appeared to indicate the significance of the tumor and the surrounding tissue in the prediction of lung injury for patients. We also found that voxels outside the tumor regions have contributed to the model predictions. Clinical verification of these findings is not possible based on the current understanding of the RP development process. Zhen et al. [35] published deep learning models for the prediction of rectal toxicity. The saliency maps from this study indicated that the highly discriminative region for the predictions was the upper region of the rectum. It is more difficult to assess the predictions with IG in lung cancer patients than in rectum due to the much larger variability of anatomical topography in lung tumors. The clinical decision-making based on these IG heat maps is challenging due to the fact that the highlighted regions seem to vary across various patient cases and are quite limited in their ability to explain the model behavior to clinical experts. Saporta et al. [40] also found that the saliency methods for localization perform worse than expert localization across multiple analyses and many important pathologies. They also reported that when these maps are used in clinical practice, they can introduce well-documented biases and erode the trust in model predictions, even when the model predictions are correct, thus limiting clinical applicability. Therefore, it is important that before these models are utilized in the clinical domain, they should be made more interpretable where detailed clinical interpretation of model predictions can be utilized to explain the salient features per output layers that are extracted and used for such toxicity predictions. We plan to build such interpretable models in our future work. At this point, while the findings on IG heat maps cannot be completely explained, showing the feasibility of generating IG maps and documenting the importance of the tumor and high dose area on these maps is promising.

One aspect we investigated was to study and build models learning from the temporal dimensions of these imaging datasets. There are recurrent neural networks (RNNs) that have the capacity to repeatedly learn based on previously remembered information for a time series-based dataset and then apply that to the current dataset. The long short-term memory-based models are one type of recurrent neural network that is very commonly used to deal with time-series based 2D datasets. Our challenge with such models has been the lack of multiple (more than 5) time points within our dataset. These techniques work well with video-based datasets that have 20-30 2D frames at a minimum for the model to gather spatial and temporal changes in the image shape and appearance. We believe that once we have a sufficient number of follow-up imaging datasets acquired on a consistent time-series basis, we should be able to build a model that can quantify the subtle image changes over time and provide more meaningful clinical insights into radiological changes in lungs after radiation treatments.

Another weakness of this study is the limited-size dataset compared to non-medical applications. The dataset lacked more samples per region of tumor location for the algorithm to make generalizable predictions. Due to the heterogeneity, size, shape, and location of the tumor regions, predicting the response to radiotherapy can be a difficult task. As a rule of thumb, deep learning models are usually

trained using tens of thousands of samples. Although we have already observed good prediction with our moderate-size cohort for the 3-class model, these models can potentially benefit from a larger cohort of datasets with imaging studies at each follow-up interval. Based on the experience from this study, as part of our future work, we therefore, plan to also include locally advanced lung cancer patients which will in turn increase our sample size and variability of tumor morphology and dose distributions, but also increase the relative number of RP cases. This strategy will be good for CNN training because it will help balance the input training datasets.

In order to gather more data, data sharing must be encouraged between multiple treating institutions. However, multi-institutional collaborations based on centrally shared patient data face privacy and ownership challenges. There are federated learning approaches [41] that utilize novel concepts of model learning leveraging where none of the available data at an institution is shared but only the model learning is shared. These types of collaborative learning approaches use the same model to train on the dataset locally without sharing any dataset (with or without patient information) to a central repository but only sharing their model updates to the central repository. The aggregation central server receives model updates from multiple institutions and combines the model weights and then sends the consensus model to all collaborating institutions for use and/or further training. Some of the challenges with these approaches are enforcing uniformity in model architecture, utilizing standard computational processes, and providing adequate techniques for data quality and benchmarking before updating the input model weights on the central aggregation servers. However, this can be one of the methods for deep learning models to be trained with large datasets from multiple institutions.

Other than the limitations with the size of the dataset, it should be noted that the deep learning algorithms and their working mechanisms remain a black box. Of course, it is useful to have the imaging features automatically extracted based on the image patterns and textures, but there are problems with the implementation of these algorithms in clinical practice because physicians are not able to gather a clear understanding of how to intuitively interpret the results obtained by such models. Thus, one of the biggest challenges with deep learning approaches is determining the reasoning behind why and where certain characteristics in the input images have a positive or negative effect on the eventual predictions. As part of our future research work, we can explore the development of interpretable model architectures specifically designed for radiation pneumonitis prediction. These models should provide detailed clinical interpretation of the extracted features and their relationship to the model's predictions. These architectures can incorporate mechanisms that explicitly encode domain knowledge or clinical insights into the model's decision-making process. By leveraging concepts such as rule-based reasoning, symbolic reasoning, or knowledge graphs, these architectures can generate explanations that are more easily understandable and meaningful to clinicians. By incorporating domain knowledge and clinical insights into the model architecture, the resulting models can offer more transparent decision-making processes.

Additionally, not having a balanced dataset with an equal number of RP and non-RP patients does not help the training process where the model is seeing a greater number of non-RP cases and learning very little about the RP cases' image patterns and features in order to provide generalizable predictions. With our study, we have balanced out the dataset by utilizing data augmentation techniques such as translation, rotation, etc. of the RP and non-RP images before the training process. There are other oversampling techniques such as synthetic minority oversampling techniques (SMOTE) [42] which

generates new samples using the combination of nearby examples of the same class. These techniques are not guaranteed to generate realistic-looking images or ones that are medically reasonable. We plan to study these techniques to make them applicable to medical imaging datasets and implement them in our future work. Furthermore, deep learning models can benefit from incorporating uncertainty estimation techniques. Uncertainty quantification methods, such as Bayesian deep learning or Monte Carlo dropout, can provide a measure of uncertainty associated with model predictions. This information can be valuable in decision-making, allowing clinicians to assess the reliability of the predictions and make informed choices based on the level of uncertainty.

In conclusion, we proposed a deep learning network to predict RP based on CT imaging scans and radiation treatment dose information. We demonstrated the model's ability to stratify patients in non-radiation pneumonitis, grade 1 and grade 2 & higher pneumonitis groups. Consequently, we looked at regions in the input images that provide the most important information to guide the model with these predictions, thus narrowing the gap between computer science techniques used for pattern recognition and precision medicine. The clinical meaning of regions of interest that are identified as important for the development of radiation pneumonitis from these models needs further investigation.

References:

1. Weiss E, Deng X, Mukhopadhyay N, Jan N. Effects of the recurrence pattern on patient survival following SABR for stage I lung cancer. Acta Oncol. 2020 Apr;59(4):427-433. doi: 10.1080/0284186X.2019.1711172. Epub 2020 Jan 12. PMID: 31928266; PMCID: PMC7060815.

2. Timmerman R, Paulus R, Galvin J, Michalski J, Straube W, Bradley J, Fakiris A, Bezjak A, Videtic G, Johnstone D, Fowler J, Gore E, Choy H. Stereotactic body radiation therapy for inoperable early stage lung cancer. JAMA. 2010 Mar 17;303(11):1070-6. doi: 10.1001/jama.2010.261. PMID: 20233825; PMCID: PMC2907644.

3. Zhao J, Yorke ED, Li L, Kavanagh BD, Li XA, Das S, Miften M, Rimner A, Campbell J, Xue J, Jackson A, Grimm J, Milano MT, Spring Kong FM. Simple Factors Associated With Radiation-Induced Lung Toxicity After Stereotactic Body Radiation Therapy of the Thorax: A Pooled Analysis of 88 Studies. Int J Radiat Oncol Biol Phys. 2016 Aug 1;95(5):1357-1366. doi: 10.1016/j.ijrobp.2016.03.024. Epub 2016 Mar 25. PMID: 27325482; PMCID: PMC5541363.

4. Guckenberger M, Heilman K, Wulf J, Mueller G, Beckmann G, Flentje M. Pulmonary injury, and tumor response after stereotactic body radiotherapy (SBRT): results of a serial follow-up CT study. Radiother Oncol. 2007 Dec;85(3):435-42. doi: 10.1016/j.radonc.2007.10.044. Epub 2007 Nov 28. Erratum in: Radiother Oncol. 2008 Feb;86(2):293. PMID: 18053602.
5. Trovo M, Linda A, El Naqa I, Javidan-Nejad C, Bradley J. Early and late lung radiographic injury following stereotactic body radiation therapy (SBRT). Lung Cancer. 2010 Jul;69(1):77-85. doi: 10.1016/j.lungcan.2009.09.006. Epub 2009 Nov 11. PMID: 19910075.

6. Chen H, Senan S, Nossent EJ, Boldt RG, Warner A, Palma DA, Louie AV. Treatment-Related Toxicity in Patients With Early-Stage Non-Small Cell Lung Cancer and Coexisting Interstitial Lung Disease: A Systematic Review. Int J Radiat Oncol Biol Phys. 2017 Jul 1;98(3):622-631. doi: 10.1016/j.ijrobp.2017.03.010. Epub 2017 Mar 15. PMID: 28581404.
7. Ma J, Zhang J, Zhou S, Hubbs JL, Foltz RJ, Hollis DR, Light KL, Wong TZ, Kelsey CR, Marks LB. Regional lung density changes after radiation therapy for tumors in and around thorax. Int J Radiat Oncol Biol Phys. 2010 Jan 1;76(1):116-22. doi: 10.1016/j.ijrobp.2009.01.025. PMID: 19406588.

8. Palma DA, van Sörnsen de Koste J, Verbakel WF, Vincent A, Senan S. Lung density changes after stereotactic radiotherapy: a quantitative analysis in 50 patients. Int J Radiat Oncol Biol Phys. 2011 Nov 15;81(4):974-8. doi: 10.1016/j.ijrobp.2010.07.025. Epub 2010 Oct 6. PMID: 20932655.

9. Mahon RN, Kalman NS, Hugo GD, Jan N, Weiss E. Lung density changes following SBRT: Association with lung dose and clinical pneumonitis. Int J Radiat Oncol Biol Phys 2019; volume 105, issue 1, supplement, E794-E795,

10. Palma DA, Senan S, Haasbeek CJ, Verbakel WF, Vincent A, Lagerwaard F. Radiological and clinical pneumonitis after stereotactic lung radiotherapy: a matched analysis of three-dimensional conformal and volumetric-modulated arc therapy techniques. Int J Radiat Oncol Biol Phys. 2011 Jun 1;80(2):506-13. doi: 10.1016/j.ijrobp.2010.02.032. Epub 2010 Jun 26. PMID: 20584582.

11. Ma J, Zhang J, Zhou S, Hubbs JL, Foltz RJ, Hollis DR, Light KL, Wong TZ, Kelsey CR, Marks LB. Association between RT-induced changes in lung tissue density and global lung function. Int J Radiat Oncol Biol Phys. 2009 Jul 1;74(3):781-9. doi: 10.1016/j.ijrobp.2008.08.053. Epub 2008 Dec 10. PMID: 19084355; PMCID: PMC4287218.

12. Liang B, Yan H, Tian Y, Chen X, Yan L, Zhang T, Zhou Z, Wang L, Dai J. Dosiomics: Extracting 3D Spatial Features From Dose Distribution to Predict Incidence of Radiation Pneumonitis. Front Oncol. 2019 Apr 12;9:269. doi: 10.3389/fonc.2019.00269. PMID: 31032229; PMCID: PMC6473398.
13. Yang X, Kwitt R, Styner M, Niethammer M. Quicksilver: Fast predictive image registration - A deep learning approach. Neuroimage. 2017 Sep;158:378-396. doi: 10.1016/j.neuroimage.2017.07.008. Epub 2017 Jul 11. PMID: 28705497; PMCID: PMC6036629.
14. Hague C, McPartlin A, Lee LW, Hughes C, Mullan D, Beasley W, Green A, Price G, Whitehurst P, Slevin N, van Herk M, West C, Chuter R. An evaluation of MR based deep learning auto-contouring for planning head and neck radiotherapy. Radiother Oncol. 2021 May;158:112-117. doi: 10.1016/j.radonc.2021.02.018. Epub 2021 Feb 24. PMID: 33636229.
15. Kim DW, Lee S, Kwon S, Nam W, Cha IH, Kim HJ. Deep learning-based survival prediction of oral cancer patients. Sci Rep. 2019 May 6;9(1):6994. doi: 10.1038/s41598-019-43372-7. PMID: 31061433; PMCID: PMC6502856.

16. Li X, Gao H, Zhu J, Huang Y, Zhu Y, Huang W, Li Z, Sun K, Liu Z, Tian J, Li B. 3D deep learning model for the pretreatment evaluation of treatment response in locally advanced TESCC: A prospective study. Int J Radiat Oncol Biol Phys. 2021 Jul 3:S0360-3016(21)00825-7. doi: 10.1016/j.ijrobp.2021.06.033. Epub ahead of print. PMID: 34229050.
17. Ibragimov B, Toesca D, Chang D, Yuan Y, Koong A, Xing L. Development of deep neural network for individualized hepatobiliary toxicity prediction after liver SBRT. Med Phys. 2018 Oct;45(10):4763-4774. doi: 10.1002/mp.13122. Epub 2018 Sep 10. PMID: 30098025; PMCID: PMC6192047.

18. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data 2019; 6: 1-48. https://doi.org/10.1186/s40537-019-0197-0

19. Chlap P, Min H, Vandenberg N, Dowling J, Holloway L, Haworth A. A review of medical image data augmentation techniques for deep learning applications. J Med Imaging Radiat Oncol. 2021 Aug;65(5):545-563. doi: 10.1111/1754-9485.13261. Epub 2021 Jun 19. PMID: 34145766

20. Huang G, Liu Z, Pleiss G, Van Der Maaten L, Weinberger K. Convolutional Networks with Dense Connectivity. IEEE Trans Pattern Anal Mach Intell. 2019 May 23. doi: 10.1109/TPAMI.2019.2918284. Epub ahead of print. PMID: 31135351.

21. He K., Zhang X., Ren S., Sun J. “Deep residual learning for image recognition” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) IEEE; 2016; pp. 770–778.

22. Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR, abs/1409.1556.

23. Zhang MJ, Lu QC, Li DX, Kim JH, Wang J. A full convolutional network based on DenseNet for remote sensing scene classification. Math Biosci Eng. 2019 Apr 18;16(5):3345-3367. doi: 10.3934/mbe.2019167. PMID: 31499617.
24. Huang, G., Liu, Z., & Weinberger, K.Q. (2017). Densely Connected Convolutional Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2261-2269.

25. Armato SG 3rd, McLennan G, Bidaut L, McNitt-Gray MF, Croft BY et.al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. Med Phys. 2011 Feb;38(2):915-31. doi: 10.1118/1.3528204. PMID: 21452728; PMCID: PMC3041807.

26. Rudin, C. (2019). Stop explaining black-box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-2015.

27. Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks." International conference on machine learning. PMLR, 2017. ArXiv, abs/1703.01365.

28. Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." arXiv preprint arXiv:1312.6034 (2013).

29. Atkinson TM, Ryan SJ, Bennett AV, Stover AM, Saracino RM, Rogak LJ, Jewell ST, Matsoukas K, Li Y, Basch E. The association between clinician-based common terminology criteria for adverse events (CTCAE) and patient-reported outcomes (PRO): a systematic review. Support Care Cancer. 2016 Aug;24(8):3669-76. doi: 10.1007/s00520-016-3297-9. Epub 2016 Jun 3. PMID: 27260018; PMCID: PMC4919215.

30. Moran A, Daly ME, Yip SS, Yamamoto T. Radiomics-based assessment of radiation-induced lung injury after stereotactic body radiotherapy. Clin Lung Cancer. (2017) 18:e425–31. doi: 10.1016/j.cllc.2017.05.014
31. Katsuta Y, Kadoya N, Mouri S, et al. Prediction of radiation pneumonitis with machine learning using 4D-CT based dose-function features. J Radiat Res. 2022;63(1):71-79. doi:10.1093/jrr/rrab097

32. Hesamian MH, Jia W, He X, Kennedy P. Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges. J Digit Imaging. 2019 Aug;32(4):582-596. doi: 10.1007/s10278-019-00227-x. PMID: 31144149; PMCID: PMC6646484.

33. Avanzo M, Stancanello J, Pirrone G, Sartor G. Radiomics and deep learning in lung cancer. Strahlenther Onkol. 2020 Oct;196(10):879-887. doi: 10.1007/s00066-020-01625-9. Epub 2020 May 4. PMID: 32367456.

34. Koçak B, Durmaz EŞ, Ateş E, Kılıçkesmez Ö. Radiomics with artificial intelligence: a practical guide for beginners. Diagn Interv Radiol. 2019 Nov;25(6):485-495. doi: 10.5152/dir.2019.19321. PMID: 31650960; PMCID: PMC6837295.

35. Zhen X, Chen J, Zhong Z, Hrycushko B, Zhou L, Jiang S, Albuquerque K, Gu X. Deep convolutional neural network with transfer learning for rectum toxicity prediction in cervical cancer radiotherapy: a feasibility study. Phys Med Biol. 2017 Oct 12;62(21):8246-8263. doi: 10.1088/1361-6560/aa8d09. PMID: 28914611.

36. Su M, Miften M, Whiddon C, Sun X, Light K, Marks L. An artificial neural network for predicting the incidence of radiation pneumonitis. Med Phys.(2005) 32:318–25. doi: 10.1118/1.1835611

37. Das SK, Chen S, Deasy JO, Zhou S, Yin F, Marks LB. Combining multiple models to generate consensus: application to radiation-induced pneumonitis prediction. Med Phys. (2008) 35:5098–109. doi: 10.1118/1.29

38. Krafft SP, Rao A, Stingo F, Briere TM, Court LE, Liao Z, et al. The utility of quantitative CT radiomics features for improved prediction of radiation pneumonitis. Med Phys. (2018) 45:5317–24. doi: 10.1002/mp.13150

39. Isaksson LJ, Pepa M, Zaffaroni M, Marvaso G, Alterio D, Volpe S, Corrao G, Augugliaro M, Starzyńska A, Leonardi MC, Orecchia R, Jereczek-Fossa BA. Machine Learning-Based Models for Prediction of Toxicity Outcomes in Radiotherapy. Front Oncol. 2020 Jun 5;10:790.

40. Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Pranav Rajpurkar, et.al. Benchmarking saliency methods for chest X-ray interpretation. medRxiv 2021.02.28.21252634; doi: https://doi.org/10.1101/2021.02.28.21252634

41. Sheller MJ, Edwards B, Reina GA, Martin J, Pati S, Kotrotsou A, Milchenko M, Xu W, Marcus D, Colen RR, Bakas S. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci Rep. 2020 Jul 28;10(1):12598. doi: 10.1038/s41598-020-69250-1. PMID: 32724046; PMCID: PMC7387485.

42. Nakamura M, Kajiwara Y, Otsuka A, Kimura H. LVQ-SMOTE - Learning Vector Quantization based Synthetic Minority Over-sampling Technique for biomedical data. BioData Min. 2013 Oct 2;6(1):16. doi: 10.1186/1756-0381-6-16. PMID: 24088532; PMCID: PMC4016036.

8. Summary

We have highlighted and emphasized the challenges faced by clinical data science researchers in accessing, integrating, and analyzing heterogeneous data and knowledge from various sources. Our research addresses these challenges by developing a scalable intelligent infrastructure that captures data from multiple sources and structures it in a knowledge base with semantically interlinked entities. This infrastructure will enable researchers to mine novel associations and gather relevant knowledge for improved clinical outcomes on a personalized basis for each cancer patient. Furthermore, the research serves as a model for implementing a learning health system not only in radiation oncology but also in other medical specialties. It has the potential of advancing personalized and data-driven medicine.

We presented the design and implementation framework of an integrated data abstraction, aggregation, and storage, curation, and analytics software; the Health Information Gateway and Exchange (HINGE), which collates data for cancer patients receiving radiotherapy. The HINGE software abstracts structured DICOM-RT data from the treatment planning system (TPS), treatment data from the treatment management system (TMS), and clinical data from the electronic health records (EHR). HINGE software has disease site specific "Smart" templates that facilitate the entry of relevant clinical information by physicians and clinical staff in the clinical workflow templates starting from initial consult to follow up, which is a part of routine clinical documentation. Radiotherapy data abstracted from all the TPS, TMS, and smart templates are processed for quality and outcome assessment. The predictive data analyses are done using evidence-based clinical and dosimetry quality measures defined by the disease site experts in radiation oncology. HINGE application software connects seamlessly to the local IT/medical infrastructure via interfaces and cloud services and performs data extraction and aggregation functions without human intervention. It provides tools to assess variations in radiation oncology practices, outcomes, and determines gaps in radiotherapy quality delivered by each provider. The design and implementation framework of HINGE was discussed in section 4.

We developed a knowledge graph-based approach to map radiotherapy data from clinical databases to an ontology-based data repository using FAIR concepts. This strategy ensures that the data is easily discoverable, accessible, and can be used by other clinical decision support systems. It allows for visualization, presentation, and data analyses of valuable information to identify trends and patterns in patient outcomes. The ETL process enables efficient and reliable data transfer while leveraging semantic interoperability. The ETL process, data model frameworks, ontologies, and presents a real-world clinical use case with mapped clinical and dosimetry records was discussed in section 5.

The efficiency and accuracy of retrieving relevant information from large clinical datasets by utilizing standard concepts defined with terminologies and ontologies. The research presents a search engine that utilizes ontology-based keyword searching and synonym-based term matching. It leverages the hierarchical nature of ontologies to retrieve patient records based on parent and children classes. To identify similar patients, a method involving text corpus creation and vector embedding models (Word2Vec, Doc2Vec, GloVe, and FastText) are employed, using cosine similarity and distance metrics. Patient similarity analysis using embedding models showed that the Word2Vec model had the highest mean cosine similarity, while the GloVe model exhibited more compact embeddings with lower Euclidean and Manhattan distances. The design framework of the data mining and keyword search tool and results from patient similarity analysis using vector embedding models were discussed in section 6.

We described the implementation of a learning health system framework for predicting radiation pneumonitis, a potential side effect following stereotactic body radiotherapy (SBRT) treatment. We investigated the use of 3D convolutional neural networks (CNN) with inputs from radiographic and dosimetric datasets of primary lung tumors and surrounding lung volumes to predict the likelihood of radiation pneumonitis (RP). Pre-treatment, 3- and 6-month follow-up computed tomography (CT) and 3D dose datasets from one hundred and ninety-three NSCLC patients treated with stereotactic body radiotherapy (SBRT). DenseNet-121 & ResNet-50 models were selected for this study as they are deep neural networks and have been proven to have high accuracy for complex image classification tasks. We also attempted to identify salient regions within the input 3D image dataset via integrated gradient techniques to assess the relevance of the tumor surrounding volume for RP stratification. These techniques appeared to indicate the significance of the tumor and surrounding regions in the prediction of RP. Overall, 3D CNNs performed well to predict clinical RP in our cohort based on the provided image sets and radiotherapy dose information. The design, methodology, and predictive results of 3D-CNN models using radiographic and dosimetric datasets were discussed in section 7.

The HINGE software and the underlying LHS framework effectively address the challenges associated with the 5 Vs of healthcare data - Volume, Variety, Velocity, Veracity, and Value. In terms of Volume, the scalable architecture of the LHS framework allows it to handle the vast amounts of data generated in healthcare, including data from EHRs, Treatment planning and dosimetry and Treatment Management Systems. The infrastructure provides the framework for robust storage and processing capabilities to manage and analyze large volumes of structured data in real-time. Regarding Variety, the HINGE software and LHS infrastructure are designed to handle diverse data formats, including structured, unstructured, and semi-structured data. The system integrates data seamlessly from various sources, standardizes and normalizes it using the disease site specific template within the HINGE software, ensuring interoperability and compatibility across systems. When it comes to Velocity, the LHS infrastructure supports data processing and analysis when the data is sent over to the HINGE central server for analysis. It enables timely capture, streaming, and analysis of data, keeping up with the continuous flow of healthcare data. For Veracity, the HINGE software and infrastructure prioritize data quality and reliability using the mapping techniques with established ontology and data standardization efforts in professional society. The use of decision trees to score the data based on the established quality measures ensures accuracy, completeness, and integrity of the data. Finally, in terms of Value, the HINGE software and LHS infrastructure focus on extracting meaningful insights and knowledge from healthcare data. This is done by efficient data analysis and mining techniques like ontology-based search tools that are designed to gather data for cohort analysis, machine learning and inform decision making with the LHS data.

Integrating new parameters into a Learning Health System (LHS) in an environment where medical information is constantly changing requires a flexible and adaptive approach. To integrate new parameters into the Learning Health System (LHS), the following steps can be taken:

- Periodic Updates of Clinical Templates: The HINGE clinical disease site templates can be periodically updated based on feedback from Subject Matter Expert (SME) panels. This updating process is planned to be closely tied to the adaptation of disease site-specific VA quality measures. When the SME panel determines the addition of new data elements in the clinical note templates, these elements will be incorporated, tested, and deployed to ensure that the LHS captures the latest relevant information.

- Updates to Interface Specifications: If the new parameters originate from the treatment planning system or management systems, the Fast Healthcare Interoperability Resource (FHIR) interface specification would need to be updated. This ensures that the HINGE software can gather new information seamlessly from these systems. By keeping the interface specifications up to date, the LHS can effectively integrate and leverage the latest data elements from various sources.
- Updating RDF Knowledge Graph and Ontology: To update the RDF knowledge graph and ontology, standardized codes must be established in the National Cancer Institute Thesaurus (NCIT) and Systematized Nomenclature of Medicine (SNOMED) terminology systems. These standardized codes serve as the foundation for the established ontologies. The D2RQ mapping script would then be updated to include the new relationships between the existing graphs and the new parameters. The advantage of using the RDF graph-based approach is that the data structures are extendable, allowing new data elements to be readily inserted or added to existing graphs without altering the structure or contents of the graphs.

By following these steps, the LHS can effectively integrate new parameters and adapt to the constantly changing medical information landscape. The periodic updates of clinical templates, interface specifications, and RDF knowledge graph ensure that the LHS remains up to date, capturing the latest data elements and providing a comprehensive view of patient information. This flexibility and adaptability enable the LHS to support evidence-based practice, facilitate research, and drive continuous improvement in patient care and outcomes.

In summary, the research and development of the IT infrastructure undertaken in this project offers a comprehensive approach to capturing data from diverse sources and organizing it into a knowledge base with interconnected entities that align with semantic meaning. This seamless integration will enable researchers to extract novel associations from multiple, heterogeneous, and diverse domain sources concurrently. By gathering pertinent knowledge, this infrastructure empowers clinical providers to offer personalized feedback, ultimately leading to improved clinical outcomes for patients. Furthermore, this project can serve as a blueprint for implementing learning health systems in other medical fields, thereby advancing the realm of personalized and data-driven medicine.

Chapter 9

9. Future Directions

Our research has resulted in the design and development of key and necessary components of a learning health system in radiation oncology. The long-term usability and ongoing development of this infrastructure are crucial considerations for ensuring its sustainability and adaptability to evolving needs. To support long-term usability, we envision the establishment of a dedicated team responsible for the maintenance, monitoring, and continuous improvement of the infrastructure. Furthermore, we anticipate fostering a collaborative and open ecosystem around the infrastructure. This involves engaging stakeholders from various domains, including healthcare providers, researchers, data scientists, and software developers. Their feedback, insights, and requirements will guide the ongoing development of the infrastructure, enabling it to address emerging challenges and meet evolving user needs. Continuous user engagement and feedback loops will facilitate the identification of new functionalities, performance enhancements, and expansion opportunities. We will actively seek opportunities to contribute to standardization efforts and participate in interoperability initiatives, ensuring alignment with industry-wide practices and promoting wider adoption. Even though we have developed a robust framework for LHS yet there are many opportunities to build on it for the future. Some of the important future directions include;

9.1 Data Sharing with Privacy Preserving Framework

In recent years, the medical community has shown great interest in leveraging Big Data to gain insights and improve patient care. However, one of the major obstacles in this pursuit is the lack of comprehensive and structured data collections. While retrospective data can serve as a valuable resource for creating large datasets, it is crucial to ensure consistent storage and organization of this data. This becomes especially challenging as individual treatment facilities typically handle a relatively small number of radiotherapy patients on an annual basis. To overcome this limitation and achieve large-scale datasets, it is necessary to aggregate data from multiple facilities.

However, transferring data outside of medical facilities poses complex challenges, primarily due to the critical importance of patient safety and privacy. Sharing patient data through anonymization has emerged as a potential solution. Anonymizing the data requires significant human effort to remove personally identifiable information and ensure the adequacy of the anonymization process. While automated solutions exist, they may not be fully trusted by privacy officers, making a semi-automated system a more feasible approach. In such a system, the majority of the data can be anonymized automatically with a high level of accuracy, and the remaining smaller portion can be manually verified to ensure proper anonymization.

Certain projects like SEER [1], TCIA [2], and TCGA [3] have partially tackled the data sharing challenge by providing expanding datasets encompassing outcomes, imaging, and genomic data. However, these databases lack cross-linkage between patients, limiting access to complete patient data and posing challenges for multi-modal research. Although these databases have grown in size and include valuable information such as outcomes, imaging, and genomic data, certain disease sites may still be underrepresented. This limitation can significantly impact the feasibility and effectiveness of certain types of studies, particularly those requiring a diverse range of patient data.

One promising direction for future research is the development and implementation of privacy preserving frameworks specifically tailored for radiation oncology data. These frameworks should encompass techniques such as data anonymization, encryption, and access control mechanisms to safeguard patient privacy while allowing for meaningful data sharing. By leveraging radiation oncology ontologies and data standards, it becomes possible to harmonize data from diverse sources, enabling seamless interoperability and integration.

A key aspect of future research should focus on the development and utilization of comprehensive ontologies as part of routine clinical practice and data recording in radiation oncology. As shown in our work, the use of these ontologies serves as knowledge representation models, capturing the complex relationships between various entities and concepts within the domain. By utilizing ontologies, researchers and clinicians are able to achieve a common understanding of terminology, facilitate data integration, and enable more efficient and accurate data sharing. These ontologies should be aligned with existing data standards to incorporate specific terminologies, such as Radiation Therapy Oncology Group (RTOG; now a part of NRG Oncology) and Digital Imaging and Communications in Medicine (DICOM), thus ensuring compatibility and interoperability across different systems and institutions.

Furthermore, future research should explore the application of advanced data standards in radiation oncology, such as the Integrating the Healthcare Enterprise (IHE) Radiation Oncology (RO) profile. This standard provides guidelines for data exchange and integration between various systems involved in radiation oncology, including treatment planning systems, imaging devices, and electronic health records. By adopting and implementing these standards, data sharing can be facilitated in a standardized and interoperable manner, promoting collaboration, research reproducibility, and improved patient care.

Another crucial research direction is the development of techniques for de-identifying and anonymizing radiation oncology data while preserving its utility for research purposes. This involves exploring innovative approaches, such as differential privacy, k-anonymity, and homomorphic encryption, to ensure that individual patient identities remain protected while allowing for meaningful analysis and knowledge extraction. It is important to evaluate the effectiveness and robustness of these techniques in the context of radiation oncology data, considering the unique challenges posed by imaging and treatment-related information.

Additionally, future research should address the legal, ethical, and regulatory considerations associated with data sharing in radiation oncology. This includes studying the impact of privacy regulations, such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR), on data sharing practices. Understanding the legal and ethical implications, as well as developing guidelines and best practices, will be essential in ensuring compliance and building trust among stakeholders involved in data sharing initiatives.

9.2 Federated Learning Framework

Federated learning has emerged as a promising approach for enhancing the accuracy of machine learning-based radiation outcome and toxicity prediction models while addressing challenges associated with data sharing and privacy. However, there are still several areas that require further research and development to fully leverage the potential of federated learning in radiation oncology. Gathering large and diverse datasets often requires collaborations between multiple institutions. The HINGE framework, for instance, collects data from 41 radiation therapy centers within the VHA. The HINGE-Central cloud collates such

data sources for model training; however, this poses additional questions on data sharing, including patient privacy, data deidentification, regulation, intellectual property, and data storage. Such challenges make centrally hosted data less practical despite advanced cloud-level security protocols for healthcare. A more viable approach is to host data locally at each center and train the model collaboratively using federated learning. In federated learning, clients independently train models on their local datasets, and a central cloud aggregates these models to create a shared global model. Communication overhead is reduced, data heterogeneity is handled, and private patient data (image/textual) do not need to be shared.

One important research direction is the exploration of advanced aggregation methods for federated learning in the context of complex convolutional neural network (CNN) architectures used in multi-modal data integration. Existing methods such as FedAvg [4], FedProx [5], AFL [6], and PFNM [7] have shown effectiveness in certain settings but may have limitations in neural network settings or when dealing with diverse data sources. It is crucial to evaluate and optimize these methods, and more recent approaches like FedMA [8], specifically for complex CNN architectures that combine multiple CNNs for text and image data. This research will contribute to improving the performance and efficiency of federated learning models for radiation oncology applications.

Furthermore, privacy-preserving federated learning is a critical area that requires attention in radiation oncology. While federated learning inherently mitigates privacy leaks by keeping patient data locally, there is still a risk of privacy breaches through gradients or model parameters [9]. Future research should focus on developing robust techniques to protect patient privacy during the federated learning process. Secure Multi-party Computation (SMC), Homomorphic Encryption (HE), and Differential Privacy (DP) are promising technologies for achieving privacy preservation [10]. However, their application in the context of complex CNN-based local models and multi-modal data (image/textual) is yet to be explored. Developing privacy-preserving federated learning models specifically designed for radiation pneumonitis prediction, considering the unique requirements and challenges of the domain, is an important avenue for future research.

In addition to privacy preservation, optimizing the communication overhead and reducing the computational burden associated with federated learning is another crucial research direction. As federated learning involves distributed training across multiple centers, communication efficiency becomes a significant concern. Investigating techniques to minimize the communication overhead and improve the efficiency of federated learning algorithms, especially in the context of radiation oncology, can greatly enhance the scalability and practicality of the approach.

Moreover, ensuring model interpretability and explainability in federated learning models for radiation oncology is an important future research direction. The ability to understand and interpret the decision-making process of federated models is crucial for gaining trust and acceptance from clinicians and stakeholders. Exploring techniques to provide insights into the model's reasoning, identifying important features, and generating meaningful explanations for predictions can enhance the clinical utility and adoption of federated learning in radiation oncology practice.

In summary, future research on the development of a federated learning framework for radiation oncology should focus on advanced aggregation methods, privacy preservation techniques, communication efficiency, and model interpretability. By addressing these research directions, we can overcome challenges related to data sharing, privacy, and model performance, leading to more accurate

and reliable prediction models that can effectively support radiation oncology decision-making and improve patient outcomes.

9.3. Ontology based Feature Selection for Machine Learning Models

Ontology graph relationships has the potential to play a crucial role in input feature selection for deep learning or machine learning models by providing valuable insights into the relationships and dependencies between different data elements within a specific domain. The ontology graph structure represents the structured representation of data elements (nodes) and their relationships (edges) in a domain-specific knowledge graph. When building a deep learning or machine learning model, feature selection is the process of choosing a subset of relevant features from the available data to train the model. This selection is essential to improve model accuracy, reduce overfitting, and enhance interpretability. Ontology graph relationships can be used to identify which data elements are closely related and have a significant impact on the target variable (e.g., pre-treatment clinical factors, patient outcomes, disease progression). By analyzing the relationships between nodes in the graph, we can pinpoint the most relevant features for the prediction task. The use of connections and hierarchies in the ontology graph capture contextual information about the data elements and this can be used to understand the semantic meaning and importance of different features, enabling better feature selection. The ontology graph can reveal redundant features—those that provide similar information to the model. By removing such redundant features, we can reduce the dimensionality of the data and prevent overfitting. The graph structure can be used to model data from multiple healthcare domains for the same patient with complex interdependencies and non-linear relationships. The ontology graph captures these complex relationships and dependencies, making it easier to select features that collectively provide the most informative and predictive power. As part of our current pneumonitis deep learning model, we considered staging, gender, age, performance status, and PTV volume as the input features. However, the ontology-based graph structure suggests that we should also consider additional features such as tobacco use history, smoking status, patient weight, height, blood pressure, oxygen levels, other comorbid conditions such as Heart disorders, lung toxicities such as Dyspnea, Bronchitis, and the use of NSAIDs. These features from the ontology graph have the potential to improve the accuracy and performance of the deep learning model.

9.4. Large Language Models for Ontology based Search tool

The potential of large language models, particularly exemplified by models like GPT-3.5, extends to the enhancement of ontology-based patient similarity search tools. These language models have exhibited remarkable capabilities in understanding and generating human-like text, making them valuable assets in healthcare research and applications. Ontologies, on the other hand, serve as structured representations of knowledge, organizing medical concepts and their relationships. However, the manual construction and maintenance of ontologies can be challenging and time-consuming. This is where large language models can play a crucial role by automatically generating and expanding ontology content through the analysis of extensive clinical text data, scientific literature, and medical records. By leveraging these models, the completeness and coverage of ontologies can be improved, facilitating more comprehensive patient similarity search.

Ontology-based search tools often rely on code-based queries or keyword matching, which may not capture the underlying context or intent of the search accurately because of the variability with search context and the use of different clinical concept terms in the clinical records. By incorporating large

language models, search tools can enhance their understanding of natural language queries and map them to relevant ontology codes/classes or concepts. This allows for more precise and context-aware patient similarity search, enabling clinicians and researchers to find patients with similar characteristics or conditions more effectively.

In addition to improving the search process, large language models can contribute to the interpretation and explainability of patient similarity results. While ontology-based search tools provide valuable insights by identifying patients who share similar ontology codes or concepts, the reasons behind the similarity may not always be apparent. Large language models can assist in generating explanations and justifications for the patient similarity based on the textual information available in medical records, research articles, or clinical guidelines. By providing interpretable explanations, clinicians and researchers can better understand the rationale behind the patient similarity results and make more informed decisions in their healthcare practices or research studies.

However, it is crucial to address the ethical and privacy implications associated with the use of large language models in patient similarity search. Privacy-preserving techniques, such as data anonymization or federated learning, should be explored to ensure the confidentiality and security of patient information during the search process. Additionally, adherence to legal and regulatory requirements, such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States, is paramount to protect patient privacy and comply with relevant guidelines and regulations. By prioritizing privacy and ethics, the integration of large language models can be carried out in a responsible and secure manner, maximizing their potential benefits in patient similarity search while safeguarding sensitive information.

References:

1. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Incidence - SEER Research Data, 9 Registries, Nov 2019 Sub (1975-2017) - Linked To County Attributes - Time Dependent (1990-2017) Income/Rurality, 1969-2017 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, released April 2020, based on the November 2019 submission.
2. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maffitt D, Pringle M, Tarbox L. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. Journal of digital imaging. 2013 Dec 1;26(6):1045-57.
3. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary oncology. 2015;19(1A):A68.
4. McMahan B, Moore E, Ramage D, Hampson S, y Arcas BA. Communication-efficient learning of deep networks from decentralized data. In Artificial Intelligence and Statistics 2017 Apr 10 (pp. 1273-1282). PMLR.
5. McMahan B, Moore E, Ramage D, Hampson S, y Arcas BA. Communication-efficient learning of deep networks from decentralized data. In Artificial Intelligence and Statistics 2017 Apr 10 (pp. 1273-1282). PMLR.
6. Li T, Sahu AK, Zaheer M, Sanjabi M, Talwalkar A, Smith V. Federated optimization in heterogeneous networks. arXiv preprint arXiv:1812.06127. 2018 Dec 14.
7. Mohri M, Sivek G, Suresh AT. Agnostic federated learning. In International Conference on Machine Learning 2019 May 24 (pp. 4615-4625). PMLR.
8. Yurochkin M, Agarwal M, Ghosh S, Greenewald K, Hoang N, Khazaeni Y. Bayesian nonparametric federated learning of neural networks. In International Conference on Machine Learning 2019 May 24 (pp. 7252-7261). PMLR.

9. Wang H, Yurochkin M, Sun Y, Papailiopoulos D, Khazaeni Y. Federated learning with matched averaging. ArXiv preprint arXiv:2002.06440. 2020 Feb 15.
10. Song C, Ristenpart T, Shmatikov V. Machine learning models that remember too much. In Proceedings of the 2017 ACM SIGSAC Conference on computer and communications security 2017 Oct 30 (pp. 587-601).
11. Young T, Hazarika D, Poria S, Cambria E (2018) Recent trends in deep learning based natural language processing. IEEE Comput Intell Magazine 13(3):55–75.
12. Wang, B, Mezlini, AM, Demir, F, Fiume, M, Tu, Z, Brudno, M, Haibe-Kains, B, Goldenberg, A. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014 Mar;11(3):333-7. doi: 10.1038/nmeth.2810. Epub 2014 Jan 26