Ontologies in radiation oncology

Citation for published version (APA):

Phillips, M. H., Serra, L. M., Dekker, A., Ghosh, P., Luk, S. M. H., Kalet, A., & Mayo, C. (2020). Ontologies in radiation oncology. *Physica Medica: European journal of medical physics*, 72, 103-113.
<https://doi.org/10.1016/j.ejmp.2020.03.017>

Document status and date:

Published: 01/04/2020

DOI:

[10.1016/j.ejmp.2020.03.017](https://doi.org/10.1016/j.ejmp.2020.03.017)

Document Version:

Publisher's PDF, also known as Version of record

Document license:

Taverne

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.umlib.nl/taverne-license

Take down policy

If you believe that this document breaches copyright please contact us at:

repository@maastrichtuniversity.nl

providing details and we will investigate your claim.

Review paper

Ontologies in radiation oncology

Mark H. Phillips^{a,*}, Lucas M. Serra^b, Andre Dekker^c, Preetam Ghosh^d, Samuel M.H. Luk^a, Alan Kalet^a, Charles Mayo^e

^a Department of Radiation Oncology, University of Washington, Seattle, WA 98195, United States

^b Department of Biomedical Informatics, University at Buffalo, 77 Goodell Street, Buffalo, NY 14260, United States

^c Medical Physics Department, Maastricht Clinic, DR. Tanslaan 12, Maastricht 6229 ET, Netherlands

^d Department of Computer Science, Engineering East Hall, Virginia Commonwealth University, Richmond, VA, United States

^e Radiation Oncology, University of Michigan, 1500 E Medical Center Dr, SPC 5010, Ann Arbor, MI, United States

A B S T R A C T

Ontologies are a formal, computer-compatible method for representing scientific knowledge about a given domain. They provide a standardized vocabulary, taxonomy and set of relations between concepts. When formatted in a standard way, they can be read and reasoned upon by computers as well as by humans. At the 2019 International Conference on the Use of Computers in Radiation Therapy, there was a session devoted to ontologies in radiation therapy. This paper is a compilation of the material presented, and is meant as an introduction to the subject. This is done by means of a didactic introduction to the topic followed by a series of applications in radiation therapy. The goal of this article is to provide the medical physicist and related professionals with sufficient background that they can understand their construction as well as their practical uses.

1. Introduction

In recent years, a number of developments in healthcare have resulted in a greater appreciation of the need for quality data about patients, their treatments and the outcomes. Progress and interest in multiple areas such as personalized medicine [1], genetic profiling [2,3], and machine learning [4,5] are all important areas that need large quantities of high quality data.

In addition to these types of data-driven developments, there is also the realization that the biochemical and physiological systems in both healthy and disease states are much more complex than previously understood. Studies of the genetics and epigenetic factors involved in cellular systems continue to highlight the fact that the development of diseases and the response to interventions involve multiple pathways that interact in myriad fashions. Progress in the treatment of disease will require that the knowledge gained in one area of medicine will need to be correlated with that in other areas.

Finally, there is the fact that research groups throughout the world are contributing to this ever-rising tide of information. Just as different specialties need to communicate with and understand others, researchers within a field need to be able to collect and understand the developments coming from other groups.

All of these factors mean that there must be some way of organizing and standardizing our knowledge in ways that facilitate sharing and responding to new knowledge. Ontologies are emerging as the most

promising approach particularly when so much of our information is being stored in computerized formats [6]. The word “ontology” comes from the Greek meaning the study of being or that which is. While the term has been used in philosophy for millennia, recent applications in the context of information science have led to a more modern and detailed definition. In short, modern ontologies are a computer-understandable description of human knowledge about a particular area, or domain, of the world. While entire books are written about the subtleties of ontologies and related concepts, the approach taken in what follows is that ontologies describe human knowledge about entities and relations that exist in the universe. In fact, they are designed to aid in scientific reasoning; they represent common, working understandings of how scientists see and interact with the world [7].

Within the radiation oncology community, as in other specialties and in medicine in general, many efforts are underway to standardize vocabularies, establish taxonomies and establish interoperability standards. Given that ontologies provide the means to accomplish many of these goals, it is natural that the relevant communities are developing them with the hope that the number of conflicting or parallel standards can be avoided or at least reconciled with one another. The goal of this paper is to describe how ontologies can be used in radiation oncology for these purposes and to provide some overview of developments to date.

The paper is intended to provide the reader with a basic understanding of the format and uses of ontologies. To this end, there is an

* Corresponding author.

E-mail address: markp@uw.edu (M.H. Phillips).

<https://doi.org/10.1016/j.ejmp.2020.03.017>

Received 20 November 2019; Received in revised form 17 March 2020; Accepted 19 March 2020

Available online 02 April 2020

1120-1797/ © 2020 Published by Elsevier Ltd on behalf of Associazione Italiana di Fisica Medica.

introductory section with more details regarding the basis and structure and rationale of ontologies. There follow a number of practical examples of ontologies that have been constructed for use in radiation oncology. The paper ends with a description of the efforts of the American Association of Physicists in Medicine, working with like-minded vendors and associated societies, to develop an ontology representing our knowledge of radiation therapy.

One additional point is appropriate: if an ontology represents our knowledge about a certain domain, how can we have multiple ontologies within a domain? Such a situation exists because it has been found efficacious to design both what are called “formal” or “top-level” ontologies and “application” ontologies [7]. The former seeks to lay the foundation of entities and relations that exist across all subdomains of the field. The application ontologies typically focus on a subset of the domain that is most appropriate to a given problem. In many cases, this facilitates the actual realization of an ontology and problem solution rather than becoming such a large all-encompassing project that progress is difficult. The formal ontology can be used as a basis for the subdomains and the subdomain knowledge is restricted to a manageable amount.

2. Introduction to ontologies

The field of ontology spans millennia and there exist multiples notions of ontology. As stated, in philosophy, ontology is the study of being and is a branch of metaphysics. Early Greek philosophers like Parmenides and Aristotle examined questions such as: What can be said to exist? What categories can we sort existing things into? [8] More modern notions of ontology have their roots in computer science where rumblings of the word could be heard during the 1980s in the field of knowledge-based systems and artificial Intelligence [9]. One of the earliest modern descriptions of an ontology and its uses can be found in the Fall 1991 issue of AI Magazine [10]. It wouldn't be until the early-mid 1990s, when a series of landmark papers including, “Toward Principles for the Design of Ontologies Used for Knowledge Sharing” would solidify the idea of using ontologies to enable libraries of reusable knowledge by proposing design criteria to guide their development [11,12]. More recently, there has been a push towards adopting an ontological realist methodology when developing ontologies [13]. A body of work exists defending this viewpoint and posits that there exists an external objective reality that is accessible to us, that we are able to form cognitive representations of these real entities and that we are able to communicate about these representations with one another [7].

This ontological realist school of thought proposes the following definition for an ontology: “A representational artifact, comprising a taxonomy as proper part, whose representations are intended to designate some combination of universals, defined classes, and certain relations between them.” [7] This definition is loaded with jargon and merits an explanation of its components. The foundation of any ontology is a taxonomic backbone which provides a tree or graph-like structure for the classification of entities. This taxonomy relies upon one of the most basic relations within an ontology: the “is a” relation. For instance, to borrow an example from biology, a feline “is a” animal and a cat “is a” feline. In this example our root node is the entity “animal”. All of the children terms stemming from the parent node inherit the properties of entities higher up on the tree. But what is the nature of these entities?

Ontologies represent universals, defined classes, and relations. Very briefly, universals are the general features of things in reality. They are responsible for the similarities of entities and represent a “natural kind”. The property of “fluffiness” is a universal or shared feature and can exist in entities like a pair of slippers or Pomeranians. Universals exist as or are instantiated by particulars (also referred to as instances or individuals). Particulars are specific entities in our world: the city of New York, your pet cat, a piece of gum, etc. Particulars exist in one place at any given time. Universals must be contrasted from “defined

classes” which are typically groups created by selection criteria defined by humans. For example, “all the smokers in New York City with pulmonary disease” is a defined class that may be an important term to include in an ontology but does not represent a “natural kind” like in the case of a universal. The last component of our definition concerns relations or the way in which two entities are connected. Within ontologies three types of relations exist: those between universals, those between a universal and a particular and those between particulars. The previously described framework is intended to give a sense of the structure of ontologies. It should be noted that the philosophical basis of ontologies is not agreed upon by all researchers in the field. Some argue that realism and alternate views, like conceptualism, can coexist, while others feel that realism is not a requisite feature for biomedical ontologies [14,15].

Ontologies deal with the meaning behind data and provide strong semantic frameworks for domains of knowledge. This is partly accomplished through the use of well-crafted definitions for entities within the ontology. Early descriptions of ontologies mention the use of both human and machine-readable definitions for terms [12]. Human-readable definitions take the form of text, which describes what the term denotes while machine-readable definitions are comprised of formal, logical axioms. The preferred form for a text definition is the so-called “Aristotelian definition” and includes a reference to the term's parent class along with a statement about the *differentiae*. The *differentiae* is a property that distinguish the term from other child terms assigned to the parent class [16]. To illustrate this form, consider the following definition: a cancer cell is a cell that invades surrounding tissue. Here, the term cancer cell is related to its parent term cell and the *differentiae*, its ability to pass the basement membrane and invade tissue, separates the cancer cell from other types of cell.

Ontologies are focused on standardizing terms and providing semantic infrastructure for areas of knowledge and research in order to enhance interoperability within data. In line with these goals, much thought has been given to standardizing ontology development. In the late 2000s, the Open Biomedical Ontologies (OBO) consortium proposed a set of shared principles that would guide ontology development. Ontologies would follow an evolving set of principles such as being open source, using unambiguous relations and definitions, having a defined scope, and using unique namespaces for terms. These principles were chosen, in part, to reduce the confusion caused by overlapping ontologies and to maximize interoperability by providing a grander schema for diverse arrays of ontologies to fit into. The resultant OBO foundry persists to this day and consists of over 128 ontologies including two of the most widely-used ontologies, the Gene Ontology and the Ontology for Chemical Entities of Biological Interest [17–19].

There have been numerous, collaborative efforts to create tools and repositories for ontologies and these endeavors represent invaluable resources for the biomedical community. BioPortal, funded by the National Center for Biomedical Ontology, is an “open repository of biomedical ontologies that provides access via Web services and Web browsers to ontologies developed in OWL, RDF, OBO format and Protégé frames” [20]. Using a web interface, users can browse ontologies, add notes, leave reviews, and view mappings between different ontologies. BioPortal now hosts over 800 ontologies¹ including the Proteomics Standards Initiative, the OBO library, and the Semantic Type Ontology of the Unified Medical Language System (UMLS) [21]. The UMLS is another resource for biomedical vocabularies and ontologies composed of three components: Metathesaurus, Semantic Network, and SPECIALIST Lexicon and Lexical Tools. The Metathesaurus includes widely-used terminologies such as the NCBI taxonomy, the Medical Subject Headings (MeSH), and the Digital Anatomist Symbolic Knowledge Base [22]. Despite its utility in integrating datasets, the UMLS has historically suffered from issues regarding coverage of terms,

¹ As reported by Bioportal, 12 March 2020

structural problems, and disregarding semantics [23–25].

3. Data and knowledge sharing

There are many avenues for improving healthcare; two that rely heavily on data-driven informatics approach are Learning Healthcare Systems and distributed learning. Both require a radiation therapy community that shares common concepts and vocabularies to be effective. For this reason, developing and applying an ontology provides a very useful backbone to the effort that minimizes redundancies and mismatches in data. In this section, the application of an ontology to these innovations is described.

A Learning Healthcare System is a system in which “science, informatics, incentives, and culture are aligned for continuous improvement and innovation, with best practices seamlessly embedded in the delivery process and new knowledge captured as an integral by-product of the delivery experience.” [26]. Projecting this onto radiation oncology, creating a learning system means that routine cancer care practice and outcome data is shared and used to derive new knowledge. This knowledge is then shared and integrated into practice rapidly and continuously.

But sharing data is hard, due to administrative, political, ethical, and administrative barriers [27]. It is proposed that a better way is to share questions that one wants to have answered using the data, rather than sharing the data itself. This concept is called federated or distributed learning, where knowledge is derived from data of, for example, radiation oncology departments without the need for data to leave the individual department [28,29].

A distributed learning healthcare system would thus alleviate many of the concerns of sharing patient data including privacy concerns. But such a system only works if there is an infrastructure (“Track”) where the holders of data such as hospitals (“Stations”) allow authenticated and authorized applications (“Trains”) to use the data in the Station to generate new knowledge. Besides having such an infrastructure, a second challenge that has to be solved is that the Trains have to be able to find and access (under well-defined conditions) a Station and should be able to understand and reuse the data of the Station. In other words,

data in Stations have to be made FAIR (Findable, Accessible, Interoperable and Reusable) [30] (See Fig. 1 below).

Ontologies are crucial in many aspects of FAIR data especially in the yellow (Rich metadata, PIDs In metadata, Metadata is always available, Linked Metadata and Metadata have multiple attributes) and red blocks (Vocabularies and Vocabularies are FAIR) of the above Figure.

In the field of radiation oncology a number of ontologies have arisen in the past years. These include the Dependency Layered Ontology for Radiation Oncology (<https://bioportal.bioontology.org/ontologies/DLORO>), the Radiation Oncology Ontology (<https://bioportal.bioontology.org/ontologies/ROO>) and the Radiation Oncology Structures Ontology (<https://bioportal.bioontology.org/ontologies/ROS>).

Efforts are now underway to merge these initial efforts of individual groups into one well-authored and formal ontology maintained by professional societies such as AAPM, ASTRO and ESTRO together with the radiation oncology community.

Besides the development of a domain specific ontology for radiation oncology, it is necessary to reuse and link to other existing and emerging ontologies adjacent to the radiation oncology field including:

- Foundational Model of Anatomy (FMA): A domain ontology that represents a coherent body of 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 into radiomics computation. (<https://bioportal.bioontology.org/ontologies/RO>)

Fig. 1. Schematic of how the principles of Findable, Accessible, Interoperable and Reusable (FAIR) data standards are applied.

- Semantic DICOM Ontology: An ontology for DICOM (<https://biportal.bioontology.org/ontologies/SEDI>)

Other adjacent artefacts include nomenclatures such as proposed by AAPM TG 263 [31] and information standards such as the minimum data element of ASTRO [32]. Also, existing and emerging syntactical transactional standards, such as DICOM, HL7 v3 CDA, HL7 v2 and HL7 FHIR, are important vehicles to communicate ontologically enriched data (see section below).

The power of ontologies, FAIR data, and distributed learning has recently been shown by the so-called 20 k Challenge [33]. In this award-winning project, data from more than 20000 patients from 8 centers from 5 countries were made FAIR and new knowledge - a prediction model - was successfully learned for overall survival in lung cancer patients [34].

It should be noted that because the data is made FAIR through the use of ontologies, the centers participating in the 20 k challenge could immediately validate the new knowledge in their own data. We expect that such rapid learning and simultaneous validation of knowledge alleviates an important concern in machine learning and AI based application, which is “Does the AI work in my patients?” and “What’s in it for me?” This will increase acceptance and clinical use of the new knowledge.

Another aspect of an ontology, especially if it is linked to other ontologies, is that it can capture the knowledge as it is known and important to the field. This knowledge can then be leveraged in data applications. As an example, an ontology or combination of ontologies may state that the class of lung cancers which have an adenocarcinoma morphology is a subclass of the class of non-small cell lung cancers. This allows a cancer center to assert the fact that a patient has lung cancer with adenocarcinoma morphology while a researcher searching for non-small cell lung cancer patient will find that patient.

Similarly, if the ontology used is linked to the FMA ontology (which holds much of the relevant anatomical information), then a query for thoracic cancers will return all esophagus, lung, heart, etc. cancers as the FMA holds the knowledge of which organs reside in the thorax. In other words, by capturing knowledge, ontologies allow data providers like hospitals to only specify the minimum amount of data elements, lessening their burden while data users can still use the rich semantics of the ontology to find and use the data.

While the development of an ontology focused on radiation oncology data captured in routine care is being embarked upon, an underdeveloped aspect is the formal description of knowledge applications. As an example, if one would like to use a knowledge application such as AI-based auto-contouring or a treatment-related morbidity model, it is important to know who made that application, on which data it was trained, on which data it was validated, for which patients/images it is suitable and in which it should not be used, and what its uncertainties are in specific cases. An ontology that helps the community systematically provide such meta-data would be a great step forward in improving the safety and reproducibility of models and related software.

4. Communications in radiation therapy

Earlier sections have noted the role that ontologies can play in formalizing domain concepts and vocabulary. An area where such agreement is critical is in the communication of information between people, between people and software, and between different software systems. Currently, there is a wide range of terminology and conceptual organization of radiation therapy procedures that are encapsulated in widely used standards and clinical software. In this section, a common example of such difficulty in communicating the details of treatment is used to highlight these issues, to place community efforts to address them in context, and to describe how the use of an ontology could help to solve the problems.

A simple situation that occurs routinely in radiation therapy is the writing of a prescription and communicating the information it contains to the relevant people and computer systems.

As an example, consider oropharyngeal cancer with suspected involvement of an ipsilateral neck node. A general prescription might be:

- {curative, target dose = 60 Gy², two phase with a field reduction}.

In more detail, the two prescriptions that are needed to carry out the intent of the general prescription might be:

- {target volume = right oropharynx + right neck nodal volumes, target dose = 56 Gy in 2 Gy fractions to be delivered daily, technique = IMRT with 9 fields, modality = 6 MV x-rays};
- {target volume = right oropharynx PTV, target dose = 14 Gy in 2 Gy fractions to be delivered daily subsequent to the initial prescription, technique = IMRT field boost, modality = 6 MV x-rays}.

The first thing that one notices is that this may be an overly detailed prescription (to a referring doctor, the patient and family) and it is not detailed enough for a dosimetrist or physicist since it does not give any information regarding the organs at risk, immobilization devices and imaging procedures. And it is not detailed enough for an oncology information system (OIS) that includes a record and verify system since there is no information regarding monitor units, beam angles, etc. Perhaps, then, we need to define exactly what a prescription is? Should one define a prescription that has as many attributes as the most detailed need requires and is a union of all the different needs? One can then filter out those attributes as needed. Or is it better to define subclasses of a prescription, each of which meets a specific need?

Another noticeable element of the prescriptions as written is the terminology that is used. Again, to different receivers of this prescription, some terms may not have much meaning. What is a fraction? A target volume? What does IMRT mean (for that matter, what does Intensity Modulated Radiation Therapy mean since a solitary wedge certainly modulates intensity)? Even experts at the same level may not be comfortable with some of the terms. OIS's as well have terminology that they require.

Let us address the latter set of questions first. Clearly, this is not the first time they have been posed and a number of different approaches have helped to reduce the confusion. The most commonly known and used one is known as DICOM-RT³. As the full name states (Digital Imaging and Communications in Medicine), communication is at the heart of this standard. The standard started out describing imaging procedures and has expanded to include radiation therapy. This is a very detailed standard that achieves a number of goals: providing an overview of a medical procedure (imaging, radiation therapy), breaking that procedure into components, and providing a detailed description of the syntax needed to describe each of those components.

Another effort to standardize the way in which medical equipment communicates is led by the organization Integrating the Healthcare Enterprise and its subgroup dedicated to radiation oncology (IHE-RO). Their approach is to define critical dataflow needs and to develop profiles of the information that is necessary for exchange. They encourage the use of standards such as DICOM and Health Level 7 (HL7). In this way, they also work to standardize semantics and syntax.

However, even two such parallel efforts can run into problems as seen below. This is an example provided by the AAPM IHE-RO committee that is working on this particular profile⁴. This example highlights the difficulties imposed by different conceptualizations of how to

² Gy: short for Gray, a unit of absorbed radiation dose.

³ <https://www.dicomstandard.org>

⁴ This IHE-RO profile is a work in progress and may not reflect the final version.

describe and deliver the radiation. Two large international organizations that are charged with reducing miscommunications have very different views of what the underlying concepts are. An ontology accepted by the community would provide both the concepts and the relationships between them so that the information can be communicated in a non-ambiguous fashion.

DICOM approach:

- A. Treatment sequence 1
 - 1. Site
 - a. site 1: R Oropharynx
 - b. site 2: Nodal volume
 - 2. Fraction sequence
 - a. 28×2 Gy
- B. Treatment sequence 2
 - 1. Site
 - a. Site 1: R Oropharynx
 - 2. Fraction sequence
 - a. 7×2 Gy
- C. Treatment total
 - 1. Site
 - a. Site 1: R Oropharynx
 - 2. Fraction sequence
 - a. 35×2 Gy

IHE-RO approach:

- A. Site
 - 1. site 1: R Oropharynx
 - a. Fraction sequence
 - i. 28×2 Gy = 56 Gy
 - ii. 7×2 Gy = 14 Gy
 - b. Total dose (all fractions) = 70 Gy
- B. Site
 - 2. site 2: Nodal volume
 - a. Fraction sequence
 - i. 28×2 Gy = 54 Gy
 - b. Total dose (all fractions) = 54 Gy

Even though the data are the same, the way in which they are organized differs substantially. This demonstrates that standardizing vocabulary, while necessary [31], is not sufficient to establishing a consistent conceptual edifice.

Recently, the American Society for Radiation Oncology (ASTRO) has formed groups to help standardize what defines a prescription [35] and what are the minimum data elements needed [32]. The former paper describes some of the difficulties at arriving at a consensus on some of these issues. In the same vein as the comparison above, they document the differences between vendors and their software for denoting a prescription. In the end, they recommend five key elements (one of which is the patient name) but no overall conceptual framework. They also explicitly state that “The key elements themselves do not form a complete directive for treatment delivery and cannot ‘stand alone’” [32].

These abovementioned efforts have been conducted by knowledgeable and dedicated experts, and yet, we can see that there are still many hurdles to overcome. Given the scope of concepts and applications that we described above, it would seem that a new approach is needed. This, then is where an ontology may provide an advantage since the taxonomy (classes and their subclasses) provides a reliable source for clearly understanding the concepts and formalizes the vocabulary (which is important for both human and machine communication). In addition, the relationships between classes that are part of the ontology add more detail that further aid in communication.

Since it is the intent of this paper to add to the overall picture of ontologies and their practical uses, the remainder of this section will concentrate on two important aspects of developing a “communications” ontology or any other application ontology. Those two aspects are: (1) making use of the data that are available, and (2) transmitting it between the parties of interest.

Generally speaking, the data that are needed is present in various databases with different purposes and made by a wide range of vendors. A short list would include oncology information systems, hospital and/or system-wide electronic medical records (EMRs), treatment planning systems, and picture archiving and communication systems (PACS). In addition, it must be added that some of these data are actually taken from other external sources, such as the International Classification of Diseases, e.g. ICD-10.

In general, these types of data storage facilities are relational databases. As such the schema of each one is likely to be unique, with different tables and columns. Some systems rely on computer file systems as well in order to organize, store, and retrieve data. All of which means that individual data extraction software will be necessary to be of use with an ontology. It also means that a mapping between these data structures and those native to ontologies will be necessary.

While this may seem, at first glance, to be prohibitive and/or to vitiate the utility of ontologies, it is good to realize that the same need to map data structures exists with DICOM and HL7 interfaces to current software. In addition, the informatics community has recognized the issue and several different software environments have been developed to aid in the extraction of data from databases for this purpose (see, for example, <https://www.w3.org/2001/sw/wiki/D2RQ>).

As the earlier sections have detailed, ontologies are comprised of both classes and properties. This structure lends itself to relationships between data and classes that can naturally be expressed in RDF format⁵. This type of data representation and the associated language for querying it (SPARQL) has been described above. The result is that there are standard methods defined within the Semantic Web framework for mapping relational databases to the concepts and vocabulary of ontologies and for querying/extracting these data as described in detail in Ref [36]. Finally, there is the necessity of transforming these structured data into different forms of messages depending on the application. These can take the form of free text or templated documents, HL7 messages, DICOM streams and others. Given that it is structured data, communicating with EMR's is one of most natural and necessary layers of communication. Historically, much of the EMR-to-EMR communications has been using the standards embodied in Health Level, Version 2 (<http://www.hl7.org>). This standard is quite old, and a new version dubbed Fast Healthcare Interoperability Resources (FHIR) is emerging. It is somewhat similar to the DICOM standard since it has a conceptual model of how healthcare is delivered and what types of processes and procedures need to be described. Although it is a bit early to determine whether this will be the standard for EMR and other communication pathways, the translation from ontologies to the FHIR resources should be able to proceed since the FHIR resources have been developed with the concepts of ontologies in mind.

In conclusion, a solid radiation therapy ontology that contains classes and properties that span the range of information needed to communicate between different institutions, physicians of various specialties and patients is needed in order to be able to systematically collect the data needed for progress and to facilitate normal workflows. Much of the information that is needed is already present in various databases, though much is also in unformatted text. Similar to the current software that extracts and reformats data for DICOM and HL7 messaging, extracting and formatting data for more modern standards and for the needs of research requires dedicated software.

5. Bayesian networks and ontological data mapping

The previous sections have provided examples of and reasons for the use of ontologies in helping to bridge gaps in communication and data transfer and usage. In this section we discuss how an ontology can be applied directly to developing probabilistic models. This illustrates the

⁵ Resource Description Framework, <https://www.w3.org/RDF/>

range of applications of ontologies and with the connections to machine learning, connects this type of approach to other emerging technologies.

One way of leveraging the advantages of ontological formalism is to utilize them in bridging machine learning (ML) models to their corresponding clinical data stores. Machine learning has been of growing interest in radiation oncology (and other fields) in recent years. Examples of applications of ML in radiotherapy include auto-segmentation for normal tissues and tumor targets, knowledge-based planning, quality assurance etc. aiming to provide automated assistance to radiation oncology workers in completing the complex workflows within radiation therapy [37]. However, the spread of ML-based applications is limited due to the challenges of extracting the relevant data to train the ML models: lack of standardized data in terms of content, format, structure, and nomenclatures. This situation is exacerbated for applications that aim to serve clinics with different hardware and software systems. The problem can potentially be alleviated in places by using an ontology that maps to both the ML models and their associated clinical data in a way that creates a reproducible, transparent link between the two.

In this section we present such an example using a dependency layered ontology for radiation oncology (DLORO). The ontology was initially designed to automate the construction of error detection Bayesian networks (BN) [38–41] for radiotherapy treatment plans but as an ontology, it also contains the semantic properties needed to map radiation oncology concepts and terms in the models to the schemas of relational databases. Here we map the DLORO to the two major oncology information systems (OIS), Aria (Varian Medical System, Palo Alto, USA) and Mosaicq (Elekta AB, Stockholm, Sweden). The ontology acts as a standardized layer that links the clinical data in different OISs to the variables in the BN, which allow automated data extraction from the OISs and thus automation on training and updating the BN.

The error detection BN model proposed by Kalet et al. [39] aims to provide automated assistance to physicists on flagging potential errors in the chart checking process. Bayesian networks are directed acyclic graphs that represent probabilistic knowledge in a compact and powerful way. A BN consists of a graphical component, with nodes that represent variables and arcs that represent the dependences between variables, and a quantitative component that specifies the strengths of dependence relations using a conditional probability table (CPT) and probability theory. Bayesian networks are well suited to applications for reasoning under uncertainty within the medical domain. In radiation oncology, BN models have been used for various purposes including diagnostic reasoning, meta-analysis of biomedical data, modeling, clinical decision support systems etc. [39,42–44]. The DLORO ontology organizes radiation oncology domain knowledge into class-subclass structures and establishes dependency among domain concepts. For example, the DLORO sets the concept *TotalFractions* as a subclass of *FractionationParameter*, which itself is a subclass of *PrescriptionParameter*, as shown in the snapshot of the DLORO viewed in the Protégé software in Fig. 2. Moreover, the DLORO establishes dependency between domain concepts that are not in the same class-subclass hierarchy using the relation “dependsOn”. Both the class-subclass and “dependsOn” relations correspond to a dependency arc in the BN, such that a BN *structure* can be constructed automatically using logical deduction when the nodes of interest are chosen. Software was developed in Kalet et al. [40] to automate this part of the process. To build those initial models to completion e.g. a usable model with full *conditional probability tables* required a manually developed and applied set of SQL statements to be made and run against the an OIS database.

A mapping between the DLORO and the schema of the Aria and Mosaicq can be performed in a way that assists the construction of data extraction. Since the relational database of Aria and Mosaicq have different schema, a mapping for each database to the ontology is needed. Fig. 3 shows a few examples of schema mapping that matches ontology classes to tables and columns in Aria and Mosaicq. These mappings are

added to the ontology file of DLORO as annotations of the mapped classes. For example, the class *BeamEnergy* has four annotations, namely *AA_Table*, *AA_VarName*, *MQ_Table* and *MQ_VarName*, which has entries of *Energy*, *TxFldPoint*, *Energy*, and *EnergyMode* respectively. Software has been developed to read these annotations and generate the corresponding SQL queries automatically for data extraction. However, there are classes in the ontology that do not have a simple one-to-one mapping to columns in one or both databases as the variables shown in Fig. 2. For example, the class *Number_of_beams* needs to count all existing fields in the same course for the one patient, which requires a relatively complicated SQL query. For these variables, the current workaround is to manually input the SQL query of these variables into the annotation and modify the software to directly use the annotation to generate SQL query.

The flow chart in Fig. 4 describes the process of constructing the complete BN from the ontology. The process involves two phases, the network construction and data acquisition phase, and the network training phase. In the network construction and data acquisition phase, the structure of the BN is constructed, and relevant clinical data are extracted from OIS relational database using SQL queries. In the second phase, the extracted data are used to train the CPTs to learn the practice in the clinic, and the CPTs are used to create an inference engine to calculate the probabilities of the parameters in new plans and flag potential erroneous plan parameters.

Combining the SQL query generator and the previously mentioned BN generator developed in Kalet et al [40], a user can choose any variables of interest in the domain of DLORO and the BN will be automatically generated and trained with clinical data that is extracted from the OIS database, either Aria or Mosaicq. Fig. 5 shows an example of the automated BN and SQL query generation for both Aria and Mosaicq.

In summary, this section demonstrates an automated data extraction approach for training and updating an error detection BN model involving a schema mapping between the relational databases in the OISs and the classes in a DLORO. The results show that ontology can facilitate model construction and automated clinical data extraction. Through transparency of terminology and mappings, model concepts can be rearranged, rebuilt, updated, or combined while retaining their semantic meaning. Thus, the ontology presents as an important potential data backbone of the domain – allowing increased transference and leveraging of models among practitioners and researchers.

6. Text extraction from clinical notes

The field of radiation oncology involves various types of unstructured and semi-structured notes such as, consult, treatment and follow-up notes, along with diagnoses or discharge summaries, and pathology or radiology reports. While structured information capture allows for the recording of predefined variables like age, sex, height, weight, medication dosages etc., the extraction of unanticipated variables from unstructured notes is still an open problem. The major source of clinical notes are the Electronic Health Records (EHRs) that are often used to collect, store and display patient information [45]. EHRs are comprised of unstructured free text that may be generated through dictation or typing. The central goal of information extraction from such free text notes in EHRs is to automatically extract and encode clinical information to assess and/or improve the quality of radiation therapy treatment, help with clinical decision support, or perform clinical research.

6.1. Methods and techniques for information extraction

Information Extraction (IE) falls under the realm of Natural Language Processing (NLP) and deals with finding both structured or unstructured objects from free text. Even before the process of information extraction, NLP is needed to analyze the text through

Fig. 2. A snapshot in the ontology software package, Protégé, of the dependency layered ontology for radiation oncology (DLORO).

methods like sentence splitting, tokenization, part-of-speech tagging, and parsing [46]. NLP also addresses other areas of IE such as speech processing, extracting predefined types of information from text, sentence detection or classification, text summarization, relationship extraction, and document categorization [47,48].

Two basic types of clinical information extraction are currently applied on EHRs: rule-based and machine learning. In the rule-based approach, typically a domain expert will identify the required knowledge and the rules for extraction of information from specific types of notes leveraging a dictionary lookup. Such approaches thus require fewer resources and can provide easier adaptation to new domains. Machine learning-based approaches on the other hand requires training data from which corpus statistics or rules are automatically derived and used to analyze free texts [48]. Rules can also be created based on a clinical domain ontology; such rules are based on ontological concepts and can be used in both named entity recognition and information extraction and perform better than those which are merely based on textual items [49].

Recognizing named medical entities from clinical text is another popular domain of information extraction and NLP. The goal of Named Entity Recognition (NER) is to identify specific words or phrases (i.e., entities) and categorize them (e.g., persons, locations, diseases, medication and so on). NER tasks are related to detection of entities and determining entity types [47] which are implemented using regular expressions or a predefined dictionary formed from ontology concepts. While most of the NER based methods are rule-based [50], other systems apply hybrid approaches combining both rules and machine learning [51–53]. Sentence identification and classification is another component of information extraction. A sentence may belong to a

specific category or section inside the clinical notes (e.g., Examination, History, or Diagnosis). Rule-based classification is generally used to categorize sentences considering the semantic categories of concepts included in them.

6.2. Tools for information extraction

According to [48], the top three most frequently used tools for information extraction in the clinical domain are cTAKES, MetaMap, and MedLEE. cTAKES is an open-source NLP pipeline that applies hybrid rule and machine learning based approaches while the latter two employs rule-based systems for information extraction from free texts in EHRs. cTAKES provides a comprehensive platform for performing many different clinical information extraction tasks including syntactic, lexical and semantic parsing [53]. Development of clinical terminologies and ontologies, such as SNOMED CT ontology graph [54] and RadLex [55] enable concept extraction from text in the medical domain; the US ONC SHARPN project [56,57] has further enabled clinical text mining by defining the semantic standards that were incorporated into cTAKES. Moreover, YTEX provides a series of extensions on cTAKES to design a generalizable framework for mapping clinical phrases from any domain ontology [52]. Amongst others, MedTAS/P provides an extensible knowledge representation model which uses different NLP techniques along with machine learning and rules to automatically extract cancer disease characteristics from free-text pathology reports [51]. MedLEE is another tool for processing clinical text that is primarily used for developing vocabulary and encoding; it was originally developed for processing radiology reports, but was later extended to several other domains. Finally, MetaMap provides a tool that can map scholarly

Ontology	Mosaiq Variable	Mosaiq Table	Aria Variable	Aria Table
Anatomic_tumor_loc	Category	Medical	Site	Prescription
Dose_Per_Fraction	Dose_Tx	Site	PrescribedDose	RTPlan
Beam_Energy	Energy	TxFIELDPOINT	Energy	EnergyMode
Table_Angle	Couch_Ang	TxFIELDPOINT	PatientSupportAngle	ExternalFieldCommon

Fig. 3. Examples of the schema mappings that match the tables and columns in Aria and Mosaiq to the classes in DLORO.

Fig. 4. Flow chart of data acquisition and network construction in building an error detection Bayesian network model.

biomedical text to the Unified Medical Language System (UMLS Metathesaurus); it uses a knowledge-intensive approach that employs NLP techniques alongside symbolic, and computational linguistic based methods [50,58]. A recent machine-learning based approach, called Dnorm, has also become popular for information extraction; it uses pairwise learning and can compute similarities between different mentions and concept names directly from the training data [59].

In the rule-based information extraction systems, the rules can be developed manually or by leveraging knowledge bases, or using hybrid systems [48] that can combine both. Such knowledge bases typically comprise of domain specific ontology. In [60], the authors developed a domain ontology called patient clinical data (PCD) pertinent to the field of radiation oncology. The PCD ontology represents the clinical data produced during the treatment activities throughout the entire process of patient visits; here, concepts were created by incorporating domain knowledge from physicians and other domain experts, examining the EHR database by formal and informal text analysis of clinical terms, and also integrating medical ontologies from the UMLS. It thus presents a comprehensive domain ontology for different information extractions tasks in radiation oncology. The PCD ontology represents concepts as classes and individuals, where each class has a relationship with other classes and may possess data properties which follow the HL7 standard. Efforts to further supplement this domain ontology are currently underway by integrating a more comprehensive set of patient records from EHRs and may benefit from crowdsourcing activities worldwide.

6.3. Outstanding challenges in information extraction

Although a lot of effort has already gone into the field of information extraction from clinical notes over the past decade, there are still several open research issues as summarized below.

Common NLP challenges relate to outstanding problems in improving the accuracies of standard NLP tasks including named entity recognition, anaphora resolution, negation detection, and acronym resolution. All of these NLP related tasks are applicable in the medical domain in general and to the field of radiation oncology in particular.

One of the most important bottlenecks in medical text mining is the lack of significant quantities of publicly available gold-standard data sets; this seriously impedes clinical NLP pipelines which are in need of annotated corpora to train their supervised classification algorithms. Such training data is expensive to generate and their validation requires equally expensive review using experienced manual labor. Hence, information extraction approaches that rely mostly on supervised learning algorithms do not scale well with the absence of methods to scale the learning process itself. Although, the cTAKES pipeline can perform automatic semantic annotation, its performance on different types of clinical notes is not reliable. In the radiation oncology domain, institutional barriers to data sharing further increases the difficulty of scaling the learning process.

In the fields of radiology and radiation oncology, identification of critical or incidental findings, body location versus anatomical structure disambiguation, proper structure name mapping in DICOM files, and extraction of abstract domain-specific concepts are all open research problems. The metadata in radiology text reports are mostly ignored or

```
select Medical.Category as Anatomic_tumor_loc,
(select COUNT(*) FROM TxField where SIT_Set_ID = Site.SIT_Set_ID) as Num
ber_of_beams,
PatPlan.Tx_Intent as Treatment_Intent,
Site.Fractions as Total_Fractions,
Site.Dose_Tx1 as PTV_dose_Rx,
Site.Dose_Tx as Dose_Per_Fraction,
Site.Technique as Plan_Technique
from Patient left join Medical on Medical.PAT_ID = Patient.PAT_ID left join
Topog on Medical.TPG_ID = Topog.TPG_ID left join PatPlan on PatPlan.MED_ID
= Medical.MED_ID left join Site on Site.PCP_ID = PatPlan.PCP_ID left join Si
tesetup on Sitesetup.SIT_Set_ID = Site.SIT_Set_ID and Sitesetup.Version0 inn
er join TxField on TxField.SIT_Set_ID = Site.SIT_Set_ID and TxField.Version0 le
ft join TxFieldPoint on TxFieldPoint.FLD_ID = TxField.FLD_ID left join Toleran
ce on TxField.Tol_Tbl_ID = Tolerance.Tol_ID where Site.Version0 and PatPlan
.Eff_DtM > '12-1-2017' and PatPlan.Eff_DtM < '1-1-2018'
```

```
select
Prescription.Site as Anatomic_tumor_loc,
(SELECT COUNT(*) FROM Radiation WHERE PlanSetupSer = PlanSetup.PlanSetupSer)
as Number_of_beams,
(SELECT Expression1 FROM Lookuptable WHERE ListSelector = 'PLAN_INTENT' AND L
ookupValue = PlanSetup.Intent) as Treatment_Intent,
RTPlan.NoFractions as Total_Fractions,
(RTPlan.PrescribedDose * RTPlan.NoFractions) as PTV_dose_Rx,
RTPlan.PrescribedDose as Dose_Per_Fraction,
(SELECT Expression1 FROM Lookuptable WHERE ListSelector = 'DEFAULT_VALUE_TECH
NIQUE' AND LookupValue = Prescription.Technique) as Plan_Technique
from Patient left join Diagnosis on Diagnosis.PatientSer = Patient.PatientSer
left join PrmrDiagnosis on PrmrDiagnosis.DiagnosisSer = Diagnosis.Diagnosis
Ser left join Course on Course.PatientSer = Patient.PatientSer left join Trea
tmentPhase on Course.CourseSer = TreatmentPhase.CourseSer left join PlanSetup
on PlanSetup.CourseSer = Course.CourseSer left join Prescription on Prescrip
tion.PrescriptionSer = PlanSetup.PrescriptionSer left join PrescriptionPropert
y on PrescriptionProperty.PrescriptionSer = Prescription.PrescriptionSer and
PrescriptionProperty.PropertyType = 2 left join RTPlan on PlanSetup.PlanSetup
Ser = RTPlan.PlanSetupSer left join Radiation on Radiation.PlanSetupSer = Pla
nSetup.PlanSetupSer left join ExternalField on ExternalField.RadiationSer = R
adiation.RadiationSer left join ExternalFieldCommon on ExternalFieldCommon.Ra
diationSer = Radiation.RadiationSer left join FieldAddn on FieldAddn.Radiat
ionSer = Radiation.RadiationSer left join Addon on Addon.AddonSer = FieldAddn
.AddonSer left join ControlPoint on ControlPoint.RadiationSer = Radiation.Ra
diationSer left join EnergyMode on ExternalFieldCommon.EnergyModeSer = Energy
Mode.EnergyModeSer left join Tolerance on Tolerance.ToleranceSer = ExternalFi
eldCommon.ToleranceSer
```

Fig. 5. An example that generates Aria and Mosaiq query automatically from a Bayesian network that is automatically generated by the DLORO after manually choosing the variables of interest.

only used implicitly due to the lack of reliable methods for integrating their latent information into the information extraction pipelines. To circumvent this problem, radiology researchers only use metadata implicitly by restricting their analysis to specific body location and modality; this obviously decreases the solution space and improves the performance of information extraction. However, such methods are strictly ad hoc and do not provide a generalized framework for explicitly using the metadata that are widely available in the EHRs. Such metadata are not just limited to body location and modality, but may also include physician information (e.g., specialty), patient information (e.g., age, sex, and history), and exam information (e.g., date, time, facility). Explicit modeling of such metadata to improve the performance of information extraction pipelines is an open research problem.

The field of word sense disambiguation research is still not mature enough specially in the field of radiation oncology which lacks a standardized ontology. Ambiguity at the word or concept level severely hampers the ability to quickly annotate a block of text that is analyzed by the large variety of specialized clinical text processing pipelines available today. While previous work in word sense disambiguation used manually created rule-sets and hand-tagged example sets, some automated methods have recently been proposed that include unsupervised learning methods and learning examples from UMLS [62], and using ontology-based measures of semantic similarity [61]. More research efforts are still needed in this area to reduce word/concept level ambiguity that are central to any information extraction task.

Additional challenges include, use of abbreviations without proper explanations, terse language, errors in usage that are hard to interpret, use of pronouns referring to prior observations in the note, and chronological comparisons made with previously observed conditions. All these challenges create the need for identification and contextualization of concepts that are beyond the traditional NLP tasks and requires addition of domain semantics and ontologies and the analysis of domain patterns of usage.

7. Bringing ontology into mainstream usage

The transformational promise for application of ontologies to healthcare data to enhance computer driven discovery and reasoning with healthcare data is compelling. Reaching that promise is challenging because of the intersecting needs of multiple stakeholders and differing priorities. In practice, when stakeholders speak about “ontologies” they often are most interested in particular components needed for practical implementation of what can be strictly defined as an ontology (Fig. 6).

To develop a formal, practically implementable, ontology a standardized data dictionary of key data elements needed for reasoning as well as a taxonomy for categorizing these elements is needed. Enumeration of a minimal set of relationships needed for clinical reasoning with these elements must also be part of that dictionary. Key data elements and relationships for radiation oncology were published as one of the findings of the 2017 Practical Big Data Workshop [20]. Together these form the basis of an ontology as originally described by Gruber [11]. Modern usage of the term implies additional steps to formalize the reasoning concepts into a computer consumable format developed using standardized design principles as identified by the OBO Foundry [7].

Practical steps to connect “real world” clinical data to concepts in the formalized ontologies require widely agreed upon, professional society endorsed nomenclatures for these data elements that can be implemented in clinical systems. AAPM Task Group – 263 represents one such nomenclature [31]. Without domain expert driven construction of data dictionaries, taxonomies, and nomenclatures, a formalized ontology will be starved for data needed to feed its potential.

In addition, contemporary usage of the term ontology also implies linkage to the emerging HL7-FHIR standard for transmission of data. This ensures ability to scale connection of clinical data to the ontology across the healthcare landscape. In addition, ontology advocates working in the radiation oncology domain are increasingly aware of the need for practical demonstrations of functionality added by the use of ontologies, which use clinical data to accomplish tasks that could not more easily be accomplished with other approaches.

The AAPM is working together with stakeholders in ASTRO, ESTRO, COMP, and with industry leaders to address each of these components. The Ontology Working Group is leading in developing an OBO Foundry based ontology through a series of tightly focused projects targeting specific subsets which can be used to demonstrate value. Other groups in AAPM are engaged with domain stakeholders in ASTRO, ESTRO, COMP and CPQR in defining taxonomies and clinically viable nomenclatures for additional data elements, e.g. disease control status, that can then be incorporated into a standardized ontology. The product of these combined efforts will be an ontology that addresses all component elements and is maintained by the professional societies that are domain experts for radiation oncology.

8. Conclusion

The ever increasing use of data for driving both healthcare advances and routine medical care has provided us with some great opportunities

Fig. 6. Critical enabling component elements of what stake holders have in mind when using the term “ontology”.

and many hurdles. The opportunities involve the use of more granular information making models, matching patients to the best treatments, and providing insights that have not been clear without so much data. Unfortunately the hurdles are impeding this progress in many cases. At one end of the spectrum, it is necessary to undertake large and expensive projects to convert data and/or digital health systems (databases and applications) that may only be applicable to one institution or applications. At the other end we have the situation where the inability for different computer systems to communicate reliably greatly handicaps the ability to provide the best healthcare possible.

Ontologies can be a key element in the effort to overcome these hurdles. In the examples provided, it has been shown that ontologies can be used to share data between institutions and countries in order to build better outcome models. Other illustrated uses are building error detection software systems, facilitating communication between patients, providers and associated software systems, and extracting information from medical records in plain text.

Radiation oncology is but one arm of healthcare but it must be well-coordinated with other aspects of patients' oncological care. In addition, as radiation therapy proves more successful in curing patients and in prolonging life, information regarding a patient's radiation therapy needs to be reliably transmitted over time and space in order to understand longterm effects and to help assess on-going healthcare needs. Therefore, we hope that the community as a whole will work towards building and using ontologies as a natural part of our ongoing efforts to advance and improve.

References

- <https://ghr.nlm.nih.gov/primer/precisionmedicine/initiative> [accessed 10/20/2019].
- Schrodi SJ, Mukherjee S, Shan Y, Tromp G, Sninsky JJ, Callear AP, et al. Genetic-based prediction of disease traits: prediction is very difficult, especially about the future. *Front Genet* 2014 Jun;2(5):162.
- Jostins L, Barrett JC. Genetic risk prediction in complex disease. *Hum Mol Genet* 2011;20(R2):R182–8.
- Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J* 2015;1(13):8–17.
- Yousefi S, Amrollahi F, Amgad M, Dong C, Lewis JE, Song C, et al. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Sci Rep* 2017;7(1):11707.
- Haendel MA, Chute CG, Robinson PN. Classification, ontology, and precision medicine. *N Engl J Med* 2018;379(15):1452–62.
- Arp R, Smith B, Spear AD. Building ontologies with basic formal ontology. *Mit Press*; 2015.
- Cohen SM. Aristotle's Metaphysics. *Metaphysics Research Lab, Stanford University*; 2016. <https://plato.stanford.edu/archives/win2016/entries/aristotle-metaphysics/>.
- Gruber T. *Ontology*. New York, NY: Springer New York; 2016. p. 1–3.
- Neches R, et al. Enabling technology for knowledge sharing. *AI Magazine* 1991;12(3).
- Gruber TR. Toward principles for the design of ontologies used for knowledge sharing? *Int J Hum Comput Stud* 1995;43(5):907–28.
- Gruber TR. A translation approach to portable ontology specifications. *Knowledge Acquisition* 1993;5(2):199–220.
- Smith B, Ceusters W. Ontological realism: a methodology for coordinated evolution of scientific ontologies. *Appl Ontol* 2010;5(3–4):139–88.
- Cimino JJ. In defense of the Desiderata. *J Biomed Inform* 2006;39(3):299–306.
- Merrill GH. Ontological realism: methodology or misdirection? *Appl Ontol* 2010;5(2):79–108.
- Rosse C, Mejino JLV. A reference ontology for biomedical informatics: the foundational model of anatomy. *J Biomed Inform* 2003;36(6):478–500.
- Smith B, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 2007;25(11):1251–5.
- Ashburner M, et al. Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat Genet* 2000;25(1):25–9.
- Degtyarenko K, et al. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res* 2008;36(Database issue):D344–50.
- Noy NF, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res* 2009;37(Web Server issue):W170–3.
- Whetzel PL, 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* 2011;39(Web Server issue):W541–5.
- Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;32(suppl_1):D267–70.
- Bodenreider O, Mitchell JA, McCray AT. Evaluation of the UMLS as a terminology and knowledge resource for biomedical informatics. *Proceedings AMIA Symposium* 2002;61–5.
- Schulze-Kremer S, Smith B, Kumar A. Revising the UMLS Semantic Network. In: *MedInfo*; 2004.
- Jiménez-Ruiz E, et al. Logic-based assessment of the compatibility of UMLS ontology sources. *J Biomed Semant* 2011;2(1):S2.
- Smith M, Saunders R, Stuckhardt L, McGinnis JM. Committee on the Learning Health Care System in America, and Institute of Medicine. *A Continuously Learning Health Care System*. National Academies Press (US); 2013. <https://www.ncbi.nlm.nih.gov/books/NBK207218/>.
- Sullivan R, Peppercorn J, Sikora K, Zalberg J, Meropol NJ, Amir E, et al. Delivering affordable cancer care in high-income countries. *Lancet Oncol* 2011;12(10):933–80. [https://doi.org/10.1016/S1470-2045\(11\)70141-3](https://doi.org/10.1016/S1470-2045(11)70141-3).
- Deist TM, Jochems A, van Soest J, Nalbantov G, Oberije C, Walsh S, Eble M, et al. Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: EuroCAT. *Clin Translat Radiation Oncol* 2017;4(June):24–31. <https://doi.org/10.1016/j.ctro.2016.12.004>.
- Kim S, Wong J. *Advanced and emerging technologies in radiation oncology physics*. Boca Raton, FL: CRC Press, Taylor and Francis Group; 2018.
- Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data* 2016;3(March):160018 <https://doi.org/10.1038/sdata.2016.18>.
- Mayo CS, Moran JM, Bosch W, Xiao Y, McNutt T, Popple R, et al. American association of physicists in medicine task group 263, standardizing nomenclatures in radiation oncology. *Int J Radiat Oncol Biol Phys* 2018;100:1057–66.
- Hayman JA, Dekker A, Feng M, Keole SR, McNutt TR, Machay, M, Martin NE, Mayo CS, Pawlicki T, Smith BD, Kudner R, Dawes S, Yu JB. Minimum data elements for radiation oncology: an ASTRO Consensus paper. *Pract Radiat Oncol*, doi: 10.1016/j.prro.2019.07.017. <https://clinicaltrials.gov/ct2/show/NCT03564457> [accessed 27 February 2020].
- Deist TM, Danks FJWM, Ojha P, et al. Distributed learning on 20000+ lung cancer patient-the personal health train. *Radioth Oncol* 2020;144:189–200.
- Evans SB, Fraass BA, Berner P, Collins KS, Nurusev T, O'Neill MJ, et al. Standardizing dose prescriptions: an ASTRO white paper. *Pract Radiat Oncol* 2016;6:e369–81.
- 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;45:e854–62.
- El Naqa I, Ruan D, Valdes G, et al. Machine learning and modeling: Data, validation, communication challenges. *Med Phys* 2018;45(10):e834.
- Kalet AM, Luk SMH, Phillips MH. Quality assurance tasks and tools: The many roles of machine learning. *Med Phys* 2019;45:2006–14.
- Kalet AM, Gennari JH, Ford EC, Phillips MH. Bayesian network models for error detection in radiotherapy plans. *Phys Med Biol* 2015;60(7):2735.
- Kalet AM, Doctor JN, Gennari JH, Phillips MH. Developing Bayesian networks from a dependency-layered ontology: a proof-of-concept in radiation oncology. *Med Phys* 2017;44(8):4350.
- Luk SMH, Meyer J, Young LA, Cao N, Ford EC, Phillips MH, et al. Characterization of a Bayesian network-based radiotherapy plan verification model. *Med Phys* 2019;46(5):2006–14.
- Meyer J, Phillips MH, Cho PS, Kalet I, Doctor JN. Application of influence diagrams to prostate intensity-modulated radiation therapy plan selection. *Phys Med Biol* 2004;49:1637–53.
- Smith WP, Doctor J, Meyer J, Kalet IK, Phillips MH. A decision aid for intensity-modulated radiation therapy plan selection in prostate cancer based on a prognostic Bayesian and a Markov model. *Artif Intell Med* 2009;46(2):119–30.
- Hargrave C, Deegan T, Bednarz T, Poulsen M, Harden F, Mengersen K. An image-guided radiotherapy decision support framework incorporating a Bayesian network and visualization tool. *Med Phys* 2018;45(7):2884–97.
- Ajami S, Bagheri-Tadi T. Barriers for adopting electronic health records (EHRs) by physicians. *Acta Inform. Med.* 2013;21(2):129.
- Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook Med Inform* 2008;17(01):128–44.
- Reese RM. *Natural Language Processing with Java*. Packt Publishing Ltd.; 2015.
- Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. *J Biomed Inform* 2018;77:34–49.
- Soysal E, Cicekli I, Baykal N. Design and evaluation of an ontology based information extraction system for radiological reports. *Comput Biol Med* 2010;40(11–12):900–11.
- Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: *Paper presented at the Proceedings of the AMIA Symposium*, 2001.
- Coden A, Savova G, Sominsky I, Tanenblatt M, Masanz J, Schuler K, et al. Automatically extracting cancer disease characteristics from pathology reports into a disease knowledge representation model. *J Biomed Inform* 2009;42(5):937–49.
- Garla V, Re III VL, Dorey-Stein Z, Kidwai F, Scotch M, Womack J, et al. The Yale cTAKES extensions for document classification: architecture and application. *J Am Med Inform Assoc* 2011;18(5):614–20.
- Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010;17(5):507–13.
- IHTSDO. SNOMED CT, Copenhagen; 2007. [Online]. Available: <http://www.ihstso.org/>.

- [55] Subcommittee RS. Radlex: a lexicon for uniform indexing and retrieval of radiology information resources. [Online]. Available: <http://www.rsna.org/radlex/>.
- [56] Pathak J, Bailey KR, Beebe CE, Bethard S, Carrell DC, Chen PJ, et al. Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPN consortium. *J Am Med Inform Assoc* : JAMIA 2013;20(e2).
- [57] Rea S, Pathak J, Savova G, Oniki TA, Westberg L, Beebe CE, Tao C, Parker CG, Haug PJ, Hu SM, Chute CG. Building a robust, scalable and standards-driven infrastructure for secondary use of {EHR} data: the {SHARPN} project. *J Biomed Inform* 2012;45(4):763–71. translating Standards into Practice: Experiences and Lessons Learned in Biomedicine and Health Care.
- [58] Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc* 2010;17(3):229–36.
- [59] Leaman R, Islamaj Doğan R, Lu Z. DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics* 2013;29(22):2909–17.
- [60] Boshnak H, AbdelGaber S, Abdo Amany, Yehia E. Ontology-based knowledge modelling for clinical data representation in electronic health records. *Int J Comput Sci Inform Sec* 2018;16(10):68–86.
- [61] Pollard SE, Neri PM, Wilcox AR, Volk LA, Williams DH, Schiff GD, et al. How physicians document outpatient visit notes in an electronic health record. *Int J Med Inf* 2013;82(1):39–46.
- [62] Wright A, Chen ES, Maloney FL. An automated technique for identifying associations between medications, laboratory results and problems. *J Biomed Inform* 2010;43(6):891–901.