Treatment data and technical process challenges for practical big data efforts in radiation oncology

CS Mayo[a)]
University of Michigan, Ann Arbor, MI, USA

M Phillips
University of Washington, Seattle, WA, USA

TR McNutt
Johns Hopkins University, Baltimore, MD, USA

J Palta
Virginia Commonwealth University, Richmond, VA, USA

A Dekker
Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Centre+, Maastricht, The Netherlands

RC Miller
Mayo Clinic, Jacksonville, FL, USA

Y Xiao
University of Pennsylvania, Philadelphia, PA, USA

JM Moran and MM Matuszak
University of Michigan, Ann Arbor, MI, USA

P Gabriel
University of Pennsylvania, Philadelphia, PA, USA

AS Ayan
Ohio State University, Columbus, OH, USA

J Prisciandaro
University of Michigan, Ann Arbor, MI, USA

M Thor
Memorial Sloan Kettering Cancer Center, New York, NY, USA

N Dixit
University of California at San Francisco, San Francisco, CA, USA

R Popple
University of Alabama at Birmingham, Birmingham, AL, USA

J Killoran
Harvard Medical School, Boston, MA, USA

E Kaleba
University of Michigan, Ann Arbor, MI, USA

M Kantor
MD Anderson Cancer Center, Houston, TX, USA

D Ruan
University of California at Los Angeles, Los Angeles, CA, USA

R Kapoor
Johns Hopkins University, Baltimore, MD, USA

ML Kessler and TS Lawrence
University of Michigan, Ann Arbor, MI, USA

(Received 12 October 2017; revised 26 June 2018; accepted for publication 26 June 2018; published 18 September 2018)

The term *Big Data* has come to encompass a number of concepts and uses within medicine. This paper lays out the relevance and application of large collections of data in the radiation oncology community. We describe the potential importance and uses in clinical practice. The important concepts are then described and how they have been or could be implemented are discussed. Impediments to progress in the collection and use of sufficient quantities of data are also described. Finally, recommendations for how the community can move forward to achieve the potential of big data in radiation oncology are provided. © *2018 American Association of Physicists in Medicine* [https://doi.org/10.1002/mp.13114]

Key words: big data, ontology, standardization, informatics, machine learning

Acronyms

AAPM	Association of Physicists in Medicine
AJCC	American Joint Committee on Cancer
API	Application Programing Interface
ASTP	As-Treated Plan Sums
ASTRO	American Society for Radiation Oncology
BDAR	Big Data Analytic Resource Systems
CARO	Canadian Association of Radiation Oncology
CDR	Clinical Data Repository
CER	Comparative Effectiveness Research
CTCAE	Common Terminology Criteria for Adverse Events
DB	Database
DICOM	Digital Imaging and Communications in Medicine
DVH	Dose–Volume Histogram
ESTRO	European Society for Therapeutic Radiation Oncology
EHR	Electronic Health Record
FAIR	Findable, Accessible, Interoperable, and Reusable
FHIR	Fast Healthcare Interoperability Standards
FIGO	International Federation of Gynecology and Obstetrics
HIPAA	Health Insurance Portability and Accountability Act
HL7	Health Level 7
ICD-O	International Classification of Diseases for Oncology
ICD9	International Classification of Diseases, Ninth Revision
ICD10	International Classification of Diseases, Tenth Revision
JSON	JavaScript Object Notation
NCCN	National Comprehensive Cancer Network
NIH	National Institutes of Health
OIS	Oncology Information System
PACS	Picture Archive and Communication Systems
PHI	Protected Health Information
PQI	Patient Quality and Improvement
PRO	Patient-Reported Outcome
PROMIS	Patient-Reported Outcomes Measurement Information System
REDCap	Research Electronic Data Capture
ROIS	Radiation Oncology Information System
RCT	Randomized Controlled Trial
ROILS	Radiation Oncology Incident Learning System
RTOG	Radiation Therapy Oncology Group
SQL	Structured Query Language
TPS	Treatment Planning System
XML	Extensible Markup Language
VHA	Veterans Health Administration

1. INTRODUCTION

To the clinician, it often seems that we have too much and too little data at the same time. We spend more time than we would like at computer terminals entering or reading data. Perhaps, it would be better stated that we would like the *data* we input to be transformed into *information* that we can use. This is the aspect of *Big Data* that this manuscript addresses. Computerized data handling has been an integral part of our field since the introduction of computerized treatment planning and record and verify systems. The question is, now that there are highly successful algorithms for using computerized data to make models for predictive purposes, can the radiation oncology community harness our data for our patients' benefit?

Pan et al. have provided a very clear picture of the difficulties that we face in collecting and using data in the clinic.[1] The questions we must answer are: (a) is it worth making an effort to improve the situation, (b) what are the details of the clinical data environment that need to be addressed, and (c) how do we accomplish our goals? An AAPM Science Council Focused Research Meeting (FOREM) meeting, jointly sponsored with vendors, was held in Ann Arbor in May of 2017, to address these questions. In this publication, we provide an overview and summary of the answers that emerged.

2. MOTIVATION FOR EMBRACING BIG DATA

2.A. Need to learn from and adapt to emerging therapies such as genetics, immunotherapy

It is now commonly understood that the explosion of data and knowledge that has resulted from genomics will have a great impact on all areas of cancer care, including radiation therapy. A patient's genetic profile may play an important role in how they will react to certain agents or in their ability to repair radiation damage.[2] The tumor's genetic profiles (since many tumors have a multitude of different mutations) are increasingly being used to determine the best therapy or combination of therapies.[3]

Immunotherapy is another area of increasing importance. The ability to use different aspects of the immune system to target tumor cells is an area of great current interest.[4]

Radiation oncology is not alone in the interest and need for better data on patients' genetic profiles. NIH has been working with a number of groups to establish a workable solution in order to avoid the current problems such as laboratorydependent formats, text-based storage, and lack of centralized storage in current electronic health records (EHR).[5]

2.B. Cancer as chronic disease and multiple care givers

As cancer therapy becomes more effective, more, and more patients are living longer. As a result, the extent and complexity of information which needs to be tracked to improve understanding of outcomes is increasing. For example, for patients who are essentially cancer free, monitoring risk for treatment-related complications when their long-term home location-based follow-up is not at the treatment center is a challenge. Parry et al. estimate that there will be 18 million cancer survivors by 2020.[6] In addition, there are the increasing numbers of patients who survive longer than ever due to improvements in targeted therapies, better imaging, and better methods for localizing dose.[7] These advances can lead to improved local control and better control of oligometastases. The upshot is that as the number of patients who suffer cancer-related health consequences increases over time, the more likely it is that they will see a wider spectrum of specialists and in a larger number of clinical settings, interacting with a large variety of recording-keeping systems.

Even just considering the electronic health records, there are no general standards for the selection and formatting of data to be recorded. Different vendors, different institutions, different departments, and even different physicians have different methods which are often not compatible. Finally, even within well-structured organizations, much of the data exist within text documents. Lack of standards for which data elements to gather, inconsistent processes for entry, and variability among commercial systems for aggregation and reporting increase the likelihood that physicians and staff will miss information or have incorrect information regarding a patient's health and/or treatments that could potentially affect decisions.

2.C. Comparative effectiveness research

In the last decade, comparative effectiveness research (CER) has been seen as an important and necessary adjunct to randomized clinical trials (RCT).[8] In CER, two different therapies or tests that are already accepted are compared, whereas RCT's focus on comparing a new to a current therapy. The Patient Centered Outcomes Research Institute cites CER as its primary method of research. Given the relatively small numbers of cancer patients that are enrolled in RCT's (approximately 3%), the need to use the information that is available through CER is understandable.

Comparative effectiveness research can be tailored along a spectrum of methods ranging from essentially an RCT to a comparison of current clinical practice with an integrated practice beyond the current norm. A recent paper by Fiore et al. looked at four different trials that sought to use only data in the current EHR's.[9] Their conclusions included: "We find that EHR-based clinical trials are feasible but pose limitations on the questions that can be addressed, the processes that can be implemented, and the outcomes that can be assessed."

Clearly, for progress to be made using CER practical methods for the easy and accurate collection of data and for the sharing of data must be available in clinics.

2.D. Quality improvement and error detection

The past few years have seen an explosion in the use of data to reduce errors in radiation therapy. ASTRO and AAPM have implemented the Radiation Oncology Incident Learning System (RO-ILS) that relies on data submitted to it to develop a shared learning platform. While this system is not "big data" in the sense that it is in text format and is a relatively small amount of data, it does count in our definition of transforming data to information. In particular, the system is setup to provide users with more knowledge about the sources of errors and how best to avoid them. Another area is in artificial intelligence applications of error detection. For example, Kalet et al. successfully mined an OIS to develop a probabilistic model of the contributing factors to errors.[10]

2.E. Modeling in radiation oncology

Perhaps the most widespread use of data in radiation oncology is in modeling. The examples are too numerous to list, but some of the most impactful models are the QUANTEC models, outcomes, tumor control probabilities, equivalent uniform dose, and biologically effective dose.[11] As construction of Big Data Analytics Resource Systems (BDARS) aggregating a wider range of health care information (e.g., laboratories, medications, genomics, demographics, patient-reported outcomes (PROs), etc.) expands, more comprehensive models are progressing beyond dose metrics alone.[12–14] In addition, heuristic type models have been constructed for automating the objectives of inverse planning and library-based contouring. A promising area for the more conventional use of big data is in machine learning for automated contouring. In this application, images that have been segmented by experts are fed into a machine learning algorithm and image features that predict the true contours are selected to produce anatomical contour models.

3. STATE OF THE DATA

One of the most important concepts is that big data, in most cases, implies more data than may be obtained by any single institution. In order to use machine learning or any modeling techniques, there must be enough data to (a) build the model, (b) test the model, and (c) validate the model. Optimally, validation (c) can be done with data from a different institution in order to account for hidden variables that may not be appreciated.[15] In addition, as our ability to differentiate patients improves, for example, genomics and radiomics, the number of patients suitable for any given model decreases, thereby increasing our need for more comprehensive capture of intrainstitutional data as well as for multi-institutional data. This has critical implications for how organizations cooperate. Whereas success in medical research in the past has favored very large single institutions that can develop a critical mass of knowledge and resources in close physical proximity, diffuse networks of institutions able to generate and share information will have an advantage in the future.

In addition to the need for broad (many patients) data sources, we also need deep (relationships among key data elements) sources. Systems promoted as big data sources may in fact be shallow, capturing only a few data elements for a large number of patients. For example, some data sources draw upon billing records or imaging records for a large number of patients, but lack depth needed to enable linkage to diagnosis, treatment, dosimetric, or outcomes details. Another impediment to obtaining the “deep” type of data is that sources often dump unstructured, “as-is,” data into data lakes where key data elements and relationships can in principle be extracted, but in practice carry a high overhead for extraction. Challenges for ensuring depth in aggregation of key data elements needed for radiation oncology fall into four categories

- Access — Staff possessing both domain knowledge of radiation oncology and of informatics need access to query data bases in source systems to construct functional big data repositories.
- Data Integrity — Data elements that may not require accurate entry to enable treatment but are vital for correctly identifying specific patient groups in practice quality improvement (PQI) and research efforts require changes in clinical processes to assure validity. This often implies a cultural shift to prioritize recording data in recoverable formats.
- Data Structure — The cost of free text is high. Lack of standardized structure for entry undermines ability to automate extraction of key data elements from text fields such as notes. To assure accurate, high volume, electronic extraction of key data elements, standardized methods for encoding key data elements need to be defined and implemented in clinical processes.
- Lack of integration among systems — Key data elements are entered and stored in a range of commercial systems that frequently do not maintain linkages needed to identify relationships between key data elements. There is no existing standard of practice to link departmental datasets with radiation oncology-specific content with large commercial and governmental datasets such as the National Cancer Database Base.

4. PROCESS AND SYSTEM CHANGES

In reviewing current practices, a number of obstacles stand in the way of obtaining the amount and quality of data needed to make substantial progress. The following outline provides a view that is geared toward identifying means of overcoming them.

1. Failure to collect necessary structured data
2. Lack of data standardization
3. Inability of different electronic data systems to communicate.

Within each of these broad categories, it is useful to provide a finer grained view of how different aspects of our clinical and electronic environments contribute to the overall difficulty in achieving the data collection and use that we seek.

4.1.A. Commercial system databases

Focus for development of commercial systems that store the range of data needed for clinical data repositories is often on the user interfaces rather than on the back-end databases. The situation is similar to a clinical focus on data required to treat the day’s patients and support billing documentation with few resources devoted to standardizations and optimizations to increase big data extractions. Individual systems may use multiple loosely connected databases, complex compound keys, lack of indexing, poorly designed schema, lack reasonable security, or use nonstandard database technologies. Vendors may also refuse to provide end-users access to extract their own data. Some commercial systems are much better than others, so end-user experience is variable.

4.1.B. Diagnosis and staging

Correct usage and quantified entry of diagnosis and staging information is central to many PQI and research efforts. For example, incorrect entry of primary disease codes (e.g., prostate 185, C61) when treating subsequent bone (C79.51), brain (C79.31), or lung (C78.00) metastasis and failure to utilize functionality in radiation oncology information systems (ROIS) to connect primary and metastatic diagnosis undermine the ability to use these codes to correctly identify patient groups by codes. Failure to utilize functionality in ROIS connecting treatment courses to these codes undermines ability to connect treatment elements (e.g., DVH metrics) to patients. The cost of not taking a few seconds to select ICD-O (International Classification of Diseases for Oncology) values linked to ICD9 (International Classification of Disease, revision 9) and ICD10 (International Classification of Disease revision 10) in the ROIS is that subsequent questions about disease site location become prohibitively expensive to answer because of the manual effort required to retrospectively revisit the chart. When survival information is obtained from EHRs, failure to utilize functionality in ROIS to enter staging information undermines ability to factor staging into survival, recurrence, and other factors. Typically, EHRs do not have functionality for quantifying diagnosis and staging information according to guidelines (e.g., AJCC, FIGO) or to connect primary and metastatic disease. On the other hand, ROIS generally do, but frequently this functionality is not utilized fully as part of clinical practice.

4.1.C. Outcomes

Patient outcomes such as toxicity and disease site status (e.g., recurrence) are frequently entered into electronic records as free text using unstandardized terminology. This renders them unavailable for accurate, automated electronic extraction. Lack of standardizations (a) for which toxicities are routinely measured, (b) how treatment site categorizations are named (e.g., breast tangents, breast tangents plus supraclavicular field, breast tangents plus supraclavicular field plus internal mammary node field, etc), (c) how categorizations for disease site status are named (e.g., no-evidence-of-disease, local recurrence) or (d) in use of regular schemas for text representation of these key data elements prevent this information from being used to its full value in routine characterization of outcomes for treated patients.

4.1.D. "As-Treated Plan Sums"

To assess correlation of outcomes with dose–volume histogram (DVH) metrics, it is necessary to first create treatment plan sums corresponding to the plans and number of fractions treated, reflecting boosts, plan revisions, and incomplete treatments. When these "as-treated" plan sums (ATPSs) are created as part of routine practice, then automated solutions for calculating DVHs metrics becomes possible. Unfortunately, often these are not created as part of routine practice, with the result that they must be constructed retrospectively, *ad hoc*, preventing systematic, automated aggregation. Currently, no major commercial system, to our knowledge, has a standard means for reporting ATPSs.

4.2.A. Prescriptions

Electronic prescription summaries that defined dose levels, target structures, number of treatments, fractionation groups (e.g., first course, plan revision, boosts, etc.) and connection to target structures, organs at risk, treated plans, and DVH metrics have been developed by a few researchers.[16,17] These custom solutions were developed to fill the void left by commercial ROISs. Recently, ASTRO has suggested a baseline set of guidelines for information that should be included in prescriptions to promote standardization.[18] Similar to ATPSs, commercial solutions and clinical processes often lack ability to retrospectively extract this key information.

4.2.B. Key treatment parameters

Ensuring ability to identify which patients were treated with special technologies and details of those treatments is important to being able to prove their efficacy. Examples include breath hold technologies, radio frequency or radio-opaque fiducials used for positioning, immobilization devices, etc. However, commercial systems and clinical approaches to utilizing those systems are frequently inadequate for retrospectively gathering this data.

4.3.A. Integration of treatment planning system with ROIS

If systems do not use a common database for treatment planning system (TPS) and ROIS, it is difficult to unambiguously move from the ROIS record of plans actually treated back to specific plans, plan sums and DVH curves in the TPS. Some vendors may even discard DICOM Unique Identifiers for plans from the TPS.

4.3.B. Integration with EHR

ROIS and TPS systems typically do not integrate with EHR's. Connections may be made through medical record numbers and inferences around dates recorded in respective systems. This is an area where Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR) could significantly improve integration.

4.3.C. Integration with specialty systems

Treatment devices other than conventional linear accelerators (e.g., brachytherapy, particles, specialty accelerators, MR-guided linacs) may provide minimal details back to the ROIS or may use specialty tables in the ROIS that do not integrate well with tables used to manage external beam therapies. This limits the range of questions around treatment details for these specialty modalities that can be addressed at large scale for all patients treated.

4.3.D. Integration with institutional registry data

Institutions with the American College of Surgeons Commission on Cancer and National Comprehensive Cancer Network (NCCN) designations are required to have medical registries that follow up on cancer patients. Registries document demographics, diagnosis, staging, survival, cause of death, and other factors. Registry data is rarely linked to radiation oncology data repositories.

4.3.E. Integration with public databases

Institutional registries supply data to state registries. Published state analyses are, unfortunately, many years behind current practice. Although state registries have high volumes of patients, there is no simple means to connect back to patients to check on the validity of the data or to investigate impact of cofactors on outcomes tracked in the registries.

5. ACCESS AND EXTRACTION ISSUES

As radiation oncology has developed, a number of structural issues have arisen that limit clinicians', caregivers', and researchers' access to the data that we do have. Access requires several key elements: knowledge of the format and schema of the stored data, software that can identify and extract the data, and permissions to view and extract the data.

Figure 1 illustrates the level of detail that is needed regarding the treatment of rectal carcinoma patients under three RTOG studies. To combine the data from these trials requires knowledge of how the problem is framed (which clinical data are needed, what are the key elements of those data), how the data are formatted (type of value, allowed values, units, standards if applicable), and the specific software needed to access the data (SQL, RDF triples, spreadsheets).

The issue of framing the medical problem is difficult but rewards are high. The DICOM standard (and its radiation therapy extension) has achieved such success in large part due to its structuring of what an imaging study (radiation treatment) *is*—what are its elements and how are they related.[19] Thus, regardless of the details of the implementation of a procedure, all partners in a communication exchange agree on the essential elements. The definition of such standards in other areas of medicine is rapidly increasing. For example, a relatively commonly used standard for data exchanges between EHR's is the standard Health Level 7 (HL7). HL7 version 2 standardized types of data and the allowed values and permitted organizations and vendors to develop software for the reliable interchange of certain data. However, it was considered to be quite limited, and version 3 was built around the Reference Information Model which was a much more robust view of healthcare in general.[20] Even more recently, they have started developing HL7-FHIR which instantiates an even more up-to-date view of medical practice, but also highlights the importance of appropriate technology. HL7-FHIR is built upon the REST specification that is the current industry standard for web-based applications.[21] Other data standards, such as the NCI thesaurus,[22] provide additional resources that facilitate the development of software for access and extraction of data.

With rare exception, major vendors of ROIS, TPS, and EHR systems, store information in relational databases. A few types of large volume objects (e.g., DICOM images) are stored in files that are referenced in the relational databases. Custom extractions from databases are carried out using structured query language (SQL). SQL queries may have dialectical variation among relational database systems (e.g., Oracle, Microsoft SQL). Ideally, relational databases are designed with categories of data grouped into tables and views (stored SQL query results) reflecting an overall view of the procedure itself. They also use normalization strategies to prevent redundant information, reduce complexity in SQL queries, and increase performance in retrieving data. Secure data retrieval requires granting read access to specific authenticated network accounts. Access may be controlled at the level of the database, table, or views. Skill with SQL is essential to any staff constructing or extracting data for a data repository.

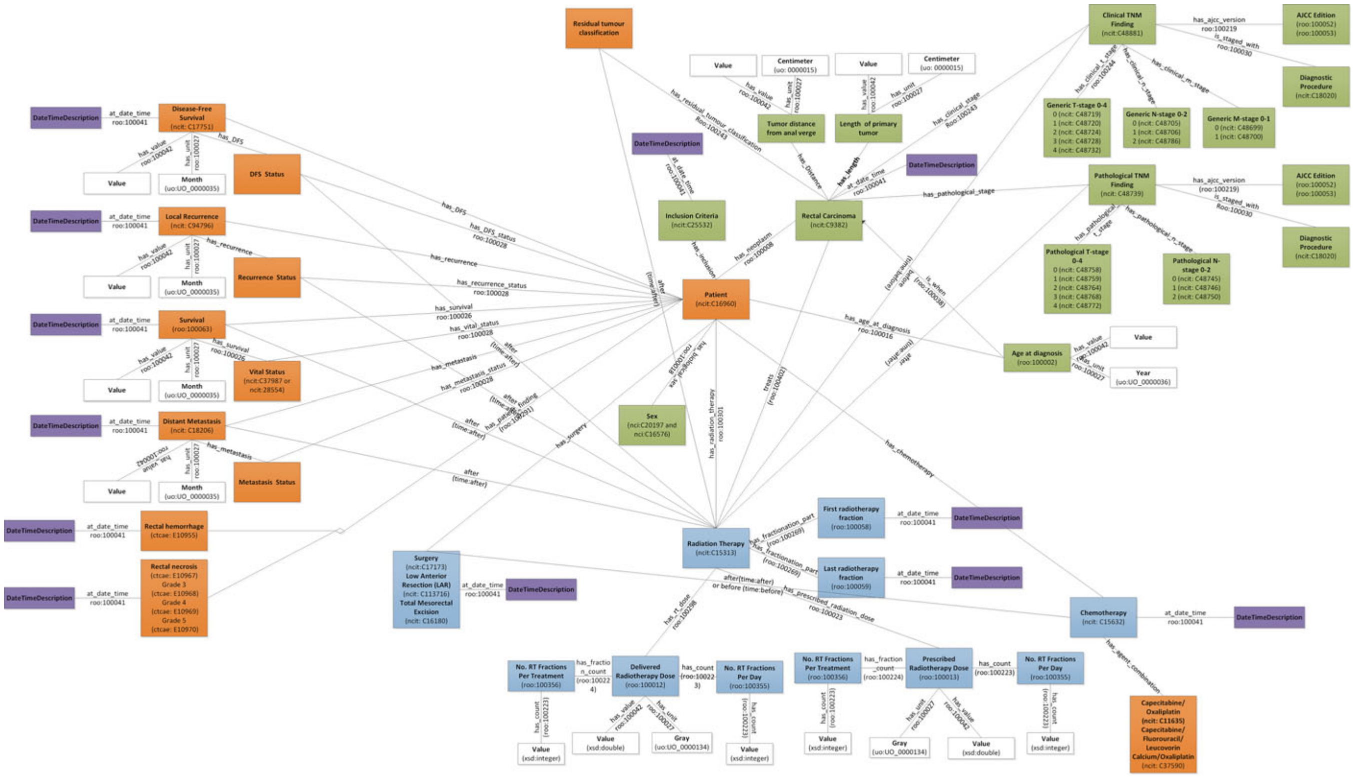
Residual tumour classification
DateTimeDescription
at_date_time roo:100041
Disease-Free Survival (ncit: C17751)
has_value roo:100042
has_unit roo:100027
has_DFS
Value
Month (uo:UO_0000035)
DFS Status
Local Recurrence (ncit: C94796)
has_recurrence
Recurrence Status
Survival (roo:100063)
has_survival roo:100026
Vital Status (ncit:C37987 or ncit:28554)
Distant Metastasis (ncit: C18206)
has_metastasis
Metastasis Status
Rectal hemorrhage (ctcae: E10955)
Rectal necrosis (ctcae: E10967) Grade 3 (ctcae: E10968) Grade 4 (ctcae: E10969) Grade 5 (ctcae: E10970)
has_DFS
has_DFS_status roo:100028
has_recurrence
has_recurrence_status roo:100028
has_survival roo:100026
has_vital_status roo:100028
has_metastasis
has_metastasis_status roo:100028
after (time:after)
after_finding
has_patient (roo:100291)
has_surgery
Surgery (ncit:C17173) Low Anterior Resection (LAR) (ncit: C113716) Total Mesorectal Excision (ncit: C16180)
Value
Centimeter (uo: 0000015)
Tumor distance from anal verge
Length of primary tumor
has_residual_tumour_classification Roo:100243
has_Distance
has_length
has_clinical_stage Roo:100243
has_pathological_stage
Inclusion Criteria (ncit:C25532)
has_inclusion
Rectal Carcinoma (ncit:C9382)
has_neoplasm roo:100008
Patient (ncit:C16960)
has_biological_sex roo:100018
Sex (nci:C20197 and nci:C16576)
has_radiation_therapy roo:100301
treats (roo:100402)
before (time:before)
is_when (roo:100038)
has_age_at_diagnosis roo:100016
has_chemotherapy
Age at diagnosis (roo:100002)
Year (uo:UO_0000036)
Clinical TNM Finding (ncit:C48881)
has_ajcc_version roo:100219
is_staged_with roo:100030
AJCC Edition (roo:100052) (roo:100053)
Diagnostic Procedure (ncit:C18020)
has_clinical_t_stage roo:100244
has_clinical_n_stage
has_clinical_m_stage
Generic T-stage 0-4 0 (ncit: C48719) 1 (ncit: C48720) 2 (ncit: C48724) 3 (ncit: C48728) 4 (ncit: C48732)
Generic N-stage 0-2 0 (ncit: C48705) 1 (ncit: C48706) 2 (ncit: C48786)
Generic M-stage 0-1 0 (ncit: C48699) 1 (ncit: C48700)
Pathological TNM Finding (ncit: C48739)
has_pathological_t_stage
has_pathological_n_stage
Pathological T-stage 0-4 0 (ncit: C48758) 1 (ncit: C48759) 2 (ncit: C48764) 3 (ncit: C48768) 4 (ncit: C48772)
Pathological N-stage 0-2 0 (ncit: C48745) 1 (ncit: C48746) 2 (ncit: C48750)
Radiation Therapy (ncit:C15313)
has_fractionation_part (roo:100269)
First radiotherapy fraction (roo:100058)
Last radiotherapy fraction (roo:100059)
after(time:after) or before (time:before)
has_prescribed_radiation_dose roo:100023
has_rt_dose roo:100298
Chemotherapy (ncit: C15632)
has_agent_combination
Capecitabine/ Oxaliplatin (ncit: C11635) Capecitabine/ Fluorouracil/ Leucovorin Calcium/Oxaliplatin (ncit: C37590)
No. RT Fractions Per Treatment (roo:100356)
has_fraction_count (roo:100224)
Delivered Radiotherapy Dose (roo:100012)
has_count (roo:100223)
No. RT Fractions Per Day (roo:100355)
Prescribed Radiotherapy Dose (roo:100013)
Value (xsd:integer)
Value (xsd:double)
Gray (uo:UO_0000134)

FIG. 1. The data from RTOG 0012, RTOG 0247, and RTOG 0822 were converted into Resource Description Framework (RDF) specifications and were uploaded onto the NRG/IROC/ACR node of the Varian learning portal. The mapping was performed according to the diagram shown above. Distributed learning is enabled for contracted institutions. The distributed learning between this node and another node on the Varian learning portal (MAASTRO Clinic, the Netherlands) was tested successfully.

Application programming interfaces (APIs) are provided by vendors of many TPSs. These may be used to gather subsets of information stored in the ROIS database or elements only calculated at run time in the TPS (e.g., DVH curves for some systems). APIs allow custom software applications to be constructed by users that interface with the TPS. Access is controlled by end-user system administrators, subject to constraints of the commercial system. Clinical staff members with coding skills are necessary for effective use of API's.

Legacy issues with vendor changes to both database and API structures are an issue for groups automating extraction from electronic records systems. Effort to re-write queries and scripts when systems are upgraded can be substantial.

Patient-reported outcomes are important outcome measures and their routine monitoring during cancer therapy has been demonstrated to improve survival.[23] However, use of paper based rather than electronic systems are more common. Electronic systems are significantly better for making the data accessible, but require substantial effort in setting up systems and arranging for staffing resources to assist patients with completing electronic surveys is required. In addition, lack of standardization in instruments to be used, redundant questions between surveys, excessive length diminishing patient willingness to participate, and question formats and logic that translate poorly to electronic systems already used in patient work flows are issues for generalized use of PROs.

Diagnostic images are stored on Picture Archive and Communication Systems (PACS) in Digital Imaging and Communication in Medicine (DICOM) format and accessed with DICOM servers. Graphic user interfaces for clinical use are not well suited to large volume, batch access of sets of patient images. The objective in utilizing these resources in connection with BDARS is not creation of a parallel PACS. Instead, when large sets of images are identified for utilization in a study, for example, developing predictive radiomics measures for a disease type, downloading a large specific set of images for batch processing is needed. Negotiating access is the primary barrier.

Finally, it is important to discuss the role that legal and commercial considerations play in limiting access to data. The Health Insurance Portability and Accountability Act of 1996 requires certain standards to be met when exchanging private health information. The standards depend on the intended use of the data, for example, clinical decisions, insurance coverage, quality improvement, and research. They also depend on the entities exchanging the information. These standards add time, effort, and new procedures to any effort to obtain data access. Intra-institutional exchange, for example, between a departmental data repository and the hospital EHR, is in general easier than between institutions, but even that type of transaction usually requires some level of administrative oversight and/or procedure. In addition, storing data in a clinical data repository for possible future research can be viewed as problematic under national ethics guidelines for human research.[24] Overall, it is difficult to make any broad statements or recommendations regarding these issues since they are, to some degree, institution- and use-specific. In addition, how the regulations are interpreted is evolving, particularly in response to some of the national healthcare programmatic initiatives such as the Affordable Care Act.

6. SELECTING TECHNOLOGIES

The objective is to use the treatment data, rather than to utilize a novel database technology. Selecting database technologies which minimize investment overhead and risk while maximizing productivity and interoperability for addressing particular tasks requires careful consideration.[25,26]

At a high level, four process steps can be considered and technology choices should be made fit-for-purpose for these steps.

6.A. Capture of treatment data

The primary use for health care data is delivery of patient care. Health care database technology is often vendor dependent and under regulatory oversight. For structured data elements (e.g., record and verify, electronic health records, outcome) relational databases are the most common technologies. Images and related objects such as treatment plans and record are generally object stores (e.g., PACS) with a relational schema containing object pointers.

6.B. Extraction

Since the primary use sources have to be taken as-is, the extraction technologies providing connectors to these primary sources should be able to handle many different sources and formats including all common relational sources. They should be able to handle nonrelational sources including "databases" that researchers and physicians often use (e.g., Excel, SPSS) and include JSON and XML support as these are common export format for more technical users. Ideally, the technology can be extendible to support common medical standards (HL7v2, HL7v3, HL7 FHIR and DICOM) as needed.

A wide range of programming languages and standard database import tools are frequently used. These have the advantage of hiding very little from the user. There are also commercial and open source software systems intended to reduce the technical skill requirements for users with the trade-off of obscuring details about the extraction, cleaning, and loading processes. Since primary sources change and extraction tools generally expand and change over time, a crucial requirement is versioning. Users of technology should be able to store different versions of the extraction scripts and configurations so that subsequent users can re-use their solutions.

6.C. Transformation, integration, and storage

For successful secondary use, the primary use sources need to be combined, integrated, and common data elements mapped on each other. An example is the combination of

ROIS/EHR data (diagnosis, comorbidities, prescriptions, treatments, follow-up), Record and Verify data (radiotherapy treatment), and DICOM data (imaging/plan). This transformation and integration is generally the most time consuming task of the process. Knowledge of the primary sources and of the secondary use data model is a requirement for staff using the tool. Again, versioning and manageability is crucial as sources change and sharing transformation scripts with others is needed for work to not be duplicated. Defining distinctions between data element categories and relationships means mapping the raw values onto a schema. For example, a schema needs to be applied so that we can inform our analytics programs if an extracted value "30" corresponds to a dose, an age, a day of the month, etc., and how that value relates to other information e.g., toxicity, survival, PROs, treatment dates, etc.

From a technology standpoint two main approaches exist.

- Schema-On-Aggregate (aka schema-on-write): Upon extraction each data element from each source is considered more or less separately, transformed, and mapped to the secondary use data model and then written in the secondary use data store. Schema-on-aggregate has as its main benefit that it often re-uses the knowledge contained in the primary use schema and forces one to decide up front how to map data items and think about transformation for each data element. The end-result is often a data store with a structured schema. Relational databases are widely used for this approach owing to their speed, ease of integration with other systems and large pool of talent for use. Nonrelational databases (e.g., object stores, graph databases and triple stores) have also been used in some research settings.
- Schema-On-Query (aka schema-on-read): The secondary use data model is applied when the secondary user requests, or queries, the data from the secondary source. In a schema-on-query system the data is stored from the primary source "as-is" and by necessity this is a nonrelational store (e.g., a data lake). An example is Apache Hive which can be used for SQL-like schema-on-query for Apache Hadoop. NoSQL databases, such as MongoDB or CouchDB, are another example. The main benefit of this approach is that the transformation and secondary use data model can be defined fit-for-purpose, and different for different use cases. Also, all primary use data can be stored immediately for later secondary use. The main drawback is that knowledge of original schema is often not available by the time the data is used and that data is stored without de-identification. Variability in nomenclature for key data elements, relationships, and formats among the various "as-is" sources requires creating and maintaining custom code for each to enable programmatic extraction. Care must be taken to ensure consistent meaning at the time of data entry so that contents of an element are internally consistent and stable.

Note that many solutions allow a combination of the above approaches, with some data elements stored in a schema generation upon aggregate and some stored "as-is" for schema at a later time point. In that case, key data elements are often duplicated into the secondary use storage.

6.C.1. Secondary use application

Secondary use of subsets of data extracted from BDARS to address specific research or clinical questions is a common use case. The secondary user usually has defined their own data model, store, and the application to analyze the data. The technology choices made by secondary users vary widely and limited influence exists especially if the secondary user is external to the primary use institution. The main job of technology here is to provide the secondary end-user with a dataset and format which he or she can use (often called a data mart). Typical requested formats include SQL database dumps, Microsoft Excel, comma (or tab) separated values (CSV), DICOM, HL7 FHIR, HL7v3, HL7v2, XML, and JSON. Additionally, data visualization and allowing the end-user to navigate the data store established in the previous step increase the efficiency and effectiveness of secondary use. The tools mentioned above generally allow such export to a variety of data formats. Figure 1 illustrates one such use case, a semantic triple store database (a.k.a. Resource Description Framework) was applied for the purpose of combining datasets from several clinical trials. Semantic triples can be used to define a range of relationships between objects (e.g., PTV → is a type of → target structure).

7. SPECIFIC RECOMMENDATIONS FOR WORK FLOWS AND STANDARDIZATIONS

1. Diagnosis and staging data should be entered into quantified fields in accessible, electronic systems that

a. have quantified fields for staging elements and overall staging, and staging guideline system used (e.g., American Joint Committee on Cancer (AJCC))
b. ensure correct selection of staging from component elements
c. provide explicit linkage to treatment courses and plans used to treat
d. link metastatic diagnosis (e.g., C79.51, Secondary malignant neoplasm of bone) to diagnosis for originating sites (e.g., C34.1, Malignant neoplasm of upper lobe, bronchus or lung).

In the current vendor landscape, the ROIS is frequently the only system in the clinical process workflow meeting these objectives.

2. Nomenclature standardizations recommended by AAPM Task Group 263 should be adopted into routine practice. These define standardized nomenclature for structure, target and DVH metric naming to promote ability to automate aggregation.[27]

3. Course cumulative as-treated plan sums should be constructed as part of routine practice. Since more than one image set may be used in the construction of the ATPS's, and relative positioning of structures may vary between sets, using the image set providing the best representation for the clinical evaluation carried out for treatment is currently the most viable approach.
4. Toxicities, recurrence, and PRO outcomes need to be routinely collected as quantified fields (instead of free-text fields) in accessible electronic systems. Standardizations for specific items and values are needed. This includes, for example, definition of recurrence nomenclature. Ability to automatically recover these values from the electronic record is important.
5. Detailing of key data elements and relationships (i.e., an ontology) is needed for a broad range of practice quality improvement and translational research efforts. An initial set, drawn from experience in constructing BDARS, is presented as an appendix to this paper. Success in gathering this information requires that clinical systems should be utilized to ensure ability to accurately aggregate these elements and relationships from the electronic record (ROIS, TPS, EHR). Ideally, professional societies such as ASTRO, AAPM, ESTRO, and CARO would combine efforts to eventually take the role of maintaining standardized ontologies to promote interoperability among institutions and commercial systems. Combining the ontology presented in the appendix with related ontologies would be a valuable step toward a common standard.[28,29]
6. In addition to demonstrating adherence to standard quality metrics, clinical entities will face increasing demands for demonstration of the value of the care they deliver as medicine in the transitions from fee for service to value based payments. Success in the value based payment environment will require the ability to conduct on-demand analysis of patient and tumor characteristics, all aspects of treatment delivery, outcomes, and cost of care.

We note that the task of creating ATPSs (item 3) needs to begin as soon as possible, guided by clinical judgment, in order to replace complete lack of data with reasonable data. In addition, further refinement is needed. Collaborations between professional societies, vendors and clinical trials groups for defining standards for the end-of-treatment dose composite are needed. Issues include means to quantify quality of the composite, identifying source images, identifying trade-off decisions in image registrations, uncertainties in structure dosimetric measures when multiple image sets are used, and realistic appraisal of the role of image deformation.

8. EXAMPLES OF CLINICAL DATA REPOSITORIES

Several groups have been actively engaged in construction of clinical data repositories (CDR), also known as data lakes and BDARS. These systems become important components for both research and clinical practice efforts in their clinics. Practical recommendations from this group have been grounded in the experience of constructing, using and sharing these systems. Brief summaries of several are highlighted to convey the scope and volume of these resources.

- The University of Michigan Radiation Oncology Analytics Resource (M-ROAR) automates aggregation of electronic data from the Treatment Planning System (TPS), Radiation Oncology Information System (ROIS), Electronic Health Record (EHR), and other databases for all patients treated. Data types include demographics, treatment and dosimetric data, chemotherapy, toxicities, comorbidities, labs, medications, encounters, and PROs. The system contains records for over 20,000 patients. Key data elements are extracted utilizing a combination of SQL queries, TPS application programming interface (API) based scripts and custom code to extract and process data from multiple source systems.[25]
- The UCLA Clinical Informatics Management System (CIMS) consists of three major modules: a physician interaction module that interacts closely with EHR, a physics parameter module that handshakes with PACS systems, treatment planning, and delivery stations for quantitative value collection and exchange, and a patient-reported outcome management system [Patient-Reported Outcomes Measurement Information System (PROMIS)] with a web/mobile portal. The physician interaction module supports comprehensive query for collection and integration of radiotherapy relevant information from other departments. The patient-reported outcome management module consists of a front-end with site-specific patient-oriented Common Terminology Criteria for Adverse Events (CTCAE) questionnaires tailored to patients. As of now, the registry contains records for 1790 definitive prostate treatment, 209 post-operative prostate treatment, 1950 breast, 663 lung, 531 brain metastasis, 484 GYN, 424 glioma, 409 meningioma, 209 rectum, 151 metastatic bone, 164 trigeminal, 111 pancreas, and over 3000 general cases.[30]
- The Ohio State University Radiation Oncology Department's "Quality Database" has been designed to serve as a data aggregation platform to capture clinical, technical, and health outcome data on all patients who receive radiation treatments. All data are stored in a REDCap database. Smart texts have been implemented in EHR to enable automated capture and extraction of discrete data elements such as adverse events from provider notes. The dosimetry data for radiation therapy are extracted via TPS's API. Demographics, diagnosis, tumor biomarkers, surgery, systemic therapy, radiation therapy, and adverse events constitute the collected data and provide means for determining effectiveness of treatment modality. The Quality Database currently contains 3385 patients and

is being populated prospectively with new patient data.

- Oncospace: Johns Hopkins University developed a comprehensive data collection and data repository system.[31] The system consists of a network of data collection systems (ROIS, clinic computer terminals, mobile devices, hospital EHR) that provides data that is transformed and loaded into a SQL database. Using a federated database approach (including University of Washington, University of Virginia, Odette Cancer Center-Sunnybrook), each institution has implemented compatible schemas so federation-wide queries will succeed. This approach has the advantages of "crowd-sourcing" ideas and technology and allowing each institution to keep control of their data while still permitting individual flexibility.
- The Veterans Health Administration (VHA) developed a pilot Radiation Oncology Practice Assessment (ROPA) program to assess the quality of radiotherapy across the entire VHA network with 40 institutions participating.[32] Data types include quality metrics targeted at workup, diagnosis, treatment planning, delivery and follow-up. The gathered quality metrics were developed by the VHA in partnership with ASTRO for locally advanced nonsmall cell lung cancer, limited stage small cell lung cancer, and intermediate and high-risk prostate cancer. Data extraction for the initial pilot project will be completed in 2018. At that time, ROPA is anticipated to contain 45,000 scores for 49 metrics aggregated from approximately 2,000 patients.

Large datasets from sources outside of radiation oncology are now available for analysis. Waddle et al. recently published utilization data derived from insurance records from a commercial warehouse (Optum Labs) to examine treatment technologies used (proton, stereotactic body radiotherapy, IMRT, 3D, other) by diagnosis code used in billing records. The data base contains utilization data on a subset of 474,533 radiation oncology patients from a larger database of over 100 million insured lives. However, connection of this data to clinical outcomes and other cofactors was pending at the time of that analysis.[33]

9. RECOMMENDATIONS FOR NEXT STEPS NEEDED TO IMPROVE DATA AVAILABILITY

9.A. Adopting national standards

As discussed above, an important aspect of data exchange is employing a generally recognized view of the medical process. HL7 FHIR is an emerging standard and one that has the crucial elements of (a) flexibility, (b) state-of-the-art technologically, and (c) widespread support.[34] As this standard is just not being formalized, this is an excellent time for the radiation oncology community to support efforts to develop radiation oncology-specific resources for this standard.[35]

9.B. Increasing multi-institutional collaborative efforts

Real, effective standards emerge from being actively engaged in exchanging data with outside groups as part of more frequent collaborations. Professional and government grant support for research efforts that develop and proof these standards as by-products are important to their emergence.

Included in this effort is the need to facilitate information exchanges that support retreatment. As patients are able to survive longer with cancer, likelihood of visiting more than one center for subsequent treatments also increases. Clinical process and data exchange standardizations needed to facilitate these exchanges should also support collaborative efforts.

9.C. Links to institutional registries

Institutions which are members of the National Comprehensive Cancer Network (NCCN) are required to have access to a registry which carries out longitudinal follow-up on a few key data elements (e.g., survival, cause of death) for treated patients. EHR database records may be substantially different from registry database records. Providing electronic access registry databases provides opportunities to synchronize data sources in constructing BDARS.

9.D. Support for public data sets

The value of producing data sets that can be publicly shared (without compromising PHI) has been heralded by several authors.[36–38] There is growing interest from funding agencies for publicly funded research to produce publically available datasets. Similarly, an increasing number of journals require publication of datasets accompanying findings. Recently Medical Physics has introduced a special publication category just for data sets. Principles for ensuring that data are findable, accessible, interoperable, and reusable (FAIR) for public access of data sets have been set out by Wilkinson et al.[39] and others.[40]

The National Cancer Institute has recently begun to implement a Cancer Research Data Commons which meet the standards of FAIR. In their announcement, they echo a number of the themes that we have set forth in this article. This is clearly a propitious time for radiation oncology to join with others in the oncology fields to make these sorts of community-wide efforts more productive.[41]

9.E. Informatics training

Clinical staff bring great value to informatics efforts because of the depth of their clinical domain knowledge with respect to key data elements, their interrelationships, clinical processes by which data is entered, end user expectations for meaning, etc. The set of clinical staff that take on expanding their informatics skills to include database, programming, statistical analysis, and machine learning also improve ability to develop practical solutions bridging needs between the larger number of specialists entirely focused in either the clinical or informatics domains.

10. CONCLUSIONS

We have laid out an argument for why it is important for the radiation oncology community to improve the means by which we can collect, share and use the data that we encounter every day. However, for various reasons, much of this data remains inaccessible to us in a format that makes it easy for us to transform data to knowledge.

The technological challenges to implementing a community-wide system of data collection, sharing and usage are formidable but the tools have been or are currently being developed. More difficult is developing the collective will to make it happen. Such a change in our clinical behavior and workflow requires buy-in from everyone, including clinic staff, physicians, and vendors. It is our hope and expectation that this sea change has already started to occur as diffuse networks grow in size and analytic power. It is necessary to do so if we are to continue to be at the forefront of harnessing technological advances to improve the treatments that we provide our patients.

ACKNOWLEDGMENTS

The authors thank Zhong, Haoyu, MS for Fig. 1.

This work was the result of collaborative efforts from participants at a meeting supported by AAPM, Varian Medical Systems, and Elekta.

CONFLICTS OF INTEREST

There are no conflicts of interest.

APPENDIX

KEY DATA ELEMENTS AND RELATIONSHIPS: A RADIATION ONCOLOGY TRANSLATIONAL RESEARCH ONTOLOGY

We have defined of a common set of key data elements and relationships important to a broad range of patient quality improvement and translational research efforts. Ranking treatment information for effectiveness requires a broad scope of information types: Radiation Treatments, Surgery, Outcomes, etc. While it is desirable to have all the data readily available, that is not a practical starting point. Our objective here is to define a minimal set of information needed to handle frequently encountered questions as a common use starting point. With that, technical and procedural efforts attempting to automate electronic aggregation supporting Big Data efforts can use these recommendations as a guide.

Optimally professional organizations (e.g., AAPM, ASTRO, ESTRO, CARO) would establish an official listing of key data elements and relationships. Our intention here is to provide a practical starting point from our experience in aggregations from multiple source systems.

The listing of key data elements and relationships define an explicit conceptualization of a body of formally represented knowledge about Radiation Oncology, that is, an ontology[42]. The listing provided here was based on the ontology developed for M-ROAR[25] and expanded as an outgrowth of discussions at the Practical Big Data Workshop. Incorporation of the ontology into a programmatic form using Ontology Web Language (OWL) is underway.

Classes (⊕) of information, list key data elements (aka properties) denoted by one of three symbols (●, ⊙, ○). Most elements (●) do not require special consideration for protection of patient health information (PHI). Elements that contain PHI (⊙) are problematic for data sharing or storage in cloud based systems. Alternatives (○), containing, reduced information, may be sufficient for a wide range of collaborative efforts or cloud based storage.

For example, dates are a type of PHI that institutional review boards (IRB) will not allow for many applications. For a wide range of investigations, detailing temporal relationships between events is important. Recording the patient's age at the event, rather than the date for the event is an alternative. For example, if the date of an event is 3/2/2013, and the patient's date of birth is 8/17/1967, then the patient's age at the time of the event, to three decimal places (Decimal F3), is 45.541. This is sufficient resolution to differentiate day on a timeline and meets requirements for protecting PHI.

Several key data elements typically are not present as distinct values in source data systems but have to be programmatically derived (⌘) from other elements. For example, the age of the patient at the time of an event is derived from date of birth and date of the event. Starred (*) items indicate particular need for recommendations of standardized values recommendations from professional societies.

When elements have only one instance they are indicated by the name of the class or element (e.g., DateOfBirth, Patient). When there may be more than one instance of an element, this is indicated by specifying a list of elements of this class (e.g., List<Course>).

Relationships among classes are categorized as Parent (⇦), Child(⇨), Sibling (⇔) or Property(■). Parent–child is a dependent relationship: a parent class object is referenced in each instance of a child class object. Sibling relationships are tracked if elements exist but do not imply dependence. Sibling relationships rather than parent–child relationships may be selected when the current state of the data will not practically support the dependent relationship. For example, Prescriptions are used in sibling relationships with respect to TreatedPlans because the current state of electronic data is inadequate to assure consistent mapping. Property relationships are used when class incorporates a set of elements grouped under a single concept.

⊕ **Patient -**

- ⊙ PatientMRN (String) -:Medical Record Number
- ○ PatientGUID (String): Generalized Universal Identifier that can be used *in cloud based storage, when PatientMR is not.*
- ⊙ DateOfBirth (Date)
- ○ YearOfBirth (Int?) ⌘
- ⊙ DateLastSurvivalCheck (Date?)
- ○ AgeAtLastSurvivalCheck (Decimal F3) ⌘
- ⊙ DateOfDeath (Date?)
- ○ AgeAtDateOfDeath (Decimal F3) ⌘
- IsAlive (Bool) – Status at last at Last Survival Check Date
- *CauseOfDeath (String) – Need for standardized list
- Gender (String)
- Race (String)
- Ethnicity (String)

Child class relationships

- ⇨ List<Radiation Therapy Course>
- ⇨ List<Prescription>
- ⇨ List<DiagnosisAndStaging>
- ⇨ List<TreatedPlan>
- ⇨ List<PatientTreatmentOutcome>
- ⇨ List<PatientReportedOutcome>
- ⇨ List<PlanningStructureSet>
- ⇨ List<HealthInformation>
- ⇨ List<Lab>
- ⇨ List<Medication>
- ⇨ List<Image>
- ⇨ List<Chemotherapy Course>
- ⇨ List<Surgical Procedure>
- ⇨ List <Pathology>
- ⇨ List <Charge>

⊕ **RadiationTherapyCourse** ⌘ – These are the treatment courses. A course Every patient has a list of courses

- CourseName (String)
- NTxSessionsInCourse (Int) ⌘ – Each treatment episode is a session, sessions used for imaging only are exclude from the count

- ⊙ DateFirstTreatment (Date)
- O AgeAtFirstTreatment (Decimal F3) ⌘
- ⊙ DateLastTreatment (Date)
- O AgeAtLastTreatment (Decimal F3) ⌘

Sibling Class Relationships

- ⇔ List<Prescription>
- ⇔ List<Chemotherapy Course>
- ⇔ List<Surgical Procedure>

Child class relationships

- ⇨ List<TreatedPlan>
- ⇨ List<DiagnosisAndStaging> – Typically only one per Course
- ⇨ List<PatientTreatmentOutcome> – Typically only one per Course
- ⇨ List<Charge>

Parent Class Relationships

- ⇦ Patient

⊕ **Prescription :** The prescription needs to fully convey the intent of the physician for the treatment plan. The Course contains a list of prescriptions

- Name (String)
- NTxSessions (Int)
- NTxPerDay (Int)
- DaysBetweenTxSessions (Decimal) ⌘
- StartOnNthDayFromCourseStart (Int) ⌘
- StartOnNthSessionInCourse (Int) ⌘
- RxDoseUnits (String) – "cGy" or "Gy" or "CGE"
- IsCourseCummulativePrescription (Bool) ⌘ – Only one value of True per Course

Class Property Relationships

- ■ List<PrescriptionDoseLevel>
- ■ List<PrescriptionDVHObjectives>

Sibling Class Relationships

- ⇔ List<TreatedPlan>

Parent Class Relationships

- ⇦ DiagnosisAndStaging
- ⇦ Patient
- ⇦ Course

⊕ **PrescriptionDoseLevel**

- RxDose (Decimal F3)
- RxStructure (String) – AAPM TG263 compliant name
- RxPointName (String)

Parent Class Relationships

- ⇦ Prescription

⊕ **PrescriptionDVHObjectives**

- Structure (String) – AAPM TG263 compliant name

- DVHMetric (String) – AAPM TG263 compliant name, e.g., Max[Gy], V20Gy[%]
- Constraint (String) – allowed values are =,<,≤,>, ≥, ALARA
- Value (Decimal F3) – null if constraint is ALARA

Class Property Relationships

■ Prescription

⊕ **DiagnosisAndStaging**

- StagingSystem (String) – e.g., AJCC 7, FIGO
- ICD9Or10 (String)
- ICD0 (String) – Defines location of disease
- Laterality (String) – Left, Right, Bilateral
- Overall Staging (String): e.g., IIa, X,
- T (String)
- N (String)
- M (String)
- P (String)
- G (String)
- OtherStagingComponents (String) – Staging components other than T,N,M,P,G
- PrimaryOrMetastatic (String) ⌘– Either "Primary" or "Metastatic"

Child Class Relationships

⇨ PatientTreatmentOutcome
⇨ DiseaseSiteStatus

Parent Class relationships

⇦ PrimaryICD9Or10? – If Metastatic, indicate Primary DiagnosisAndStaging element
⇦ Course
⇦ Patient

⊕ **DiseaseSiteStatus**

- DateOfStatus (Date)
  ⊙ AgeAtDateOfStatus (Decimal F3) ⌘
- *Status(String) – Need standardized list, e.g., No Evidence of Disease, Local Recurrence, Regional Recurrence, Distant Recurrence

⊕ **TreatedPlan** : Every course has a list of treated plan objects. One table for all types of plans defining key elements to track. This simplifies mixed modality tracking e.g., External + Brachy and handling of individual plans vs plan sums. Only plans actually treated are tracked. Details of actual vs number of fractions delivered are tracked.

- PlanName (String): Corresponds to PlanID in ARIA
- *TreatmentAreaClassifier (String): e.g., Head and Neck, Lung_L, Breast_R+SC
- TPSSourceSystem (String) ⌘
- IsCourseCummulativePlan (Bool): The plan or plan sum(ATPS) represents all plans treated in the course
- IsPlanSum (Bool): The dose associated with the plan is created by summing dose from other plans

⊙ DateOfFirstPlanTreatment (DateTime)
○ AgeAtFirstPlanTreatment ⌘
⊙ DateOfLastPlanTreatment (DateTime)
○ AgeAtLastPlanTreatment ⌘

- PrimaryTxDeliveryFacility (String) – Facility where most of plan fractions were delivered
- PrimaryTxDeliveryMachine (String) – Machine on which most of the plan fractions were delivered
- NFractions_Planned (Int)
- NFractions_Delivered (Int)
- TotalDose_Planned (Decimal) – Dose planed for highest dose structure, e.g., PTV_High
- TotalDose_Delivered (Decimal) – Dose delivered for highest dose structure, e.g., PTV_High
- TotalDose_Units (String) – Gy, cGy, CGE
- UsedFiducials (Bool) ⌘
- FiducialType (String) – Gold, Calypso, Carbon
- UsedBreathMotionControl (Bool) ⌘
- BreathMotionControlType (String): SDX, ABC, Compression
- MeanSessionTimeMinutes(Int) ⌘
- MeanSessionBeamOnTimeMinutes (Int) ⌘
- MeanSessionImagingTimeMinutes (Int) ⌘
- NImages_MV (Int) ⌘ - Total number of MV images for all sessions treating this plan
- NImages_kV (Int) ⌘ - Total number of kV images for all sessions treating this plan
- NImages_CBCT (Int) ⌘ :Total number of CBCT for all sessions treating this plan
- NImages_MR (Int) ⌘: Total number of MR images for all sessions treating this plan
- List<SupplementalTreatmentDetail>

Sibling Class Relationships

⇔ Prescription
⇔ List<Images> – Image Class Objects related to the TreatedPlan, e.g., CBCT, kV

Child Class Relationships

⇨ PlanningStructureSet
⇨ List<DVHCurve>
⇨ List<DVHMetric>
⇨ List<PatientPositioningDevice>
⇨ TreatmentPlanDetails_XRT

⇨ TreatmentPlanDetails_Brachy
⇨ TreatmentPlanDetails_Particles
⇨ PlanningStructureSet

Parent Class Relationships

⇦ Patient
⇦ Course
⇦ ComponentOfATPS (TreatedPlan) – Plans that are components of ATPS link back to the ATPS

⊕ **PlanningStructureSet**

- StructureSetName (String)
- ImageModality (String) ⌘ : e.g., CT, MR

⊙ DateOfImageAcquisition (Date)
○ AgeAtImageAcquisition (Decimal F3) ⌘
⊙ DICOMImage_UID (String) DICOM_UID of image use for the plan. In the Image list attached to the patient.
⊙ DICOMPlan_UID (String)
⊙ DICOMStructure_UID (String)
⊙ DICOMDose_UID (String)

- PatientPosition (String)

Parent Class Relationships

⇦ Patient
⇦ TreatedPlan

⊕ **PatientPositioningDevice**

- *DeviceCategory (String) – Need standardized list
- DeviceName(String)
- SetupDetails (String)

⊕ **TreatmentPlanDetails_XRT**

- List<EnergyModality>
- TotalPlanMU (Decimal)
- UsedIMRT (Bool) ⌘
- UsedVMAT (Bool) ⌘
- UsedFIF (Bool) ⌘
- UsedWedges (Bool) ⌘
- UsedBolus (Bool) ⌘
- UsedNonCoplanarBeams (Bool) ⌘
- NBeams (Int) ⌘
- NFractionsPlanned (Int)
- NFractionsDelivered (Int)
- List<SupplementalTreatmentDetail>

Parent Class Relationship

⇦ TreatedPlan

⊕ **TreatmentPlanDetails_Brachy**

- List<EnergyModality>
- NSourcesTotal (Int)
- TotalActivity (Decimal)
- *TotalActivityUnits (String)- Need standardized list, e.g., MBq, Ci, mCi, GBq
- UsedRadiopharm (Bool)
- UsedApplicator (Bool)
- TotalHDRDwellTimeMin (Decimal)
- TotalPDRDwellTimeMin (Decimal)
- TotalLDRImplantTimeMin (Decimal)
- List<SupplementalTreatmentDetail>

Child Class Relationships

⇨ List<Applicator>

Parent Class Relationship

⇦ TreatedPlan

⊕ **Applicator**

- *ApplicatorType (String) Need standardized list e.g., Needle, BrachyCath, TandemAndOvoid, Cylinder, Mamosite, Savi
- NApplicatorsInserted (Int) ⌘
- NApplicatorsUsedInTx (Int) ⌘

Parent Class Relationships

⇦ TreatmentPlanDetails_Brachy

⊕ **TreatmentPlanDetails_Particles**

- List<EnergyModality>
- UsedPassiveScattering (Bool)
- UsedSpotScanning (Bool)
- UsedEndOfRangeToSpareCriticalOAR (Bool)
- List<SupplementalTreatmentDetail?>

Parent Class Relationships

⇦ TreatedPlan

⊕ **EnergyModality**

- Energy (String) – Need standardized list, e.g., X06, X06FFF, X10, X10FFF, E06, E09, E12, E16, E20, Ir192, I125, P70, C250
- *Modality (String) – Need standardized list, e.g., XRT, HDR, LDR, Proton, CyberKnife, Gamma-Knife

Parent Class Relationship

⇦ TreatedPlanDetails_XRT
⇦ TreatedPlanDetails_Brachy
⇦ TreatedPlanDetails_Particles

⊕ **SupplementalTreatmentDetail**

- Name (String)
- Value (String)
- ValueType (String)

Parent Class Relationships

⇦ TreatedPlanDetails_XRT
⇦ TreatedPlanDetails_Brachy
⇦ TreatedPlanDetails_Particles
⇦ TreatedPlan

⊕ **Image**: Information about image objects relevant to patient's treatment

- ImageName (String)
- DICOM_UID (String)
- ImageModality (String), e.g., CT, kV, CBCT, MR-T1w, MR-T2w, PET, etc.
- SourceSystem (String) ⌘ Where to find the image and how to get it, e.g., ARIA, Velocity, Hospital PACS, etc

⊙ AccessionNumber (String)

- StudySeries (String)
- BodySite (String)

⊙ DateOfImageAcquisition (Date)
○ AgeAtImageAcquisition (Decimal F3) ⌘

- RelevanceComment (String?), e.g., TumorResponse

Sibling Class Relationships

⇔ List<ImageDataFeature>
⇔ TreatedPlan
⇔ Course

Parent Class Relationships

⇨ Patient

⊕ **DVHCurve**: Store the DVH curve for as treated (i.e., number of fractions delivered) plans and plan sums. Every Treated Plan has a list of DVH curves

- StructureName (String) – Use TG263 Standardization
- Volume[cc] (Decimal)
- Min[Gy] (Decimal)
- Max[Gy] (Decimal)
- Mean[Gy] (Decimal)
- Median[Gy] (Decimal)
- Stdev[Gy] (Decimal)
- DVHCurve (String) ⌘ – Dose, Volume tuples separated by semi-colons. Dose is in units of Gy, Volume is in units of percent of structure volume, e.g., 0,100; 50,100; 50.5,99.5; etc.

Sibling Class Relationships

⇔ List<DVHMetric>

Parent Class Relationships

⇦ TreatedPlan

⊕ **DVHMetric**: Metrics provide quick look up of most important values. Sibling relationship to DVH curves is maintained so that they can be reported separately if needed.

- StructureName (String) – Use standard nomenclature from TG263
- MetricName (String) – Use standard nomenclature from TG263
- Value

Sibling Class Relationships

⇔ List<DVHCurve>

Parent Class Relationships

⇦ TreatedPlan

⊕ **ImageDataFeature**: specific values associated with the image that e.g Radiomics values.

Every Image has a list of image data features

- *FeatureName(String) – Need for a standardized list of defined feature names and acceptable values
- Data Type (String): text, number, datetime, bool
- Value (String) ⌘

⊙ DateOfImageDataFeature (Date)
○ AgeAtImageDataFeature (Decimal F3) ⌘

Parent Class Relationships

⇦ Image
⇦ Patient

⊕ **PatientTreatmentOutcome**

- *DiseaseStatus (String) – Need standardized list, e.g., Local Recurrence, NED, BiochemicalFailure

⊙ DateOfStatus (Date)
○ AgeAtStatus (Decimal F3) ⌘

Class Property Relationship

■ DiagnosisAndStaging

Parent Class Relationships

⇦ Patient
⇦ Course

⊕ **PatientReportedOutcome**

- *SurveyInstrumentName (String) – Need for standardized list
- *ElementName (String) – Need for standardized list

⊙ DateOfPRO (Date)
○ AgeAtPRO (Decimal F3) ⌘

- Value (String)
- ValueType (String) – e.g., Bool, Date, Number

Sibling Class Relationship

⇔ Course

Parent Class Relationship

⇦ Patient

⊕ **ProviderReportedToxicity**

- *ToxicityName – Use standard names from CTCAE or other standards
- ToxicityStandard (String), e.g., CTCAE

⊙ DateOfReportedToxicty (Date)
○ AgeAtReportedToxicity(Decimal F3) ⌘

- Value (String)
- ValueType (String) – e.g., Bool, Date, Number
- Attribution (String)

Sibling Class Relationship

⇔ Course

Parent Class Relationship

⇦ Patient

⊕ **HealthInformation**: Used to record data elements relevant to patient status, e.g., smoker, rock climber, diabetes, etc.

- *HealthInformationItemName (String) –Need for standardized list, e.g., HasDiabetes, IsCurrentSmoker, SmokingPackYears

⊙ Date (Date)
○ AgeDate (Decimal F3) ⌘

- Value (String) – e.g., True, 20
- ValueType (String) – Decimal, Bool, Date, String

Sibling Class Relationships

⇔ List<Course>

Parent Class Relationships

⇦ Patient

⊕ **Lab**

- LabName (String)
- LOINCShortName (String)
- LOINCCodeName (String)

⊙ Date (Date)
○ AgeAtDate (Decimal F3) ⌘

- Value (String)
- Units (String)
- ValueType (String) – Decimal, Bool, Date, String

Sibling Class Relationships

⇔ Course

Parent Class Relationships

⇦ Patient

⊕ **Medication**

- MedicationType (String)
- MedicationName (String)
- DosageValue (Decimal)
- DosageUnit (String)
- Frequency (String)

⊙ DateOfMedicationRecord
○ AgeAtMedicationRecord (Decimal F3) ⌘

Sibling Class Relationships

⟺ Course

Parent Class Relationships

⟸ Patient

⊕ **ChemotherapyCourse**: Set of Chemotherapy administrations

- *Protocol (String) – Need standardized list
- Agent (String)
- Facility (String)
- IsNeoAdjuvant (Bool)
- IsConcurrent (Bool)
- IsAdjuvant (Bool)

⊙ DateFirstTreatment (Date)

○ AgeAtFirstTreatment (Decimal F3) ⌘

⊙ DateLastTreatment (Date)

○ AgeAtLastTreatment (Decimal F3) ⌘

Sibling Class Relationships

⟺ Radiation Therapy Course
⟺ Surgical Procedure

Child Class Relationships

⟹ List<Chemotherapy Administration>

Parent Class Relationships

⟸ Patient
⟸ DiagnosisAndStaging

⊕ **ChemotherapyAdministration**

- Agent (String)
- Dosage (String)

⊙ DateOfAdministration (Date)
○ AgeAtAdministration (Decimal F3) ⌘

⊕ **SurgicalProcedure**

- Facility (String)
- *Purpose (String) – Need for standardized list
- *Margins (String) – Need for standardized values
- *BiopsyStatus (String) – Need for standardized values
- Is PreIrradiation (Bool)

⊙ DateOfSurgery (Date)
○ AgeAtSurgery (Decimal F3) ⌘

Sibling Class Relationships

⟺ Radiation Therapy Course
⟺ ChemoTherapy Course

Parent Class Relationships

⟸ Patient
⟸ DiagnosisAndStaging

⊕ **Pathology**

- *ElementName(String) – Need standardized list
- *ElementValue (String)
- *ElementType (String)

⊙ DateOfPathology (Date)
○ AgeAtPathology (Decimal F3) ⌘

Sibling Class Relationships

⟺ DiagnosisAndStaging

Parent Class Relationships

⟸ Patient

⊕ **Charge**

- CPTCode (String)
- NCodeInstances(Int)
- DateStartRange (Date)
  ○ AgeAtStartRange (Decimal F3) ⌘
- DateEndRange (Date)
  ○ AgeAtEndRange (Decimal F3) ⌘

Parent Class Relationships

⟸ Patient
⟸ Course

[a] Author to whom correspondence should be addressed. Electronic mail: cmayo@med.umich.edu

REFERENCES

1. Pan HY, Mazur LM, Martin NE, et al. Radiation oncology health information technology: is it working for or against us? *Int J Radiat Oncol Biol Phys*. 2017;98:259–262.
2. Kerns SL, Ostrer H, Rosenstein BS. Radiogenomics: using genetics to identify cancer patients at risk for development of adverse effects following radiotherapy. *Cancer Discov*. 2014;4:155–165.
3. Stevens EA, Rodriguez CP. Genomic medicine and targeted therapy for solid tumors. *J Surg Oncol*. 2015;111:38–42.

4. Jaffee EM, Dang CV, Agus DB, et al. Future cancer research priorities in the USA: a lancet oncology commission. *Lancet Oncol*. 2017;18: e653–e706.
5. Shirts BH, Salama JS, Aronson SJ, et al. CSER and eMERGE: current and potential state of the display of genetic information in the electronic health record. *J Am Med Inform Assoc*. 2015;22:1231–1242.
6. Parry C, Kent EE, Mariotto AB, et al. Cancer survivors: a booming population. *Cancer Epidemiol Biomarkers Prev*. 2011;20:1996–2005.
7. Oeffinger KC, Argenbright KE, Levitt GA, et al. Models of cancer survivorship health care: moving forward. *Am Soc Clin Oncol Educ Book*. 2014;4:205–213.
8. Witt CM, Herman PM, Tunis S. Comparative effectiveness research in integrative oncology. *JCNI*. 2017;52:Igx013.
9. Fiore LD, Lavori PW. Integrating randomized comparative effectiveness research with patient care. *N Engl J Med*. 2016;74:2152–2158.
10. Kalet AM, Gennari JH, Ford EC, Phillips MH. Bayesian network models for error detection in radiotherapy plans. *Phys Med Biol*. 2015;60:2735–2749.
11. Bentzen SM, Constine LS, Deasy JO, et al. Quantitative analyses of normal tissue effects in the clinic (QUANTEC): an introduction to the scientific issues. *IJORBP*. 2010;76:S3–S9.
12. Mayo CS, Matuszak MM, Schipper MJ, Jolly S, Hayman JA, Ten Haken RK. Big data in designing clinical trials: opportunities and challenges. *Front Oncol*. 2017;7:187.
13. Luo Y, El Naqa I, McShan DL, et al. Unraveling biophysical interactions of radiation pneumonitis in non-small-cell lung cancer via Bayesian network analysis. *Radiother Oncol*. 2017;123:85–92.
14. El Naqa I, Kerns SL, Coates J, et al. Radiogenomics and radiotherapy response modeling. *Phys Med Biol*. 2017;62:R179–R206.
15. Lambin P, van Stiphout RGPM, Starmans MHW, et al. Predicting outcomes in radiation oncology–multifactorial decision support systems. *Nat Rev Clin Oncol*. 2013;10:27–40.
16. Mayo CS, Pisansky TM, Petersen IA, et al. Establishment of practice standards in nomenclature and prescription to enable construction of software and databases for knowledge-based practice review. *Pract Radiat Oncol*. 2016;6:e117–e126.
17. Matuszak M, Anderson C, Lee C, et al. An integrated application for radiation therapy treatment plan directives, management, and reporting (SU-G-TeP4-06). *Med.Phys*. 2016;43:3686.
18. Evans SB, Fraass BA, Berner P, et al. Standardizing dose prescriptions: an ASTRO white paper. *Pract Radiat Oncol*. 2016;6:e369–e381. https://doi.org/10.1016/j.prro.2016.08.007Epub 2016 Aug 24 PubMed PMID: 27693224.
19. https://www.dicomstandard.org/; accessed 1/26/18.
20. http://www.hl7.org; accessed 1/26/18.
21. https://www.hl7.org/fhir/DSTU1/http.html; accessed 1/26/18.
22. https://ncit.nci.nih.gov/ncitbrowser/; accessed 1/26/18.
23. Basch E, Deal AM, Kris MG, et al. Symptom monitoring with patient-reported outcomes during routine cancer treatment: a randomized controlled trial. *J Clin Oncol*. 2016;34:557–565. https://doi.org/10.1200/JCO.2015.63.0830
24. https://www.hhs.gov/ohrp/international/ethical-codes-and-research-standards/index.html; accessed 1/26/18.
25. Mayo CS, Kessler ML, Eisbruch A, et al. The big data effort in radiation oncology: data mining or data farming? *Adv Radiat Oncol*. 2016;1:260–271.
26. Bailis P, Hellerstein JM, Stonebraker M. *Readings in Database Systems*, 5th edn. Burlington, MA: Morgan Kaufmann.http://www.redbook.io/pdf/redbook-5th-edition.pdf
27. Mayo CS, Moran JM, Bosch W, et al. AAPM TG-263 standardizing nomenclatures in radiation oncology. *IJORBP*. 2018;100:1057–1066.
28. http://bioportal.bioontology.org/ontologies/ROO, accessed 2/30/2018
29. https://bioportal.bioontology.org/ontologies/DLORO, accessed 2/30/2018
30. Wang P, Kupelian P, Ruan D, et al. Implementation of a comprehensive radiation therapy registry: focus on feasibility and reliability. *IJORBP*. 2012;83:S664.
31. Robertson SP, Quon H, Kiess AP, et al. A data-mining framework for large scale analysis of dose-outcome relationships in a database of irradiated head and neck cancer patients. *Med Phys*. 2015;42:4329–4337.
32. 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. *IJROBP*. 2017;99:E391–E392.
33. Waddle MR, Kaleem TA, Niazi S, et al. Cost of acute and follow up care in patients with pre-existing psychiatric diagnoses undergoing radiation therapy. *IJROBP*. 2017;99:1231.
34. http://argonautwiki.hl7.org/index.php?title=Main_Page; accessed 1/26/18.
35. Phillips M, Halasz L. Radiation oncology needs to adopt a comprehensive standard for data transfer: the case for HL7 FHIR. *Int J Radiat Oncol Biol Phys*. 2017;99:1073–1075.
36. Skripcak T, Belka C, Bosch W, Baumann M, et al. Creating a data exchange strategy for radiotherapy research: towards federated databases and anonymised public datasets. *Radiother Oncol*. 2014;113:303–309.
37. Nyholm T, Olsson C, Montelius A, et al. A national approach for automated collection of standardized and population-based radiation therapy data in Sweden. *Radiother Oncol*. 2016;119:344–350.
38. Roelofs E, Dekker A, Lambin P, et al. International data-sharing for radiotherapy research: an open-source based infrastructure for multicentric clinical data mining. *Radiother Oncol*. 2014;110:370–374.
39. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR guiding principles for scientific data management and stewardship. *Nat Sci Data*. 2016;3:160018.
40. https://ncip.nci.nih.gov/blog/face-new-tragedy-commons-remedy-better-metadata/; accessed 1/26/2018
41. https://ncip.nci.nih.gov/blog/towards-cancer-research-data-commons/
42. Gruber TR. A translation approach to portable ontology specifications. *Knowl Acquis*. 1993;5:199–220.