Article Type: Special Issue Paper

CORRESPONDING AUTHOR EMAIL: marthamm@med.umich.edu

Performance/Outcomes Data and Physician Process Challenges for Practical Big Data Efforts in Radiation Oncology

Matuszak MM[1], Fuller CD[2], Yock TI[3], Hess CB[3], McNutt T[4], Jolly S[1], Gabriel P[5], Mayo CS[1], Thor M[6], Caissie A[7], Rao A[1], Owen D[1], Smith W[8], Palta J[9], Kapoor R[9], Hayman J[1], Waddle M[10], Rosenstein B[11], Miller R[10], Choi S[2], Moreno A[2], Herman J[2], and Feng M[13]

1) University of Michigan, Ann Arbor, MI, 2)MD Anderson Cancer Center, Houston, TX, 3)Massachusetts General Hospital, Boston, MA, 4)Johns Hopkins University, Baltimore, MD , 5) University of Pennsylvania, Philadelphia, PA, 6) Memorial Sloan Kettering Cancer Center, New York, NY, 7) Dalhousie University, Halifax, Nova Scotia, 8) University of Washington, Seattle, WA, 9) Virginia Commonwealth University, Richmond, VA, 10) Mayo Clinic, Jacksonville, FL, 11) Icahn School of Medicine at Mount Sinai, New York, NY, 12) Mayo Clinic, Phoenix, AZ, 13)University of California at San Francisco, San Francisco CA

Abstract

It is an exciting time for big data efforts in radiation oncology. The use of big data to help aid both outcomes and decision making research is becoming a reality. However, there are true challenges that exist in the space of gathering and utilizing performance and outcomes data. Here, we summarize the current state of big data in radiation oncology with respect to outcomes and discuss some of the efforts and challenges in radiation oncology big data.

Introduction

The promise and potential of "big data" in radiation oncology cannot be overstated. There is tremendous excitement regarding the ability to learn about the efficacy of treatment, discover new interactions, and overall being able to offer our patients improved and tailored treatments based on the experience of many. There is also the hope of shared decision making between providers and patients using informed tradeoffs between cancer control and side effects. However, genuine challenges are to be faced before this can become a reality and to meet those challenges, one must first examine the nature of this "big data." There is a tendency to use the term "data mining" when thinking about informatics, when in fact, data farming is a more accurate term, reflecting the reality that the entire process, from planting the seeds of data in organized rows, watering and tending the growth of data, then harvesting it, is critical to understand and plan for (1).

Our ability to provide patients with answers about their best course of treatment relies on our a priori knowledge of how patients with similar disease, demographics, preference, and clinical characteristics were treated, and how they responded to treatment including both tumor control and treatment-induced toxicities. This data must be captured in a useable way so that it can be extracted and analyzed, with user-friendly predictive models created so that treatment can be customized for each patient.

In radiation oncology, there are two critical general issues, which must be addressed: 1.) Since radiation oncology data is different than medical/surgical oncology data, data platforms which have been designed with this in mind (many of which already exist) must be utilized. 2.) Existing standards where possible should be utilized to meet the big data needs of the multiple stakeholders (current and future patients, physicians, registries, insurance companies, the informatics community and many other groups) in radiation oncology in order to avoid duplication of work. We herein summarize the clinical aspects of big data collection in radiation oncology, and highlight the challenges and future work needed so that we can realize the potential of big data.

Radiation Oncology Big Data is Unique

An essential point that must be embraced for radiation oncology big data to reach its potential is, as mentioned under 1.) above, that its format and nature is inherently different from other disciplines. Fortunately, radiation oncology has recognized this leading to a number of existing specialized data structures in its arsenal, including DICOM-RT structure and dose files. Archiving treatment images, structures and doses in DICOM format is a relatively easy first step toward ensuring that radiation oncology treatment data is captured. It also provides a great step toward future quality assurance of that data. However, some features of treatment are not captured in DICOM format, including, for example, motion management and use of bolus (if not included in the simulation). Recreating delivered dose requires the integration of additional information (e.g. CBCT, log files from the treatment machine) in addition to the treatment plan.

Standardizing nomenclature and definitions are crucial to our efforts to believe and understand aggregated data (2). There is a recognized, but currently unmet need in radiation oncology to standardize naming and delineation procedures of normal structures as well as targets. Standardization includes not only naming structures, but consistency of anatomic borders and instructions on the extent of normal organs to be contoured. For example, naming every esophagus "esophagus" rather than "eso" or "esoph" and contouring it from the cricoid to the stomach is imperative if we hope to better understand dose-volume response-relationships. If every "esophagus" in a big data set must go through independent quality assurance, then the effort will not get very far. This is where planting the seeds correctly in the first place pays off. Even with the best intentions, the complete OAR delineation can be compromised by a treatment planning scan of limited extent, so standard nomenclature, as suggested in TG263, of partial structures is recommended for clarity (2). Another often overlooked element in radiation oncology big data is encoding of spatial information, especially with recurrence. It is essential to know the spatial location of recurrence and its relationship to the delivered dose, not just planned dose. Further, understanding why a marginal recurrence occurred (e.g. variable patient positioning, inadequate GTV/IGTV delineation, poor image registration, inadequate PTV margin) requires analysis of information from many steps of the process. These are examples of data rarely available outside a research study, but essential to determining tumor dose-response relationships.

Use case examples

Radiation oncology has a number of early adopters of the big data paradigm that can help guide the field into best practices for successful capture of patient outcomes data. One well-known example is the euroCAT infrastructure (3). Below are several other examples that were presented or discussed as part of a breakout session at the 2017 Practical Big Data Workshop. In each example, a successful workflow has been implemented to capture outcomes and performance data. The benefits and limitations of each use case are given below. It should be noted that this is a list of examples and not an exhaustive list of all of the excellent big data initiatives that are ongoing in the radiotherapy community. Table 1 attempts to summarized the use cases presented here for quick reference.

M-ROAR – University of Michigan

The University of Michigan has developed the Michigan Radiation Oncology Analytics Resource (M-ROAR) to aid in practice patterns and outcomes analyses in Radiation Oncology. This effort involved a multi-faceted strategy of requiring entry of critical elements as discrete data, building a database platform, which pulls data from the oncology information systems (OIS) and electronic health records (HER), and creating a self-service interface. On the data-entry size, everyone in the clinic made a commitment to entering tumor staging, diagnosis code, pain scores, patient reported outcomes, and Common Terminology Criteria for Adverse Events (CTCAE) scores so that this data would be available for future analysis. Also, structure nomenclature was standardized. The MS SQL database aggregates data for >17,000 patients treated in the department since 2002, including information from both the radiation oncology and hospital information systems. The self-service interface allows users to easily create and optimize reports for cohort discovery in minutes rather than waiting to get to the top of a report-writer's queue with each request or iteration.

With implementation of this strategy, the M-ROAR database can be used to answer innumerable clinical questions, such as what factors predict patient risk of hospitalizations, decline in patient function, and treatment-related complications, so that patient treatment protocols can be adjusted in advance. As an example, for head and neck cancer, the association between radiation dose and toxicity can be stratified based on HPV status. Information to optimize clinical operations can also be gathered, such as: How long does a certain treatment plan take to deliver vs. another one so that therapy time slots can be scheduled properly, and What patients are at risk for dehydration so that nutrition consults can be requested or outpatient hydration appointments scheduled in advance? These are only a few examples of practice-changing queries, which are currently possible. This database is primarily to inform and guide quality improvement, with IRB approval needed when used for research.

Challenges remaining in M-ROAR are consistent and standardized assessment of physician and patient-reported toxicities, as well as recurrence scoring.

MD Anderson

A vision of optimizing electronic health record (HER) utilization is currently being investigated at MD Anderson Cancer Center in a multiphase process. Initiated within the Radiation Oncology department, a thorough evaluation of user performance and available toolsets within EPIC was performed in order to determine suboptimal practices that were limiting efficiency within the clinic workflow. A general consensus of a need for standardized documentation and consistent nomenclature for the purposes of improving quality and safety measures, accurate staging and billing, and decreasing duplication of data entry led to the development of over 40 specialty-specific templates for note generation. These templates “pull in” discrete data elements entered into EPIC by a single person (such as a nurse, midlevel, or primary referral service) so that the need for dictation/manual data entry by other providers generating notes is minimized. The patient’s existing medical conditions, cancer stage, performance status, symptoms/ROS, laboratory values, and radiologic imaging information are all structured fields which are now automatically populated into specific locations within each template. Furthermore, these templates utilize the Smartlist function in EPIC, which are lists of customizable text that can also be retrieved at a later date as structured data. Smartlists have therefore been used to define specialty-specific treatment options, protocol descriptions, and structured CTCAE grading systems. Another advantage of EPIC is the ability for patient-related outcome (PRO) forms to be sent to the patient electronically. When patients fill out these forms, the results are then sent back and saved in EPIC as discrete data, which is then incorporated into templates and allows for more rapid documentation.

Overall, these templates offer additional advantages including increased patient screening for protocol enrollment and user-friendly, electronic functionality for various research endeavors. By having the variables listed above as structured, extractable data, every aspect of clinical research becomes optimized. Patients can be quickly assessed and evaluated for protocol eligibility, and once the patient is undergoing treatment under protocol, the collection and reporting of clinical response and toxicity become more automated. Protocol-specific templates have been created in order to ensure that all required data collection per individual protocol is recorded in a uniform manner. Since completing phases I and II of template creation and implementation within the Radiation Oncology department, there have been ongoing efforts to expand standardized EHR documentation methods within other departments, beginning with GI Medical Oncology and GI Surgery. So far, these services are adapting the templates to maintain a similar data entry structure while tailoring sections such as the impression and plan to suit their documentation needs. Our ultimate goal is to have the entire institution adopt the use of standardized templates and structured data entry to 1) improve the efficiency of documentation for providers and decrease the risk of provider burn-out, 2) improve patient coordination within a multidisciplinary clinic setting, and 3) create an institution-wide system of patient data collection for research purposes and assessment of clinical outcomes.

Pediatric Proton Registry Consortium

The Pediatric Proton Consortium Registry (PPCR) was established in 2012 to expedite proton outcomes research in children and to better define the role of proton radiotherapy in the pediatric cancer population (4). Approximately 1800 pediatric patients have been enrolled in the PPCR across 13 participating pediatric proton centers. The PPCR is a consented registry built upon the NIH supported free web-based data collection/repository platform, REDCap and is currently open to any U.S. proton center that would like to participate. The PPCR collects information on demographics, diagnosis and staging, baseline health status, chemotherapy and surgery, radiation details, diagnostic imaging, and follow-up (5). Radiation plans are centrally archived in the universal DICOM-RT format. Due to funding issues and required manual effort, there is limited participation and variable data entry. Thus, there is an urgent need to improve efficiency of data collection through automation.

The major challenges within the PPCR also present opportunities. Given that there are a limited number of OIS and EHR platforms, there exists an opportunity to leverage the data already contained within these platforms if appropriate programming bridges can be constructed. An upfront investment of time and resources from technical personnel is needed and standard interface should be created with standard basic information mapped from stable locations in each OIS to minimize the need for additional customization at multiple sites.

Another opportunity exists with the general EHR. Given the critical mass of EPIC users in the PPCR, we may be able to leverage collaboration to streamline data input and extraction. A start could be the sharing and use of electronic templates and automation of population of certain (standardized) fields in the database. It is key that templates must be efficient and user-friendly with minimal free text so that clinicians will use them routinely and must be convinced in the overall mission or be given timesaving in another area to counter-balance the extra work of discrete data input.

The final component of PPCR is aggregation of plan information, which is eventually used to help make the link between radiation dose and treatment outcomes. To facilitate this, a partnership has been put in place with MiM Software (MiM Software Inc, Cleveland, OH) to allow web-based archival for each participating institution. The partnership has led to the development of a faster anonymization procedure and a script for automated nomenclature standardization using TG263 (2).

In summary, the PPCR is an established and successful registry that has met some hurdles along the way. As it has grown out of its funding source, it requires that we look into electronic efficiencies that will help PPCR and other Radiation Oncology-related Big Data efforts. Sufficient funding is critical to success of data collection. Mild funding pressure can spur technological advances that can improve efficiencies, but these also need an upfront investment in order to achieve them. Given the relatively few electronic radiation charts and the few EHRs, we are better poised than ever to start to realize the goal of automation in data entry.

Oncospace

The Oncospace program at Johns Hopkins began with the design of a relational analytical database that housed the treatment planning data in a form for fast query. The database schema includes the full 3D dose for multiple radiation therapy sessions as well as the 3D anatomy including relevant structures (5). The system also houses features of the dose such as the dose-volume histograms (DVHs) and shape relationships in the overlap volume histograms (OVHs) (6). In the earlier work, the database was used for the development of shape-based automated treatment planning where one could rapidly query the OVHs to determine all prior treatments with critical organ that were “harder” to plan and use it to predict the best achievable dose metric from DVHs (7-10). This method is in use today for both plan quality evaluation and automated planning.

For outcomes, the Oncospace philosophy was that prospective structured data collection should be integrated with the clinical workflow. Since 2007, a website enabling tablet devices to be used in the clinic for data capture is available (11). Critical to the adoption is the ability to generate clinical notes from the collected structured data and additional patient-related information queried from the OIS. Using the same technology, electronic patient-reported outcomes have been successfully captured for more than 8 years. Currently, there are >5000 patients (prostate, head and neck, thoracic, breast and pancreas) in the database with full treatment planning data, patient reported outcomes, clinician assessments on-treatment and in follow-up, disease response as well as diagnosis, and lab data interfaced from clinical systems. Data are currently included from Johns Hopkins, the University of Washington, the University of Virginia, and the University of Toronto Sunnybrook.

The rapid access to the treatment data enables data science models to be explored (12). The Oncospace group is now building predictive models for specific clinical decisions using classification and regression tree models for weight loss and xerostomia prediction in head and neck cancer and surgical candidacy in pancreatic cancer. The challenge in clinical prediction is to focus on the decision to be made and what information truly informs it. For weight loss, the decision is around the appropriate symptom management for improved nutritional support such as feeding tube placement. In other cases, modifications to the treatment plan may reduce risks if it does not compromise on target coverage. Additionally, the impact of the spatially distributed radiation dose beyond DVHs to better understand how the patterns of dose may impact the treatment related toxicities could be explored (13). The continued data growth will allow continuous learning to fulfill the concept of a learning health system in the future (14).

University of Pennsylvania

The Penn Medicine Oncology Research and Quality Improvement Datamart (ORQID) aggregates data from multiple source information systems, including Penn's enterprise EHR, ROIS, TPS, Cancer Registry, and Center for Personalized Diagnostics. ORQID focuses on organizing cancer patients' demographics, vital status, disease stage and prognostic indicators, genomic variants, details of systemic therapy and external-beam radiotherapy, and physician-reported toxicities.

Outcomes have been among the most challenging data elements to capture. Penn implemented structured, site-specific templates for documenting physician-reported toxicities within the EHR in 2011. The templates are based on the CTCAE grading system, and clinical teams selected the toxicities of focus for each disease site. To maximize opportunities for data capture by providers at all levels, only clinically symptomatic toxicities (e.g. pain) not requiring diagnostic interpretation (e.g. radiation pneumonitis) were included. Nurses have embraced the effort and capture rates have been as high as 95% for on-treatment visits, which they routinely staff. Physician adoption has been more challenging, and for follow-up visits (which have less nursing support) capture rates have been below 50% of visits. Nevertheless, Penn has amassed over 2 million toxicity observations on over 28,000 unique patients in the datamart. Efforts are currently underway to implement widespread patient-reported outcome collection as routine standard of care to help augment and complement the physician-reported toxicities.

For other outcomes, progression is tracked via the institutional cancer registry, which only documents the timing and nature of the first progression event after initial treatment. Deaths are identified from the EHR, cancer registry, and social security death masterfile, but remain a challenge, with many deaths not documented or without accurate dates.

US Veterans Health Administration (VHA) Radiation Oncology Practice Assessment

The National Radiation Oncology Program (NROP) office of VHA, with an oversight of 40 radiation therapy treatment centers treating over 15,000 patients annually has launched a pilot program initiative in which patient-specific radiotherapy data is collected for quality assurance assessment and comparative analysis of many treatment modalities and other factors at their centers (15). The NROP office collaborated with the American Society of Radiation Oncology (ASTRO) disease site expert committees to define clinical measures. These clinical measures are based on established clinical guidelines, patterns of care assessment done by the American College of Radiology's Quality Research in Radiation Oncology program (16), and expert consensus opinions. These measures have formed the basis for assessing the quality of treatments and practice variations and identification of the care gaps in the VHA. Although dosimetry data was automatically abstracted from treatment planning systems (TPS), clinical data had to be manually abstracted from the electronic health records (EHR) for the pilot project.

The NROP office has embarked on a project to automatically extract all data for ROPA from heterogeneous data sources that include EHR, TPS and Treatment Management Systems (TMS) for clinical practice assessment, outcomes, and prospective decision support analytics. An integrated data curation, storage and analytics portal, titled as HINGE (Health Information Gateway and Exchange), was built that can extract and aggregate data from TPS and TMS, physician clinical notes and DICOM-RT files. HINGE integrates data from these disparate sources coherently and standardizes it for quality assessment and predictive analytics. The HINGE database is based on well-defined quality measures defined by radiation oncology disease site experts. HINGE has (i) tools to aggregate data from physician note templates (ii) a built-in DICOM-RT parser to extract DVH based dose constraints, (iii) a natural language processing (NLP) module to extract relevant physician assessments from the clinician notes, and (iii) a decision-support and genomics module to provide supplementary insight to treatment predictions, treatment outcomes and research hypotheses. The HINGE application would reside at each VHA radiation oncology treatment site and transmit information to a centralized database server thus making big data analytics possible. HINGE is capable of seamlessly connecting to local IT/medical infrastructure via network and performs data extraction and aggregation. The built-in modules (TMS extraction, DICOM parser, NLP) extract defined clinical data and are easily extendable. The modules of decision-support and genomics provide preliminary insights into a patient's treatment and health profile. Automatic data abstraction with HINGE will enable real time assessment of clinical practices and determine care gaps.

Mayo Clinic Florida

The Mayo Clinic Florida Department of Radiation Oncology has leveraged Mayo Clinic’s unique cost warehouse to aggregate data on the cost of radiation therapy and other associated healthcare costs in the first two years after radiotherapy on approximately 3,000 patients over a five year period incurred. The Mayo cost data warehouse is a unique resources consisting of linked EMR data and administrative data from Mayo Clinic’s hospital and clinics in Florida, Minnesota, and Wisconsin (17). These costs were linked to other sources of institutional data, such as departmental treatment records captured through its radiation oncology information system, demographic, tumor specific, and outcomes data obtained through Mayo’s tumor registry, adverse events recorded in the EMR, and other disease specific registries containing non-oncological diagnosis data, such as psychiatric comorbidities. Waddle *et al* have used this cost warehouse to demonstrate that patients with co-existing psychiatric morbidities utilize the emergency department and inpatient hospitalization at rates greater than patients without psychiatric co-morbidities at 6 months and two years after radiotherapy. (18) It should be noted that even with many successes, toxicity capture remains challenging.

The Radiogenomics Consortium (RGC)

The hypothesis that genetic/genomic alterations may function as surrogate biomarkers of disease response or normal tissue toxicity represents the basis of the field of radiogenomics (19). A principal goal of research in the field of radiogenomics is to identify the genomic markers associated with the development of adverse outcomes resulting from cancer radiotherapy. However, in order to accomplish this goal and definitively discover and validate the critical genomic markers, access to the radiotherapy treatment information and long-term longitudinal follow-up data reporting details as to adverse outcomes must be obtained for large numbers of patients. In order to enable the creation of large cohorts of patients who received radiotherapy, the Radiogenomics Consortium (RGC) was created in 2009, which is a cancer epidemiology consortium through the Epidemiology and Genomics Research Program of the NCI of the NIH (20). The RGC now has 225 investigators at 132 institutions in 31 countries. Although the RGC has successfully assembled large cohorts to perform adequately-powered studies, data harmonization remains a problem when multiple cohorts involve patients treated with a variety of radiotherapy techniques and evaluated using multiple grading systems. Nevertheless, a number of large studies have been accomplished in which substantial amounts of radiotherapy data have been gathered for studies that typically comprise over a thousand patients.

Four large studies involving the use of Big Data are currently in progress whose main goal is to discover new SNPs and validate previously identified genetic biomarkers predictive of susceptibility for the development of adverse effects resulting from radiotherapy. The first project involves roughly 6,000 men treated for prostate cancer, which encompasses multiple cohorts created by RGC investigators. DNA samples from all of these men have been genotyped and detailed clinical data are available with a minimum of two-years of follow-up.

The second large multi-center study developed by RGC members is REQUITE (Validation of predictive models and biomarkers of radiotherapy toxicity to reduce side-effects and improve quality-of-life in cancer survivors)(21). REQUITE addresses the challenge of data heterogeneity that, as for other big data projects, requires harmonization of the different outcome measures and confounding variables used in multiple cohorts. This study does not stipulate the radiotherapy protocols to be used but involves standardized case report forms across centers and countries to ensure data in identical categories are collected. A key aspect of REQUITE is the centralized database that includes pre-treatment DICOM and DVH files.

A third study involves three large cohorts comprising roughly 4,500 breast cancer patients treated with radiotherapy for which blood samples and detailed clinical information are available. These samples and data are available from three large groups of patients: (1) 1,500 patients treated under a series of breast cancer clinical protocols performed at New York University School of Medicine (22-25); (2) ~2,000 breast cancer patients enrolled though the REQUITE study and (3) ~1,000 women who receive breast cancer treatment through participation in RTOG 1005 (26).

The fourth effort being made is to create a biorepository with linked clinical data for patients treated with charged particle therapy (CPT). With the increasing use of CPT, there is a need to establish cohorts for patients treated with these advanced technology forms of radiotherapy. In recognition that the formation of patient cohorts treated with CPT for radiogenomic studies is a high priority, efforts are underway to establish collaborations involving institutions treating cancer patients with protons and/or carbon ions as well as consortia, including the Proton Collaborative Group, the Particle Therapy Cooperative Group and the Pediatric Proton Consortium Registry.

State of the data

As noted by the varied workflows highlighted in the use cases, hospital-wide and radiation oncology-specific EHR systems are not often designed to facilitate collection of key data elements for subsequent extraction and use. Typically, when a patient is referred to radiation oncology, the diagnosis for that patient has been entered to the hospital EHR system. Most radiation oncology-specific EHRs can link to the hospital EHR via HL7 FHIR (27) to sync the diagnosis information. However, linking the specific diagnosis relevant to a given treatment plan is often a manual process requiring physician input. In addition, there is generally not a mechanism to input the staging information into the radiation oncology EHR or link metastatic sites to the original diagnosis, which are in general of interest for outcome analyses. Thus, curation of the diagnosis and staging information that comes into radiation oncology can be cumbersome. Apart from simple diagnosis information, data elements from pathology, radiology, surgery, internal medicine and medical oncology that may be relevant for radiation oncology outcomes are seldom entered in discrete fields or even templated free-text formats, and are, therefore, often inaccessible for automatic extraction and use.

As the patient goes through treatment, physicians typically see the patient weekly for on treatment visits. However, the documentation of these visits, including routine toxicity assessments relies on each individual institution creating their own clinical practice, datasheets and custom tools for reporting. While many institutions are beginning to recognize the importance of standardized toxicity assessments and PROs and are putting mechanisms in place to track this data, there is still inconsistency, which can lead to missing data. Further, once institutions have these tools in place, it can be challenging to share personalized templates across the varying platforms and clinical workflows that exist at different institutions. Adding this to the lack of standardized key data elements and time points to track for different treatment sites, multi-institutional datasets are rarely comprehensive.

While some existing standards can be leveraged, it is important to evaluate if these standards take into account the needs of all stakeholders and if not, determine if new standards or perhaps simply minor amendments can be suggested to minimize the need to start at the ground up. One must recognize that efforts to standardize common data elements is a complex and time-consuming endeavor, but one that is ultimately worthwhile. An excellent published discussion and proposed set of standard patient-reported outcomes within oncology shows the complexity of these issues (28).

Once collected, Big Data will perform a crucial role by providing accurate outcome data in order to build clinical decision support systems (CDSS) (29). Conversely, decision models themselves can be used to guide the selection of data elements to include. In a recent work, for example, a decision cost-model in the form of an influence diagram was constructed to model the choice between photons and protons for the treatment of locally advanced non-small cell lung cancer (30). By including the monetary cost of managing acute toxicities, it was possible to determine the ROC characteristics of a biomarker for radiosensitivity that a physician would need in order to select patients for proton radiotherapy when their total expected cost for protons is below that of photons. As this cost-model example illustrates, models can guide data farming efforts by establishing outcomes that are important for clinical decision making, and by placing requirements on how accurately these outcomes need to be known. In this case, the required sensitivity and specificity were established for a novel test for radiosensitivity for the decision to lower treatment costs. This use of models may be especially important when resources (e.g. cost of human labor) for populating databases are limited, allowing efforts to be directed towards collecting the data that is most likely to lead to improved clinical decision making.

This in turn highlights an important issue in constructing data standards for capturing outcome data, namely, the standards need to be easily expandable. As big data results are applied in the clinic, used for clinical decision support, or new interactions are discovered within the data, these efforts will inevitably – and rapidly – call for the collection of different types of data. Adaptability is emerging as a feature of data and communication standards throughout healthcare, as recognition grows that developing a standard which attempts to include everything will fail to do so, and in the process will become unwieldly. HL7 FHIR, for example, is a communication standard which follows an 80/20 directive, whereby 80% of the elements which are implemented are included in the specification itself (31). These core elements are referred to as resources, and the remaining elements, called profiles, are definable by individual institutions or groups in order to alter or add properties to resources. Single institution databases can attempt to cover a greater proportion than 80%, although the principle remains. By embedding adaptability within a database initially intended to capture, for example, only traditional treatment planning data, the database may later be populated with patient reported outcomes, "omics" data, or patient preferences in the form of utilities, rendering it useful in significantly more applications.

Collection and Curation

In order for the promise of big data to be realized in more than just individual radiation oncology departments or networks of systems, standardized key data element lists and input schemas are required. For example, the connection of diagnosis information to treatment courses should be automated within vended systems and reviewed for quality on an ongoing basis as part of a routine workflow, such as chart rounds. In addition, the relevant staging, pathology, and histology information should be automatically extracted from the EHRs into appropriate fields within the radiation oncology information system. Free-text searches or simple natural language processing will be necessary for scanned outside hospital reports and for other information not entered in discrete fields for easy extraction, particularly for information not generated in radiation oncology and thus beyond our immediate control.

Collection of standardized key data elements related to toxicity, disease status, and patient reported outcomes requires the definition of standards, as discussed above. However, even with standard elements and data entry tools, there must be a culture shift in the radiation oncology community to recognize the importance of comprehensive entry of the data as part of the standard care for each patient. It is our responsibility to the field and future patients to make collection of key data elements related to outcomes a priority.

Access and Extraction

Accessibility and extraction of the clinical data entered by the physician and patients, in the case of patient-reported outcomes, is essential. The data storage infrastructure must provide a mechanism for end users to extract the key data elements and aggregate the data with other related data, such as dosimetric information. The system should be designed with accessible application programming interfaces enabling user data extraction in the most suitable and meaningful way. However, data extraction should not be performed on a project-by-project basis. Rather, institutional information technology groups, especially those housed in radiation oncology, should make it a priority and be proactive in supporting the construction of big data analytics resource systems (BDARS). This may require a partnership between radiation oncology users and the IT managers so that domain knowledge can be shared and the BDARS designed in such a way that the information is in a complete and usable format. The development and use of a radiation oncology-specific ontology will be a key development in ensuring that individual BDARS can be combined into true sets of big data.

Specific Recommendations for Standardizations

While there is clear work ahead in the community to reach a point where standard key data elements are recorded routinely for all patients in radiation oncology, there are first steps that can be taken. Summarized in Table 2 are example standard key data elements that could be collected and thus should begin to be supported by vended systems. Note that many such elements would be collected at various timepoints including baseline, during treatment, end of treatment, and at follow-up. Therefore, properly capturing dates and being consistent with relative dates is essential.

While Table 1 serves as a starting point for standardization of requested data elements, collection of the data requires:

1. Creation of a standardized workflow that enables collection of proper data, at the right time for the right patient.
2. Initiation of a working group to develop standards for classifying recurrence in radiation oncology that includes spatial and dose information.

Recommendations for Next Steps Needed to Improve Data Availability

The current climate is such that "big data" is becoming a known term and fills one with the promise of solving mysteries of care with a lot of data and computer. There is a focus on data mining, as if the data is sitting waiting to be taken and analyzed. However, it is clear that the data must be created and structured in a way to make it possible to harvest and answer important and relevant clinical questions. As more providers buy into the need to standardize for the sake of quality and process improvement, they will become more committed to inputting essential common data elements related to outcomes. Vendors must also allow the data to be accessed in a variety of ways, maintaining HIPAA compliance but no longer being a major barrier to quality assurance. Improved automation in both capturing and accessing data within vended systems is recommended to improve efficiency and accuracy in capturing outcomes data. Engagement with all stakeholders, including physicians, legislators, patients and patient advocates is essential to design modern approaches to handling protected health information and drafting policies and legislation regarding how health care data can be used in a safe way so as to maximize healthcare value and efficiency while maintaining security.

References

1. Mayo CS, Kessler ML, Eisbruch A, Weyburne G, Feng M, Hayman JA, Jolly S, El Naqa I, Moran JM, Matuszak MM, Anderson CJ, Holevinski LP, McShan DL, Merkel SM, Machnak SL, Lawrence TS, Ten Haken RK. The big data effort in radiation oncology: Data mining or data farming? Adv Radiat Oncol. 2016 Oct 13;1(4):260-271.
2. Mayo, C. et al. AAPM Task Group 263: Tackling Standardization of Nomenclature for Radiation Therapy. International Journal of Radiation Oncology • Biology • Physics , Volume 93 , Issue 3 , E383 - E384
3. Deist, T. M., et al. "Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT." Clin Transl Radiat Oncol 4: 24-31. (2017)

4. Kasper HB, et al. The pediatric proton consortium registry: A multi-institutional collaboration in u.S. Proton centers. International Journal of Particle Therapy 2014;1:323-333

5. McNutt T, Wong J, Purdy J, Valicenti R, DeWeese T. OncoSpace: A New Paradigm for Clinical Research and Decision Support in Radiation Oncology Proceedings of the XVIth International Conference on the Use of Computers in Radiation Therapy, Amsterdam, 2010, Editor Jan-Jakob Sonke, Published by Het Nederlands Kanker Instituut - Antoni van Leeuwenhoek Ziekenhuis ISBN: 978-90-75575-29-3

6. Michael Kazhdan, Patricio Simari, Todd McNutt, Binbin Wu, Robert Jacques, Ming Chuang, and Russell Taylor, "A Shape Relationship Descriptor for RadiationTherapy Planning" Medical Image Computing and Computer-Assisted Intervention 5762/2009(12), 100–108 (2009)

7. Binbin Wu, Francesco Ricchetti, Giuseppe Sanguineti, Misha Kazhdan, Patricio Simari, Ming Chuang, Russell Taylor, Robert Jacques, Todd McNutt, "Patient Geometry-Driven Information Retrieval for IMRT Treatment Plan Quality Control", Medical Physics, 2009 Dec;36(12):5497-505

8. Wu, B., Ricchetti, F., Sanguineti, G., Kazhdan, M., Simari, P., Jacques, R., Taylor, R., McNutt, T.: "Data-driven approach to generating achievable dose-volume histogram objectives in intensity modulated radiotherapy planning". International Journal of Radiation Oncology, Biology, Physics 2011 Mar 15;79(4):1241-7. Epub 2010 Aug

9. Steven F. Petit, Binbin Wu, Michael Kazhdan , André Dekker, Patricio Simari, Rachit Kumar, Russel Taylor, Joseph M. Herman, Todd McNutt," Increased organ sparing using shape-based treatment plan optimization for intensity modulated radiation therapy of pancreatic adenocarcinoma", Radiotherapy and Oncology, 102 (2012) 38–44.

10. Binbin Wu, Todd McNutt, Marianna Zahurak, Patricio Simari, Dalong Pang, Russell Taylor, Giuseppe Sanguineti, "Fully Automated Simultaneous Integrated Boosted-Intensity Modulated Radiation Therapy Treatment Planning Is Feasible for Head-and-Neck Cancer: A Prospective Clinical Study". International journal of radiation oncology, biology, physics 2012 Dec 1;84(5):e647-53

11. W Y Yang, J Moore, H Quon, K Evans, A Sharabi, J Herman, A Hacker-Prietz and T McNutt, “Browser Based Platform in Maintaining Clinical Activities – Use of The iPads in Head and Neck Clinics,” Phys.: Conf. Ser. 489 012095 doi:10.1088/1742-6596/489/1/012095 Int’t Confernce on Computers in Radiotherapy, Melbourne AUS 2013

12. Robertson S, Quon H, Kiess, A., Moore J, Yang W.,Cheng Z, Afonso S., Allen M., Richardson M., Choflet A., Sharabi A., McNutt T, “A Data-Mining Framework for Large Scale Analysis of Dose-Outcome Relationships in a Database of Irradiated Head and Neck (HN) Cancer Patients”, Medical Physics 42(7), 4329 (7/2015)

13. Marungo F, Robertson S, Quon H, Rhee J, Paisley H, Taylor R, McNutt T, “Creating a Data Science Platform for Developing Complication Risk Models for Personalized Treatment Planning in Radiation Oncology.” 48th Hawaii International Conference on System Sciences (HICSS). Kauai, HI USA: IEEE; 2015

14. McNutt T., Moore K., Quon H. “Needs and Challenges for Big Data in Radiation Oncology,” Int’l J. of Radiation Oncology, Biology, Physics. Published online: November 27 2015. Print . July 1, 2016 Volume 95, Issue 3, Pages 909–915

15. D Caruthers, S Brame, JR Palta, MP Hagan, E Wilson, C Cowan, L Yun, S Brown, LM DeBerry, S Mutic, et al. Development and implementation of quality measures for the survey based performance assessment of radiation therapy in the VHA. International Journal of Radiation Oncology, Biology, and Physics, 99(2): E391-E392, 2017.

16. Owen, J., White, J., Zelefsky, M. and Wilson, J. (2009). Using QRRO™ Survey Data to Assess Compliance with Quality Indicators for Breast and Prostate Cancer. Journal of the American College of Radiology, 6(6), pp.442-447

17. Visscher SL, Naessens JM, Yawn BP, Reinalda MS, Anderson SS, Borah BJ. Developing a standardized healthcare cost data warehouse BMC Health Serv Res. 2017 Jun 12;17(1):396. doi: 10.1186/s12913-017-2327-8.

18. M.R. Waddle, T. Kaleem, S.K. Niazi, L.J. White, J.M. Naessens, T.A. Rummans, D.I. Aljabri, J.Y. Habboush, R.C. Miller. Cost of Acute and Follow-Up Care in Patients with Pre-Existing Psychiatric Diagnoses Undergoing Radiation Therapy. IJROBP 2017, 99(5): 1321.

19. Rosenstein, B.S. (2017) Radiogenomics: Identification of Genomic Predictors for Radiation Toxicity. Seminars in Radiation Oncology 2017 Oct;27(4):300-309.

20. West, C. and B.S. Rosenstein (2010) Establishment of a radiogenomics consortium. Radiotherapy and Oncology 94(1):117-118.

21. West, C., D. Azria D et al. (2014) The REQUITE project: validating predictive models and biomarkers of radiotherapy toxicity to reduce side-effects and improve quality of life in cancer survivors. Clin. Oncol. (R. Coll. Radiol.) 26(12):739-742.

22. Constantine C., P. Parhar et al. (2008) Feasibility of accelerated whole-breast radiation in the treatment of patients with ductal carcinoma in situ of the breast. Clinical Breast Cancer 8(3):269-274.

23. Formenti S.C., D. Gidea-Addeo et al. (2007) Phase I-II trial of prone accelerated intensity modulated radiation therapy to the breast to optimally spare normal tissue. Journal of Clinical Oncology 25(16):2236-2242

24. Freedman, G.M., J.R. White et al. (2013) Accelerated fractionation with a concurrent boost for early stage breast cancer. Radiotherapy and Oncology 106(1):15-20.

25. Raza S., S.C. Lymberis et al. (2012) Comparison of acute and late toxicity of two regimens of 3- and 5-Week concomitant boost prone IMRT to standard 6-week breast radiotherapy. Frontiers in Oncology 2:44.

26. Cooper, B.T., S.C. Formenti et al. (2016) Prospective randomized trial of prone accelerated intensity modulated breast radiation therapy with a daily versus weekly boost to the tumor bed. International journal of radiation oncology, biology, physics. 95(2):571-578.

27. https://www.hl7.org/fhir/

28. Reeve BB, Mitchell SA, Dueck AC, et al. Recommended patient-reported core set of symptoms to measure in adult cancer treatment trials. J Natl Cancer Inst 2014;106

29. Gaebel, Jan, Mario A. Cypko, and Heinz U. Lemke. "Accessing patient information for probabilistic patient models using existing standards." Proceedings of the 10th eHealth2016 Conference. Vol. 223. 2016.

30. Smith, W.P., Richard, P.J., Zeng, J., Apisarnthanarax, S., Rengan, R., Phillips, M.P., "Decision Analytic Modeling for the Economic Analysis of Proton Radiotherapy for NSCLC." Translational Lung Cancer Research, in press.

31. https://www.hl7.org/fhir/overview-arch.html.

Table 1. Examples of Big Data Use Cases in Radiation Oncology

Institution/Entity	Type of Database/Project	Source of Data/Tools	Magnitude	Key Features	Key Challenges
M-ROAR/University of Michigan	tumor staging, diagnosis code, pain scores, patient reported outcomes, and CTCAE scores	Oncology Information Systems, Treatment Planning System, and Electronic Health Record	>17,000 Patients since 2002	Microsoft SQL Database; Self-service report building interface	Consistent/standardized physician and patient reported toxicities and reccurrence scoring
MD Anderson	Creation of Radiation Oncology Site Specific Templates for Data Input	Electronic Health Record (EPIC)	>40 specialty specific templates in Radiation Oncology with expansion into other departments	Specialty specific templates for standardized note generation	High level of customization in each site and department limits standardization in some elements
Pediatric Proton Registry Consortium	demographics, diagnosis and staging, baseline health status, chemotherapy and surgery, radiation details, diagnostic imaging, and follow-up	Oncology Information Systems, Treatment Planning Systems, and Electronic Health Record	>1800 patients from at least 13 centers	RedCap Tools; Collection of DICOM plan data	Funding; Data input efficiency
Oncospace	treatment planning data, patient reported outcomes, clinician assessments, disease response, diagnosis, and lab data	Oncology Information Systems, Treatment Planning System, and Electronic Health Record	>5000 patients from 4 centers	Tablet and web based data capture; Generation of notes from structured data entry;	Multi-institutional data standardization; Funding for maintenance and expansion
University of Pennsylvania	demographics, vital status, disease stage and prognostic indicators, genomic variants, details of systemic therapy and external-beam radiotherapy, and physician-reported toxicities	Oncology Information Systems, Electronic Health Record, Treatment Planning System, Cancer Registry, and Center for Personalized Diagnostics	>28,000 patients	Structure, site-specific templates; Only capture clinically symptomatic toxicities; Strong adoption by nurses	Physician adoption; Gathering of detailed progression information; Accurate identification of death events

US Veterans Health Administration (VHA) Radiation Oncology Practice Assessment	clinical measures, treatment planning information	Oncology Information Systems, Electronic Health Record, Treatment Planning System	Development is being finalized	novel tools to extra data including note processing; secure environment where data is housed locally	Development of custom tools to minimize manual data entry and support heterogeneous data sources
Mayo Clinic Florida	institutional data, demographics, tumor specific data, outcomes data, adverse events recorded in the EMR, and non-oncological diagnosis data	Electronic health record, administrative data, oncology information system,tumor registry, other disease specific registries	>3,000 patients	Includes administrative component with healthcare cost data capture	Toxicity reporting and data capture
The Radiogenomics Consortium	genomic data, treatment data, toxicity and outcomes data	Electronic health record, treatment planning systems	132 institutions; > 6000 prostate patients and >4500 breast patients in specific projects	combined captured of genomic and treatment data	Data harmonization across different techniques and reporting methods

Table 2. Example Key Data Elements for Radiation Oncology

Key Data Element Category	**Diagnosis = breast cancer**	**Diagnosis = lung cancer**	**Diagnosis = bone met**
ICD-10 code	All, including laterality info	All, including laterality info	All, including location(s)
TNM staging	TNM staging	TNM staging	N/A
Performance Status	KPS	KPS	KPS
Toxicity Data Elements with CTCAE grade	Dermatitis	Dermatitis	Dermatitis
	Pain	Pain	Pain
		Esophagitis	
		Pneumonitis	

Recurrence Data Elements	Local recurrence	Local recurrence	Local recurrence
	Regional recurrence	Regional recurrence	
	Distant recurrence	Distant recurrence	Distant recurrence
Generic Data Element **{name=____, description=____}**	Custom	Custom	Custom