Practical data collection and extraction for big data applications in radiotherapy

Todd R. McNutt,[a)] Michael Bowers, Zhi Cheng, and Peijin Han
School of Medicine, Radiation Oncology, Johns Hopkins University, Baltimore, MD 21231, USA

Xuan Hui
Epidemiology, University of Chicago, Chicago, IL 60637, USA

Joseph Moore
School of Medicine, Radiation Oncology, Johns Hopkins University, Baltimore, MD 21231, USA

Scott Robertson
Radiation Oncology, Wellspan York Hospital, York, PA 17403, USA

Charles Mayo
Radiation Oncology, University of Michigan, Ann Arbor, MI 48109, USA

Ranh Voong and Harry Quon
School of Medicine, Radiation Oncology, Johns Hopkins University, Baltimore, MD 21231, USA

(Received 9 November 2017; revised 15 January 2018; accepted for publication 25 January 2018; published 24 August 2018)

The capture of high-quality treatment data and outcomes is necessary in order to learn from our clinical experiences with big data analytics. In radiotherapy, there are several practical challenges to overcome. Practical aspects of data collection are discussed pointing to a need for a culture change in clinical practice to one that captures structured patient-related data in routine care in a prospective manner. Radiation dosimetry and the contoured anatomy must also be captured routinely to represent the best estimate of delivered radiation. The quality and integrity present in the data are critical which poses opportunities to introduce electronic validity checking to improve them. Similarly, data completeness and methods and technology to improve the efficiency and sufficiency of data capture can be introduced. In the manuscript, the types of clinical data are discussed including patient reports, images, biospecimens, treatments, and symptom management. With a data-driven culture, the realization of a learning health system is possible unlocking the potential of big data and its influence on clinical decision-making and hypothesis generation. © *2018 American Association of Physicists in Medicine* [https://doi.org/10.1002/mp.12817]

Key words: big data, decision support, learning health system, machine learning

1. INTRODUCTION

The practice of modern oncology is complex and multifaceted. It requires the coordination of a multidisciplinary group of providers to help each individual patient choose and complete an optimized course of treatment. Furthermore, treatment regimens span weeks or months, requiring frequent interaction with health-care providers. The patient is then faced with months to years of follow-up appointments to assess tumor control and quality of life. Throughout such a long care path, there are many time points at which patient data must be collected to inform physician decisions. Ideally, the collection of encounters, treatments, and outcomes within a learning health system should provide a mechanism for feedback and evaluation of the impact of any treatment decisions.

Medicine is currently not practiced in a way that supports a learning health system.[1,2] The transcribed medical records serve three primary goals. The most important is to convey the status and well-being of the patient to other care providers that may interact with the patient in the future. Secondly, it is important to have an accurate record of what procedures were performed and possible complications that might have arisen for legal and billing requirements. Last but not least, it is the primary source for the evaluation of care quality or retrospective research questions.

Medical research, on the other hand, is evaluated and presented in a quantitative way helping us understand how to progress our field of medicine. User interfaces that incorporate data collection into the clinic workflow can bring that research-level analysis to the point of patient care and will catalyze the implementation of learning health systems in the future. These interfaces must be more directed to the physicians and nurses that use them. They must also enable complete and efficient structured data collection while maintaining clinician presence with patients. Such user interfaces can align with specific encounter types and employ displays that trend the state of the patient while enabling updates and integrity checks with minimal interaction.

In radiation oncology, our treatment is four-dimensional in nature and we know quite accurately where it is distributed in the patient and when it is delivered. Extracting and

organizing the dose distribution, the related anatomy and imaging information in a form that is easily accessed and processed facilitates the rather complex analysis utilized in machine learning approaches.[3–5] Coupling the treatment information with the patient outcomes and complications' measures provides a rich environment to learn from our clinical data.

For big data applications, we are limited to the types of data that are captured or able to be captured in the standard clinical practice.[6] The goal is to best capture these data with the highest integrity, greatest efficiency, and high completion rates.

2. TYPES OF CLINICAL DATA AND HOW THEY ARE CAPTURED

2.A. Clinician assessments

Conventionally, clinical assessments are documented in free-text form of clinical notes, sometimes cursory due to the nature of the clinic. These might be sufficient for the healthcare providers when they only need to interact with the patients for a limited number of encounters. However, this is not always the case, especially for cancer patients. During physicians' regular course of management, unstructured documentation may cause unwanted transcription errors, limited information capture, physician recall bias, and difficulty in retrospective chart review. All of these could potentially lead to severe consequences such as clinical inertia.[7]

There is interest in using in natural language processing of unstructured notes, which has limited success.[8,9] These methods fall short when the information simply is not in the note. Clinical notes highlight the critical aspects of the patient and fail to identify the absence of a toxicity or complication or the presence of toxicity at a level not requiring medical intervention. Practices that have moved to prospective collection of structured clinician assessments using electronic forms that can be used to generate clinical documentation will undoubtedly have more control over what is captured with higher compliance and completion rates. These methods have to be carefully adopted and tailored to the specific type of patient and visit. Improving human–computer interfaces to facilitate this practice will assist in changing the culture to include the structured capture of clinician assessments.

In clinics that have adopted structured forms, clinicians assess the patients in consult, during on-treatment visits, and in follow-up to manage the patient's symptoms and monitor the disease control. During these patient–physician interactions, a new form is constructed at the first visit with patients' general information and clinical conditions; thereafter, an integrated data review is presented during each on-treatment and follow-up visits for validation and modification; multidisciplinary assessments are also attached with the assessment date and time that are securely saved in the medical record. This entire process, depicted in Fig. 1, seamlessly streamlines the clinic with an electronic tablet, which overcomes the interruptive nature of the clinic, substantially improving efficiency.

2.B. Patient reported

Patient-reported outcomes (PROs) are collected by providing patients with questionnaires and other structured instruments that assess everything from a patient's ability to cope with their disease and symptoms to their quality of life. These seek to quantify the extent of any complication and its impact on the patient from the patient's perspective. Many of these instruments have been validated in the literature and provide individual questions and scoring models for the various measures of interest.[10–12] Electronic medical records (EMRs) are improving the ability to capture patient-reported outcomes on mobile devices in the clinical setting and through patient portal's in the home or mobile setting.[13] Key among these efforts is the validity of the construction of electronic PROs compared to their original paper-based versions. Work to date suggests that minimal changes in the construction of the equivalent electronic PRO do not invalidate the instrument.[14]

Additionally, health-monitoring mobile devices are becoming more prevalent and can track activity and a myriad of health conditions. Interfaces for these devices to enable physician access to the data through the EMR are in their infancy. Ultimately, patient-reported data can be the most longitudinally complete data, as the patient is continually present throughout their own experience. However, it is important to recognize that incorporating these "out-of-clinic" strategies needs to similarly consider the equivalent of "in clinic" data workflow collection issues by considering the normal routines of the patient.

2.C. Biospecimen

Laboratory data are well structured and are currently transferred electronically between electronic medical systems through the Health Level Seven (HL7) standard.[15] HL7, however, does not define naming standards, so incoming data from different laboratory sources are likely to have different naming. Although these data are well suited for analysis, there remain difficulties in adherence to standard nomenclature[16] and units such as those defined by LOINC and SNOWMED creating translational problems when data are aggregated.

Pathology reports are often unstructured though there are evolving standards and a strong trend toward using more structured reporting. In fact, the College of American Pathologists have provided cancer-reporting templates and have also developed a list of specific features that define a synoptic reporting format that lends itself to data curation. Pathology reports standardized information such as AJCC tumor staging[17] and the International Classification of Diseases for Oncology (ICD-O) coding, margin-positive/negative status, lymph node involvement, and disease-specific coding such as the Gleason grading system. These are standardized and may also be processed with natural language processing with some level of success.

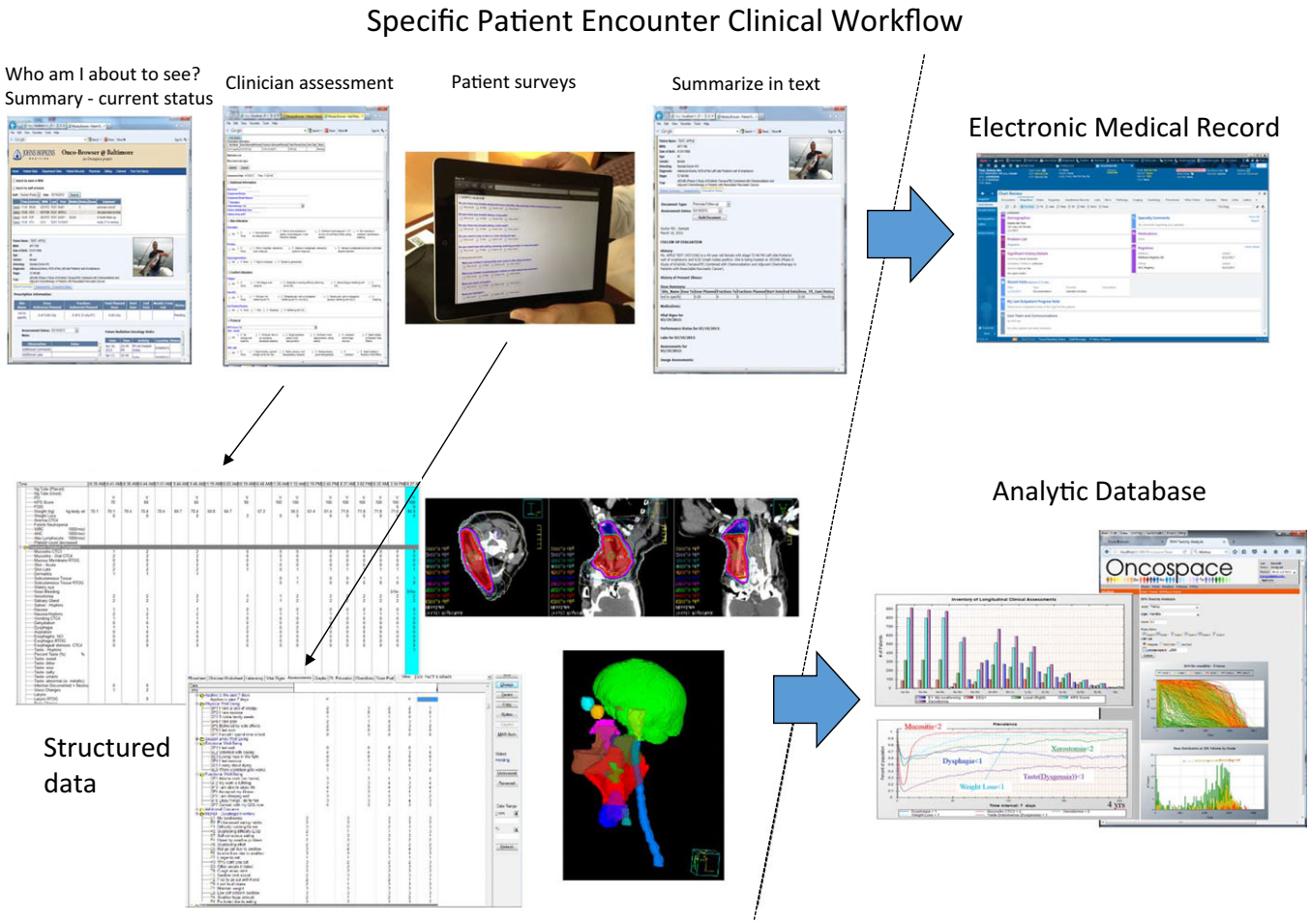

FIG. 1. A depiction of a clinical workflow using web forms that fulfills both clinical documentation requirements while also capturing structured data suitable for analytics. The survey forms allow the clinician and patient to fill out questionnaires specific to a type of patient and encounter. The results can be combined with information about the patient already in the system to create (compute) a text-based summary of findings that can be transferred to the Electronic Medical Record where any additional physician impressions may be added. The captured data can be presented as trends for that individual and also added to an analytic database for aggregation.

2.D. Image-derived features

By definition, medical images are a form of structured data, containing both metadata describing the image and its acquisition as well as data arrays containing the pixel/voxel data. Algorithmic assessments both within defined regions and of the image globally can generate additional metadata quantifying the image in terms of countless different features. For radiation oncology, the routine workflow, whereby the cancer and surrounding normal structures are routinely contoured/segmented, lends itself to algorithmic assessments with radiomics. At this point, image datasets are fully electronic at most institutions and are stored in PACS systems. The main issue now is in the extraction of features from the images. Simple features such as tumor dimensions or density of the lung remain challenging as radiologists may report the anatomical location and maximum dimension of a tumor or presence of emphysema, but do not delineate it on the images. Radiomics is the process to convert digital medical images into minable high-dimensional data.[18–22] This process will be facilitated by robust and automated deformable registration and anatomical image segmentation methods. Advanced data analysis approaches, such as machine learning methods, can be performed with the raw image voxel values to illustrate how segmented regions can contribute to the toxicity. However, automated feature extraction will be the key to linking those findings to physiological processes and may serve as guidance for laboratory science.

2.E. Treatment

The treatment of cancer is multifaceted and very complex. Traditionally, patients may undergo surgery, chemotherapy, and/or radiation therapy depending on their disease and medical status. Further complexity comes from hormonal treatments that may be used in combination, such as androgen suppression or estrogen blockers, or more recently immunotherapies. All of these treatments and their specific timing over the course of care effect both the disease control and the potential toxicity risks and quality of life experienced

by the patient. For practical big data applications, the goal is to capture the information from treatment in a quantitative way to better understand the nuances in the future.

Of the three main cancer treatment modalities, radiation is highly quantitative and lends itself for data curation and analysis. Radiation dosimetry is readily calculated; however, current practice does not robustly determine the "delivered" dose distribution for every patient. Patients may have modified fractionation or changes in the tumor and surrounding anatomy that cause the actual delivered dose to deviate from the original treatment plan. The volume and shape of the tumor may change due to patient-related factors during treatment necessitating midcourse modifications to the treatment plan. These impose challenges to obtaining the actual delivered dose. Therefore, current platforms should aim to account for the capture of the best possible representation of the delivered dose. In many cases today, our best estimate of delivered dose is the original treatment plan. However, proper attention to changes and a workflow to generate a final composite of a patient's treatment would improve data accuracy.

Surgery information is typically known at the procedure level. Detailed information about the location, extent of the surgery, and complications is more difficult to quantify and is documented in clinical notes. Guidance imaging and robotics tracking during surgery are possible, but currently limited in its availability. The success of surgery in terms of disease resection is documented in the pathology from the surgical specimen.

Chemotherapy is known at the regimen level and also at the level of medicinal administration. Timing of chemotherapy, treatment response, and interruption reasons in relation to radiation treatments may be critical to the patient's outcomes, so capturing the times and amount of each administration is critical in evaluating outcomes related to the timing of drugs or concomitant therapies. Standardized naming of drugs is available with RxNorm which is also used by several of the commercial drug knowledge bases used for ordering that provide decision support on drug–drug interactions.

While departments for each of these treatment modalities is solely responsible for their respective data collection efforts, perhaps the greatest challenge is to coordinate data sharing and analysis across departments. This includes agreeing in advance to adopt standardized clinical assessments whenever possible. The increased utilization of multidisciplinary clinic models should enhance efforts to coordinate treatment-related data collection over the entire course of patient care.

2.F. Symptom management

Documenting symptom management is also a key in understanding the treatment-related toxicities in the big data applications. The extent of symptom management depends on the patient's use of the intervention more than on the physician's prescriptions. For example, a prescription for pain relief does not mean that the patient is taking the medication as prescribed. Likewise, feeding tube placement does not indicate use. Tracking patients' adherence to prescribed medications is critical in understanding their impact and how they contribute to meaningful clinical outcomes. This data can only be collected through clinician assessments and patient-reported methods unless they are electronically monitored. Furthermore, as patients are often referred to other specialists (e.g., physical therapists and speech pathologists), an effort must be made to centralize progress reports in a structured format.

2.G. Technology

The primary goal of an informatics platform is to provide efficient access to the vast amounts of data for broad-based analytics and research. In general, the clinical informatics environment consists of transactional databases meant to support workflows, scheduling, and encounter-based data capture and a secondary data warehouse for building a data science environment facilitating analytics. In radiation oncology, the transactional databases also incorporate treatment planning, image guidance, and oncology and medical information systems. Therefore, the push now is to develop and populate data warehouses that can aggregate these high-dimensional data in an analyzable form.

The data warehouse should adhere to data standards whenever possible, to insure common and published meaning of each data element. It should also be designed for efficient queries for processing. Most data warehouses that deal with structured data are Structured Query Language (SQL)-based relational databases. Well-designed SQL databases can be very powerful for advancing knowledge and are in widespread use with support in many analytics software platforms. SQL is limited in that tables are defined and a simple link is used to define a relationship between them. Resource Description Framework (RDF or triplestore) databases offer the opportunity to link data together with more meaning to the link in a subject–predicate–object model where the predicate defines the nature of the link. More recently, the emerging of "Not only SQL" (NoSQL) databases enable storage of raw data, unprocessed data (e.g., clinical notes), which can be extracted further with tools, such as natural language processing.

Data transformation is data being translated from the clinical systems into the data warehouse. There are two different forms of transformation: (a) transforming the raw data directly and (b) creating derived features that may be of interest. In this process, data are subject to integrity checks such as verification of delivery of treatment or consistency of measurements over time. Data integrity checking with adherence to data standards is critical to the success of transformation and may further enable large-scale programs in data sharing and advanced data analysis in radiation oncology.

2.H. Clinical implementation

When it comes to the clinical implementation, there are three fundamental and critical components: (a) having the recognition that high-quality data collection in a prospective

TABLE I. This table begins to identify what information is used for care coordination and documentation of the encounter, and also what data can be interpreted into a structured form or directly measured. The specific data depend on the type of patient, their disease, and treatment. Ultimately, a smart system should determine, based on each type of patient and visit, which data can be captured and adhere it to standards, and capture it in the clinical workflow.

Encounter	Radiology	Consult	Surgery	Pathology	Medical Oncology	Radiation	On-Treatment	Follow-up
Care Coordination	• Diagnosis	• How to treat • Referral • Candidate for surgery	• Location • Extent • Margin status	• Disease stage and classification • Margin status • Histology	• Care plan modification • Resistance monitoring • Disease response	• RT schedule	• RT schedule • Referral for symptom management	• Disease status • Late toxicity management
Clinical Documentation	• Radiologist read	• History of present Illness • Review of systems • Past, family, and social history • Physical examination findings • Impression and plan	• Description of procedure • Specimen	• Type of specimen • Processing method • Description of findings	• Drug quantity and delivery method • Toxicity • Laboratories • Vital	• RT Rx and technique • Treatment plan	• Toxicities • Symptom management	• History of present illness • Disease status • Late toxicities • Review of systems • Physical examination findings
Interpreted Data	• Diagnosis • Disease extent • RECIST • Nodal involvement	• Diagnosis • Baseline quality of life • Baseline toxicity • Examinations • Family history • Social behaviors • Medical history	• Location • Extent • Pathologic results	• Disease stage • Histology • Margin status • Nodal involvement	• Drug quantity • Delivery method • Toxicities • Patient survey • Disease response • Symptom management	• Contoured targets • Treatment technique	• Patient survey • Acute toxicities • Disease response • Symptom management	• Patient survey • Late toxicities • Disease status
Measured Data	• CT images • MR images • Nuclear medicine	• Laboratories • Vitals • Imaging		• Digital slides • Genomics • Biopsy cores	• Vitals • Laboratories	• Imaging • 3D dosimetry • Contoured anatomy • Treatment calendar	• Imaging • Vitals	• Imaging • Vitals • Laboratories
Relevant Naming Standards	• ICD10 • DICOM	• CTCAE • LOINC • SNOWMED CT		• AJCC • ICD-O	• CTCAE • RxNorm	• AAPM TG263 • FMA	• CTCAE • RxNorm	• CTCAE • LOINC • SNOWMED

fashion will lead to a data resource supporting an acceleration in the advancement of medical knowledge and improved decision-making; (b) having an environment that recognizes the evolving culture change and enables the time and effort as well as customization of technology to seamlessly integrate with the clinical practice; and (c) development of early, effective and understandable tools to support clinical decision-making in the clinic.

Each patient encounter needs to be clearly documented to support coordinated care and provides an opportunity to capture data necessary for future analysis. Table I begins to identify what information is used for care coordination and documentation of the encounter, and also what data can be interpreted into a structured form or directly measured. Each specific visit type in the course of care provides that opportunity and the more understanding and standardization of the data; the more technology solutions can be created to breakdown current barriers to workflow related to efficiency and ease of use. Ultimately, a smart system should determine, based on each type of patient and visit, which data can be captured efficiently and adhere it to standards.

The hardest data to capture are the clinician- and patient-reported assessments of the patient condition. EMRs have the ability to create structured data items as part of the clinical documentation. Starting with a minimal set of data items specific to a disease site and type of visit (consult, on-treatment, or follow-up), begin to integrate them into the clinical visit with the physician and nursing staff. The informatics support personnel can help order the data elements and produce interfaces that minimize the complexity for the clinicians. Once started, continue to add to the structured data and allow the cultural change to evolve. As the clinician's take more interest, they can add new data that align with their interests or concerns for their patients. Moreover, as they see value in the analysis of that data, it will perpetuate.

The critical component for the informaticist is to fully understand the available data standards such as SNOWMED CT[23] or CTCAE to align any data collection to those standards. Significant effort is being put into improving the data standards such as these and the radiation oncology community should be engaged in the effort. In addition, informaticists need to respect that clinicians are trying to care for their patients and the technology should be as integrated into the clinical workflow as possible, ultimately enhancing patient care. Given the limitations of current clinical information systems, the current models often include home built tools to better streamline the workflow requiring some software development experience. Ultimately, such experimental user interface development should guide improvements and modifications to commercial systems for broader utilization in the community.

3. CONCLUSION

Our ability to use medical records to advance medical knowledge and improve experience-based decision through big data is arguably an eventuality. The current state of electronic medical records is not sufficient. A cultural change toward prospective data collection is necessary, and informaticists and medical physics can play a large role in catalyzing this evolution. Radiation oncology is perhaps the most quantitative medical field where we know where our dose goes and when we deliver it. We are very well poised as a field to lead medicine in the use of big data. Persistence and dedication to improving our data collection in an unobtrusive way will lead to advancements in medicine in the future.

[a)]Author to whom correspondence should be addressed. Electronic mail: tmcnutt1@jhmi.edu.

REFERENCES

1. McNutt TR, Moore KL, Quon H. Needs and challenges for big data in radiation oncology. *Int J Radiat Oncol Biol Phys*. 2016;95:909–915.
2. Marungo F, Robertson S, Quon H, et al. Creating a data science platform for developing complication risk models for personalized treatment planning in radiation oncology. *2015 48th Hawaii Int Conf Syst Sci*. 2015:3132–3140.
3. McNutt T, Wong J, Purdy J, Valicenti R, DeWeese T. OncoSpace: a new paradigm for clinical research and decision support in radiation oncology. *Proc XVIth Intl Conf Comput Radiother*; 2010;1.
4. Moore KL, Kagadis GC, McNutt TR, Moiseenko V, Mutic S. Vision 20/20: automation and advanced computing in clinical radiation oncology. *Med Phys*. 2014;41:010901.
5. McNutt T, DeWeese T, Herman J, Quon H, Moore J, Wong J, inventors; Johns Hopkins, assignee. *System and method for medical data analysis and sharing*. patent 20160378919; December, 2016.
6. 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.
7. Phillips LS, Branch WT, Cook CB, et al. Clinical inertia. *Ann Intern Med*. 2001;135:825–834.
8. Burger G, Abu-Hanna A, de Keizer N, Cornet R. Natural language processing in pathology: a scoping review. *J Clin Pathol*. 2016. https://doi.org/jclinpath-2016-203872
9. Lin FP, Pokorny A, Teng C, Epstein RJ. TEPAPA: a novel in silico feature learning pipeline for mining prognostic and associative factors from text-based electronic medical records. *Sci Rep*. 2017;7:6918.
10. National Cancer Institute. Patient-reported outcomes version of the common terminology criteria for adverse events (PRO-CTCAE™); 2017. https://healthcaredelivery.cancer.gov/pro-ctcae/.
11. Wei JT, Dunn RL, Litwin MS, Sandler HM, Sanda MG. Development and validation of the expanded prostate cancer index composite (EPIC) for comprehensive assessment of health-related quality of life in men with prostate cancer. *Urology*. 2000;56:899–905.
12. EORTC. EORTC quality of life; 2018. http://groups.eortc.be/qol/quality-life.
13. Yang W, Moore JA, Quon H, et al. Browser based platform in maintaining clinical activities – use of the iPads in head and neck clinics. *J Phys*. 2014;489:012095. (XVII International Conference on the Use of Computers in Radiation Therapy (ICCR 2013) 6–9 May 2013, Melbourne, Australia). http://iopscience.iop.org/article/10.1088/1742-6596/489/1/012095/pdf.
14. Gwaltney CJ, Shields AL, Shiffman S. Equivalence of electronic and paper-and-pencil administration of patient-reported outcome measures: a meta-analytic review. *Value Health*. 2008;11:322–333.
15. Health level seven international, 2018. www.hl7.org/implement/standards/. Accessed 10/2/2017.
16. Passiment E, Meisel J, Fontanesi J, Fritsma G, Aleryani S, Marques M. Decoding laboratory test names: a major challenge to appropriate patient care. *J Gen Intern Med*. 2013;28:453–458.
17. Amin MB, Edge S, Greene F, et al. (eds.). *AJCC cancer staging manual*. New York, NY: Springer Science+Business Media; 2016.

18. Aerts HJ, Velazquez ER, Leijenaar RT, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun*. 2014;5:4006.
19. El Naqa I, Grigsby P, Apte A, et al. Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. *Pattern Recognit*. 2009;42:1162–1171.
20. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer*. 2012;48:441–446.
21. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol*. 2017;14:749–762.
22. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology*. 2016;278:563–577.
23. Nikiema JN, Jouhet V, Mougin F. Integrating cancer diagnosis terminologies based on logical definitions of SNOMED CT concepts. *J Biomed Inform*. 2017;74:46–58.