Published in final edited form as:

Int J Radiat Oncol Biol Phys. 2016 July 1; 95(3): 895–904. doi:10.1016/j.ijrobp.2015.11.009.

How Will Big Data Improve Clinical and Basic Research in Radiation Therapy?

Barry S. Rosenstein, PhD[1,2], **Jacek Capala, PhD**[3], **Jason A. Efstathiou, MD, DPhil**[4], **Jeff Hammerbacher**[5], **Sarah Kerns, PhD**[6], **Feng-Ming (Spring) Kong, MD, PhD, FACR**[7], **Harry Ostrer, MD**[8], **Fred W. Prior, PhD**[9], **Bhadrasain Vikram, MD**[3], **John Wong, PhD**[10], and **Ying Xiao, PhD, FAAPM**[11]

[1]Departments of Radiation Oncology, Genetics and Genomic Sciences, Dermatology and Preventive Medicine, Icahn School of Medicine at Mount Sinai, New York, New York [2]Department of Radiation Oncology, New York University School of Medicine, New York, New York [3]Clinical Radiation Oncology Branch, National Cancer Institute, National Institutes of Health, Bethesda, Maryland [4]Department of Radiation Oncology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts [5]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York [6]Department of Radiation Oncology, University of Rochester Medical Center, Rochester, New York [7]Department of Radiation Oncology, GRU Cancer Center and Medical College of Georgia, Georgia Regents University, Augusta, Georgia [8]Departments of Pathology and Genetics, Albert Einstein College of Medicine, Bronx, New York [9]Mallinckrodt Institute of Radiology, Washington University School of Medicine, St Louis, Missouri [10]Department of Radiation Oncology, The Johns Hopkins University School of Medicine, Baltimore, Maryland [11]Department of Radiation Oncology, University of Pennsylvania, Philadelphia, PA

Abstract

Historically, basic scientists and clinical researchers have transduced reality into data so that they might explain or predict the world. Because data are fundamental to their craft, these investigators have been on the front lines of the Big Data deluge in recent years. Radiotherapy data are complex and longitudinal data sets are frequently collected to track both tumor and normal tissue response to therapy. As basic, translational and clinical investigators explore with increasingly greater depth the complexity of underlying disease processes and treatment outcomes, larger sample populations are required for research studies and greater quantities of data are being generated. In addition, well-curated research and trial data are being pooled in public data repositories to support large-

Reprint requests to: Barry S. Rosenstein, PhD, Department of Radiation Oncology, Box 1236, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, New York, NY 10029. Tel: (212)82-8960; barry.rosenstein@mssm.edu.

Conflicts of Interest: F.-M.K. received a speaker’s honorarium and research grants from Varian Medical System, F.P. is a Siemens shareholder.

Disclaimer: B.V. contributed to this work in his personal capacity and the views expressed herein may not represent the position of the NIH or DHHS.

scale analyses. Thus, the tremendous quantity of information produced in both basic and clinical research in radiation therapy can now be considered as having entered the realm of Big Data.

1. WHAT is the Big Data Question?

Advances in computer storage, computing power, statistical methods and the ability to electronically associate multiple types of data from disparate sources (e.g. demographic, genetic, imaging, treatment and outcomes) have enabled "Big Data" research in radiotherapy. Rather than setting a certain minimum number of computer memory bytes to define what is meant by Big Data, since this threshold would be continuously increasing with technological advances and the size of databases, it is most reasonable to refer to Big Data simply as volumes of large, complex and linkable information (1–2). As an example, the cost of genotyping used for analysis of DNA derived from the tumor and normal tissue of cancer patients has decreased multiple orders of magnitude over the past 20 years (3). Patients with cancer may soon have both their germline and tumor DNA sequenced as part of routine clinical care to individualize and optimize the treatment protocol, including radiotherapy. As a result, it is widely believed that the use of information technology can reduce the cost and increase the quality of healthcare.

Within radiation oncology, Big Data may include detailed 3D dosimetric and imaging data, as well as changes in images over time. One of the major promises of Big Data research is to identify the multiple clinical, biological, and treatment variables that determine clinical outcomes so that we can define better predictive models. This will enhance our ability to tailor therapies for each individual patient; e.g. delivering more aggressive therapies where needed and less-aggressive treatments when appropriate. Thus, Big Data can increase accuracy and support the development of precision methods for healthcare (4).

2. WHY is this Big Data issue important in radiation oncology and cancer care?

Our ability to extract new knowledge from the huge volumes of data produced by 21st century research is essential for advances in cancer diagnosis, treatment, and follow-up care. There is a clear need in radiation oncology, across multiple cancer sites, to use Big Data for research aimed at both improving cancer cure rates and reducing the incidence of normal tissue toxicity (Figure 1).

High-throughput screening using genomics, transcriptomics, proteomics, metabolomics, and other "-omics" techniques provide an unprecedented opportunity to apply systems biology approaches and advanced bioinformatics for basic and translational research in radiation biology (5). Analysis of the changes in cell signaling pathways in response to radiation will provide insight into crucial questions including the: (i) mechanisms regulating radio-sensitivity of different tissues, species, and individuals; (ii) response to different types of radiation and radiation schedules; (iii) low-dose effects including a possible safety threshold for radiation-induced cancer; (iv) bystander effect; and (v) biomarkers of radio-sensitivity and predictors of the response to radiation therapy.

3. How is the current emerging Big Data state to evolve with non-Radiation Oncology approaches?

One example is the Commons pilot experiment proposed by NIH aimed to efficiently store, manipulate, analyze, and share research output from a variety of public and private institutional facilities (each referred to as a Commons provider). A Commons provider agrees to support the rules of the Commons and to provide or permit services that facilitate the use of the Commons. Research objects within the Commons will be cataloged in an index being developed as part of the Big Data to Knowledge (BD2K) Initiative (6) and hence identifiable/searchable to users regardless of physical location. NIH established the BD2K program with the goal to help the biomedical research community benefit from the potential of Big Data.

Radiogenomics represents another important example of a non-radiation oncology approach enabling the use of Big Data from genetic/genomic studies to accelerate discovery and uncover information that smaller-scale studies could not reach (7–9). Radiogenomics uses methods that were developed in genomics science to uncover the genetic basis for numerous diseases and traits, and application of these approaches to radiation oncology has led to the discovery of genetic variants predictive of radiation-associated toxicity. Studies have identified multiple loci tagged by single nucleotide polymorphisms (SNPs), which are associated with late toxicity resulting from radiotherapy (10–12). Most of the genetic markers identified in these studies implicate genes not previously known to be involved in cellular radiation response, highlighting the role that Big Data is having in our understanding of radiobiology and clinical risk for toxicity. For example, SNP rs7582141 was identified in a genome-wide association study of late genitourinary and gastrointestinal toxicity among 1,740 prostate cancer radiotherapy patients (10). This SNP tags a haplotype block on chromosome 2q24.1 and lies within the *TANC1* gene and affects expression of this gene. The TANC1 protein plays a role in regenerating damaged muscle, a pathway that may be important in normal tissue damage during radiation exposure that has not previously been explored. This observation highlights the fact that Big Data allows investigators to take an agnostic look at the genome in order to uncover new information with clinical relevance. Big Data approaches can also be used to consider genome-wide gene expression, DNA methylation, histone modification, as well as the vast clinical data within medical records and clinical trials (Table 1a and 1b).

4. WHERE will this growth and development in radiation oncology occur and be evaluated?

NCI reformulated the National Clinical Trials Network (NCTN) to more efficiently conduct cancer clinical trials focused on state-of-the-art research questions (13). One of the service groups established was the Imaging and Radiation Oncology Core (IROC). This group's mission is to provide integrated radiation oncology and diagnostic imaging quality control as well as clinical, scientific, and technical expertise in support of the NCTN. A Center for Innovation in Radiation Therapy (CIRO) has also been established within NRG Oncology for implementation of state of the art radiotherapy technologies into NCTN clinical trials. A cloud-based informatics infrastructure (Figure 2) was established for data transfer, quality assurance evaluation and data integration with NCI systems and standards. The infrastructure includes interface with analysis applications that are centrally available to the research community.

The NCI-funded cancer genome atlas (TCGA) (14) has made genomic data on a variety of cancers commonly treated by radiotherapy publicly available and enabled genomic studies. The American Society for Radiation Oncology (ASTRO) could facilitate linkages among the various kinds of data (pathology, genomics, images, treatment planning data, drug therapy, outcomes data) in coordination with other organizations such as Radiological Society of North America (RSNA) and American Society for Clinical Oncology (ASCO), thereby creating a critical data repository useful for improving the diagnosis and treatment of patients.

Other examples of the open resources community and data sharing are the Biomedical Informatics Research Network (BIRN) (15) and the National Cancer Informatics Program (NCIP) (16), which has incorporated the Cancer Biomedical Informatics Grid (caBIG) (17). The NCIP represents interoperable biomedical information systems built on community-driven data standards.

The Cancer Imaging Archive (TCIA) (18) is a Big Data resource to support NCI-funded research activities and the cancer research community (Figure 3). The TCIA team defined standard operating procedures for data acquisition, de-identification, and curation, and invented mechanisms for reliable, clustered hosting of multiple National Biomedical Imaging Archive software instances on an open source private cloud infrastructure to provide increased performance and reliability (19–21). The TCIA team provides expert curation and quality control of incoming data sets, extension of the software/technology used to meet production deployment standards of service, and dissemination of knowledge to the wider research community in areas of Digital Imaging and Communications in Medicine (DICOM) de-identification and open image archives (21). TCIA has been expanded to support all DICOM format radiotherapy information objects as well as radiology and pathology images, clinical trial data and image annotations (22). Information in TCIA is cross-referenced at the subject level with information in TCGA (23) and via the TCIA programmatic interface (24) supports cross-information system query and retrieval of information. In addition, the TCIA is an official repository for the Nature Publishing Group and a growing number of additional journals. Data sets are published using a DOI that can either be stand-alone, published as a Nature Data Descriptor or included as a reference.

Another potential major source of Big Data for future research is through the Veterans Affairs (VA) system and the Veterans Health Administration (25) since roughly 40,000 new patients are treated using radiotherapy within the VA system each year. In particular, this data source could be used to perform retrospective studies to explore scientific hypotheses, and patterns of care summaries could serve as a useful means to explore utilization and ensure a minimum quality of care. It should also be noted that all data for veterans is housed in the VA Informatics and Computing Infrastructure (26).

Oncospace is an informatics model (Figure 4) that is intended to be a distributed analytic database for structured information, forgoing export/import of data (27). The vision is to distribute Oncospace(s) to the collaborating clinics where the adherence to the Oncospace common database schema and data standards would enable more efficient and effective queries and analyses. It is important to integrate data collection as part of the clinical workflow, such as during patient encounter, to minimize the need of repeating data extraction and entry. Equally important is the implementation of efficient ETL (Extract, Transform, and Load) processes from different data sources, such as the Oncology Information System or the radiation treatment planning system to populate each Oncospace and to allow web-based access to the distributed Oncospace(s). This approach avoids the burden of data export and provides a rich forum for data mining to address many research questions or data sharing to support clinical decision-making. An Oncospace Consortium consisting of four radiation oncology clinics has been formed to investigate the merits and pitfalls of the approach.

EUROCAT represents an important example of a global distributed learning network representing 18 centers from 9 countries (28–29). EUROCAT strictly adheres to a no-data-leaves-the-hospital paradigm in which learning applications are distributed to the hospitals rather than transferring the data, which alleviates many of the administrative, political and ethical barriers associated with data sharing. The EUROCAT open source solution relies on the W3C vision of the Semantic Web (30) in which features required to answer research questions are published on a Semantic Web enabled intranet page (also known as a SPARQL endpoint) with all partners using the same Radiation Oncology Ontology (31). In this way, fully semantic interoperable data are created, which is a condition for distributed applications to learn from the data despite differences in languages, vendors and multiple additional factors between centers. Another example of effective data aggregation and syntactic/semantic inter-operability is use of the Text Information Extraction System (TIES), which is an an open-source computer-based system that uses natural language processing methods to automate annotation of tissue samples using text-based electronic medical records. This is the basis of the TIES Cancer Research Network (TCRN), which is a multi-institutional, collaborative research network providing de-identified clinical data and associated biospecimens to investigators from member institutions (32).

Data collection for use in larger multi-institutional environments requires directly integrating local clinical data sources through an oncology specific workflow management system. Such systems embed the clinical data within the greater context of patient and departmental process. Such effort is formalized in the new DICOM RT 2nd generation standards and in intelligent workflow management systems such as the xxx Whiteboard Informatics Platform that directly support them. The xxx platform uses an enterprise data bus that captures all data transactions and creates and maintains an up-to-date patient record as specified by the DICOM RT 2nd generation RT Course object. The RT Course object thus becomes the repository and reference for all patient data records produced over the course of treatment.

The National Radiation Oncology Registry (NROR) prostate pilot (33) is an example of ASTRO and the Radiation Oncology Institute (ROI) pursuing a real-world registry to define practice patterns aligned with Big Data objectives and aiming to lead to improvements in quality care delivery. Preliminary results highlight the importance of integrating data collection into clinical workflow. In addition, developing a standardized taxonomy (data dictionary) with high-yield data elements, defining quality metrics of interest, providing quality feedback to providers and establishing consensus on documentation practices and semantics will help facilitate the interpretation of Big Data.

5. WHEN can we expect results?

Early results are promising and it is anticipated that the findings of Big Data used in basic and clinical research will be translated into the clinic within 5–10 years. For example, the Radiogenomics Consortium (RGC) will complete a genome wide assessment for approximately 7,000 men treated with radiotherapy for prostate cancer, and an assay to predict clinical toxicity outcomes is expected to be available for testing in studies of clinical utility within ≈5 years (8–9). Another ongoing RGC study aims to validate predictive models of normal tissue toxicity by enrolling 5,300 patients irradiated for prostate, breast, and lung cancer. Comprehensive data regarding clinical outcomes are being prospectively collected before, during and after treatment (34). This study, termed REQUITE (Validating Predictive Models and Biomarkers of Radiotherapy Toxicity to Reduce Side-Effects and Improve Quality of Life in Cancer Survivors), uses various Big Data methods to build predictive models incorporating dosimetric parameters, SNPs, readout from cellular assays and clinical risk factors. As part of this study, predictive models will be developed, validated and refined and it is anticipated these instruments will be available for testing in the broader clinical radiation oncology community by 2018. Similarly, genomic markers (e.g. SNPs) can be used to define predictive models for tumor response (e.g. an individual’s tumor radiation dose response) to potentially guide personalized radiotherapy (35).

6. WHO will be impacted, and who will provide resources?

This work will broadly impact the entire radiotherapy community; e.g. researchers, providers, insurers, and most importantly patients. For example, it will be incumbent upon the radiation oncology research community to support the development of young investigators with expertise in bioinformatics, molecular epidemiology and biostatistics. Providers will need to become versed in these methods so that they can effectively incorporate this information into their treatment planning software and additional tools will be required to facilitate the translation of this knowledge to the clinic. However, the largely unanswered “$64 billion question” is: Who will provide the necessary resources?

The impact on patients might be enormous. Better stratification of patients based on their own expected tumor and normal tissue factors will enable therapy to be highly tailored. Patients with low risk disease (e.g. women with low risk breast cancer or men with low risk prostate cancer) might be able to largely avoid treatment. We can better define the individual patient subgroups that are benefiting from specific components of radiotherapy. For example, assume a particular intervention has been shown to increase the long-term cure rate (e.g. from 50% to 55%) in a broad group of patients. However, it might be that only 5% of patients will benefit from the intervention, while 50% of the patients do not require the intervention and 45% would derive no benefit. Identifying those 5% of patients affected by the intervention through Big Data research can alter therapy for the other 95% of patients.

7. WHAT are the potential major obstacles and HOW can these be addressed?

Modern research has evolved rapidly toward large-scale experiments and trials producing massive volumes of complex data. The challenges posed by such Big Data research include capture, curation, storage, search, sharing, transfer, analysis and visualization. It is also important to note that fundamental characteristics of Big Data including, volume, variety, velocity, variability and veracity (36) are also critical aspects of the Big Data produced in the context of radiotherapy research. Some of specific challenges in the use of Big Data for research in radiation oncology include; (1) information extraction and cleaning, (2) data integration, aggregation, and representation and (3) query processing, data modeling and analysis.

Beyond standard clinical, molecular, and imaging data, cancer researchers and clinicians have access to new data sources like consumer sensors and interactions on patient social networks such as Inspire (37). However, there are many new challenges when working with data collected from multiple sites outside the usual scope of clinical trials. For example, researchers must ensure that patients provide appropriate consent. The Portable Legal Consent (38), developed for use with Apple's Research Kit, is one approach. In addition, on-line data sharing platforms have been developed to host and share genotypes, electronic health records and family history that have been uploaded by individuals (39).

A major hurdle confronting the effective use of Big Data is that its enormous volume impedes analyses and data exchange. The storage, management and processing of Big Data have become onerous and create a situation where healthcare data and information are inefficiently and ineffectively disseminated or even lost. In radiation oncology, there exist long standing NCI funded cooperative groups, such as the Radiation Therapy Oncology Group (RTOG), now part of NRG Oncology, which manage randomized trial research with established quality assurance (QA) standards, credentialing processes and stringent randomization procedures to ensure data integrity and valid conclusions. As our treatment has evolved from 3D conformal radiation therapy to intensity modulated radiation therapy to image guided radiation therapy, the data burden exposes deficiencies of the current cooperative research process that must be addressed if we are to take full advantage of the potential.

Besides cooperative research, it is also useful to examine how clinical data and information are utilized currently for clinical decision support. Some standardization, such as DICOM, has enhanced imaging data exchange. However, significant amounts of important information such as routine and non-standard laboratory results, radiation dose deposition data, treatment outcomes, etc. are stored in unstructured formats and incompatible databases. The volume, complexity and inaccessibility of the data limit the depth of information that is retrieved for decision support. Indeed, at present, only a small fraction of the information is retrieved for decision making by the individual provider or a multidisciplinary group. Rarely, if ever, is the information captured in a structured format for systematic update and re-use by individual physicians and certainly not by other physicians who might benefit by that experience in the management of an individual, newly diagnosed patient. In this regard, CancerLinQ™ (40), which is being developed by the ASCO Institute for Quality, is an important health information technology platform that will obtain data from millions of individual patients living with cancer nationwide. The radiation oncology community should initiate a similar effort.

SUMMARY

8. Is the future bright for radiation oncology?

Cancer research performed over the past decades demonstrates that patients and their cancers respond in a variable fashion to radiotherapy and therefore treatment should be tailored to each individual patient (41). It is therefore anticipated that the day may not be far off when a radiation oncologist will be able to enter clinical, imaging, dosimetric, genomic and other panomic data into a software package that will analyze the information and provide treatment recommendations into an approach that could be termed “Knowledge Guided Radiotherapy” (KGRT) (Figure 5). Although numerous diagnostic and therapeutic tools exist, it is at the moment extremely difficult for a physician and patient to decide on what is best in an individual case. The availability of molecular tools now enables a large-scale, parallel, quantitative, and inexpensive assessment of molecular states. The data from these tools are also increasingly being made publicly available. There is a culture of open sharing in molecular biology, genetics/genomics and bioinformatics that continues to grow. However, the data, in their current fragmented form, creates a major barrier for research integrating the biological, physical and clinical science. Improved infrastructure for development of outcomes data and a rapid learning health-care system, where we can be informed from each patient to guide practice, is increasingly viewed as crucial. With the tremendous growth in the rate of increase at which new data are generated, it is essential to employ bioinformatics algorithms and strategies for analysis and modeling that incorporates biological, physical and clinical data forms.

The future is promising for radiation oncology, but only if we engage all of our stakeholders; i.e. academia, medical centers, funding organizations, equipment vendors, pharmaceutical companies and professional societies, in an efficient and concerted effort to use the explosion in the availability of Big Data to improve clinical and basic research in radiation therapy.

Acknowledgements

This work was supported by grants to B.S.R. from the National Institutes of Health (1R01CA134444), the Department of Defense (PC074201) and the American Cancer Society (RSGT-05-200-01-CCE), by grants to F.M.K. from the National Institutes of Health (R01CA142840), by grants to Y.X. from the National Institutes of Health (U24CA12014, U24CA180803, U10CA180868, U10CA180822) and PA CURE and by contracts to F.W.P. from National Cancer Institute, Contract No. HHSN261200800001E and 1U01CA187013. Dr. Andre Dekker and Dr. Jatinder Palta are thanked for the information concerning EUROCAT and VA system, respectively.

References

1. Khoury MJ, Ioannidis JP. Big data meets public health. Science. 2014; 346:1054–1055. [PubMed: 25430753]
2. Toga AW, Foster I, Kesselman C, et al. Big Biomedical data as the key resource for discovery science. J Am Med Inform Assoc. 2015 (Epub ahead of print).
3. Hayden EC. Technology: The $1,000 genome. Nature. 2014; 507:294–295. [PubMed: 24646979]
4. Chaussabel D, Pulendran B. A vision and a prescription for big data-enabled medicine. Nat Immunol. 2015; 16:435–439. [PubMed: 25898187]
5. Shrager J, Tenenbaum JM. Rapid learning for precision oncology. Nat Rev Clin Oncol. 2014; 11:109–118. [PubMed: 24445514]
6. Margolis R, Derr L, Dunn M, et al. The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. J Am Med Inform Assoc. 2014; 21:957–958. [PubMed: 25008006]
7. Rosenstein BS, West CM, Bentzen SM, et al. Radiogenomics: radiobiology enters the era of big data and team science. Int J Radiat Oncol Biol Phys. 2014; 89:709–713. [PubMed: 24969789]
8. 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. [PubMed: 24441285]
9. Kerns SL, West CM, Andreassen CN, et al. Radiogenomics: the search for genetic predictors of radiotherapy response. Future Oncol. 2014; 10:2391–2406. [PubMed: 25525847]
10. Fachal L, Gómez-Caamaño A, Barnett GC, et al. A three-stage genome-wide association study identifies a susceptibility locus for late radiotherapy toxicity at 2q24.1. Nat Genet. 2014; 46:891–894. [PubMed: 24974847]
11. Kerns SL, Stock RG, Stone NN, et al. Genome-wide association study identifies a region on chromosome 11q14.3 associated with late rectal bleeding following radiation therapy for prostate cancer. Radiother Oncol. 2013; 107:372–376. [PubMed: 23719583]
12. Kerns SL, Ostrer H, Stock R, et al. Genome-wide association study to identify single nucleotide polymorphisms (SNPs) associated with the development of erectile dysfunction in African-American men after radiotherapy for prostate cancer. Int J Radiat Oncol Biol Phys. 2010; 78:1292–1300. [PubMed: 20932654]
13. http://www.cancer.gov/research/areas/clinical-trials/nctn
14. Hanauer DA, Rhodes DR, Sinha-Kumar C, et al. Bioinformatics approaches in the study of cancer. Curr Mol Med. 2007; 7:133–141. [PubMed: 17311538]
15. Helmer KG, Ambite JL, Ames J, et al. Enabling collaborative research using the Biomedical Informatics Research Network (BIRN). Journal of the American Medical Informatics Association: JAMIA. 2011; 18:416–422. [PubMed: 21515543]
16. https://cbiit.nci.nih.gov/ncip
17. Buetow KH. An infrastructure for interconnecting research institutions. Drug Discov Today. 2009; 14:605–610. [PubMed: 19508923]
18. Kirby J, Tarbox L, Freymann J, et al. TU-AB-BRA-03: The Cancer Imaging Archive: Supporting Radiomic and Imaging Genomic Research with Open-Access Data Sets. Med Phys. 2015; 42:3587.
19. Prior, FW.; Clark, K.; Commean, P., et al. TCIA: an information resource to enable open science. Engineering in Medicine and Biology Society (EMBC), 2013; 35th Annual International Conference of the IEEE; IEEE; 2013. p. 2
20. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. Journal of Digital Imaging. 2013; 26:1045–1057. [PubMed: 23884657]
21. Moore SM, Maffitt DR, Smith KE, et al. De-identification of Medical Images with Retention of Scientific Research Value. RadioGraphics. 2015; 35:727–735. [PubMed: 25969931]

22. Commean PK, Rathmell JM, Clark KW, et al. A Query Tool for Investigator Access to the Data and Images of the National Lung Screening Trial. Journal of Digital Imaging. 2015:1–9. [PubMed: 25416467]
23. Jaffe CC. Imaging and genomics: is there a synergy? Radiology. 2012; 264:329–331. [PubMed: 22821693]
24. Kalpathy-Cramer J, Freymann JB, Kirby JS, et al. Quantitative imaging network: Data sharing and competitive algorithm validation leveraging The Cancer Imaging Archive. Translational Oncology. 2014; 7:147–152. [PubMed: 24772218]
25. Shelton JB, Skolarus TA, Ordin D, et al. Validating electronic cancer quality measures at Veterans Health Administration. Am J Manag Care. 2014 Dec.20:1041–1047. [PubMed: 25526392]
26. http://www.hsrd.research.va.gov/for_researchers/vinci/
27. Bowers M, Robertson S, Moore J, et al. SU-E-P-26: Oncospace: A Shared Radiation Oncology Database System Designed for Personalized Medicine, Decision Support, and Research. Med Phys. 2015; 42:3232.
28. Lambin P, Zindler J, Vanneste B, et al. Modern clinical research: How rapid learning health care and cohort multiple randomised clinical trials complement traditional evidence based medicine. Acta Oncologica. 2015:1–12.
29. https://urldefense.proofpoint.com/v2/url?u=https-3A__www.youtube.com_watch-3Fv-3DZDJFOxpwqEA&d=AwIFAw&c=4R1YgkJNMyVWjMjneTwN5tJRn8m8VqTSNCjYLg1wNX4&r=bwjNX3_CobeoDdW0MZSXlYIqDuq-h75uvIJGqE75Aug&m=oK-tSbdtdk7jyVH14V4FId1gYFkzz25bOYZtXG1krrM&s=hvXWva76BJjz0_UXlkwn3FVBmzNeeXOy6KbVptdJO7Q&e=.
30. https://urldefense.proofpoint.com/v2/url?u=http-3A__www.w3.org_standards_semanticweb_&d=AwIFAw&c=4R1YgkJNMyVWjMjneTwN5tJRn8m8VqTSNCjYLg1wNX4&r=bwjNX3_CobeoDdW0MZSXlYIqDuq-h75uvIJGqE75Aug&m=oK-tSbdtdk7jyVH14V4FId1gYFkzz25bOYZtXG1krrM&s=bKNSnwfle85IXh0wsm2Pf7g9Gb4F6c9gW14d9M5ZfX4&e=.
31. https://urldefense.proofpoint.com/v2/url?u=http-3A__bioportal.bioontology.org_ontologies_ROO&d=AwIFAw&c=4R1YgkJNMyVWjMjneTwN5tJRn8m8VqTSNCjYLg1wNX4&r=bwjNX3_CobeoDdW0MZSXlYIqDuqh75uvIJGqE75Aug&m=oK-tSbdtdk7jyVH14V4FId1gYFkzz25bOYZtXG1krrM&s=nBJrAnZqs65I_jB32aMM9VU-cVI-NAfk4CZHhlgY__A&e=.
32. Jacobson RS, Becich MJ, Bollag RJ, et al. A Federated Network for Translational Cancer Research Using Clinical Data and Biospecimens. Cancer Research. (in press).
33. Efstathiou JA, Nassif DS, McNutt TR, et al. Practice-based evidence to evidence-based practice: building the National Radiation Oncology Registry. J Oncol Pract. 2013; 9:e90–e95. [PubMed: 23942508]
34. West C, Azria D, Chang-Claude J, et al. 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). 2014; 26:739–742. [PubMed: 25267305]
35. Jin J-Y, Wang W, Ten Haken RK, et al. Use a survival model to correlate single-nucleotide polymorphisms of DNA repair genes with radiation dose response in patients with non-small cell lung cancer. Radiother Oncol. 2015 Aug 4. pii: S0167-8140(15)00378-3. [Epub ahead of print]).
36. Gandomi A, Haider M. Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management. 2015:35137–144.
37. https://www.inspire.com/.
38. http://sagebase.org/e-consent/.
39. Shabani M, Borry P. Challenges of web-based personal genomic data sharing. Life Sci Soc Policy. 2015; 11:22.
40. http://www.instituteforquality.org/cancerlinq.

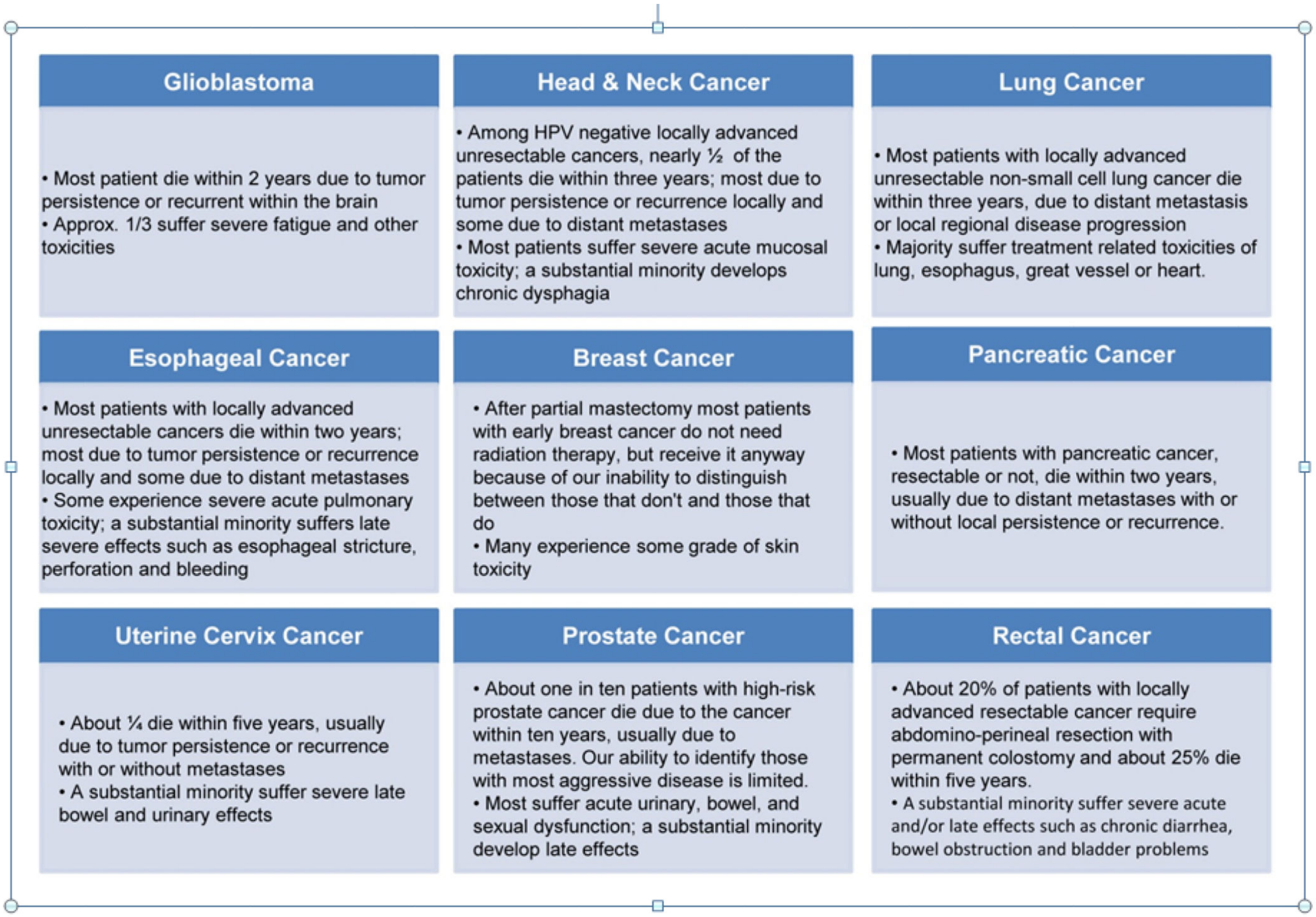
41. Kong FM, Ao X, Wang L, Lawrence TS. The use of blood biomarkers to predict radiation lung toxicity: a potential strategy to individualize thoracic radiation therapy. Cancer Control. 2008; 15:140–150. [PubMed: 18376381]

42. Bennett W, Matthews J, Bosch W. SU - GG - T - 262: Open - Source Tool for Assessing Variability in DICOM Data. Medical Physics. 2010; 37:3245.

43. Jin JY, Kong FM. Personalized Radiation Therapy for Lung Cancer, Chapter #10, "Lung Cancer and Personalized Medicine: Novel Therapies and Clinical Management", as part of the series 'Advances in Experimental Medicine and Biology' by Springer Publishers. 2015 (in press).

SUMMARY

Data derived from biomedical research are diverse and complex as this includes imaging, phenotypic, genetic/genomic, molecular, exposure, health, behavioral, demographic, treatment, outcomes, toxicity and many other types of information. Big Data methods allow researchers to maximize the potential of existing data and enable new directions for research. In radiation oncology, the main challenge is for investigators to harness the power of Big Data to advance basic and clinical research whose goal is to discover novel ways to increase cancer cure rates and/or decrease treatment toxicity.



Glioblastoma

• Most patient die within 2 years due to tumor persistence or recurrent within the brain
• Approx. 1/3 suffer severe fatigue and other toxicities

Head & Neck Cancer

• Among HPV negative locally advanced unresectable cancers, nearly ½ of the patients die within three years; most due to tumor persistence or recurrence locally and some due to distant metastases
• Most patients suffer severe acute mucosal toxicity; a substantial minority develops chronic dysphagia

Lung Cancer

• Most patients with locally advanced unresectable non-small cell lung cancer die within three years, due to distant metastasis or local regional disease progression
• Majority suffer treatment related toxicities of lung, esophagus, great vessel or heart.

Esophageal Cancer

• Most patients with locally advanced unresectable cancers die within two years; most due to tumor persistence or recurrence locally and some due to distant metastases
• Some experience severe acute pulmonary toxicity; a substantial minority suffers late severe effects such as esophageal stricture, perforation and bleeding

Breast Cancer

• After partial mastectomy most patients with early breast cancer do not need radiation therapy, but receive it anyway because of our inability to distinguish between those that don't and those that do
• Many experience some grade of skin toxicity

Pancreatic Cancer

• Most patients with pancreatic cancer, resectable or not, die within two years, usually due to distant metastases with or without local persistence or recurrence.

Uterine Cervix Cancer

• About ¼ die within five years, usually due to tumor persistence or recurrence with or without metastases
• A substantial minority suffer severe late bowel and urinary effects

Prostate Cancer

• About one in ten patients with high-risk prostate cancer die due to the cancer within ten years, usually due to metastases. Our ability to identify those with most aggressive disease is limited.
• Most suffer acute urinary, bowel, and sexual dysfunction; a substantial minority develop late effects

Rectal Cancer

• About 20% of patients with locally advanced resectable cancer require abdomino-perineal resection with permanent colostomy and about 25% die within five years.
• A substantial minority suffer severe acute and/or late effects such as chronic diarrhea, bowel obstruction and bladder problems

Figure 1.

Most pressing issues in contemporary radiation oncology that could benefit from Big Data research. For each tumor site, areas in need of improvement in tumor and normal tissue outcomes are highlighted.

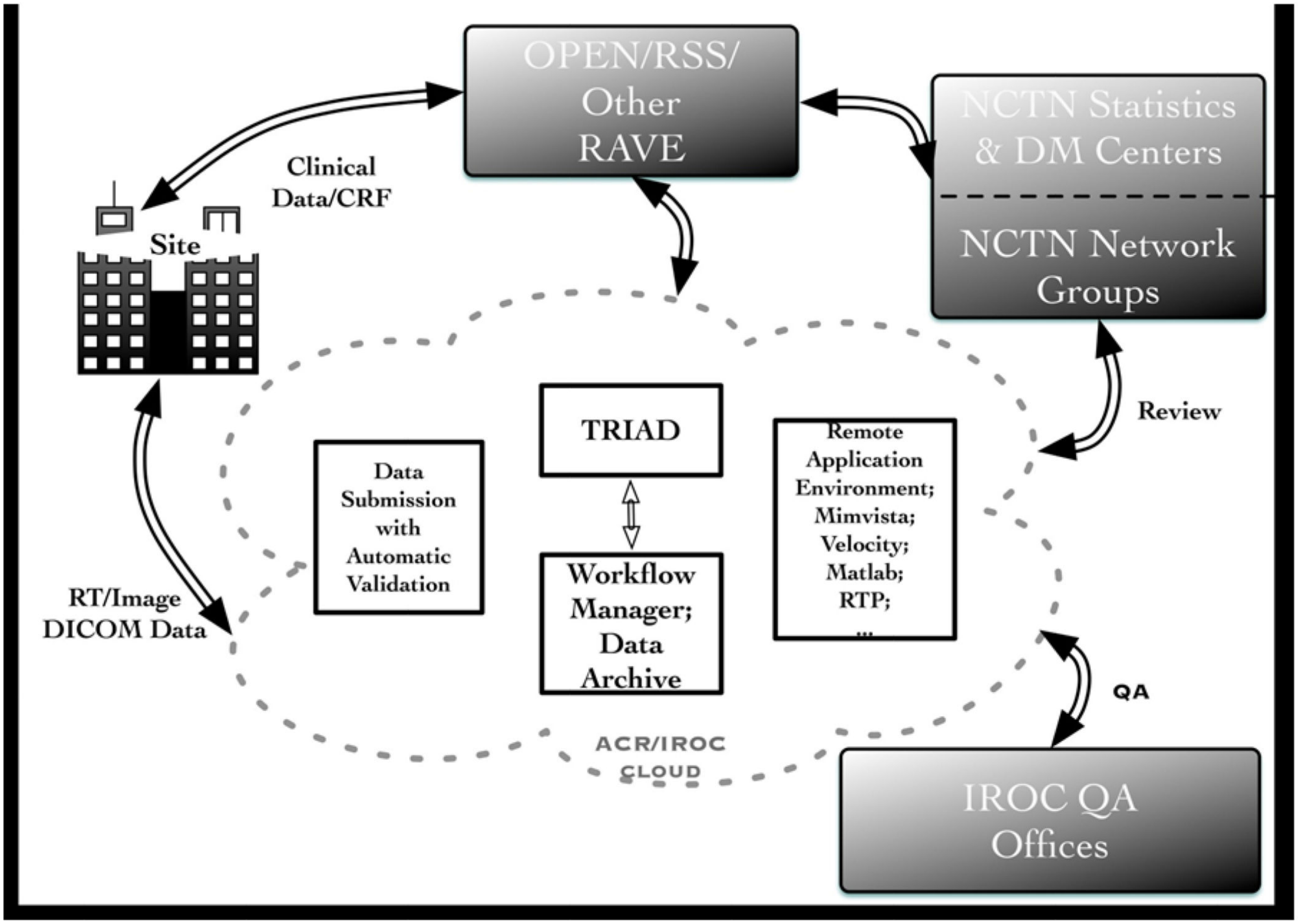


Figure 2.

A cloud-based informatics infrastructure established for data transfer, quality assurance (QA) evaluation, data integration with NCI systems and standards. NCTN systems: OPEN - Oncology Patient Enrollment Network; RSS - Regulatory Support System; IROC – Image and Radiation Oncology Core. American College of Radiology (ACR) system: TRIAD – Transfer of Image and Data.

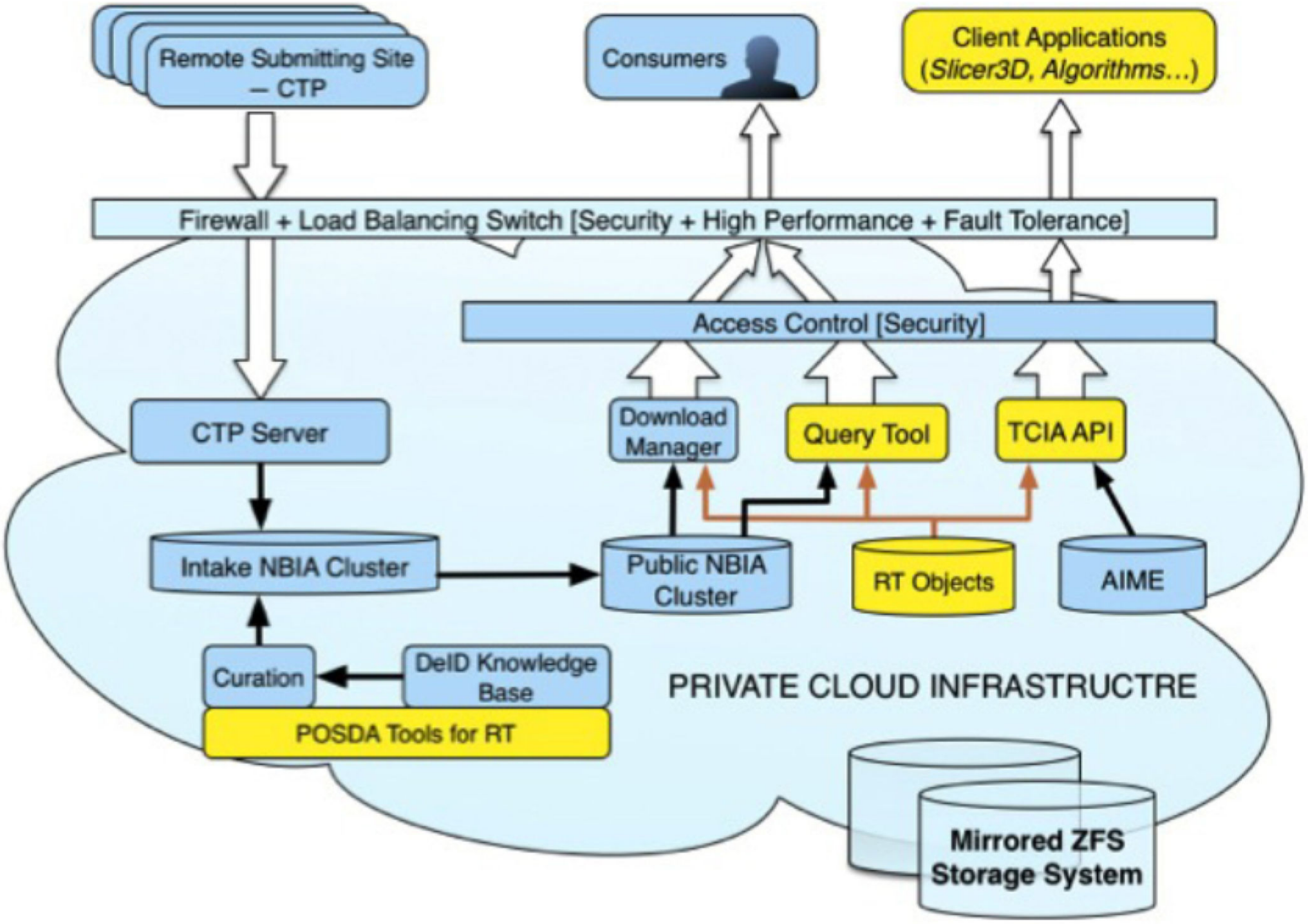

Remote Submitting Site – CTP
Consumers
Client Applications (Slicer3D, Algorithms…)
Firewall + Load Balancing Switch [Security + High Performance + Fault Tolerance]
Access Control [Security]
CTP Server
Download Manager
Query Tool
TCIA API
Intake NBIA Cluster
Public NBIA Cluster
RT Objects
AIME
Curation
DeID Knowledge Base
POSDA Tools for RT
PRIVATE CLOUD INFRASTRUCTRE
Mirrored ZFS Storage System

Figure 3.

The Cancer Imaging Archive. (TCIA) is an NCI funded information repository that aggregates images (radiology, pathology), Radiation Therapy (RT) Information objects, annotations, clinical trial data, and information derived from quantitative image analysis to support Big Data analytics. TCIA has been extended to support curation, quality assurance, management and distribution of advanced RT information objects (illustrated in yellow) by incorporation of the POSDA (Perl Open Source Dicom) tool set (42) and associated processes.

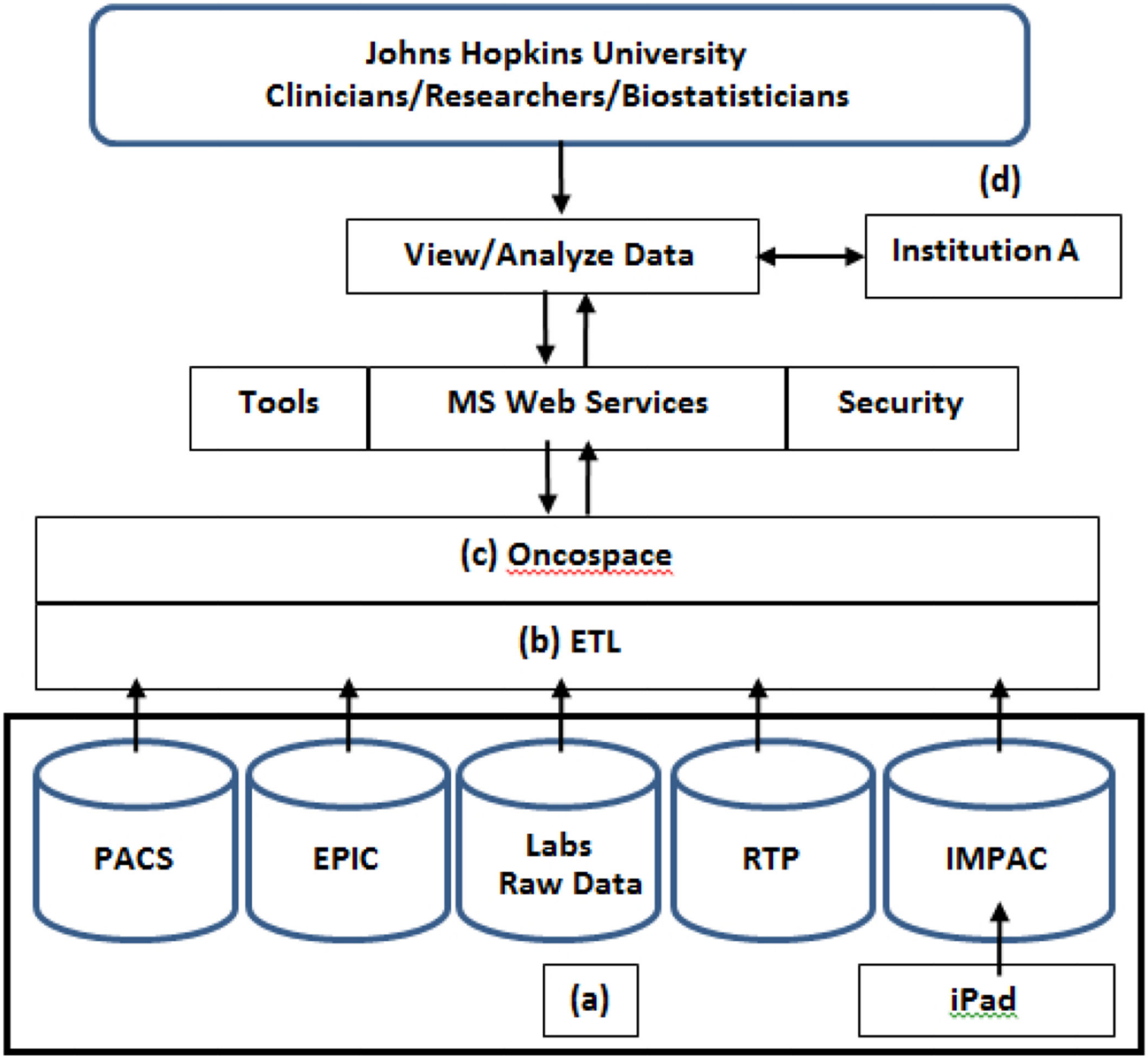
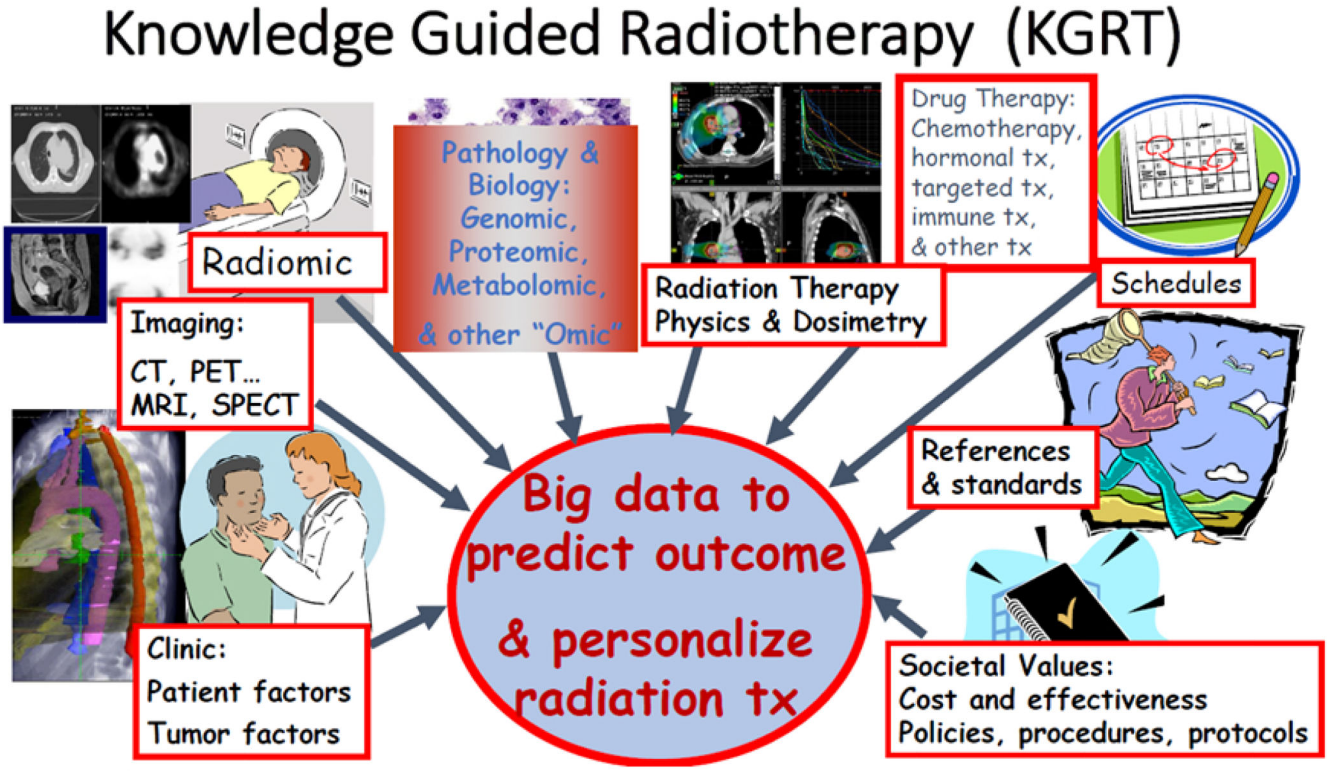


Figure 4.

Schematic of utilization of Oncospace at xxx. (a) Sources of raw data that are (b) "extracted, transformed and loaded" (ETL) into (c) Oncospace as metadata for efficient viewing and analysis by web-based query from (d) an outside institution as a member of the Oncospace consortium.



Knowledge Guided Radiotherapy (KGRT)

Radiomic

Imaging:
CT, PET…
MRI, SPECT

Pathology & Biology:
Genomic,
Proteomic,
Metabolomic,
& other "Omic"

Radiation Therapy
Physics & Dosimetry

Drug Therapy:
Chemotherapy,
hormonal tx,
targeted tx,
immune tx,
& other tx

Schedules

References
& standards

Clinic:
Patient factors
Tumor factors

Big data to predict outcome
& personalize radiation tx

Societal Values:
Cost and effectiveness
Policies, procedures, protocols

Figure 5.

Knowledge Guided Radiotherapy. Big data will provide comprehensive knowledge for each patient from clinic to the laboratory. Such a knowledge-guided radiotherapy (KGRT) provides a potential to predict treatment outcome for each individual patient and guide personalized care for a maximized therapeutic gain for that individual and society (43). xxx is thanked for providing the initial formulation of this figure.

Table 1

a: Big Data sources from medical records linked to outcomes and/or genomics		
Source	**Description**	**Link**
Electronic Medical Records and Genomics (eMERGE)	A national network organized and funded by the National Human Genome Research Institute (NHGRI) that combines DNA biorepositories with electronic medical record (EMR) systems for large scale, high-throughput genetic research in support of implementing genomic medicine.	http://www.genome.gov/27540473
The Phenotype Knowledgebase (PheKB)	A collaborative environment to building and validating electronic algorithms to identify characteristics of patients within health data. PheKB was functionally designed to enable such a workflow and has purposefully integrated tools and standards that guide the user in efficiently navigating each of these stages from early stage development to public sharing and reuse. PheKB has tools to enable cross-site collaboration for algorithm development, validation, and sharing for reuse with confidence.	https://phekb.org/
Clinical Data Research Networks (CDRNs)	CDRNs will develop the capacity to conduct randomized comparative effectiveness studies using data from clinical practice in a large, defined population. These established or newly developed networks involve two or more healthcare systems, with plans to function as integrated research network.	http://www.pcornet.org/clinical-data-research-networks/
VA Informatics and Computing Infrastructure (VINCI)	VINCI is an initiative to improve researchers' access to VA data and to facilitate the analysis of those data while ensuring Veterans' privacy and data security. VINCI welcomes all researchers in the VA community to explore the environment and tools available.	http://www.hsrd.research.va.gov/for_researchers/vinci/

b: Big Data sources from consortia linked to exposures and/or genomics		
Source	**Description**	**Link**
NCI Cohort Consortium	The NCI Cohort Consortium includes investigators responsible for more than 50 high-quality cohorts involving more than 7 million people. The cohorts are international in scope and cover large, rich, and diverse populations. Extensive risk factor data are available on each cohort, and biospecimens including germline DNA collected at baseline, are available on approximately 2 million individuals. Investigators team up to use common protocols and methods, and to conduct coordinated parallel and pooled analyses.	http://epi.grants.cancer.gov/Consortia/cohort.html
International Cancer Genome Consortium	ICGC Goal: To obtain a comprehensive description of genomic, transcriptomic and epigenomic changes in 50 different tumor types and/or subtypes which are of clinical and societal importance across the globe.	https://icgc.org/
Clinical Sequencing Exploratory Research (CSER)	The CSER consortium represents a diverse collection of projects investigating the application of genome-scale sequencing in different clinical settings including pediatric and adult subspecialties, germline diagnostic testing and tumor sequencing, and specialty and primary care.	https://cser-consortium.org/projects