Article Type: Special Issue Paper

Genomics, Bio specimens and other Biological Data: Current status and future directions

Barry S. Rosenstein[1,2], Arvind Rao[3], Jean M. Moran[3], Daniel E. Spratt[3], Marc Mendonca[4,5], Bissan Al-Lazikani[6], Charles S. Mayo[3], Corey Speers[3‡]

[1]Department of Radiation Oncology, Icahn School of Medicine at Mount Sinai, New York, NY,10029, USA; [2]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA; [3]Department of Radiation Oncology, University of Michigan Health System, Ann Arbor, MI, 48109, USA; [4]Departments of Radiation Oncology, Radiation and Cancer Biology Laboratories, Indiana University School of Medicine, Indianapolis, Indiana, 46202, USA; [5]Department of Medical & Molecular Genetics, Indiana University School of Medicine, Indianapolis, Indiana, 46202, USA; [6]Cancer Research UK Cancer Therapeutics Unit, Division of Cancer Therapeutics, The Institute of Cancer Research, London, SW7 3RP, UK

‡ To whom correspondence should be made:
Corey Speers, M.D., Ph.D.
Department of Radiation Oncology
University of Michigan, UH B2 C490
1500 E. Medical Center Dr., SPC 5010
Ann Arbor, MI 48109-5010
Phone: 734-936-4300
E-mail: cspeers@med.umich.edu

Keywords: genomics, data storage, informatics, radiogenomics, biospecimens

Conflict of interest disclosure: The authors have no conflicts to disclose

Introduction

Recent technological advances have allowed for a more sophisticated understanding of the biology of tumors and an ability to generate massive data at an unprecedented pace. These advances now routinely allow for the assessment of genomics (DNA mutations and copy number alterations), transcriptomics (RNA expression levels), methylation profiles and protein and phosphoprotein abundance to unravel the biologic underpinnings of various types of cancers. This information is now being combined with various imaging modalities, histopathology, clinical and patient characteristics, and treatment information to allow for a systems-based approach to understanding and characterizing cancer. With these advances, however, new challenges have emerged in how to acquire, store, catalog, analyze, and integrate these varying types of biologic data. This article will review examples of successful integration of genomic and biologic data, the current state of this research, issues surrounding access, extraction, collection, and curation of the genomic and biospecimens data. It will also suggest recommendations for standardizations and next steps to improve data availability.

With the completion of the human genome project and the subsequent inception and completion of The Cancer Genome Atlas (TCGA) project, the acquisition, storage, and subsequent availability of large-scale genomic, transcriptomic, and proteomic data has led to an accelerated pace of discovery and understanding of cancer[1-4]. This data has led to the development of new, effective targeted agents, and ascertainment of this genomic and biospecimen data is now making its way into routine clinical practice[5]. Indeed, multiple groups have recently published the findings of molecular tumor boards and this molecular data is now beginning to be used, including in the NCI-sponsored MATCH and IMPACT trials, the AACR sponsored GENIE project, and ASCO sponsored TAPUR trial, to inform clinical decision making as it relates to disease prognosis, effectiveness, therapeutic benefit, and mechanisms of treatment resistance[6-8]. Other examples of the successful capture and annotation of genomic and biospecimen data

includes the Encyclopedia of DNA elements (ENCODE) project, and the International Cancer Genome Consortium (ICGC) project.

While the benefits of this molecular data in areas such as targeted drug development are increasingly clear, it's utility to predict radiation treatment toxicity and therapeutic response remains uncertain. While there are many reasons for this disparity, multiple initiatives including the REQUITE, RAPPER, Gene-PARE, RadGenomics, and canSAR projects are currently underway to collect, catalog, and make available this information[9-13]. The success of these radiation-associated databases, and subsequent projects, however, will depend on the ability for these databases to be accessed, annotated, integrated, and updated.

State of the research

The acquisition and storage of genomic and biospecimen data is currently the exception, not the rule in radiation oncology clinical practice. When this information is gathered, it is usually for research purposes with variable translatability into clinical practice. While the reasons for this lack of sample collection are numerous, a major limitation to specimen collection is the requirement that it be prospectively incorporated into research and non-research protocols. The collection and analysis of patient derived biospecimens requires institutional review board (IRB) approval. This approval, in turn, is dependent on a clearly formulated rationale for collecting the information, and safeguards regarding the utilization of the information and protection of potentially identifiable information. This requires foresight, resources (monetary, staff, and space-related resources), and patient and physician buy in. Some groups have begun to address this challenge by creating "boilerplate" language that can be incorporated into the standard consenting process for any patient undergoing radiation treatment. This consent includes language that allows for the de-identified patient genomic data to be used for research purposes, and is easily included in prospective trials as well as the regular clinic workflow [14]. Efforts to make this language and consent template more widely available are already underway.

An important area of radiation oncology research utilizing big data is radiogenomics, whose goal is the identification of genomic markers that are predictive

for the development of outcomes resulting from cancer treatment with radiation[15]. Work in radiogenomics has greatly benefited from creation of the Radiogenomics Consortium (RGC). The RGC was created in 2009 and is a cancer epidemiology consortium through the Epidemiology and Genomics Research Program of the NCI of the NIH[16]. The RGC now has 225 member investigators located at 131 medical centers in 32 countries. The common goal of the RGC membership is to share biospecimens and data so as to achieve large scale studies with increased statistical power to enable identification of relevant genomic markers. However, in order to accomplish this work and definitively discover and validate the critical genomic markers, access to the radiotherapy treatment information and long-term longitudinal follow-up data reporting details such as outcomes must be obtained for large numbers of patients. The RGC does not maintain a centralized biorepository, but serves to facilitate the development of collaborations between investigators with similar research goals who have assembled cohorts and collected data that can be synthesized into one large study. Although the RGC has successfully assembled large cohorts to perform adequately-powered studies, data harmonization remains a challenge for studies involving multiple patient cohorts treated with a variety of radiotherapy techniques and evaluated using multiple grading systems [15]. Although a proposed set of reporting requirements have been promulgated for research in radiogenomics, it would be advantageous if identical, or at least similar, case report forms were utilized for all radiogenomic research[17].

An important example of the research projects launched by RGC investigators is the large multi-center REQUITE study (Validation of predictive models and biomarkers of radiotherapy toxicity to reduce side-effects and improve quality-of-life in cancer survivors) [10]. 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. The objectives of REQUITE are to: (1) Perform a multi-center, observational cohort study in which epidemiologic, treatment, longitudinal toxicity and quality-of-life data are collected from approximately 5,000 patients treated with radiotherapy for either breast, prostate or

lung cancer. (2) Produce a centralized biobank in which DNA is isolated from patients enrolled in the observational study and create a centralized data management system for secure collection, integration, mining, sharing and archiving of all project data. A key aspect of the centralized database is that it includes pre-treatment DICOM and DVH files. (3) Validate published SNP biomarkers of radiosensitivity and discover new variants associated with specific outcomes following radiotherapy, (4) Validate clinical/dosimetric predictors of radiotherapy toxicity and incorporate SNP biomarker data. (5) Design interventional trials to reduce long-term adverse cancer treatment effects. (6) Deliver interventional trial protocols using validated models incorporating biomarkers to identify patient sub-populations likely to benefit from interventions. (7) Serve as a resource exploitable for future studies exploring relationships between genetics and radiotherapy outcomes using developing technologies such as next generation sequencing. Those interested in becoming a member investigator of the RGC should contact Barry Rosenstein via email at barry.rosenstein@mssm.edu.

Access and Extraction

With the decreasing costs and increasing availability of DNA and RNA sequencing, protein expression and metabolite assessment, the amount of data generated per patient continues to increase. Despite the increase in the availability of this data, the subsequent capture of this information for anything other than to answer a specific research question or direct a clinical treatment decision remains extremely limited. While the reasons are myriad, the amount and complexity of the data is a major factor. It is now common for germline and somatic testing of patients and tumors to include at least some of the following assessment: DNA sequencing assessing germline and/or somatic mutations and copy number variation, single nucleotide polymorphism (SNP) assessment, RNA expression (either through RNA sequencing or gene expression microarrays, epigenomic assessment, proteomic assessment (through mass spectrometry or reverse phase protein lysate arrays), metabolomic assessment, and pathologic assessment of tumor samples (through immunohistochemical staining, flow cytometry assessment, or though the creation of institutional or multi-institutional tissue microarrays (TMAs). In addition to the sheer volume of biological molecules being

assessed, the methods to analyze, interpret, and report the data is also varied. Issues of DNA sequencing read depth and sequence mapping for sequencing data continue to confound analyses of such data. Variation in algorithms utilized for DNA sequence mapping, variance and allelic calls also contributes to the complexity and heterogeneity in type and quality of the data. One must decide whether to collect and store pre-processed vs. raw data (i.e. normalized expression data vs. CEL files for expression microarrays or FASTQ vs. BAM files for DNA sequencing data). As an example, when raw DNA sequencing information is obtained, it usually comes from the DNA sequencer as FASTA or FASTQ files. FASTA and FASTQ format is based on simple text and contains the raw data of each sequence read. For FASTA files, each sequence starts with a ">" followed by the sequence name, a space and, optionally, the description. In addition, a separate FASTA file will include the quality information of the given read sequences. FASTQ files were developed to provide a convenient way of storing the sequence and the quality scores in the same text-based file. It bears noting that depending on the sequencing technique (Sanger vs. Illumina sequencing), different FASTQ files are generated based on the different ways in which quality is assessed between Sanger and Illumina sequencing. Because of this difference, the source of FASTQ data should be noted when storing the data as the encoding for the quality scores is diferent between Sanger and Illumina sequencing. In addition, paired reads are now routinely generated in which two reads are generated from the same single molecule to aid in sequence alignment. In this case (paired-reads data) two FASTQ files are created, one for the first read of the pairs and another one for the second, and the files should hold the reads exactly in the same order. Moving beyond simple read sequences FASTA or FASTQ files, alignment (SAM/BAM) and variation (VCF) files can also be created. Sequence Alignment Map (SAM) files were first created to store not only sequence and quality data (like in FASTQ files), but also mapping information for the sequences (i.e where does each sequence align on the genome). To capture this more complex data, SAM files are tab-based and include 11-12 fields that fill one line and may include a header. SAM can express the same information as FASTQ, but also includes mapping information (see https://samtools.github.io/hts-specs/SAMv1.pdf for more information). SAM is rarely used as the format for data storage, instead, files are

stored in binary alignment mapping (BAM) format, which is a compact binary representation of SAM. It stores the same information, just more efficiently, and in conjunction with a search index, allows fast retrieval of individual records from the middle of the file. Because of its binary nature, BAM files are also much more compact than compressed FASTQ or FASTA files. Thus, when considering storage of genomic data once must decide upon file format storage (raw data in FASTQ vs. processed and mapped data in BAM). Finally, there are data access (and limitations to access to preserve data security and patient anonymity) and extraction issues that have made the wide-spread availability of this information a challenge.

While there are no quick or easy solutions to these challenges, many groups have already grappled with these questions and found useful solutions. For example, in the case of the REQUITE trial, standardized case report forms were developed for data collection of epidemiological and patient characteristics. Collection of clinical/pathologic, physics and treatment data were also standardized. Of critical importance, the full radiotherapy dose volume histogram was obtained for each subject, which provides substantial detailed dosimetric data. Data collection forms were provided in the different languages of the patients located in the multiple countries where they were enrolled into the study. Paper and web-based submission methods were provided in parallel. Submitted data underwent centralized quarterly quality control and plausibility checks for quality assurance according to a standardized quality assurance protocols. The database was enhanced to enable sample tracking in conjunction with the biobank information system, and empowered with user friendly interfaces to enable flexible data mining and data downloads in various formats. The database is only accessible to authorized persons via network and database passwords.

Collection and Curation

Important lessons can be learned from the challenges facing data extraction from health system-wide electronic medical records (EMR). Natural language processing (NLP) is part of a solution to extract data that is only available in free-text fields in EMRs. We are at a junction where development of a standardized format for collection

and storage of genomics data (i.e all BAM format) could potentially save significant resources in connecting genomics data to patient outcomes and dosimetric data. Independent validation is essential in the path towards the robust use of genomics data in clinical practice. By standardizing our clinical data collection, we can accelerate the discovery of which data are the most beneficial for specific classes of patients. One potential opportunity for increased capture and curation is to integrate with commercial (Flatiron) or organizational (CancerLinQ) platforms. These groups are already invested in data integration from EMR systems, and the increasing amount of clinically and commercially available genomic and biospecimen testing results may be extracted using these platforms. While these commercial and organizational platforms are still in their infancy, early integration into these groups may eliminate some of the challenges with later-stage integration. In addition, initial discussions with these groups could lead to standardization of collection and storage that could lessen, if not eliminate, subsequent challenges when the data is accessed/prepared for analysis. Certainly issues to consider in these initial discussions include: which format should be used to store data and whether raw or normalized data should be collected; should the data be normalized and if so which technique will be used; can EMRs be reconfigured to host and handle this genomic and biospecimen data; should biospecimen data be built into EMRs from radiation treatment unit vendors including Varian and Elekta (with Aria and Mosaiq) and how do we limit redundancy or discrepant data in these biospecimen data sets. While the answers to these questions are not immediately obvious, working group consensus and advocacy will allow for a clearer path forward as we seek to collect and curate genomic and biospecimen data.

Specific Recommendations for standardizations

The utility of genomic and biospecimen data collection and utilization will depend heavily on the quality and completeness of the data collected. In an ideal world, this data would be automatically collected and seamlessly integrated into other databases of collected data (patient outcomes, comorbidities, toxicity, dosimetric and treatment related information, etc.). While this is unlikely to be reality in the near term, the

following recommendations will allow for the gradual transition to this new reality. These recommendations include:

1. Pool genomics and bio-specimen analysis templates among centers active in genomics for clinical research so that common features are universally captured and similarly named for ease of extraction in the future. Appropriate batch effect corrections across sample acquisiton and preparation sites would be necessary prior to data collation[18] .
2. Develop a standard nomenclature for data collection. Similar to the TG-263 task group, the formation of a similar task group to standardize genomic and biospecimen data nomenclature and reporting would significantly aid in this process.
3. Harmonize the preferred format for standard fields to store genomics data within the hospital EMR with appropriate patient privacy safeguards built in. When housing within the EMR is not practical/feasible, uploading of genomics data with clinical outcomes and de-identified patient information into cBioPortal should be done (http://www.cbioportal.org)
4. Publish the recommendations such that individual institutions can request the major EMR vendors implement those standard fields. Concentrated and consistent pressure by end-users is likely to be more effective in implementing change than scatter shot, disjointed requests.
5. Identify institutions that would be able to perform validation of another institution's results through a standard data query. Data standardization and completeness is a key limitation on integrating this more globally, and quality assurance measures and standardized operating procedures that are universally available and implemented will be key to subsequent data utilization and integration.
6. Following examples like TCGA, the Sharing of standardized analysis pipelines enable the communciaton of "best practices" for concordant, reproducible and rigorous data analysis. Methods that enable the models to learn across institutional cohorts (i.e distributed learning), rather than requiring the data to be centrally stored can create viable alternatives for effective data learning and interpretation while being cognizant of potential privacy concerns[19].

Recommendations for next steps

Develop and conduct a survey to determine the state, quantity and quality of genomics and bio-specimens data in hospital EMRs

In order to successfully fix a problem, one must first effectively identify and define said problem. Before the integration of genomic and biospecimen data can become a reality, one must first understand the present barriers and limitations in real-world terms. A survey that includes academic and industry participants, data generators and end users is critical to further identifying and then understanding the problem. The results of this initial survey will provide the basis for subsequent task group's efforts as they seek to assist the integration of genomic and biospecimen data into the radiation oncology space. Involvement of health ethicists and geneticists as well as health policy experts will also be key in navigating issues surrounding the housing of genomic data within an EMR (including health insurance and employer privacy concerns as well as protocols for notification should actionable germline mutations be identified.

Share institutional best practices in data collection and storage and identify institutions, organizations, and companies who are willing to share current data templates

While an effort that begins by trying to capture all data at all institutions on all patients is unlikely to be successful in the near term, an effort that includes multi-institutional and multi-tiered (i.e. academic, organizational, and industry) collaborations is likely to help move the field towards this greater goal. Critical to the successful advocacy for the integration of this data is receiving the support of large organizations already operating within this space. This includes dialogue with and endorsement from the American Society for Radiation Oncology (ASTRO), the American Society of Clinical Oncology (ASCO), the NRG, the National Institutes of Health (NIH), the National Cancer Institute (NCI) and the Global Alliance for Genomics in Health. As the sources for genomic and biospecimen data becomes increasingly available and complex, the inclusion of all parties associated with the data generation and usage in these

collaborations will be important. By identifying those groups that both have an interest in the integration of this data and who "touch" the data on a daily basis, potential pitfalls will be more readily identified, and avoided, as this process continues.

Develop and publish the harmonized template to standardize data collection, generation and analysis to facilitate connection to patient outcome and dosimetric data

As was noted earlier in the article, the formation of a task group to address issues pertaining to standardized collection and nomenclature will be crucial to the successful integration of genomic and biospecimen data into radiation oncology treatment paradigms. One of the mandates of this working group will be the publication of recommended templates and nomenclature standardization that will allow for the data to be universally accessed and utilized. The publication of uniform access requirements and sharing of "Best Practices for Data Collection" will be critical to the success of this project. Similar efforts for establishing standardized analysis templates (for variant interpretation, gene expresison analysis etc), will be essential to create datasets amenable to sharing and joint mining in the context of corresponding imaging and outcome data.

Final considerations

In addition to the previously noted "next steps", integration "discovery" and "validation" pipelines into the workflow will enable the more effective utilization of this data in the future. By carefully considering the collection and partitioning of these "discovery" and "validation" cohorts, subsequent integration of findings utilizing genomics and biospecimen data will be expedited. Critical to this collection, curation, and storage is the need for data housing standardizations that are HIPAA compliant and removes patient identifiable information. This also includes the need to incorporate our medical ethicist colleagues to consider the ethical issues surrounding the acquisition, storage, and reporting of genomic and biospecimen data.

Conclusions

As we continue to translate the use of genomics data to guide treatment decisions for individual patients, we have an opportunity to accelerate this translation by developing and applying standard templates for data collection. By standardizing, they can be used to more robustly connect genomics and bio-specimen data directly to patient outcomes and dosimetric data. There is a lot of enthusiasm for how standardized nomenclature for organs-at-risk and targets will accelerate the analysis of dose and patient outcomes (AAPM TG-263)[20]. Similar potential exists within the collection, annotation, and storage of genomic and biospecimen space. Initial steps should include: standardizing nomenclature for data collection and harmonizing format of data collection and entry, pressuring EMR-vendors to build genomic and biospecimen data collection into the EMR platform, and establishing a task-group to generate specific guidelines governing the collection, analysis and reporting of genomic data. Successful completion of these steps will allow genomic and biospecimen data to be integrated into future data analysis as we seek to improve treatment efficacy and limit normal tissue toxicity.

References:

1. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature.* 2001;409(6822):860-921.
2. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature.* 2008;455(7216):1061-1068.
3. Comprehensive molecular portraits of human breast tumours. *Nature.* 2012;490(7418):61-70.
4. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012;489(7414):57-74.
5. Bombard Y, Bach PB, Offit K. Translating genomics in cancer care. *J Natl Compr Canc Netw.* 2013;11(11):1343-1353.
6. TAPUR: Testing the Use of Food and Drug Administration (FDA) Approved Drugs That Target a Specific Abnormality in a Tumor Gene in People With Advanced Stage Cancer

(TAPUR). 2017; https://clinicaltrials.gov/ct2/show/NCT02693535. Accessed 9-25-2017, 2017.

7. Coyne GO, Takebe N, Chen AP. Defining precision: The precision medicine initiative trials NCI-MPACT and NCI-MATCH. *Current problems in cancer.* 2017;41(3):182-193.
8. AACR Project GENIE: Powering Precision Medicine through an International Consortium. *Cancer discovery.* 2017;7(8):818-831.
9. Burnet NG, Barnett GC, Elliott RM, et al. RAPPER: the radiogenomics of radiation toxicity. *Clinical oncology.* 2013;25(7):431-434.
10. 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. *Clinical oncology.* 2014;26(12):739-742.
11. Ho AY, Atencio DP, Peters S, et al. Genetic predictors of adverse radiotherapy effects: the Gene-PARE project. *Int J Radiat Oncol Biol Phys.* 2006;65(3):646-655.
12. Iwakawa M, Imai T, Harada Y, et al. [RadGenomics project]. *Nihon Igaku Hoshasen Gakkai zasshi Nippon acta radiologica.* 2002;62(9):484-489.
13. Tym JE, Mitsopoulos C, Coker EA, et al. canSAR: an updated cancer research and drug discovery knowledgebase. *Nucleic Acids Research.* 2016;44(D1):D938-D943.
14. Roychowdhury S, Iyer MK, Robinson DR, et al. Personalized Oncology Through Integrative High-Throughput Sequencing: A Pilot Study. *Science translational medicine.* 2011;3(111):111ra121-111ra121.
15. Rosenstein BS. Radiogenomics: Identification of Genomic Predictors for Radiation Toxicity. *Semin Radiat Oncol.* 2017;27(4):300-309.
16. West C, Rosenstein BS. Establishment of a radiogenomics consortium. *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology.* 2010;94(1):117-118.
17. Kerns SL, de Ruysscher D, Andreassen CN, et al. STROGAR - STrengthening the Reporting Of Genetic Association studies in Radiogenomics. *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology.* 2014;110(1):182-188.
18. Leek JT, Scharpf RB, Bravo HC, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature reviews Genetics.* 2010;11(10):733-739.
19. Jochems A, Deist TM, van Soest J, et al. Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital - A real life

proof of concept. *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology.* 2016;121(3):459-467.

20. Mayo C, Moran JM, Xiao Y, et al. AAPM Task Group 263: Tackling Standardization of Nomenclature for Radiation Therapy. *International Journal of Radiation Oncology • Biology • Physics.*93(3):E383-E384.