Community science and reaching the promise of big data in health care

> Alone we can do so little; together we can do so much - Helen Keller

At its core, the concept of "big data" in health care embraces the promise of creating transcendent knowledge generation systems using the power of information gathered from routine processes for all patients. By harnessing large-scale aggregation of information generated during routine healthcare delivery to the speed and capacity of machine learning (ML) and artificial intelligence (AI) algorithms, "big data" pioneers believe we can reduce research costs, deepen our understanding of factors affecting patient outcomes and improve patient care. Ability of ML and AI to identify and characterize interactions among sets of variables and cohorts of patients, much larger than humans are capable of analyzing, should deepen our understanding key factors affecting outcomes and quality of care. These are early days; with substantial amounts of foundational work needed to reach that promise. Part of that foundation includes improving standardizations for quantifying data elements and building systems to increase the volume of quality data that may be consumed by these data hungry ML and AI algorithms.

By analogy, consider our love of fast, powerful and convenient cars. Without the foundational work of constructing roads, bridges, fuel stations and other supports, connections and logistics plus development of science and engineering underlying design, the reality of driving would not be possible. A myriad of standardizations (eg, diameter of nozzle on fuel pump at gas station, traffic laws, road design, regulatory standards) enable us to just focus on driving, without having to also grapple with endless variations in key details. These elements did not emerge quickly and fully formed from the minds of a handful of people. Instead, they evolved, gradually out of the combined trial and error iterations of communities of enthusiasts to find a common solution.

Similarly making the promise of "big data" a practical, routine part of our clinical reality will be an outgrowth of what we are able to do together as a community to build needed core concepts (eg, clinically linked measures of ML/AI algorithm reliability) and standardizations (eg, nomenclatures, ontologies, toxicity measures, disease site status/recurrence categorizations). Practice standardizations (eg, how recurrence information is entered into a treatment note) enable our electronic systems to make distinctions among data elements that can then be fed accurately, rapidly and in large volume to learning algorithms. In our cars, standardizations let us take for granted the ability to drive up to any gas station to fuel our travels. By contrast, lack of standardized categorizations and entry processes in our clinics, means we cannot take for granted the ability to electronically extract accurate information on treatment outcomes, treatment variables, and relevant patient host variables from available electronic health records to fuel our treatment outcomes modeling.

Furthermore, if we aspire to eventually understand global patterns of care and treatment outcomes for all cancer patients treated by our healthcare systems, as opposed to outcomes of limited patient cohorts accrued at a relatively small number of centers, then our communities and the foundational work required of them, must expand beyond the scale of a few institutions. It requires that we move toward community science, where collaborations spanning multiple institutions, clinic sizes and national borders are recognized as key factors for success in supporting creation of practical enabling standardizations, ontologies, algorithms and processes. While this principle is recognized in clinical trials the magnitude of cooperation needed to scale "big data" to the majority of patient treated is different. Furthermore, comparable funding, as well as institutional and academic supports needed for these foundational "big data" efforts often are not evident. Using our analogy, it is often easier to get support for designing a futuristic car than for constructing roads and bridges.

One recent example of community science is the American Association of Physicists in Medicine's (AAPM's) Task Group 263 (TG-263) on Radiation Oncology Nomenclature.[1] This task group worked with a large and diverse group of 57 physicians, physicists, industry representatives and others, drawn from large clinics and small, academic, and non-academic centers, the AAPM, the American Society of Therapeutic Radiation Oncology (ASTRO), European Society for Radiation Oncology (ESTRO), NRG and IHE-RO and other stakeholder groups. The Task Group created and piloted a proposed set of nomenclature standardization recommendations designed to improve the ability to electronically extract and use large data sets of dosimetric data to support "big data" efforts. For example, when analyzing the history of treatment plans at an institution one often finds dozens of character combinations that are used to represent each organ at risk (eg, left optic nerve). Once the nomenclature is in place, with just one recommended naming for each structure, it is possible to automate accurate, routine extraction of large volumes of dosimetric data on treatment planning structures for analysis. An important lesson from that effort was the vital role that professional societies play in supporting and endorsing these efforts.

This special issue of *Medical Physics* is another example of a community effort to bring the promise of "big data" closer to reality. The first Practical Big Data Workshop (PBDW), held at the University of Michigan in Ann Arbor in 2017, was an effort to promote coalescing the nascent community of builders and users of "big data" systems for cancer care. Shared recognition of a common set of challenges and need for consensus solutions gave rise to a set of papers to share these perspectives with the larger community. Because cancer care is positioned at a cross roads with many health care specialties, the approaches taken have applicability beyond Radiation Oncology and Imaging.

As healthcare builds slowly toward a reality where "big data" and analytics become more routine elements of clinical practice, there are significant implications for the training and credentialing of healthcare professionals. A lesson from the PBDW is that physicians, physicists and others must blend their Radiation Oncology domain knowledge with skill sets from other domains (eg, informatics, application development, machine learning, ethics, genomics, radiomics, etc.). New combinations of skill sets, for example, physician-ethicists, physician-informaticists, physicist-data scientists, and physicist-database designers, played vital roles at the meeting in identifying challenges, formulating solutions and effectively communicating these challenges and solutions to the wider community.

Realizing the promise of "big data" in health care will be most effectively approached if we work together as communities to transcend boundaries that separate institutions, professional identities, and differently structured clinical service lines. Embracing both the need for expanding the range of skill sets outside our traditional health care training curricula and the importance of building networks of collaborators will enable us to build the strong foundation needed for knowledge generating systems to emerge.

The papers in this special issue span a wide range of subject areas encountered in meeting specific challenges, solutions and collaborative efforts which are part of reaching the potential of "big data" in radiation oncology.

In "Treatment Data and Technical Process Challenges for Practical Big Data Efforts in Radiation Oncology", Mayo et al. address several factors affecting many key data elements.[2] These include: need for process and system changes to improve quality and availability of key data elements and relationships, access and extraction issues for obtaining data from various source systems used in patient treatment, selection considerations for database technologies, review and comparison of clinical data repositories, specific recommendations for workflows and standardizations, examination of next steps needed to improve data availability. In addition, the appendix of this manuscript details a translational research ontology that specifies core data elements and relationships important to a broad range of patient quality improvement and translational research efforts. Their recommendations for improving clinical process include: more complete and consistent utilization of diagnosis and staging tools in radiation oncology information systems, implementation of TG-263 standardizations for nomenclature, routine creation of as treated plan sums in treatment planning systems that reflect all dose delivered in the treatment course and standardized entry of patient reported outcomes and provider reported toxicities into the electronic record.

Matuszak et al. focus on the efforts and challenges for aggregation of outcomes information in their manuscript, "Performance/Outcomes Data and Physician Process Challenges for Practical Big Data Efforts in Radiation Oncology".[3] Building from a detailed examination of the "big data" projects of 8 groups, they examine common issues affecting data availability, access, and quality. They provide specific recommendations for improvements through standardized workflows and discuss need for multi-institutional consensus based standards for classifying recurrence categorizations.

In "Genomics, Bio specimens and other Biological Data: Current status and future directions“ by Rosenstein et al., challenges for the use of genomic and bio-specimen data are examined.[4] Acquisition and storage of this key data element is currently the exception. They examine the state of large-scale research efforts to using this data element, challenges for access and extraction, issues for collection and curation and provide specific recommendations for standardizations aimed at reducing barriers to more wide spread, routine use of this data to support modeling patient outcomes. Recommendations include developing a standardized nomenclature to reduce variability in collecting genomics and bio-specimen data, developing standardizations through multi-institutional and vendor collaborations to improve interoperability, increasing the frequency of multi-institutional data pooling, and harmonizing approaches for encapsulating this information in the EHR.

Mackie et al. deal with challenges for "big data" in aggregation of imaging information, radiomics measures and analysis of quantitative images to find biomarkers of disease, in "Opportunities and Challenges to Utilization of Quantitative Imaging: Report of the AAPM Practical Big Data Workshop".[5] They address challenges in information curation that stand as obstacles to medicine transforming itself into a "knowledge-based" discipline, carefully referencing impacts on clinical trials and NCI funded imaging consortia. Highlights from their recommendations include need: for Radiology practices embrace the needs of Oncology for more detailed quantitative imaging features, to include more quantitative measurement data in images, to improve standardized radiology and oncology workflows, to add more quantitative information on image features as part of routine practice, to improve quantitative imaging reproducibility, accuracy and curation, and to examine approaches to regulation of imaging biomarkers.

In "Machine Learning and Modeling: Data, Validation, Communication Challenges", El Naqa at el highlight the potential of ML and AI for clinical advancement, while also addressing pitfalls when applying these powerful analytic tools.[6] They discuss common issues requiring careful consideration including proper use of analysis metrics, sufficient volume and quality of data in training sets, parsimony and generalizability of models, quality assurance, and clinical interpretability of results. Recommendations include, establishing standardized clinically relevant objective criteria for evaluating ML results, constructing publically available benchmark data sets to validate and cross check models, using resampling techniques to estimate model performance, and benchmarking changes in predictive performance of ML models using new biomarkers against with comparison to standard clinical factors.

The move to construction of learning health systems requires careful consideration of ethical obligations to patients, construction of informed consent and addressing inconsistencies and variable interpretations of the regulatory environment. Spector-Bagdady and Jagsi provide much need guidance and perspective for addressing these challenges in “Big Data, Ethics, and Regulations: Implications for Consent in the Learning Health System”.[7]

Traverso et al. discuss their extensive experience with multi-institutional data sharing practices in “The Radiation Oncology Ontology (ROO): publishing linked data in radiation oncology using Semantic Web and Ontology techniques”.[8] Use of standardizations and design of scalable “big data” systems are important principles for making data sets Findable, Accessible, Interoperable and Reusable (FAIR). They discuss their use of the ROO with semantic web technologies to meet these goals.

United States based clinicians and researchers may be unfamiliar with the extensive efforts in Canada to improve the landscape for “big data”, as part of improving safety and practice quality initiatives. In “Improving Patient Outcomes and Radiotherapy Systems: A Pan-Canadian Approach to Patient Reported Outcome Use”, Caissie et al. provide a review of these efforts.[9] Among them, the Canadian Partnership for Radiation Therapy (CPQR) combines the work of several groups including Association of Radiation Oncology (CARO), Canadian Organization of Medical Physicists (COMP), and the Canadian Association of Medical Radiation Technologists (CAMRT). The work of CPQR in promulgating standardizations (eg, TG-263) and key quality indicators is discussed. In addition, their innovative work developing standardized approaches to administration and use of patient reported outcomes (PROs) across Canada is presented.

In “Practical data collection and extraction for Big Data applications in radiotherapy”, McNutt et al. discuss their experience with overcoming practical challenges to capture of high quality treatment and outcomes data.[10] Detailed examinations of factors affecting: clinician assessments, PROs, bio-specimen, imaging, treatment and symptom management are discussed. They review approaches to technology and clinical implementation they have used to address these challenges.

In “Perspectives on potential research benefits from big data efforts in Radiation Oncology”, Vikram discusses several themes that frequently emerge in research studies that may be positively affected by these “big data” efforts.[11] Specific challenges facing radiation oncology and areas that “big data” researchers should try to address are discussed.

Wei et al. examine the impact of “big data” efforts and supporting standardizations on clinical trials in “Implementation and enforcement of the standardization for radiotherapy with protocol guidelines, libraries and software systems assure the clinical trial data quality”.[12] They share their in-depth perspective on implementation details of management tools in the several network groups of the National Clinical Trials Network (NCTN). From this perspective, they underscore the significant overlap of standardization recommendations highlighted throughout this special issue with NCTN objectives.

We hope that you will find the manuscripts in this special issue helpful in your personal journey and entry into this growing community of practical big data in health care.

Charles Mayo
University of Michigan, Ann Arbor, MI, USA

REFERENCES

1. Mayo CS, Moran JM, Bosch W, et al. American Association of Physicists in Medicine Task Group 263: standardizing nomenclatures in radiation oncology. *Int J Radiat Oncol Biol Phys*. 2018;100:1057–1066.
2. Mayo CS, Phillips M, McNutt T, et al. Treatment data and technical process challenges for practical big data efforts in radiation oncology. *Med Phys* 2018 (In Press).
3. Matuszak MM, Fuller CD, Yock T, et al. Performance/outcomes data and physician process challenges for practical big data efforts in radiation oncology. *Med Phys* 2018 (In Press).
4. Rosenstein B, Rao A, Moran JM, et al. Genomics, bio specimens and other biological data: current status and future directions. *Med Phys* 2018 (In Press).
5. Mackie TR, Jackson1 EF, Giger M. Opportunities and challenges to utilization of quantitative imaging report of the AAPM Practical Big Data Workshop. *Med Phys* 2018 (In Press).
6. El Naqa I, Ruan D, Valdes G, et al. Machine learning and modeling: data, validation, communication challenges. *Med Phys*. 2018 https://doi.org/10.1002/mp.12811.
7. Spector-Bagdady K, Jagsi R. Big data, ethics, and regulations: implications for consent in the learning health system. *Med Phys*. 2018 https://doi.org/10.1002/mp.12707.
8. Traverso A, van Soest J, Wee L, Dekker A. The Radiation Oncology Ontology (ROO): publishing linked data in radiation oncology using semantic web and ontology techniques. *Med Phys*. 2018 https://doi.org/10.1002/mp.12879.
9. Caissie A, Brown E, Olson R, et al. Improving patient outcomes and radiotherapy systems: a pan-Canadian approach to patient reported outcome use. *Medl Phys*. 2018 https://doi.org/10.1002/mp.12878.
10. McNutt TR, Bowers M, Cheng S, et al. Practical data collection and extraction for big data applications in radiotherapy. *Med Phys* 2018 https://doi.org/10.1002/mp.12817.
11. Vikram B. Perspectives on potential research benefits from big data efforts in radiation oncology. *Med Phys*. 2018 https://doi.org/10.1002/mp.13109.
12. Zou W, Geng H, Teo BK, Finlay J, Xiao Y. NCTN clinical trial standardization for radiotherapy through IROC and CIRO. *Med Phys*. 2018 https://doi.org/10.1002/mp.12873.