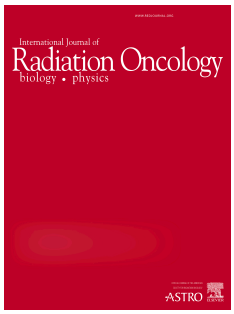
Accepted Manuscript



Using Big Data Analytics to Advance Precision Radiation Oncology

Todd R. McNutt, Stanley H. Benedict, Daniel A. Low, Kevin Moore, Ilya Shpitser, Wei Jiang, Pranav Lakshminarayanan, Zhi Cheng, Peijin Han, Xuan Hui, Minoru Nakatsugawa, Junghoon Lee, Joseph A. Moore, Scott P. Robertson, Veeraj Shah, Russ Taylor, Harry Quon, John Wong, Theodore DeWeese

PII: S0360-3016(18)30327-4

DOI: 10.1016/j.ijrobp.2018.02.028

Reference: ROB 24806

To appear in: *International Journal of Radiation Oncology • Biology • Physics*

Received Date: 30 November 2017

Revised Date: 13 February 2018

Accepted Date: 20 February 2018

Please cite this article as: McNutt TR, Benedict SH, Low DA, Moore K, Shpitser I, Jiang W, Lakshminarayanan P, Cheng Z, Han P, Hui X, Nakatsugawa M, Lee J, Moore JA, Robertson SP, Shah V, Taylor R, Quon H, Wong J, DeWeese T, Using Big Data Analytics to Advance Precision Radiation Oncology, *International Journal of Radiation Oncology • Biology • Physics* (2018), doi: 10.1016/j.ijrobp.2018.02.028.

Using Big Data Analytics to Advance Precision Radiation Oncology

Todd R McNutt[1], Stanley H Benedict[2], Daniel A Low[3], Kevin Moore[4], Ilya Shpitser[5], Wei Jiang[1], Pranav Lakshminarayanan[1], Zhi Cheng[1], Peijin Han[1], Xuan Hui[6], Minoru Nakatsugawa[7], Junghoon Lee[1], Joseph A Moore[1], Scott P Robertson[1], Veeraj Shah[1], Russ Taylor[5], Harry Quon[1], John Wong[1], Theodore DeWeese[1]

[1]Department of Radiation Oncology and Molecular Radiation Sciences, Johns Hopkins University, Baltimore, MD

[2]Department of Radiation Oncology, University of California-Davis, Sacramento, CA

[3]Department of Radiation Oncology, University of California-Los Angeles, Los Angeles, CA

[4]Radiation Medicine and Applied Sciences, University of California-San Diego, La Jolla, CA

[5]Department of Computer Science, Johns Hopkins University

[6]Department of Public Health Sciences, University of Chicago

[7]Toshiba Medical Systems Corporation

Please send communications and reprint requests to:

Todd R. McNutt
Department of Radiation Oncology
Harry & Jeanette Weinberg Bldg.
401 North Broadway, Suite 1440
Baltimore , MD 21231---2410
Tel: (410) 614---4594
Email: tmcnutt1@jhmi.edu

Conflicts of interest

Todd McNutt has funded collaborations with Philips and Toshiba/Cannon

Kevin Moore has funded collaborations with Varian Medical Systems

Abstract:

Big Clinical Data Analytics as a primary component of precision medicine is discussed,] identifying where these emerging tools fit in the spectrum of genomic and radiomic research. A learning health system (LHS) is conceptualized that utilizes clinically acquired data with machine learning to advance the initiatives of precision medicine. The LHS is comprehensive and can be used for clinical decision support, discovery, and hypothesis derivation. These developing uses can positively impact the ultimate management and therapeutic course for patients. The conceptual model for each use of clinical data, is however different, and an overview of the implications is discussed. With advancement in technologies and culture to improve the efficiency, accuracy and breadth of measurements of the patient condition, the concept of a LHS may be realized in precision radiotherapy.

1. Introduction

The goal of precision medicine is to improve overall patient care and determine when and how to personalize patients' treatments. Currently, this is guided by a physician's understanding of the patient's condition by drawing from their experience to align the specifics of care to the patient. Guidelines[1,2] assist in the overall pathways for specific diseases, but, for the most part, precision medicine is performed with finer granularity than the guidelines provide.

A learning health system[3-5] (LHS) is a concept where quantifiable diagnostic, treatment and outcome data are captured from a continuous stream of patients and placed in a knowledge base. Knowledge is accessed by analytical tools that employ statistical and machine learning algorithms to present trends and make predictions and causal inferences on outcomes. As more

patient data are accumulated, the system continues to learn and improve on its models and ability to make specific predictions for individual patients.

Evaluating the possibilities of a LHS, it is important to recognize the difference between predictive modeling to assist in clinical decisions and knowledge discovery of the underlying mechanisms or causes of particular outcomes. In decision making, we decide on the most appropriate intervention for the patient which may or may not be guided by complete knowledge of the underlying biological mechanisms. New discovery however, must uncover the biological understanding or derive hypotheses that may be further validated under more controlled studies. Clinical data complements pathology, genomics and radiomics by providing details of the treatments and outcomes of patients for the advancement of precision medicine.

2. What is Big Clinical Data?

The ability of bigclinical data[6],, to represent the real world with minimal bias, to accumulate assessments over time, to be linked with other databases, to be used and reused, and to enable a multidimensional understanding, should all be considered to unlock the potential. Clinical data represent prior experience from patients and are captured through a multitude of methods, but limitations of our current protocols and pathways result in only a small fraction being used to make clinical decisions. For machine learning and statistical algorithms to take advantage of the entirety of the available data, medical records must adapt to support continuous feature extraction. Clinical data generally have a number of complications not found in typical cross-sectional study datasets. For example, clinical data exists in forms of free text to three-dimensional volumes to structured data elements all with longitudinal sampling. Clinical data also suffer from selective sampling, missingness, and measurement error.

Aside from lifestyle covariates, clinical data contains patient and disease status, treatment and symptom management, clinical and quality of life (QoL) outcomes, adverse effects, and survival. The key for enabling access is to extract meaningful information or features and store them in standardized ways.[7]

Naturally, the level of precision in measuring outcomes dictates the quality of subsequent clinical conclusions. For instance, in current practice, a recurrence of a patient's cancer may be recorded, but often without the specific location. This limits our understanding of whether or not the recurrence was coincident with the radiation treatment. Also, the measurement of a patient's clinical condition depends on available time and resources. For example, xerostomia can be scored by the clinician, evaluated through patient questionnaires or measured with controlled stimulation methods, each with a corresponding increased time and cost.

Longitudinal assessment of patient status requires careful feature extraction. One can evaluate acute changes in toxicity such as taste disturbance or mucositis during treatment to understand a patient's ability to cope with treatment. Alternatively, evaluating longer term toxicities provides a measure of permanent damage. Time to recovery of a particular function may also be measurable, as initial injury likely has different causal attributes than recovery of various radiation related toxicities.

Unlike standard cross-sectional studies, where treatments are binary and represent case and controls groups, radiation therapy involves a three-dimensional dose delivered over multiple days, yet protocol standards extract simplistic dose-volume features as efficient measures of treatment plan quality. Dose Volume Histograms (DVHs) leave out useful information, and thus are insufficient on their own to support precision medicine.[8] A DVH assumes each location

within a region is equally sensitive to radiation and equally responsible for biological function. Advanced methods of extracting dose features and patterns would enable a better understanding of the impact on patient outcomes.[9]

3. The learning health system and predictive modeling

A common goal of traditional statistical modeling is the discovery of the underlying mechanisms or cause of outcomes. Breiman[10] compares a "data model" where a statistical model is assumed to describe a relationship and validated with the data to an "algorithmic model" where the mathematical model that relates the input variables to the outcomes is computationally determined through machine learning. Both approaches have benefits and flaws: the "data models" are usually hypothesis-driven, yet may not reflect the complexity of the true process, but nonetheless enable improved understanding of the system. The "algorithmic models", on the other hand are hypothesis-generating presenting superior predictive accuracy, yet make it challenging to uncover the dominant input variables and/or causal attributes.

Medical information is very complex and often aggregated into features that can mask important underlying details. Such dimension reductions are necessary, but risk being insufficient. A good example is the selected points on a DVH, where we have essentially reduced three-dimensional dose in a region to a single value of dose or volume. This data reduction may have a negative impact on the ability to build a model to predict organ function or disease control after treatment that may have spatial dependence. It is not easy to proactively determine whether this type of ad hoc feature will preserve or discard useful relationships between the features and outcomes. Developing and applying dimension reduction strategies that usefully preserve true relationships in the data may improve normal tissue complication models.[11]

Considerations for predictive models must include the purpose of building them, whether they are to be used for decision support or for discovery of new knowledge. There is more than one tool and selecting the right one to apply to the clinical question and purpose will be critical for making more precise patient care decisions.

3a. Decision support

The goal of decision support is to provide the most appropriate intervention for the patient[12] and not necessarily to discover new knowledge. This begs the question of which outcome prediction models should be selected with what accuracy requirement.

The key to selecting the more performative model is understanding the decision and intervention to be made. For example, if the intervention is to use a feeding tube to prevent weight loss for head and neck cancer patients undergoing treatment with radiation and chemotherapy, then it may not be necessary to know what combinations of toxicity caused the weight loss since the intervention is intended to treat the symptom instead of its underlying cause. Alternatively, if it is understood that, for a particular patient, taste disturbance would likely cause excessive weight loss, then the intervention may be to modify the radiation treatment to minimize the taste disturbance, or to refer to a nutritionist to consult on nutritional support.

Figure 1 depicts a framework for decision support[3] where at some time point in the care of a patient a decision needs to be made. The inputs to the predictive model include the facts about the patient and potential interventions. Outcome predictions such as risk of a particular toxicity or probability of local disease control are presented to the clinician and patient with the specific attributes most influential to the prediction. These outputs could then be used to assist the

decision making whether it be selection or change in the treatment course or an intervention to improve symptoms.

An evaluation of the dominant attributes of a specific prediction must consider an understanding of the individual patient and the existing knowledge of underlying causes. Predictive models often do not separate causation from association. Thus, interventions that depend on treating a causal attribute must consider the limitations of the predictive models.

3b. Discovery and hypothesis derivation

A LHS also provides the opportunity to extend knowledge through discovery and hypothesis derivation. In essence, the goal is to both understand features most predictive of outcomes and uncover the underlying causes.

Figure 2 depicts a framework for discovery using the LHS. The process is to find features of the patients that most influence an outcome by generating predictive models and cross validating them with the available data to maximize prediction accuracy. In this approach, iterative exploration of an unlimited set of features seeks out those that maximize the predictive accuracy. Post-validation, a review of the relevant features can support hypothesis derivation and help uncover discoveries that can be further studied.

Aside from predictive modeling, cause and effect relationships between features and outcomes are important types of hypothesis and are often the most scientifically relevant. These types of hypotheses are most relevant for decision support, since making decisions based on purely associational criteria amounts “to an irrational policy of managing the news, and results, in practice, in replication failures and poor recommendations.”(D. Lewis)[13] Identifying cause-effect relationships entails systematically adjusting for selection effects and confounding bias,

using methods such as g-computation,[14] propensity score matching,[15] and inverse probability weighting.[16] In addition, under strong assumptions inferences about causal directionality underlying associational relationships between multiple variables are possible.

Though there is a large effort in machine learning and statistics to identify cause and effect relationships from observational data, all causal hypotheses generated by such methods must ultimately be validated by formal randomized controlled trials.

3c. What's missing?

Both decision support and discovery are limited by the knowledge contained in the database. For example, one institution may have ancillary care pathways that differ from another institution's such that these differences impact patients' outcomes. If institution A utilizes a speech pathologist for routine swallow therapy and institution B does not, their outcomes for swallow function may be different. If the details on a patient's adherence to swallow therapy are not contained in the database for either institution, then the treating institution would be an aggregate variable that might correlate with a swallow function outcome.

This missing of data also manifests itself when models are validated between institutions. If a model is built from only institution A's data and validated with institution B, the unknown information may dominate and the validation will fail. Alternatively, if a model is built with both institutions' data, and institution selection is the most dominant variable, there may be little difference relative to having two models, one for each institution, since the prediction will mostly depend on the treating institution.

This has implications.When using the LHS for decision support the goal is to have the most accurate prediction, and that may happen with models built using only patients treated at the

institution where the patient is to be treated. For discovery, however, the goal is to uncover underlying mechanisms, and for this, inter-institutional validation becomes important and completing missing information in the data is crucial to uncovering this new knowledge.

In addition to outright missing information, the knowledge base is limited within the norms of clinical care. With radiation treatments, for example, only the variability of the dose distributions present in the knowledge base is available.[17] If a particular anatomical region of every patient received the same dose, then there is no possibility of learning the impact on the outcome of varying dose to that region. Since patients are treated with similar dose goals in planning, the data will inherently subdue the importance of the known dose goals, while potentially being unethical to deviate from them. In essence, "Without deviation from the norm, progress is not possible."(F. Zappa)[18] As the effects of radiation on patients are explored, consideration of the existing knowledge and how much of it is inherently included, and thus subdued, is critical in any interpretations. Furthermore, as we trend towards standardized clinical guidelines, we risk further limitations on the knowledge contained in the data and on our ability to personalize care in the context of improved quality and safety.

4. Examples

4a. Treatment plan quality prediction

An early example of using big data tools in radiotherapy is the concept of geometry-driven or knowledge-based treatment planning (KBP).[19-24] KBP aligns with the LHS model in that it provides actionable predictions of dose goals for planning and continuously learns as more treatment planning data is accumulated.

KBP analyzes a plurality of prior treatments to discover patient-specific anatomical features that precisely correlate to high quality radiation dose delivery. With the model-based dose predictions, KBP can be used for treatment plan quality control (QC) or outright plan automation. In its generalized form, KBP makes use of established machine learning techniques such as supervised inference engines to discover relevant geometric variables and their correlation to patient-specific dose prediction.

While KBP is already in routine clinical use at some institutions for the purposes of automated planning,[25] one of the most important contributions from KBP has come in the combination of knowledge-based plan QC with a cooperative group clinical trial to assess the frequency and clinical severity of sub-optimal treatment planning in a diverse multi-institutional data set.[26]

4b. Incorporating toxicity outcomes and clinical intervention

The prediction of toxicities is also critical to a patient's ability to tolerate treatment and their long-term QoL. An example is weight loss prediction using a classification and regression tree for head and neck cancer patients.[27] Two predictions at different time points were developed to predict weight loss at 3 months post-treatment: 1) during planning using patient demographic and dosimetry data, and 2) at the end of treatment using additional on-treatment toxicities and patient-reported QoL data. During planning, the top two predictors of weight loss were tumor site and higher doses to the masticatory muscle, a potentially modifiable factor. By the end of the treatment, when radiation-induced toxicities started to present, patient-reported oral intake, tumor site, and dose to combined parotid were more predictive. Early identification of high risk for excessive weight loss may inform interventions such asfeeding tube placement, referral to speech pathologist for swallow function evaluation and exercise, or frequent monitoring early after treatment.

Another example is in the prediction of radiation-induced xerostomia for head and neck cancer patients. A wide range of clinical, demographic, and dosimetric factors wereevaluated by the algorithm and subsequently cross-validated by the accruing data. In this example, low dose bath to combined parotids, and intermediate level irradiation to submandibular glands alongside clinical factors of chemotherapy, HPV infection, feeding tube use, baseline BMI were identified as crucial for patients prone to severe xerostomia. Downstream conditional predictive factors including age, alcohol use, age and smoking are also attributable.

The LHS allows a comprehensive exploration of predictors for a variety of treatment related toxicities beyond the single organ DVH and simple normal tissue complication models, and further, bridging all other clinical and patient factors into an all-encompassing prediction model. Evidence from such models warrant the foundation for clinical decision support for the prevention and/or management of toxicities.

The exploration of dose distribution patterns and the inter-organ dependencies may be critical to precision medicine. Early studies show the spatial dependence of dose on xerostomia in the parotid glands,[28-30] anddysphagia across the swallowing muscles.[9,31] Further exploration of these spatial and multi-organ dependencies will be enabled by the LHS and may improve our knowledge of the impact of radiation on normal function.[32]

5. Genomics, pathology and radiomics

At a higher level, Radiomics, Genomics and Pathology are patient-specific data that are subjected to feature extraction in clinical practice and for research. [33-35]

Radiomics is a clear example where a portion of a diagnostic image is identified and features of the voxel values: density, texture and gradients are calculated and presumed to reflect characteristics of the specific tissue being analyzed. These features are used to predict disease response to treatment or toxicity. The features themselves do not necessarily reflect underlying mechanisms or status of the tissue, but they might sample characteristics that reflect underlying differences between patients.

In contrast, Pathology has had a long history of feature extraction where cell type, grade levels and differentiation are characterized from biopsy slides, and the disease is classified with staging and grading models.[36] This history has provided a means of communicating complex image and tissue characteristics and is used to classify patients for both research analysis and clinical decision pathways.[37]

Genomics is another very complex data set used to seek out known features that are associated with particular outcomes. The dominant research looks to discover genetic predisposition to disease or response to treatment. Other work suggests there may also be a predisposition to radiation toxicity based on genomic signatures.[38] The LHS offers an opportunity to explore genomicsin much greater detail and assist in uncovering genomic patterns that influence outcomes which would otherwise be impossible to discover.

Uncovering how the features derived from images, pathology and genomic signatures can inform clinical practice or discovery but ultimately relies on accurate measures of outcomes and treatment information. Thus it is the combination of the clinical data with these measures that will advance their ability to provide precise treatment options for our patients.

6. Discussion

Just as machine learning is being used to drive autonomous vehicles, is it reasonable to expect similar successes in radiation oncology? At this point, self-driving cars focus on the rules of the road and respond to immediate detection of obstructions in their local environment. They, however, exhibit difficulty in defensive driving where they must weigh the risks of the unknown and anticipate what might happen. Radiation oncology, though precise in the treatment, presents a similar situation with a few rules of the road and acute observations, but may be dominated by unknowns and patterns of defensive practice. As such, our expectation for the foreseeable future should be one of improved risk/outcome prediction as a supplement to physician-based clinical decision making.

The key to success is to uncover and measure as many of the unknowns as possible. Is a future possible in which we accurately measure the critical aspects of patient's outcomes and treatment? Computerization of healthcare is advancing rapidly and the societal culture evolving from having smartphones amplifies the likelihood that good measures of the continuous patient condition will only advance.As outcome measures improve, radiation oncology must do its part to accurately archive treatments in easily retrievable form, adhering to standard nomenclatures. It should be possible to query features of the patient's history, physical exam, radiographic studies, laboratory tests, and "delivered" dose for any patient from our clinical archive without significant processing. It should be part of the practice to be good stewards of the data and accurately record three-dimensional delivery while capturing the clinical data, appreciating that it ultimately will contribute to the LHS.

Presentation of a patient's condition is currently conveyed in mostly text and is typically presented in the absence of population based information. Disease and patient specific presentation through modern human computer interfaces coupled with population based statistics

can highlight how well a patient is doing in the context of their disease peers, and in itself aid in individualized decision making. Advances in such patient presentations offers the framework to present risk and outcome predictions in a form that is actionable.

The vision is a future where data is instinctively collected and each patient is provided a prediction of their disease outcomes and complications on the backdrop of their peer populations with treatments tailored to an individual's needs and sensitivities. Continuous learning of this LHS will open insights that involve patterns in data far more complex than our traditional evidence-based methods can uncover and will open the flood gates of knowledge.

References

1. Clinical practice statements --- american society for radiation oncology (ASTRO). https://www.astro.org/Clinical---Practice---Statements.aspx. Accessed Nov 21, 2017.

2. Jazieh AR, McClure JS, Carlson RW. Implementation framework for NCCN guidelines. *J Natl Compr Canc Netw*. 2017;15(10):1180---1185. doi: 10.6004/jnccn.2017.7020 [doi].

3. McNutt TR, Moore KL, Quon H. Needs and challenges for big data in radiation oncology. *Int J Radiat Oncol Biol Phys*. 2016;95(3):909---915. doi: S0360---3016(15)26770---3 [pii].

4. Chen RC, Gabriel PE, Kavanagh BD, McNutt TR. How will big data impact clinical decision making and precision medicine in radiation therapy? *Int J Radiat Oncol Biol Phys*. 2016;95(3):880---884. doi: S0360---3016(15)26649---7 [pii].

5. Friedman CP, Allee NJ, Delaney BC, et al. The science of learning health systems: Foundations for a new journal. *Learn Health Sys*. 2017;1(1):n/a. http://onlinelibrary.wiley.com/doi/10.1002/lrh2.10020/abstract. Accessed Nov 21, 2017. doi: 10.1002/lrh2.10020.

6. Bellazzi R. Big data and biomedical informatics: A challenging opportunity. *Yearbook of medical informatics*. 2014;9:8. http://www.ncbi.nlm.nih.gov/pubmed/24853034.

7. Tran T, Luo W, Phung D, et al. A framework for feature extraction from hospital medical data with applications in risk prediction. *BMC bioinformatics*. 2014;15(1):425. http://www.ncbi.nlm.nih.gov/pubmed/25547173. doi: 10.1186/s12859---014---0425---8.

8. Moiseenko V, Dyk JV, Battista J, Travis E. Limitations in using dose---volume histograms for radiotherapy dose optimization. In: *The use of computers in radiation therapy.* Springer, Berlin, Heidelberg; 2000:239---241. https://link.springer.com/chapter/10.1007/978---3---642---59758---9_90. Accessed Nov 10, 2017.

9. Serena Monti, Giuseppe Palma, Vittoria D'Avino, et al. Voxel---based analysis unveils regional dose differences associated with radiation---induced morbidity in head and neck cancer patients. *Scientific Reports (Nature Publisher Group)*. 2017;7:1. https://search.proquest.com/docview/1957144863. doi: 10.1038/s41598---017---07586---x.

10. Breiman L. Statistical modeling: The two cultures. *Statistical Science*. 2001;16(3):199---231.

11. Razieh Nabi Computer Science Department Johns Hopkins University rnabiab1@jhu:edu. Semi---parametric causal sucient dimension reduction of high dimensional treatments. *Methodology*. 2017. https://arxiv.org/abs/1710.06727. Accessed Nov 21, 2017.

12. Lambin P, Stiphout RGv, Starmans MH, et al. Predicting outcomes in radiation oncology---multifactorial decision support systems. *Nature Reviews Clinical Oncology*. 2012;10(1):27---40. http://www.narcis.nl/publication/RecordID/oai:repository.ubn.ru.nl:2066%2F109112. doi: 10.1038/nrclinonc.2012.196.

13. Lewis D. Causal decision theory. *Australasian Journal of Philosophy*. 1981;59(1):5---30. https://doi.org/10.1080/00048408112340011. Accessed Nov 21, 2017. doi: 10.1080/00048408112340011.

14. James M. Robins, "A new approach to causal inference in mortality studies with sustained exposure periods ----- application to control of the healthy worker survivor effect," Mathematical Modeling, Vol. 7, pg. 1393---1512,1986

15. Paul R. Rosenbaum and Donald B. Rubin, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," Biometrika, vol 70, issue 1, 1983, pg 41---55

16. D. G. Horvitz and D. J. Thompson, "A generalization of sampling without replacement from a finite universe,"Journal of the American Statistical Association, vol 47, pg 663---685, 1952

17. Robertson SP, Quon H, Kiess AP, et al. A data---mining framework for large scale analysis of dose---outcome relationships in a database of irradiated head and neck cancer patients. *Med Phys*. 2015;42(7):4329---4337. doi: 10.1118/1.4922686 [doi].

18. Zappa F, Occhiogrosso P. *The real frank zappa book.* New York, New York USA: Poseidon Press; 1989:185.

19. Wu B, McNutt T, Zahurak M, et al. Fully automated simultaneous integrated boosted---intensity modulated radiation therapy treatment planning is feasible for head---and---neck cancer: A prospective clinical study. *Int J Radiat Oncol Biol Phys*. 2012;84(5):647.

20. Appenzoller LM, Michalski JM, Thorstad WL, Mutic S, Moore KL. Predicting dose---volume histograms for organs---at---risk in IMRT planning. *Med Phys*. 2012;39(12):7446---7461. Accessed Nov 21, 2017. doi: 10.1118/1.4761864.

21. Petit SF, Wu B, Kazhdan M, et al. Increased organ sparing using shape---based treatment plan optimization for intensity modulated radiation therapy of pancreatic adenocarcinoma. *Radiother Oncol*. 2012;102(1):38---44. Accessed Nov 21, 2017. doi: 10.1016/j.radonc.2011.05.025.

22. Wang Y, Zolnay A, Incrocci L, et al. A quality control model that uses PTV---rectal distances to predict the lowest achievable rectum dose, improves IMRT planning for patients with prostate cancer. *Radiother Oncol*. 2013;107(3):352---357. Accessed Nov 21, 2017. doi: 10.1016/j.radonc.2013.05.032.

23. Wu B, Ricchetti F, Sanguineti G, et al. Data---driven approach to generating achievable dose---volume histogram objectives in intensity---modulated radiotherapy planning. *Int J Radiat Oncol Biol Phys*. 2011;79(4):1241---1247.

24. Wu B, Ricchetti F, Sanguineti G, et al. Patient geometry---driven information retrieval for IMRT treatment plan quality control. *Med Phys*. 2009;36(12):5497---5505.

25. Li N, Carmona R, Sirak I, et al. Highly efficient training, refinement, and validation of a knowledge---based planning quality---control system for radiation therapy clinical trials. *Int J Radiat Oncol Biol Phys*. 2017;97(1):164---172. doi: S0360---3016(16)33279---5 [pii].

26. Moore KL, Schmidt R, Moiseenko V, et al. Quantifying unnecessary normal tissue complication risks due to suboptimal planning: A secondary study of RTOG 0126. *Int J Radiat Oncol Biol Phys*. 2015;92(2):228---235. doi: 10.1016/j.ijrobp.2015.01.046 [doi].

27. Cheng Z, Nakatsugawa M, Hu C, et al. Evaluation of classification and regression tree (CART) model in weight loss prediction following head and neck cancer radiotherapy. *Advances in Radiation Oncology*.

28. Deasy JO, Moiseenko V, Marks L, Chao KSC, Nam J, Eilsbruch A. Radiotherapy dose---volume effects on salivary gland function. *Int J Radiat Oncol Biol Phys*. 2010;76(3 0):S63. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4041494/. Accessed Nov 10, 2017. doi: 10.1016/j.ijrobp.2009.06.090.

29. F. Marungo, S. Robertson, H. Quon, et al. Creating a data science platform for developing complication risk models for personalized treatment planning in radiation oncology. *2015 48th Hawaii International Conference on System Sciences*. 2015:3132---3140. doi: 10.1109/HICSS.2015.378.

30. Quon H, Park S, Plishker W, et al. Preliminary clinical evidence of parotid subvolume radiosensitivity and the risk of radiation---induced xerostomia in head and neck cancer (HNC) patients. *International journal of radiation oncology, biology, physics*. 2016;96:E341. Accessed Nov 10, 2017. doi: 10.1016/j.ijrobp.2016.06.1484.

31. Kumar R, Madanikia S, Starmer H, et al. Radiation dose to the floor of mouth muscles predicts swallowing complications following chemoradiation in oropharyngeal squamous cell carcinoma. *Oral Oncol*. 2014;50(1):65---70. doi: 10.1016/j.oraloncology.2013.10.002 [doi].

32. Blinded for Review

33. Aerts HJ, Velazquez ER, Leijenaar RT, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun*. 2014;5:4006. doi: 10.1038/ncomms5006 [doi].

34. 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(6):1162---1171. doi: 10.1016/j.patcog.2008.08.011 [doi].

35. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images are more than pictures, they are data. *Radiology*. 2016;278(2):563---577. doi: 10.1148/radiol.2015151169 [doi].

36. *AJCC cancer staging manual.* New York, NY: Springer Science+Business Media; 2016.

37. Coley RY, Zeger SL, Mamawala M, Pienta KJ, Carter HB. Prediction of the pathologic gleason score to inform a personalized management program for prostate cancer. *Eur Urol*. 2017;72(1):135---141. doi: S0302---2838(16)30472---9 [pii].

38. West CM, Barnett GC. Genetics and genomics of radiotherapy toxicity: Towards prediction. *Genome medicine*. 2011;3(8):52. http://www.ncbi.nlm.nih.gov/pubmed/21861849. doi: 10.1186/gm268.

The conceptual model for decision support versus discovery with big clinical data analytics is different, and an overview of the implications of these differences is discussed in the context of precision medicine.

Figure 1: Decision support framework to make predictions of outcomes of individual patients. The models, derived from the knowledge base, use the facts and clinical options/variables in making the predictions which are presented to the physician to assist in decision making.

Figure 2: Hypothesis generation utilizes predictive modeling to maximize the prediction accuracy to uncover specific features of the patients and their treatments most correlated with an outcome. The features are derived from the raw clinic data.

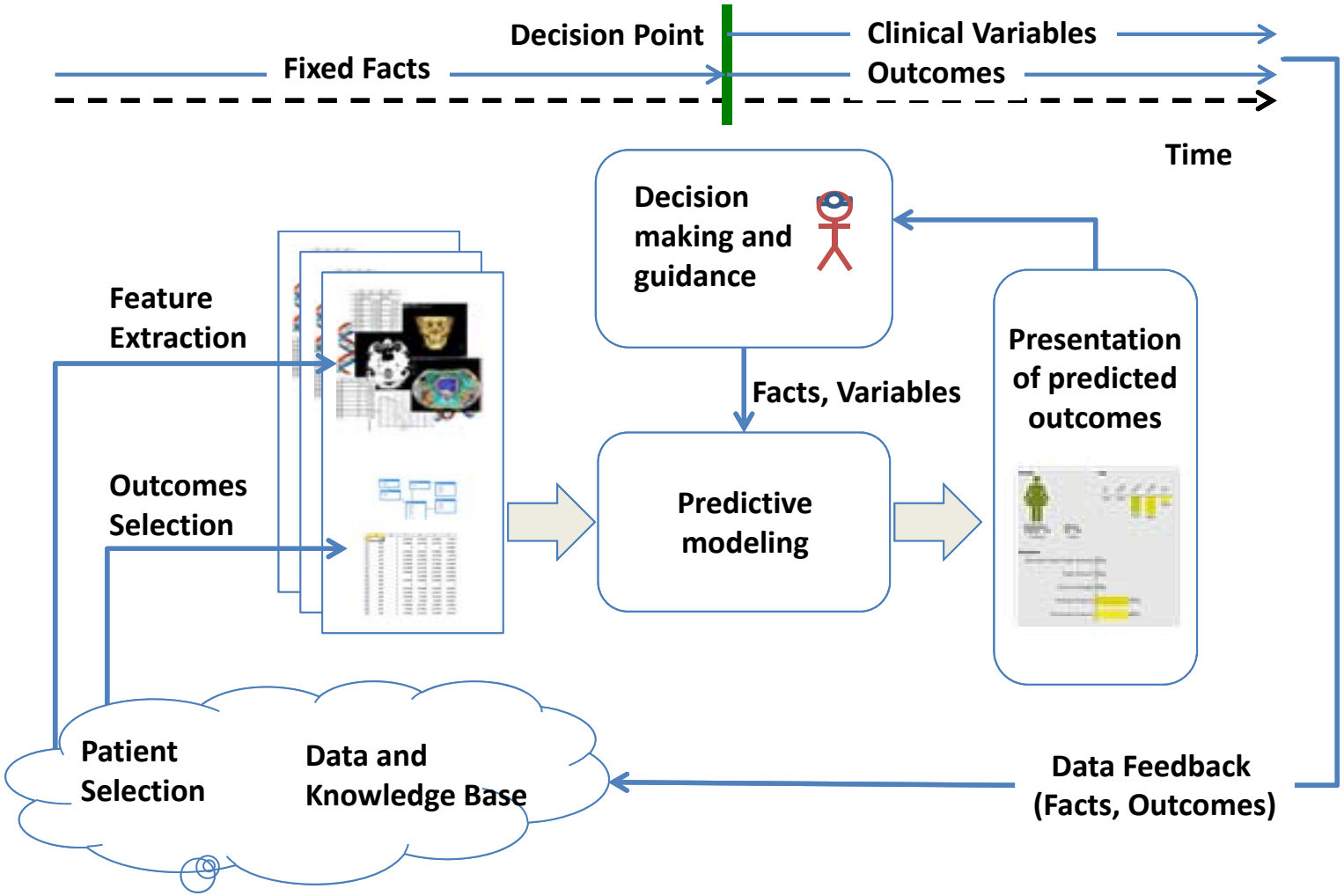


Figure 1: Decision support framework to make predictions of outcomes of individual patients. The models, derived from the knowledge base, use the facts and clinical options/variables in making the predictions which are presented to the physician to assist in decision making.

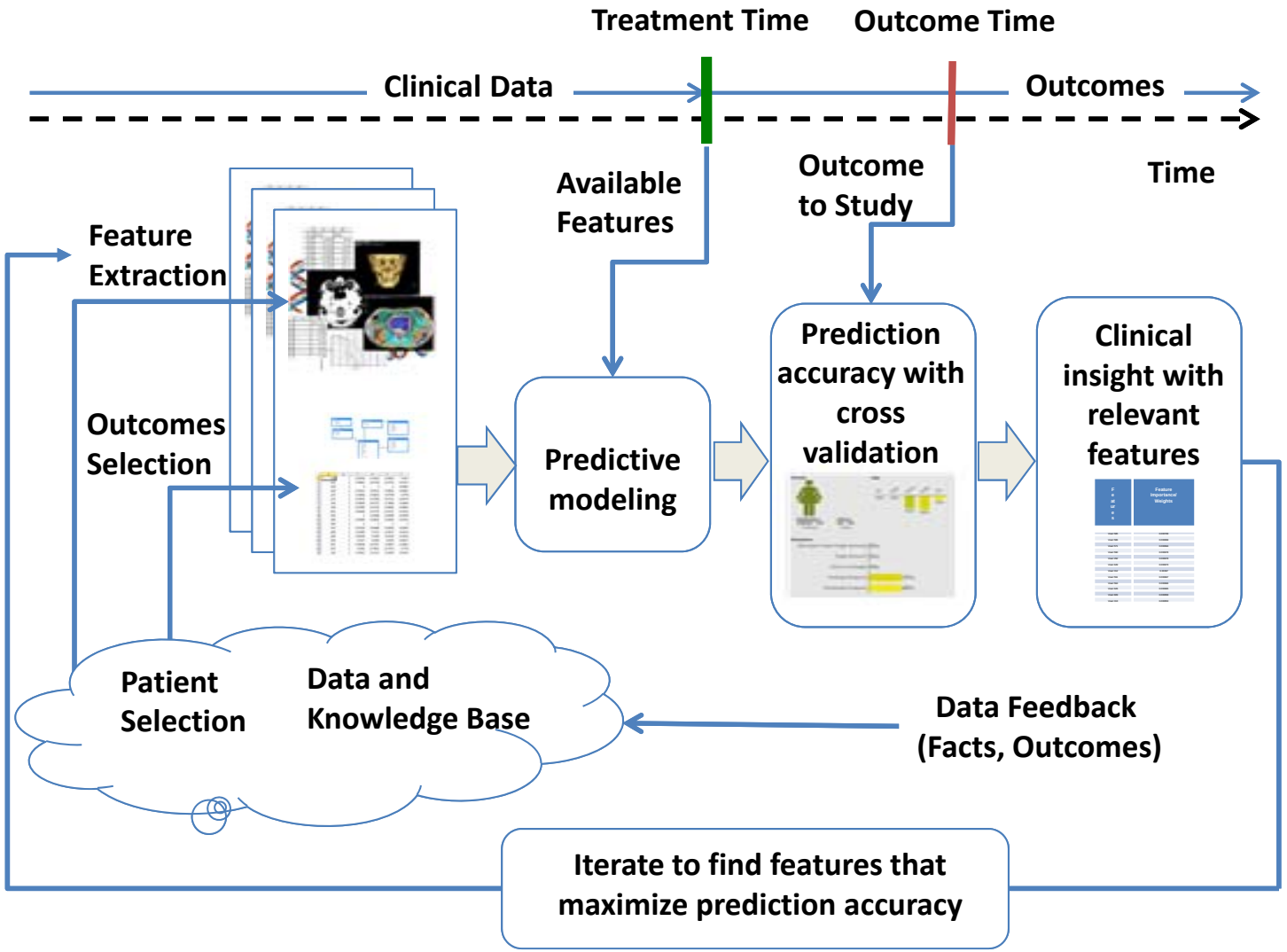


Figure 2: Hypothesis generation utilizes predictive modeling to maximize the prediction accuracy to uncover specific features of the patients and their treatments most correlated with an outcome. The features are derived from the raw clinic data.

The conceptual model for decision support versus discovery with big clinical data analytics is different, and an overview of the implications of these differences is discussed in the context of precision medicine.