Critical Review

Project 8: UI for Radiation Therapy Cohort Selection

Keefer Chern | CIS II | 3/12/2019

Paper Name: Using Big Data Analytics to Advance Precision Radiation Oncology

Citations:

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.[1]

Mcnutt, Todd R., et al. "Needs and Challenges for Big Data in Radiation Oncology." *International Journal of Radiation Oncology*Biology*Physics*, vol. 95, no. 3, 2016, pp. 909–915., doi:10.1016/j.ijrobp.2015.11.032.[2]

The goal of our project is to develop a User Interface that will allow researchers and clinicians the ability to select a patient cohort based upon any number of the variables and view historic patient outcomes. The reason I choose this paper that was authored by our mentor, Dr. McNutt, is because the results of our project is a small part of the larger concept of the learning health system this paper develops. This paper would also provide us information about the different types of variables intended for our project to utilize.

The authors start the paper off by providing some background knowledge. Precision medicine is a method of tailoring the treatment to the characteristics of the patient[2]. Therefore, the patient would be able to obtain the quality of care that would most benefit them with less side effects and costs. The goal behind precision medicine "is to improve overall patient care and determine when and how to personalize patients' treatments." [1] The current method for precision medicine is heavily dependent on the physician and the knowledge the physician possesses for the treatment. Guidelines have been established, which provide "overall pathways" [1] to treat specific diseases, but precision medicine requires a "finer granularity" [1] than general guidelines. I support the author's point of requiring more detail and information than general guidelines for precision medicine. I see that there are intrinsic issues with depending on the physician and the physician's knowledge for precision care. The method is subjectively based on the physician's decision and is difficult to translate care between multiple physicians due to this variability.

This paper discusses how using big clinical data analytics can provide this "finer granularity" and shows how these tools can benefit patient care and treatment. The authors of this paper established the concept of a learning health system (LHS), where "quantifiable diagnostics, treatment, and outcome data are captured from a continuous stream of patients and

placed in a knowledge base." [1] The stored knowledge can then be analyzed with statistical and machine learning tools to reveal nonobvious trends. The comprehensiveness of the LHS can be used for supporting clinical decisions, discovery of important factors behind treatment, and help with deriving experimental hypotheses. The authors believe that the LHS can positively impact the therapeutic care of patients, but there exist many implications that need to be resolved before the benefits of the LHS can be fully realized. I agree with the author, big clinical data is a significant resource that has yet to be tapped into, which can greatly improve quality of patient care.

The LHS is based on the requirement and need of big clinical data, which can be defined as a large database containing clinical information of patients. This clinical information include be lifestyle covariates, disease status, symptom management, quality of life outcomes, adverse side effects, and survival. There are many advantages to big clinical data as it has minimal bias, can be reused, provide multidimensional understanding, and can improved by linking with other databases. Yet, the usage of big clinical data is severely limited by the quality of clinical data in the database, which suffer from selective sampling, missingness, and measurement error. An example the authors provide is that recurrence of cancer in a patient may be recorded, but without any detail regarding location of recurrence. This makes it hard to determine whether recurrence is due to radiation treatment. Another example of this limitation is the method of gathering xerostomia data, which can be scored by a physician, from patient questionnaires, or controlled stimulation studies. I see these issues and examples presented by the authors as one of the biggest negative factors that would prevent the LHS from achieving maximum efficiency. Any researcher or data analyst would be heavily restricted if the data they have is not good.

The authors believe that even with the limitations due to clinical data, the LHS can provide the important tool of predictive modeling. Predictive modeling in terms of a data model is meant to describe a relationship and validate the relationship with the data. The authors state that before the usage of a predictive model, one must have a reason for building the model and must know whether they are planning on using it for decision support or knowledge discovery.

The authors state that the main goal of decision support is to provide the best intervention for the patient. To do this a decision support framework would try to predict the outcomes of individual patients before undergoing the therapy. The models developed using this framework take in facts and clinical options from the clinicians and output the possible outcomes. The key with choosing the best model would be to have a clear picture of what kind of decision you want to make and what intervention is being considered. The authors provide an example of deciding whether to use a feeding tube to prevent weight loss for patients undergoing head and neck radiation treatment. This intervention focuses on treating the symptom and not the cause, so knowing how combinations of toxicity can cause weight loss would not help the model. In contrast when the intervention wants to modify toxicity levels to prevent taste disturbance, knowing underlying causes is important. In the general case, interventions that wish to focus treating causes are hindered because predictive models can not differentiate association from causation.

The authors state that the main goal of using the LHS for knowledge discovery is to expand knowledge by understanding what features best predict outcomes, discover underlying causes, and helps deriving hypotheses. The hypotheses that are derived are cause-and-effect relationships present between features and outcomes. These hypotheses significantly help with decision support predictive modeling, which have issues with determining causation. The

methods the authors suggest for determining cause-effect relationships "entails systematically adjusting for selection effects and confounding bias, using methods such as G-computation, propensity score matching, and inverse probability weighting." [1] Using even stronger assumptions, showing relationships between multiple variables is even possible. Though despite all the effort to computationally identify cause and effect relationships, they still need to be validated through formal controlled trials with real subjects.

The authors then show how a significant limitation to decision support and discovery is the knowledge within the database. The example that the authors provide is about institutions having different systematic care procedures, which would impact patient outcomes and therefore the data. Such as if institution A provides therapy for swallowing, while institution B does not, the patients capability to swallow can be different. Even worse, whether this therapy is provided may not be in the database, which would be a significant confounder when performing data analysis. This is especially true for validation studies where the researcher would like to validate the model created using institution A's data with institution B's data. The implication of this confounder varies. For decision support, the implication is not that big since the goal is the have the best model for the prediction of outcomes for the institution the patient would receive treatment from. The model would be built with data of patients from that institution. On the other hand, if the goal is to discover underlying mechanisms, resolving missing data is a crucial factor. Another problem that would limit the knowledge in the database is the normalization of care restricts deviation. Since care is normalized, meaning all patients tend to receive a similar dosage, it can be unethical to deviate from guidelines just for the purposes of gathering data. Though many think dosages should never deviate, the effects of irradiation in treatment is difficult to explore with limited knowledge. We have many of the same issues regarding data as

the authors have talked about. Not only can data gathering be a problem, the data labels in the database for our project makes it difficult to analyze. For us, much of the same information could be under different labels.

The last topic the authors move towards regards the feature extraction of radiomics, genomics, and pathology data. In radiomics, part of the diagnostic image is identified and features such as density, texture, and gradient are calculated and a presumed to be the representation of the tissue analyzed. [1] In pathology, features such as cell type, grade level, and differentiation is extracted from biopsy slides. Furthermore, disease information can be categorized with staging and grading models. For genomics, the authors talk about extracting features from genomic information to show genomic predisposition to the disease.

In summary this paper introduces the concept of a Learning Health System (LHS), which uses big clinical data to produce relevant and useable outcomes. Two of the largest proposed usages of the LHS is decision support, getting treatment outcome from patient characteristics, and knowledge discovery, understanding the underlying cause and effect relationships. One of the biggest necessities that the LHS still does not have is well documented and collected data.

My evaluation of this paper is that it is very clear about what topic it is talking about. For each topic authors provide enough examples to support their ideas. The paper is also very well organized as it always mentions both the positives and negatives of each concept they go over. One negative aspect of this paper is that they do not go into detail how the LHS should be implemented. In addition, they do not provide many solutions to the problems they talk about. Lastly, the paper does not describe the cost of implementing a LHS and the struggles it would have in regards to patient permission and confidentiality for gathered data.