Domonique Carbajal (dcarbaj1@jhu.edu)

Project 8: User Interface for Radiation Therapy Cohort Selection

Mentors: Dr. Todd McNutt, Pranav Lakshminarayanan

Computer Integrated Surgery II Paper Critique

Project Overview:

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 in available patient information and view historic patient outcomes. We are working with a system already in existence called Oncospace and utilizing C# and SQL to extract and display information. The levels of difficulty include displaying simple static variables to displaying variables that require some derivation from the original data source. In creating this patient cohort selection, we hope to allow researchers the ability to fully utilize the big data available and select historic examples to better improve patient care. In addition, we hope to implement a method of saving queries and reloading previous query calls so as to see new patients that may have been added.

Papers Selected:

The paper I chose is "The big data effort in radiation oncology: Data mining or data farming?" by C.S Mayo et al and originally published in the journal *Advances in Radiation Oncology* October to December 2016. This paper was chosen from a relatively small list of papers as the specificity of the field of applications of big data to radiation oncology is large. This paper was chosen due to the fact that it addressed the general difficulties with fully utilizing big data in radiation oncology and a vision for using only the most clinically relevant data within

the set. It also addressed warnings for handling data errors and processing so as allow for the best possible cohort selection.

The Big Data Effort in Radiation Oncology:

The main efforts of the article follow an extended metaphor of farming (data farming) and contrast it to our understanding of data mining. Radiation Oncology and healthcare specific data have more complex features and messier data connections than the typical data mining data. The emphasis is placed on the fact that in "mining" the resource is there to be found, while in "farming" one must cultivate the resource and put forth effort to result in something harvest worthy. The Figure 1 below summarized aptly this idea as beginning with standardized inputs and the most correct data possible which allows for maximal extraction, transferring, and loading (ETL). The data is then used with strategic technology that has capabilities they laid forth in the document (ability to perform queries, ability to integrate into development of clinical applications, etc.).

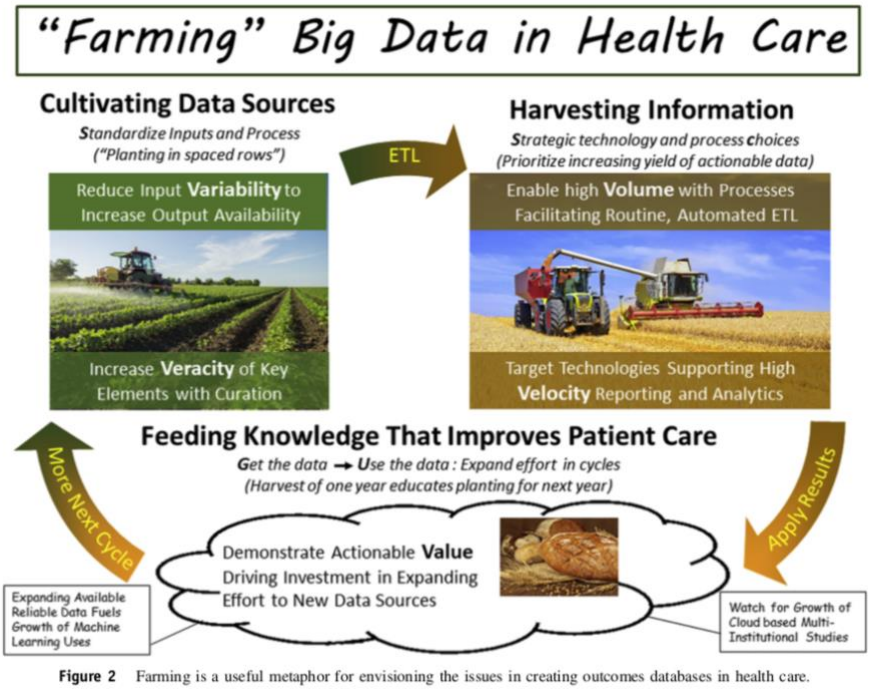

Figure 2 Farming is a useful metaphor for envisioning the issues in creating outcomes databases in health care.

C.S Mayo et al

The process of ETL is hopefully aided by automated sources to produce the largest volume of actionable data and the technologies ideally support the best possible velocity of data actions. Within this discussion of the extraction of the databases from the sources was the comment on the difficulties provided by Electronic Health Records (EHR) implementation of a free text data collection which is difficult to standardize and process when it holds valuable toxicity and grade information. This is a problem we currently face in our data as the toxicities are taken from a free text source meaning that any variations in the term (taste versus taste-bitter) result in unrelated classes of toxicity. The table below shows the paper's summary of key data elements, its priority as a data value, and the difficulties faced in obtaining those values. As it can be seen there are a considerable number of important data sources that create issues due to free text implementation, lack of standardization, and missing data prevalence.

Table 1 Categorization of key data element categories and summary of our experience of challenges to extract, transform, and load (ETL) of data from source systems to aggregation tier.

Key element category	Demand ranking	ETL difficulty	Typical source systems	Access	Multiple source systems	Use or used free text entry	Missing data	Data accuracy	Lack of standar dization	PHI constraints limit access	Legacy formats or systems	Require process changes	Extensive transformation	Other
Demographics ●	1	L	EHR	×										E
Health status factors	2	L	EHR	×										E
Pathology ⊙	3	M to H	EHR	×		×	×		×		×	⊠		E, X
Surgery ⊙	2	M to H	EHR	×		×	×		×		×	⊠		E, X
Chemotherapy ●	2	M	EHR, ODB	×										E
Encounter details ● Office, emergency room, hospitalization	3	L	EHR	⊠									×	R
Diagnosis ●,▲,⊕	1	M	EHR, ROIS	×	×			×			×	⊠		R, E
Staging ●,▲,⊕	1	H	EHR, ROIS	×	×	×		×			×	⊠		E
Prescription ▲,◆	1	H	ROIS, ODB						⊠			×		E, X, R
As-treated plan details ●	1	M	ROIS										×	
DVH ●,□,◆	1	M	TPS				×		×		×	⊠	×	ATPS
Survival ●	1	M	EHR, XLS, ODB	×						⊠				UD, E
Recurrence ▲,⊕	1	H	EHR	×		×	×			×	×	⊠		E, X
Toxicity ●,▲	1	H	EHR, ROIS	×		×	×			×	×	⊠		E, X
Patient-reported outcomes ▲	2	H	EHR, P	×			×			×	×	⊠		E, X
Laboratory values ●	2	M	EHR	⊠				×					×	E
Medications ●	2	M	EHR	⊠				×					×	E
Height, weight, BMI ●	2	M	EHR	⊠				×					×	E
Treatment imaging: Timeline details ●	3	H	ROIS										×	R

C.S Mayo et al

The paper primarily recommends implementation of advanced technologies capable of better extraction, and a process of understanding which variables are most helpful to clinicians in

creating the methods of collecting data. The recommendations include cultivating the data in a way that reduces input error by having a system in place that checks the values and properly handles missing data (not by assigning a default value of 0). Often the recommendations would alter to some degree the documentation load of clinicians which are already very high.

In Review:

Positive points include the use of very relevant information to radiation oncology, consistent usage of language and level of analysis, and specific examples of undermined cohort selection due to data error. The paper was thorough in looking at all related aspects of data collection and representation for the use of cohort selection. Negative aspects include an underutilization of examples from the system developed by the authors (the University of Michigan instance of a Radiation Oncology Analytics Resource) and discussion of their solutions to the problems presented. Additionally, the solutions they propose to increasing the volume of actionable data are extremely expensive, not fully developed tools, and would often have negative impacts on the documentation responsibilities of physicians making them almost infeasible in nature. In summation, this paper had direct relevance to my project and will help direct us in the improvements we are making to the system and the data that we are extracting from the database.

Resources:

Mayo, Charles S., et al. “The Big Data Effort in Radiation Oncology: Data Mining or Data Farming?” Advances in Radiation Oncology, vol. 1, no. 4, 2016, pp. 260–271., doi:10.1016/j.adro.2016.10.001.