Big Data Meets Medical Physics Dosimetry

Fumbeya Marungo
Team Member

Hilary Paisley
Team Member

John Rhee
Team Member

Todd McNutt, PhD
Mentor

Scott Robertson, PhD
Mentor

Russell H. Taylor, PhD
Instructor

May 6, 2014

Abstract

1 Introduction

Medical physicists face a trade-off when planning oncology radiotherapy treatments. The dosimetrist must deliver sufficient dose to kill the diseased tissue while managing the risk of a variety of possible toxicities that may arise from damage to normal tissue.

While advances in radiotherapy allow for sophisticated three dimensional treatment plans. The models for assessing the complication probability lag well behind. For example, the Lyman-Kutcher-Burman (LKB) model for assessing toxicity probability neither accounts for treatment placement, nor differentiates between higher doses over small volumes and lower doses over large volumes.

The course instructor and mentor have conducted previous work in using three dimensional shape descriptors of the location of diseased tissue relative to organs as tools in treatment planning[**?**]. This work explores using data within the clinical record to calculate the probability of complications. The ultimate goal is to use the knowledge discovered to allow previous experience to inform treatment planning (Figure 1.1), and to obtain insights into the factors leading to treatment complications.

2 Background

2.1 Introduction

In this section we discuss the following in turn: the current complication probability models in radiotherapy; motivations for applying big data techniques to the area; the knowledge discovery process; and approaches for predicting medical outcomes using data mining.

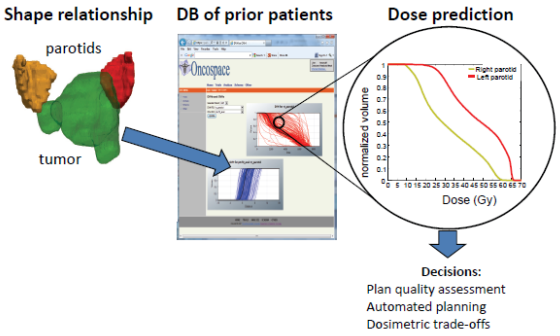


Figure 1.1: Using data on the clinical outcomes of previous patients provides insights into current patient treatment planning and assessment (courtesy of Todd McNutt)

2.2 Current complication risk assessment modeling

The common measure of toxicity risk in radiotherapy is normal tissue complication probability (NTCP). The Lyman-Kutcher-Burman (LKB) model is the most oft-cited method for is the calculating NTCP. LKB builds upon prior work by Rubin and Cassaretti [**?**] that a tolerance dose (TD_{50}) that, when delivered uniformly to an organ's entire volume, results in a a 50% risk of a given toxicity. LKB assumes that the effect on NTCP of delivering a uniform dose over a partial volume of an organ is related to the effect of delivering the same dose to the entire organ by a power function [**?**], that is:

$$TD_{50}(V) = \frac{TD_{50}(1)}{V^n} \tag{2.1}$$

Using a normal distribution with $\mu = TD_{50}$, we get:

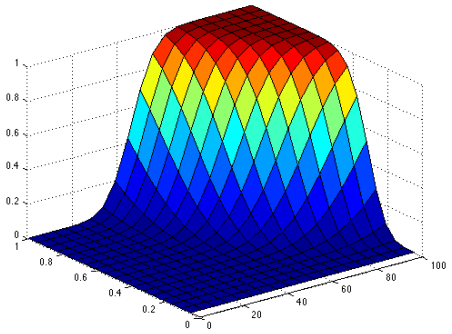
$$NTCP(V) = \Phi(t), \text{ where} \tag{2.2}$$

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{\frac{-t}{2}} \mathrm{d}x$$

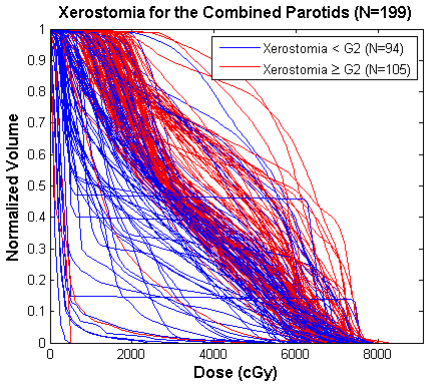
$$t = \frac{D - TD_{50}(V)}{\sigma(V)},$$

$$\sigma(V) = mTD_{50}(V).$$

Ermani, et al. [**?**] and Burman et al. [**?**] provide widely cited values for m, n, and TD_{50} over a variety of organs and toxicities. Given the parameters, and assuming a constant dose, we can calculate NTCP from equations 2.1 & 2.2. In order to account for dose variation, LKB assumes a second power relationship



(a) LKB NTCP surface for a parotid using parameters $n = 0.70$, $m = 0.18$, and $TD_{50} = 4600$[**?**]. x = Dose in Gy, y = Volume proportion, and z = NTCP



(b) Dose volume histograms (DVH) of parotid gland exposure (courtesy of Todd McNutt).

Figure 2.1: Visualization of the LKB model.

between dose and volume [**?**]:

$$\Delta V_{eff} = \Delta V_i \left(\frac{D_i}{D} \right)^{\frac{1}{n}} \tag{2.3}$$

In practice, $D = D_{max}$, the maximum treatment dose the patient receives, and D_i corresponds to doses assigned to bins in a histogram. We can therefore calculate:

$$V_{eff} = \sum_i \Delta V_{eff} \left(\frac{D_i}{D} \right)^{\frac{1}{n}} \tag{2.4}$$

Figure 2.1a display the interaction of dose and volume in LKM. Parameter m represents the range of responsiveness to dose. That is the steepness along the dose plane — a low m implies a sharp increase in risk near TD_{50} a higher m implies a gradual increase in risk that begins at low doses.

In the clinical setting, medical physicists visualize treatment plans using dose volume histograms (DVH). Figure 2.1b is a collection of parotid DVH's. Each curve represents an individual patient's treatment plan; patients whose subsequently experience to xerostomia are in red. The DVH can be read as "this percent of the organ volume (y-axis) received at least this dose level (x-axis)." While NTCP's are often not used directly in the clinical setting, the LKB value can be computed directly from the DVH using Equation 2.4.

2.3 Motivation for refining risk assessment

Marks, et al.[**?**] present a number of limitations in the LKB model that restrict its direct clinical applicability. The model reduces a DVH to a single pair of dose and volume values; the two numbers are the sole factors in risk assessment. LKB does not evaluate factors specific to the patient, such as, dose placement, other treatments, general health, etc.

Twenty years of clinical experience in three-dimensional approaches to radiotherapy combined with increases in cost effective storage capacity and computing processing power present future directions for refining NTCP modeling. Bentzen et al.[**?**] presents a number of trends, including: addressing the more diverse spectrum of treatments modern oncology patients receive; personalized risk-benefit assessments; and focusing the developing methods based on "more data" as opposed to creating "more [analytical] models."

Improvements in NTCP calculation can increase the statistic's applicability in the clinic in a number of ways. For example, accurate NTCP values can be a factor in optimizing treatments. Additionally, the values can provide insights into the biological effect of different approaches of treatment on normal tissue.

2.4 Knowledge discovery in health-care

Clinical patient data is collected in the course of treatment and stored in health information systems. The data, therefore, are not in a format that is immediately conducive to analysis.

Fayyad, et al.[**?**] introduced what is generally considered the fundamentals of the process for knowledge discovery in databases (KDD). Typically the vast majority of data analyzed was not collected for that purpose, but rather in the course of an institution conducting its general activities. In the case of health-care, data is generally from electronic health records (EHR), or other components within hospital information system.

[**?**] divides knowledge discovery into nine steps (Figure 2.2): (1) understanding the problem domain and the previous work in the area; (2) selecting a target dataset; Data cleaning and preprocessing; (3) data reduction and projection; (4) matching the knowledge discovery goals with a data mining approach; (5) exploratory analysis with hypothesis and model testing; (6) data mining; (7) interpreting results; (8) interpreting mined patterns; and (9) acting on discovered knowledge.

Cios & Moore[**?**] presents issues in KDD specific to health-care, such as: the heterogeneity of medical data; the multidisciplinary nature of the process due to the need for both clinical and computational expertise; ethical issues arising from the nature of the problem; etc. In addition, [**?**] places emphasis on KDD's iterative nature (Figure 2.3).

2.5 Data mining: techniques and assessment

Random forests[**?**] are classifiers built of independently trained trees. Each tree has some mechanism for random determination of the features under consideration. For example: different features may be selected at each node; each node select a fixed number of inner products of the input vector and randomized weight vector; each tree may randomly select features and use the subset across all its nodes; etc. Training individual

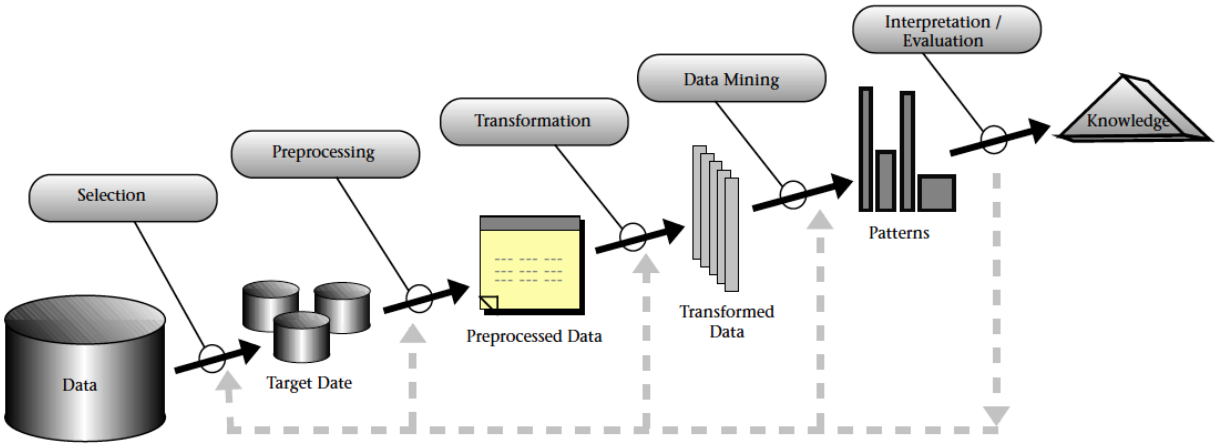


Figure 2.2: The process of extracting knowledge from data (copied from [**?**]).

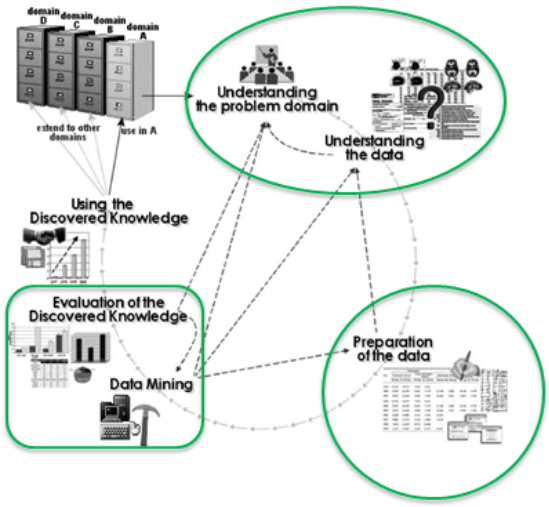


Figure 2.3: KDD areas covered by Project IX (copied from [**?**], selections added).

trees uses separate bagged samples. The forest's classification algorithm is to return a summary (generally the mean), of each tree's zero or one classification vote.

Tree structures have a number of advantages. Many splitting methods not only do not require scaling and centering, but also can process both data containing both numerical and categorical features. Moreover, random forests converge as the number of trees increases, with accuracy that is a function of the accuracy of the underlying trees and the level of independence between the trees [**?**].

The characteristics above often make random forest classifiers strong candidates for application to medical data. Many features in medicine neither lend themselves naturally to numerical representation nor are strong signals of outcome[**?**].

In assessing results Cios & Moore[**?**] provides four levels of validity:(1) face validity — clear and obvious inconsistencies with between the model and known fact are not present; (2) internal validity — in classification, for example, a metric such as the receiver operating characteristic curve supports the model's predictive performance; external validity — the model is resistant to over-fitting and maintains performance against other external datasets; and (4) clinical utility — the ability for clinicians to interpret the results in the context of treatment.

2.6 Project management

A "Surgical Team" surgical team methodology[**?**, **?**] to software project management entails a chief programmer who is ultimately responsible for the project's software development. Other project roles naturally naturally divide into administrative and technical functions. The project team normally consists of a total ten members, including the chief programmer.

2.7 Summary

In assessing normal tissue complication probability (NTCP), the conventional Lyman-Kutcher-Burman (LKB) model, approach assumes: a TD_{50} value that represents a constant dose distributed uniformly over the organ resulting in a 50% probability of a given complication[**?**]; a power function equivalences between organ volume and dose that transform a dose volume histogram (DVH) into two values — maximum dose (D_{max}) and effective dose(V_{eff})[**?**, **?**]; and parameterization that fits dose to D_{max} and V_{eff} to a normal distribution[**?**, **?**].

LKB limits analysis to the DVH; it does not account for the multi-factor nature of predicting radiation oncology complications. The lack of robustness limits the model's applicability in the clinic [**?**]. In refining NTCP calculation focus is shifting from improving modeling to exploiting the increasing amounts of data

now available within the EHR, such as treatment plans, previous chemotherapy, medical history, etc.[**?**]

Extracting the information and performing the analysis necessary to create sophisticated data-driven NTCP models requires the KDD process. KDD, both generally and in medicine, is a multi-step, interdisciplinary, and iterative process[**?**, **?**]. The process requires a team of individuals with a combination of domain knowledge, technical expertise, and project management skills. The result of a successful KDD process is a pipeline that transforms data within a database into a format that is suitable for data-mining; analysis of the data-mining results then yield new knowledge.

In medicine, the data-mining algorithms and assessment must be applicable to a broad mixture of numerical, and categorical information. The random forest algorithms often are highly suitable due to their underlying tree based structure and convergence properties. For assessing classification effectiveness, the receiver operating characteristic's area under the curve is a frequently used metric because it can capture both specificity and sensitivity.

The "Surgical Team" approach to managing a technical project, such as a KDD task, organizes a team around a chief programmer. The chief programmer is responsible for the entire technical implementation. The other nine team members provide support; their roles generally separate into managing either administrative or technical functions that are necessary for successful project completion.

3 Methodology

3.1 Project management

3.1.1 Team members

Fumbeya Marungo, a first-year PhD student in Computer Science served as Project IX's team lead (TL); Hilary Paisley, an undergraduate student double majoring in Biomedical Engineering and Applied Mathematics served as the project manager (PM); John Rhee, an undergraduate student double majoring in Biomedical Engineering and Computer Science served as the software engineer (SE).

Dr. Todd McNutt, and Dr. Scott Robertson, both of the Radiation Oncology Department, served as mentors and provided domain expertise in medical physics.

Assessments of the mandated artifacts and suggested improvements were provided by course instructor Dr. Russell Taylor of the Computer Science Department.

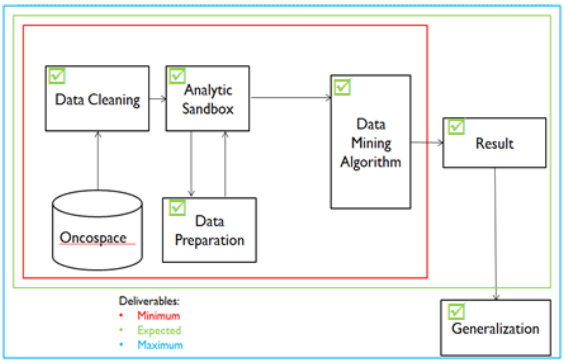


Figure 3.1: Project IX deliverables

3.1.2 Team member responsibilities — who did what

Project IX's used an abridged version of the "Surgical Team" approach (see Section 2.6). The PM's responsibilities generally covered the team's administrative requirements, including: delivering 1st drafts of presentation slides; maintaining the project plan in ProjectLIBRE (an open source Microsoft Project compatible project management program); creating custom calendars of the planning schedule; and maintaining the team's web site. In addition, Hilary requested technical tasks. Her technical responsibilities included: researching the random forest algorithm; reviewing the Weka's[**?**] random forest implementation; and both reviewing and commenting the team's source code.

The SE's responsibilities generally covered the Project IX's technical requirements, including: creating the 1st draft of that software connects to the Oncospace database and reads the ROI mask and

As the chief programmer[**?**, **?**], the TL's responsibilities entailed: deciding upon the technical approach; mapping the KDD steps[**?**, **?**] (see Section 2.4) to a tangible project plan; assigning tasks; transforming the "1st drafts" provided by the team into the finished deliverables; and data mining.

3.1.3 Deliverables

Project IX's minimum deliverables were: (1) an analytical pipeline consisting of an analytic sandbox (Section 3.2), software to for cleaning and preparing Oncospace's data, and a data mining algorithm for calculating xerostomia NTCP using parotid gland data; and a report on the results. The expected deliverables were to find a data mining algorithm that provided better prediction performance than the LKB model. The maximum deliverable was to demonstrate the general nature of the platform by performing analysis on a second organ and complication (Figure 3.1).

No.	Task	Start	End	Critical Dependencies
1	Select Project	28-Jan-14	30-Jan-14	None
2	Maintain Wiki	28-Jan-14	9-May-14	None
3	Project Planning Presentation	11-Feb-14	11-Feb-14	None
4	Project Planning Report	17-Feb-14	17-Feb-14	None
5	Project Planning	3-Feb-14	17-Feb-14	None
6	Setup Development Environment	6-Feb-14	20-Feb-14	None
7	Literature Review	11-Feb-14	28-Feb-14	Input from mentors
8	IRB	14-Feb-14	19-Feb-14	None
9	Database Access	20-Feb-14	27-Feb-14	Task 8, Mentor action, Support JHH IT
10	Target Database Access	20-Feb-14	20-Feb-14	Task 8, Mentor action, Support JHH IT
11	Meeting with mentors	20-Feb-14	20-Feb-14	
12	Develop Target Database	20-Feb-14	11-Mar-14	Input from mentors
13	Begin Preparing Paper Seminar	20-Feb-14	5-Mar-14	Task 7, Input from mentors
14	Data Cleansing and Preprocessing	24-Feb-14	6-Mar-14	Task 12, Input from mentors
15	Meeting with mentors	27-Feb-14	27-Feb-14	None
16	Paper Presentation	6-Mar-14	6-Mar-14	Task 13
17	Data Reduction and Transformation	6-Mar-14	25-Mar-14	Task 14
18	Meeting with mentors	10-Mar-14	10-Mar-14	None
19	Meeting with mentors	14-Mar-14	14-Mar-14	None
20	Data Mining	13-Mar-14	27-Mar-14	Task 17, Input from mentors
21	Check Point Presentation	18-Mar-14	18-Mar-14	
22	Assess Models	20-Mar-14	10-Apr-14	Task 20, Input from mentors
23	Writing Report	20-Mar-14	9-May-14	Task 22
24	Integrate Software	10-Apr-14	2-May-14	Task 22
25	Work on Poster	11-Apr-14	9-May-14	Task 22
26	Poster Day	9-May-14	9-May-14	Task 23, Task 25

Table 3.1: Original tasks and critical dependencies

3.1.4 Management planning

Table 3.1 provides Project IX's initial task list. The tasks incorporate the course requirements as well as the steps in the KDD process. The TL an PM monitored task progress using ProjectLIBRE. As the need arose tasks were added.

In addition to the project plan, the team held semiweekly on Mondays at 10pm and Thurdays at 3pm. When necesary, generally every two weeks, the Thursday meeting would double as a mentor meeting.

Critical dependencies — iėdependencies that threaten to delay the schedule — were addressed with schedule changes and task assignments during the team meetings.

3.2 The analytic sandbox

An analytic sandbox is a staging area for performing exploratory data analysis without impacting Oncospace[**?**]. The KDD data cleansing and preprocessing step (see Section 2.4) occurs as data is extracted from the Oncospace and stored in the analytic sandbox.

The red rectangles in Figure 3.2 surround the tables in Oncospace that contain data moved to the analytic sandbox. The 3D data consists of region of interest (ROI) masks — stored in the *RegionsOfInterest* table

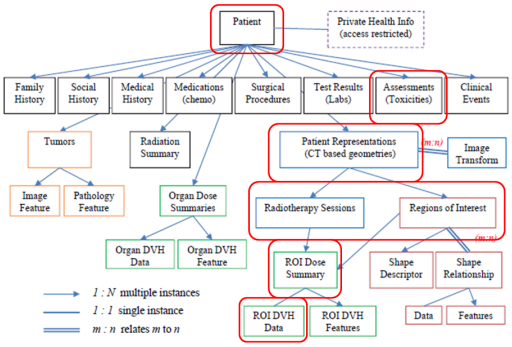


Figure 3.2: Project IX deliverables

— and the dose grids — stored in the *RadiotherapySessions* table. Data for both are from the treatment planning tools. The ROI masks is a binary grid of voxels for a given ROI (eġthe left parotid ROI).

Dose grids are stored in Oncospace as binary large objects (BLOBs) as an ordered list of dosages (in cGy), with each element representing a voxel. Each dose grid averages 3.5MB, as noted in previous work, compression can offer significant improvements in both storage requirement and communication latency with little additional processing time[**?**]. During as part of preprocessing, data grids are stored in the sandbox as gzipped BLOBs.

During the initial uploading of data into Oncospace, dose grids were stored in both big endian and little endian format; as part of data cleansing, the dose grids are stored in the sandbox exclusively in big endian format.

The data preprocessing also removed need for the intermediate *PatientRepresentationsId*. All data is directly linked to a patient via the *PatientId*. Ground truth data is in the *Assessments* table. Figure 3.3 is the data model for the resulting analytic sandbox.

3.3 The software

Project IX's software platform is primarily written in Java 7, with some Microsoft Transact-SQL queries, and Groovy scripting. The 3-D parotid visualizations are generated by Matlab.

3.4 Data mining and evaluation

For each patient i in the dataset we use the DVH to calculate $NTCP_i$ using Equation 2.2.

We then grouped dose into five bands in (cGy): (1) 500 - 2499; (2)2500 - 3999; (3) 4000 - 5499; (4) 5500 - 6999; (5) $\geq$ 7000

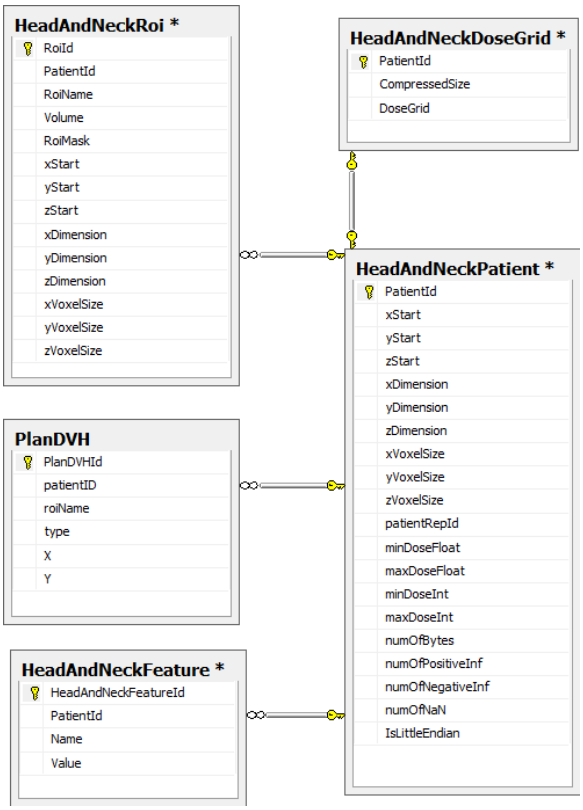


Figure 3.3: Project IX deliverables

4 Results

5 Discussion

6 Conclusion

A Management plan