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 9, 2014

1 Introduction

1.1 Motivation and goals

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.

In attempting to address this trade-off, many questions naturally arise. Are there critical sections within organs that should be spared? More generally, what are the likely complication profiles for a variety of treatment regiments? These questions are ever more pressing as oncology are multi-modal, often involving combinations chemo and radiation therapy, and as increased survival rates translates into patients who, while treated successfully, will permanently suffer decreased quality of life due to complications[1].

A natural approach to answer the questions above is to create a multi-factored method to assess the probability of various complications, and use the method to assess a variety of treatment plans. However, traditional models for calculating what is known as normal tissue complication probability (NTCP) are not sufficient for such a task. For example, the Lyman-Kutcher-Burman (LKB) model only uses two factors — the maximum dose delivered, and the variation of dosage uniformly across the volume. The model accounts for neither a patient's medical history, nor other treatments the patients receives. Further, by assuming an equivalence between higher doses over small volumes and lower doses over large volumes, LKB does not account for dose placement.

The goal of Project IX is to create a platform that allows for the development of data-driven multi-factor models that can improve NTCP calculation and address the questions above. Such models can allow previous experience to inform treatment planning (Figure 1.1), and to obtain insights into the factors leading to treatment complications.

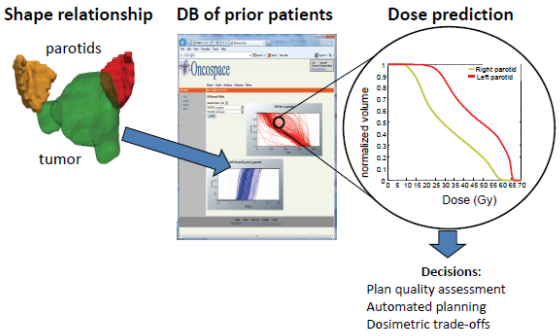


Figure 1.1: Using data on the clinical outcomes of previous patients provides insights into current patient treatment planning and assessment (courtesy of Todd McNutt)

Project IX seeks to build the platform around the Oncospace database, a cloud-based database of clinical information developed by Johns Hopkins Hospital's Radiation Oncology Physics Department. Along with other information Oncospace contains three dimensional treatment planning data. The department has employed this data to create shape descriptors to predict the nature of treatment plans[2]. This project seeks to apply the spatial data to NTCP calculation.

1.2 Contents

The background section presents: the foundational works of the LKB model; the nine-step knowledge discovery in databases process Project IX used to develop the platform; the data mining and assessment applied to the data; and the project management approach used.

The methodology section opens with a description of the project management methods directly applied to the project. The descriptions includes who did what, the required deliverables, and the management plan. Next the analytic sandbox and software developed are described. Finally, the methods of data mining, feature selection and evaluation are presented.

The results first provides a management report on how well the plan performed in the course of executing the project. Next the analytic sandbox and software implementations are presented. Finally, the outcome of data mining results are provided.

The discussion section assesses the implications of the data mining results. The future work examines the project in the context of the outstanding steps in knowledge discovery that were not addressed in this project. Finally, the conclusion highlights Project IX's deliverables, how they were satisfied, and our opinions as to what the project has demonstrated.

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.

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 [3] 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 [4], that is:

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

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

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

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

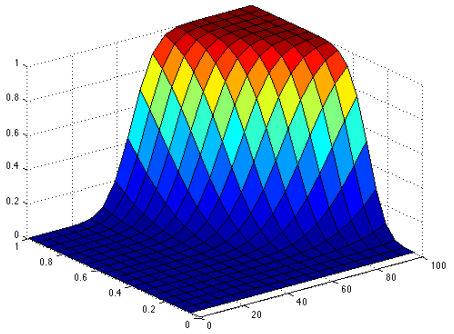
$$t = \frac{D - TD_{50}(V)}{\sigma(V)},$$

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

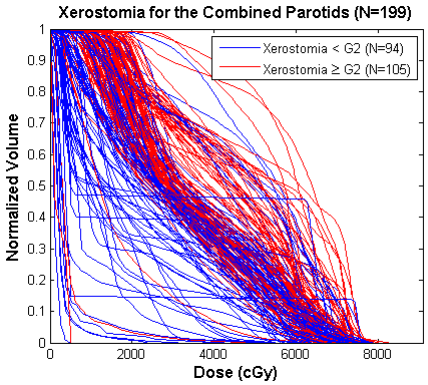
Ermani, et al. [5] and Burman et al. [6] provide widely cited values for m, n, and TD_{50} over a variety of organs and toxicities. Given the parameters, and assuming a constant dose, we can calculate NTCP from equations 2.1 & 2.2. In order to account for dose variation, LKB assumes a second power relationship between dose and volume [7]:

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

In practice, $D = D_{max}$, the maximum treatment dose the patient receives, and D_i corresponds to doses



(a) LKB NTCP surface for a parotid using parameters $n = 0.70$, $m = 0.18$, and $TD_{50} = 4600$[6]. 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.

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.

Marks, et al.[8] 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.[1] 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

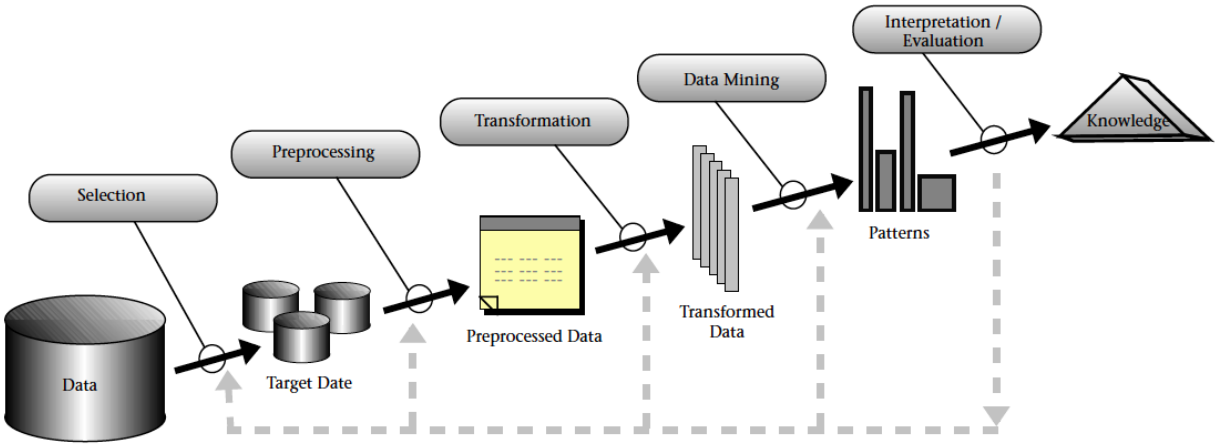


Figure 2.2: The process of extracting knowledge from data (copied from [9]).

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.

2.3 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.[9] 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.

[9] 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; (3) Data cleaning and preprocessing; (4) data reduction and projection; (5) matching the knowledge discovery goals with a data mining approach; (6) exploratory analysis with hypothesis and model testing; (7) data mining; (8) interpreting results and mined patterns; and (9) acting upon the discovered knowledge.

Cios & Moore[10] 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, [10] places emphasis on KDD's iterative

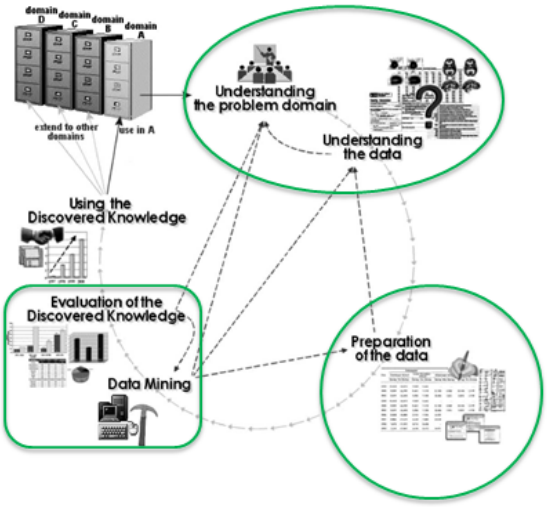


Figure 2.3: KDD areas covered by Project IX (copied from [10], selections added).

nature (Figure 2.3).

2.4 Data mining: techniques and assessment

Random forests[11] 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 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 [11].

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[10].

In assessing results Cios & Moore[10] 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; (3) external validity — the model is resistant to over-fitting and maintains perfor-

mance against other external datasets; and (4) clinical utility — the ability for clinicians to interpret the results in the context of treatment.

2.5 Project management

A "Surgical Team" surgical team methodology[12, 13] 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.

2.6 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[3]; 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})[4, 7]; and parameterization that fits dose to D_{max} and V_{eff} to a normal distribution[5, 6].

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 [8]. 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.[1]

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[9, 10]. 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.

3.1.2 Team member responsibilities — who did what

Project IX's used an abridged version of the "Surgical Team" approach (see Section 2.5). 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[14] 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[12, 13], the TL's responsibilities entailed: deciding upon the technical approach; mapping the KDD steps[9, 10] (see Section 2.3) to a tangible project plan; assigning tasks; transforming the "1st drafts" provided by the team into the finished deliverables; and data mining.

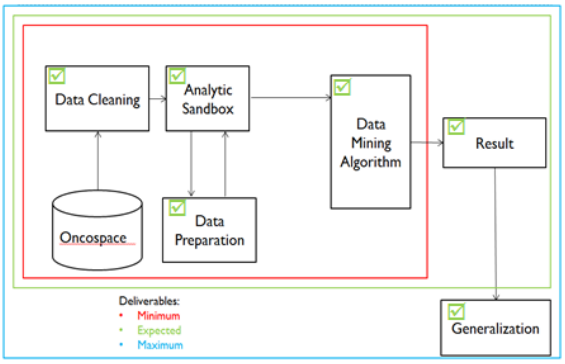


Figure 3.1: Project IX deliverables

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).

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 — ie̊dependencies 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[15]. The KDD data cleansing and preprocessing step (see Section 2.3) 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

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

— 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[16]. 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.

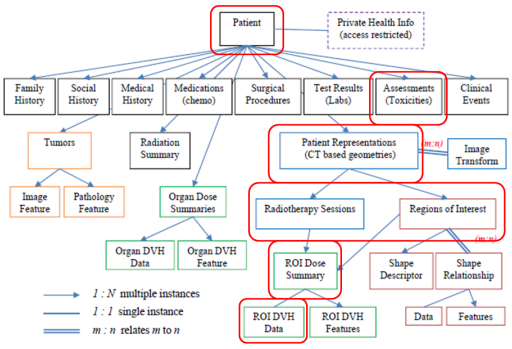


Figure 3.2: Project IX deliverables

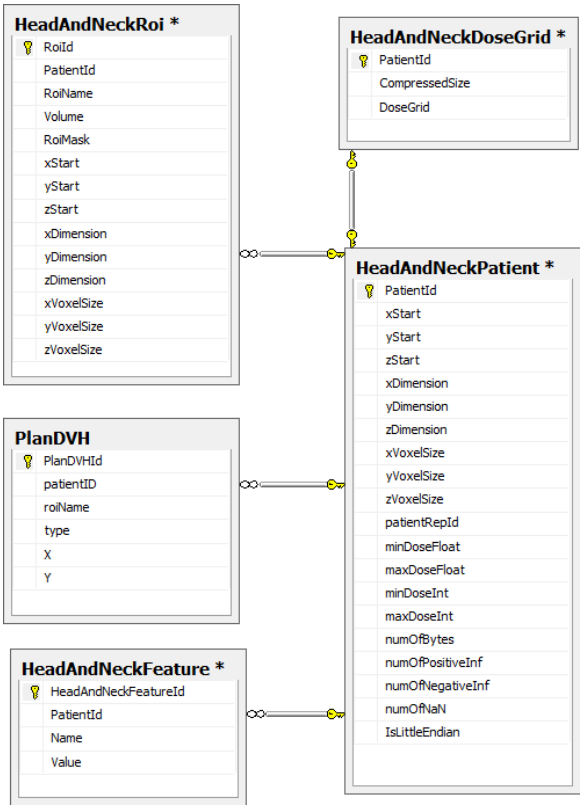


Figure 3.3: Analytic sandbox.

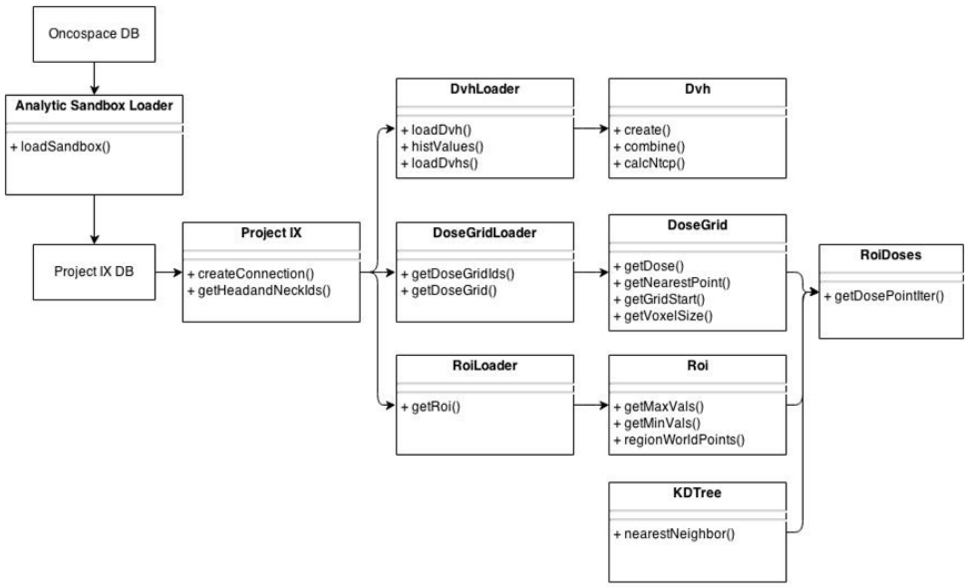


Figure 3.4: Software design.

3.3 The software

Project IX's software platform is written primarily in Java 7. Figure 3.4 displays the data cleansing and preprocessing steps that load the analytic sandbox. The *ProjectIX* class provides connectivity to the sandbox database. Loader classes use the connection to read the DVH's, dose grids and the ROI masks. The dose grids and the ROI masks have differing voxel sizes and boundaries. The *ROI* and *DoseGrid* classes convert data to a common world coordinate frame. *DoseGrid* uses a *KDTree* in order to find the nearest dose voxel for a given point in world coordinates. The *RoiDoses* class applies an ROI mask to a dose grid; provides an iterator over the points in the ROI mask that returns four-dimensional real number vectors containing a the world coordinate point from the ROI mask and a corresponding dose values from the dose grid.

3.4 Data mining and evaluation

3.4.1 Initial parotid gland features

For each patient in the dataset the V_{eff}, and D_{max} values both individually the left and right parotid glands as well as for the combined organ were calculated from the DVH's using Equation 2.4. Combined values were weighted for by organ volume. $NTCP$ values were calculated using Equation 2.2. The LKB parameters were $m = 0.18$, $n = 0.70$, $TD_{50} = 4,600$[5, 6].

Based on clinical experience of the mentors, dose levels were grouped into five bands (in cGy): (1) 500 - 2499; (2)2500 - 3999; (3) 4000 - 5499; (4) 5500 - 6999; (5) $\geq$ 7000. Using the DVH data, we calculated the percentage of total dose within the band delivered to the combined organ each gland received. We refer to

these values as "gross dose histograms" for a band.

Along each spatial dimension, the left and right parotid glands were divided into fifths using the maximum and minimum values for the dimension; this yielded 15 subregions per gland. We then calculated the percentage of the total dose within the band delivered to each subregion. We refer to these values as "dose grid distributions" for a band, on a gland's i^{th} dimension's j^{th} bin.

These calculations provided 169 features: the NTCP; three V_{eff} and D_{max} values; each gland's volume; 10 gross dose histograms; and 150 dose grid distributions.

A clinical history containing an Xerostomia grade greater than one represented a true positive complication outcome.

We created similar features for voice changes in the larynx using $m = 0.16$, $n = 0.45$[17]. However, we did not perform data analysis due to the low number of positive cases ($n = 99$, $n_+ = 8$).

3.4.2 Feature selection, classification, and evaluation

All feature selection, classification, ROC AUC calculations were performed using the Weka[14] suite. Using 10 fold sampling, we selected 18 features based on each features improvement of information entropy. Information entropy is defined as the following:

$$H = -\sum_i p(x_i) \log p(x_i), \text{ Information gain is: } H - H(f).$$

Information gain represents the decline in information entropy based on the conditional class distribution given a feature as compared to the unconditional distribution.

Linear regression and random forest classifiers were used with 10 fold stratified cross validation. Results for both classifiers, as well as the LKB model, were evaluated using ROC AUC.

3.5 Summary

Project IX's methodology employed a "surgical team" approach to manage a KDD project. The team delivered an analytic pipeline that can clean and preprocess data within Oncospace; store the results in an analytic sandbox; and transform selected data into a format appropriate data mining. Further, we performed and evaluated linear regression and random forest classification using spatial information derived from clinical treatment plans for xerostomia. We prepared larynx data for mining voice complications, but did not proceed due to the low number of positive cases.

No.	Task	Start	End
1	Select Project	28-Jan-14	30-Jan-14
2	Maintain Wiki	28-Jan-14	9-May-14
3	Project Planning Presentation	11-Feb-14	11-Feb-14
4	Project Planning Report	17-Feb-14	17-Feb-14
5	Project Planning	3-Feb-14	17-Feb-14
6	Setup Development Environment	6-Feb-14	18-Feb-14
7	Literature Review	11-Feb-14	4-Mar-14
8	IRB	14-Feb-14	19-Feb-14
9	Database Access	27-Feb-14	27-Feb-14
10	Meeting with Mentors	20-Feb-14	20-Feb-14
11	Begin Preparing Paper Seminar	20-Feb-14	12-Mar-14
12	Define Initial Data Preprocessing Interface	27-Feb-14	3-Mar-14
13	Target Database Access	6-Mar-14	6-Mar-14
14	Data Cleansing and Preprocessing	6-Mar-14	31-Mar-14
15	Meeting with Mentors	6-Mar-14	6-Mar-14
16	Develop Target Database	6-Mar-14	20-Mar-14
17	Data Reduction and Transformation	13-Mar-14	15-Apr-14
18	Seminar Presentation - Fumbeya	13-Mar-14	13-Mar-14
19	Meeting with Mentors	27-Mar-14	27-Mar-14
20	Plan for Check Point	24-Mar-14	31-Mar-14
21	Data Mining	18-Apr-14	2-May-14
22	Check Point Presentation	1-Apr-14	1-Apr-14
23	Assess Models	25-Apr-14	2-May-14
24	Seminar Prep	2-Apr-14	14-Apr-14
25	Seminar Presentation - Hilary/John	15-Apr-14	15-Apr-14
26	Writing Report	15-Apr-14	9-May-14
27	Work on Poster	15-Apr-14	9-May-14
28	Prep Final Checkpoint	21-Apr-14	28-Apr-14
29	Final Checkpoint Presentation	29-Apr-14	29-Apr-14
30	Code Review and Documentation	28-Apr-14	6-May-14
31	Poster Day	9-May-14	9-May-14

Table 4.1: Final tasks.

4 Results

4.1 Management report

Project IX was able to meet its maximum deliverables, and progress was generally orderly (Figure 4.1). Comparing the initial task list with Table 4.1, the major delays surrounded data preprocessing and transformation, and feature extraction. These problems led to data mining beginning approximately one month later than anticipated. However, once the features were ready, data mining progressed relatively quickly.

A key factor driving Project IX's successful completion of its deliverables is consistent meetings with and input from the team's mentors. In addition, early in the project, the team's development and research environment was set. This allowed for easy coordination of the teams work.

Consistent team meetings, with only one cancellation throughout the semester, coupled with the surgical team approach led to rapid turnaround of work and the ability to quickly adapt to delays impacting dependencies.

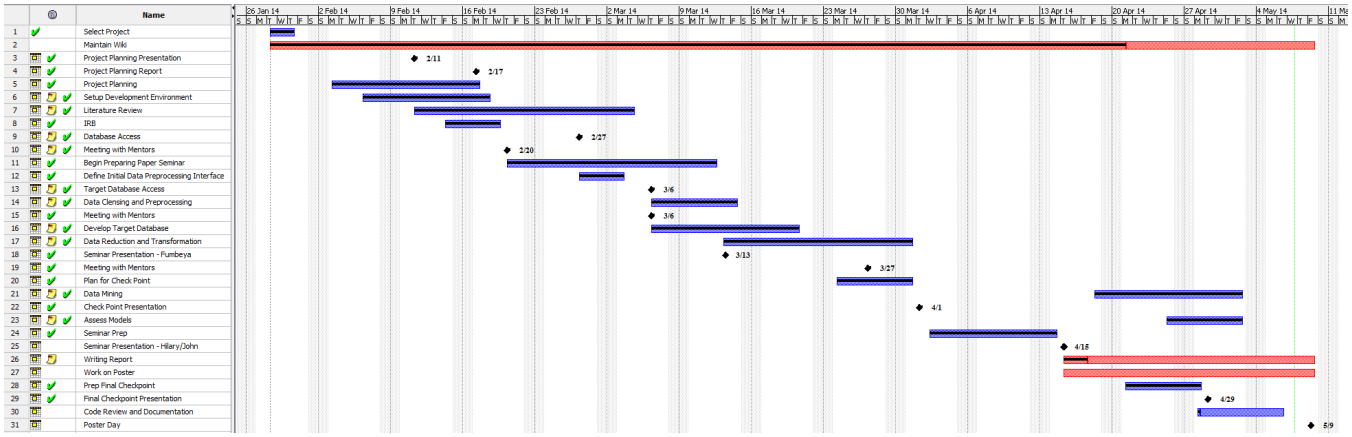


Figure 4.1: Project IX's Gantt chart

4.2 The analytic sandbox

A key benefit of the analytic sandbox is the decoupling of KDD operation from the underlying data source. The decoupling permitted not only the prototyping of different data models for analysis, but also the safe modification of data storage. As noted in Section 3.2, the data for dose grids in the sandbox is stored in a gzip format. The resulting 73.8% compression rate reduced the average dose grid storage size from 3.22MB per patient to 864.37KB and the total storage from 1.09GB to 292.91MB.

The reduction is size is particularly important in KDD where it may be necessary to transmit, the entire dataset over network connections.

4.3 The software

The data cleansing and preprocessing Java software manages transformations that are not practical to perform using SQL. For example the software performs the compression described previously. In addition, the consistent use of big endian encoding is also implemented in Java.

The data reduction and projection Java software can read the encoded dose grids and ROI masks (see Section 3.3). The *RoiDoses* simplifies the task of combining spatial dose data with organ locations. This not only simplifies feature extraction, but also allows for informative visualization. The images were used for verification by examining unusual dosage patterns and confirming the result against the original treatment plan (see Figure 4.2). Additionally, reviewing the entire dataset provided insights on feature design.

4.4 Data mining

Tables 4.2 & 4.3 lists the features selected for data mining using 10 fold stratified cross validation. With Xerostomia, the LKB model is the top ranked feature; however in voice the model provides essentially no

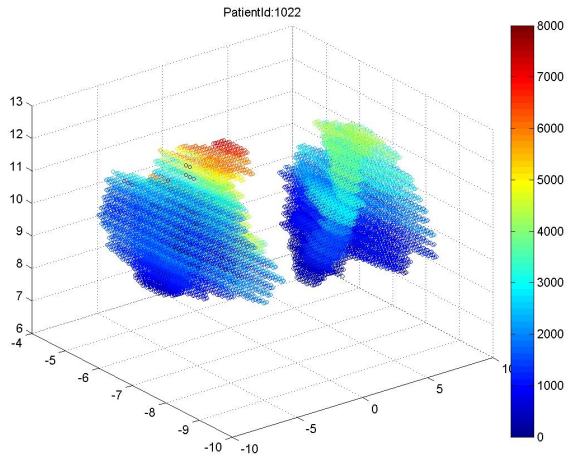


Figure 4.2: Spatial dose visualization for the parotid glands. The top fifth of the right parotid gland received 100% of the planned dose above 7,000 cGy.

information gain (0.002 ± 0.016).

Linear regression outperformed random forest classification in both toxicities; and in both cases, bagging improved the linear regression classifier's performance (Table 4.4).

4.5 Summary

Using the KDD process and project management approach described in Sections 2.3 and 2.5, respectively; Project IX successfully created a medical physics data mining pipeline for the Oncospace database.

The team delivered a analytic sandbox that both allows for ad hoc, exploratory analysis as well as features data compression and standardization. The software allows for the straight forward development of feature vector for data mining, as well as visualizations.

The system was applied to assess xerostomia and voice change NTCP. Linear regression outperformed LKB, with further improvement after bagging.

5 Discussion

The data mining analysis presents a number of interesting insights. That linear regression outperforms the LKB demonstrates that improving prediction performance is possible. That the high bias linear regression classifier outperforms random forest suggests that there may be considerable predictive gains from more data. We can be relatively confident in the linear regression's performance due to the increase provided by bagging. The improvement over a large sampling of data implies the general stability of the original result.

average merit	average rank	attribute
0.11 ± 0.011	1 ± 0	Parotid_NTCP
0.084 ± 0.014	3.6 ± 2.46	GrossDoseHist_4000-5500_parotid_r
0.08 ± 0.013	4.5 ± 2.06	MaxDose_parotid_l
0.077 ± 0.012	5.6 ± 1.91	GrossDoseHist_>_7000_parotid_r
0.077 ± 0.011	6 ± 1.95	MaxDose_parotid_r
0.076 ± 0.012	6.9 ± 2.3	GrossDoseHist_5500-7000_parotid_l
0.073 ± 0.008	7 ± 4.02	Veff_parotid_all
0.075 ± 0.012	7 ± 2.45	GrossDoseHist_>_7000_parotid_l
0.075 ± 0.012	7.1 ± 2.3	GrossDoseHist_5500-7000_parotid_r
0.07 ± 0.009	8.8 ± 2.48	GrossDoseHist_4000-5500_parotid_l
0.067 ± 0.011	10.2 ± 1.89	MaxDose_parotid_all
0.054 ± 0.006	13.5 ± 1.12	DoseGridDist_2500_4000_parotid_r_Z-2
0.052 ± 0.005	14.1 ± 1.3	DoseGridDist_5500_7000_parotid_r_Z-3
0.046 ± 0.007	17.9 ± 2.95	DoseGridDist_4000_5500_parotid_l_Z-3
0.05 ± 0.006	18.3 ± 7.85	Veff_parotid_r
0.046 ± 0.008	19 ± 2.72	DoseGridDist_2500_4000_parotid_r_Z-3
0.045 ± 0.005	20 ± 3.79	DoseGridDist_4000_5500_parotid_r_Z-2
0.043 ± 0.007	23.1 ± 5.82	DoseGridDist_2500_4000_parotid_l_X-5
0.043 ± 0.015	27.1 ± 30.08	GrossDoseHist_2500-4000_parotid_l

Table 4.2: Xerostomia (parotid gland) features selected based on information gain, using 10 fold cross-validation.

average merit	average rank	attribute
0.104 ± 0.005	1.4 ± 0.59	DoseGridDist_4000_5500_larynx_Z-0
0.076 ± 0.011	4.7 ± 5.07	DoseGridDist_7000_2147483647_larynx_Y-0
0.076 ± 0.011	5.1 ± 1.36	DoseGridDist_500_2500_larynx_Y-0
0.095 ± 0.026	5.1 ± 8.63	DoseGridDist_2500_4000_larynx_Z-1
0.098 ± 0.027	5.6 ± 13.37	Veff_larynx

Table 4.3: Voice (larynx) features selected based on information gain, using leave-one-out validation.

Method	Xerostomia (Parotid Glands) ($n = 364$, $n_+ = 275$, $n_- = 89$)	Voice (Larynx) ($n = 99$, $n_+ = 8$, $n_- = 91$)
LKB	0.700	0.596
Bagged linear regression (1000 bags)	0.732	0.916
Linear regression	0.730	0.893
Random forest (1000 trees)	0.701	0.870

Table 4.4: Data mining ROC AUC results.

The performance of random forest algorithms converge based on the Law or Large Numbers; the convergence rate is a function of the independence and predictive power of the individual trees[11]. Given that we used a reasonable large number of trees, we believe that it is the lack of data that is leading to the random forest's over weakness.

In evaluating the absolute ROC AUC values, it is important to emphasize that the need for larger, well balanced datasets — the voice data in particular merely contains eight positive cases. Nonetheless, we do observe the considerable differences in the results from xerostomia versus voice change. In the former, the LKB model provide the highest information gain, and performs respectably in comparison to other models; The xerostomia LKB almost matches the random forest's performance. In voice complications LKB provides little information gain; the model significantly underperforms the other models.

Many high information gain xerostomia features are also aggregates over either of the two glands or both. Save one, the small number of high information gain voice features are location specific.

These results may suggest that dose placement may a greater factor in larynx irradiation than in parotid gland irradiation. LKB makes a uniform volume assumption (Section 2.2) that does not account for placement. The existence of critical sections where, perhaps lower doses, can have a large effect, would not be captured by LKB.

6 Future work

Project IX's work consists of one iteration of the first seven steps in the KDD process (Section 2.3)[9]. The two remaining steps are interpretation and implementation. Those tasks necessitates placing the data mining results in a clinical context. One approach in addressing the issues would be extending the dataset both internally as well as with data from other treatment centers. This would allow for external validity checking (Section 2.4[10], verification of the results presented, and further insight into the complication causes.

If there are critical sections in the larynx, deformable registration of the volume, that is creating a mathematical model of the organ's shape may provide more specific dose placement information. We also hope to apply Project IX's platform to a number of other toxicities and organs within Oncospace.

7 Conclusion

Project IX's original deliverables were: at a minimum to provide a pipeline for analyzing the Oncospace dataset; with an expectation of using the pipeline to create a useful model for NTCP prediction of xerostomia due to parotid gland irradiation; and ideally to prove the pipeline's generalizable nature by applying the

platform to a second toxicity and organ combination.

We delivered on the minimum requirement by using a combination of the knowledge discovery process[9] and IBM's surgical team approach to project management[12, 13].

To reach our expected an maximum deliverables, we performed classification using linear regression and random forests. We compared the results of both algorithms to the conventional LKB model using ROC AUC.

At the conclusion of the Project IX, we believe we have demonstrated: (1) a successful approach to applying Big Data techniques to medical data collected in the clinic; (2) the potential of a data-driven approach to improve outcome prediction in radiation oncology; (3) the likelihood that larger datasets will allow for more robust models; (4) the possibility that irradiation of critical sections within the larynx may be the determining factor in voice complications.

References

[1] Søren M Bentzen, Louis S Constine, Joseph O Deasy, Avi Eisbrunch, Andrew Jackson, Lawrence B Marks, Randall K Ten Haken, and Ellen D Yorke. Quantitative analysis of normal tissue effects in the clinic (QUANTEC): An introduction to the scientific issues. *International Journal of Radiation Oncology Biology Physics*, 76(3):S3–S9, February 2010.

[2] Binbin Wu, Francesco Ricchetti, Giuseppe Sanguineti, Michael Kazhdan, Patricio Simari, Robert Jacques, Russell Taylor, and Todd McNutt. Data-driven approach to generating achievable dose–volume histogram objectives in intensity-modulated radiotherapy planning. *International Journal of Radiation Oncology* Biology* Physics*, 79(4):1241–1247, 2011.

[3] P Rubin and G Casarett. Direction for clinical radiation pathology. the tolerance dose. Technical report, Univ. of Rochester, NY, 1972.

[4] John T Lyman. Complication probability as assessed from dose-volume histograms. *Radiation Research*, 104(2s):S13–S19, 1985.

[5] Bahman Emami, J Lyman, A Brown, L Cola, M Goitein, JE Munzenrider, B Shank, LJ Solin, and M Wesson. Tolerance of normal tissue to therapeutic irradiation. *International Journal of Radiation Oncology* Biology* Physics*, 21(1):109–122, 1991.

[6] Chandra Burman, GJ Kutcher, B Emami, and M Goitein. Fitting of normal tissue tolerance data to an analytic function. *International Journal of Radiation Oncology* Biology* Physics*, 21(1):123–135, 1991.

[7] Gerald J Kutcher and C Burman. Calculation of complication probability factors for non-uniform normal tissue irradiation: The effective volume method gerald. *International Journal of Radiation Oncology* Biology* Physics*, 16(6):1623–1630, 1989.

[8] Lawrence B Marks, Ellen D Yorke, Andrew Jackson, Randall K Ten Haken, Louis S Constine, Avraham Eisbruch, Søren M Bentzen, Jiho Nam, and Joseph O Deasy. Use of normal tissue complication probability models in the clinic. *International Journal of Radiation Oncology* Biology* Physics*, 76(3):S10–S19, 2010.

[9] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37, 1996.

[10] Krzysztof J Cios and G William Moore. Uniqueness of medical data mining. *Artificial intelligence in medicine*, 26(1):1–24, 2002.

[11] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[12] Frederick P Brooks Jr. *The Mythical Man-Month, Anniversary Edition: Essays on Software Engineering*. Pearson Education, 1995.

[13] F. Terry Baker. Chief programmer team management of production programming. *IBM Systems journal*, 11(1):56–73, 1972.

[14] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[15] Bill Schmarzo. *Big Data: Understanding how Data Powers Big Business*. John Wiley & Sons, 2013.

[16] Fumbeya Marungo and Paul Taylor. Storage and breast region segmentation for a non-distributed approach to clinical scale content-based image retrieval in mammography. In *SPIE Medical Imaging*, pages 86740H–86740H. International Society for Optics and Photonics, 2013.

[17] Tiziana Rancati, Marco Schwarz, Aaron M Allen, Felix Feng, Aron Popovtzer, Bharat Mittal, and Avraham Eisbruch. Radiation dose—volume effects in the larynx and pharynx. *International Journal of Radiation Oncology* Biology* Physics*, 76(3):S64–S69, 2010.