HIG-2016-01

Health Informatics Group

COMP 396

Analyzing Prediction for Radiotherapy Waiting Time and Identifying Outliers

Di Zhang

December 2016

Approved for public release; distribution is unlimited.

COMP 396 HIG-2016-01
December 2016

Analyzing Prediction for Radiotherapy Waiting Time and Identifying Outliers

Di Zhang

School of Computer Science
McGill University, Montreal
di.zhang2@mail.mcgill.ca

COMP 396 Final Report

Supervised by: Laurie Hendren, Computer Science

Abstract: For patients, waiting time before the start of any treatment can be uncertain and stressful. With the increasing use of an electronic database to store patient information in a hospital setting, the radiation oncology department at the McGill University Health Center has developed an algorithm to predict treatment planning waiting time. This project analyzes the possible outliers from the machine learning algorithm in an attempt to understand the waiting time behavior. Factors considered include total waiting time, waiting time between each stage of the treatment planning process, the type of cancer and the month when treatment starts. By testing the prediction algorithm, it is seen that the more patient information are used, the lower the chance of getting prediction errors. Furthermore, the research also suggests that the waiting time outliers should be identified separately within each step of the planning process; this is done by finding if the waiting time falls into the confidence interval defined for each stage. Finally, a program to find outliers was defined. In conclusion, the report discuss the possible application of the results and ways to improve the outlier algorithm. Analysis was done using Python, R and Excel.

Table of Contents

Figures and Tables

Figures

Tables

Acknowledgements

This work was supported by the Health Informatics Group at the MUHC. I would like to sincerely thank to all the leaders of the HIG; Prof. Laurie Hendren, Dr. John Kildea and Dr. Tarek Hijal. Your passion, encouragement and enthusiasm is what made this project possible. Special thanks to Ackeem Joseph and Robert Maglieri for setting up and explaining the structure of the database and to Marc Palaci for helping me understand his prediction algorithm. Finally, thanks to all the McGill HIG undergraduate students for the insightful feedback and advice. Working with everyone was a wonderful experience.

Abbreviations

EHR: Electronic Health Records

HIG: Health Informatics Group

MUHC: McGill University Health Center

1 Introduction

Patients come to hospitals for medical and surgical treatments, usually with an above average anxiety level and a desire to meet a health professional in the shortest time. Previous research has shown that uncertainty about waiting time duration in a hospital setting can create distrust between a patient and the staff within the hospital [1]. Well aware of this situation, health professionals however need proper time to prepare each patient's file. In particular, as one of the main specialties to treat cancer, radiation oncology is a science that needs a scrutinized planning process to increase the success rate of the treatment. It is believed that understanding and predicting the waiting time for radiotherapy treatment would greatly decrease the stress factor for patients and maintain the quality of the service.

The goal of the Health Informatics Group (HIG) led by Prof. Laurie Hendren, Dr. John Kildea and Dr. Tarek Hijal from the McGill University Health Center (MUHC) is to create tools that can cater to the needs of hospital patients by using electronic health records (EHR) in clinical setting. In particular, M.Sc. candidate Marc Palaci has developed a machine learning algorithm to tackle the lack of accuracy for appointment scheduling by predicting the start date of radiotherapy treatment [2]. This algorithm has an accuracy of around one to two days in predicting waiting time duration, but has a number of unexplained behavior. The goal of this project is to test this algorithm, understand the behavior of the waiting time based on various features and be able to identify the outliers from the predictions. In line with the prediction algorithm, we only looked at patient history scheduled between January 2012 to September 2016 and that has priority of 3 or 4[1]. We expect the amount of information about the patient, the waiting time between stages, the type of diagnosis and the period of the month when treatment can start will all be important variables influencing the presence of outliers. The result from this report will be crucial to accompany the predicted waiting time for treatment and improve the patient's overall experience.

This report goes through the motivation of the paper, previous related work, discusses the results of the prediction algorithm, finds possible variables influencing the waiting time

[1]Each patient is giving a priority value between 1 to 4, with 1 meaning the patient's case is urgent and has priority over others.

and suggests a method to identify outliers. Finally, the report ends with a discussion on the application of the finding and possible future works to be done.

2 Background and Related Work

Machine learning is becoming a crucial tool to analyze large amounts of data and draw predictions on statistical trends. In 2015, a book on the specific usage of machine learning for radiation oncology has been published to advance precision medicine in oncology [3].

Within the HIG, multiple students took on COMP396 and COMP401 projects to better understand the waiting time before starting radiotherapy treatment. The work of students Alvin Leung and Marc Palaci especially were great introductions to the problem at hand. The former focused on understanding radiotherapy waiting time by stages of planning, diagnosis and oncologist and started a waiting time prediction algorithm. The latter improved on the machine learning algorithm by testing different models and ways to filter patient data [2, 4]. Furthermore, the process of radiotherapy waiting time has been a hot topic for many researchers. In 2015, the Cancer Quality Council of Ontario showed that the median number of days from first radiotherapy consultation to first treatment is 23 days in Ontario, with a heavy dependence on the cancer type [5]. In the work by Benk et al. also conducted in Ontario, it is observed that the waiting time vary significantly by health regions, age, diagnosis and gender [6]. Currently, in Canada, the target of "Receiving care within 28 days" after initial consultation for radiotherapy is met 98% of time in Quebec, 99% in Ontario and 93% in British-Columbia [7]. Other works focused on waiting times in emergency room setting, showing the importance of outlier studies. In the United States, a study of laboratory test turnaround time in 11 community hospitals concludes that there is a heavy correlation between the percentage of outliers for test turnaround time and the average length of stay in the emergency department waiting room [8]. This present project is hence a continuation from previous works and focuses on the role of outliers when predicting radiotherapy waiting time.

3 Description of Treatment Planning

Radiation oncology is one of the three main specialties involved in the curative or palliative treatment of cancer. The work of the oncologist consists in prescribing radiation dose, monitoring radiotherapy planning and following the patient treatment progress.

The hospital receives an application from a patient's referring doctor containing the patient information, the cancer type and any relevant information from previous diagnosis. From there, the radiation oncology department opens a new patient file and sets an initial consultation appointment. After the consultation, if both the patient and doctor decides to proceed with treatment, a requisition form is filled. The waiting time for treatment is officially defined to be the period between the first "CT-Sim" appointment and the time when the patient is "Ready for Treatment". If a patient officially decides to go through the radiation oncology treatment, a CT simulation scan is done to assess the physical condition of the patient. This step includes waiting to be matched to both scan and radiation machines. Afterward, at the "MD Contour" step, hospital staff contour the tumor regions on the CT images. Depending on whether the region to be treated is around the neck or the head, this step may also include the preparation of a thermoplastic mask. The next stage is "Dose Calculation"; measuring the proper quantity of radiation for tumor treatment. When all preparations are done, the plan is then approved by the doctor who writes a prescription note and by the physicists who check the feasibility of the plan, "Physics QA". If all steps are met, the patient's file is marked as "Ready for Treatment".

Scheduled Planning Step	Scheduled Start Time
Ct-Sim	2015-09-22 08:00:00
Ct-Sim	2015-09-24 11:30:00
READY FOR MD CONTOUR	2015-09-25 09:55:00
READY FOR DOSE CALCULATION	2015-09-25 10:56:00
READY FOR DOSE CALCULATION	2015-09-28 13:26:00
READY FOR PHYSICS QA	2015-10-05 10:14:00
READY FOR TREATMENT	2015-10-05 13:23:00

Table 1. Example of Treatment Planning Process

Table 1. is an example of the planning schedule before the first treatment for an average patient where all steps are done in order. It is important to note that even if this plan is the most common path, each patient has different situations and needs, changing the planning process accordingly.

4 Database Description

4.1 ARIA Database

The data for this project comes from ARIA, a relational database used by the MUHC Radiation Oncology department to document EHRs. To ensure traceability, each data information is identified by an ARIA serial number.

The ARIA database has a large number of fields and tables. For the purpose of this project, only tables relevant to waiting time were used: Patient, Diagnosis, Priority, Alias, Appointment, Task, and PatientDoctor. The Alias table was needed to identify the type of cancer and the appointment and task description. An important distinction should be made about task and appointment. While both includes time stamps and all important scheduling information, appointment entries specifically imply a meeting with a doctor.

4.2 Data Filtering

The data analyzed in this project comes from a static SQL file of all patient information from 2006 up until September 2016. This represents 43159 different patient serial numbers. We decided to only look at data that was collected after January 1st, 2012 as prior records may not be representative of the overall hospital performance. Also, the algorithm only focuses on patients with priority 3 or 4. The reason is that patients with priority 1 or 2 are seen within a time frame of 1 to 3 days, making the use of a prediction algorithm redundant. To eliminate all values that may cause errors, filtering is done such that the patient must have at least one registered appointment or task in the database. With these criteria, there is 4982 patients remaining that can go through the prediction algorithm successfully.

As the goal of this project is to look for waiting time duration, we did a time difference calculation between scheduling starting time such that only working days are included (holidays and week-end are not accounted). Hence, any reference to days implies working days. Furthermore, we noticed that all CT-Sim appointments are scheduled in 30-minute interval. For example, a CT-Sim appointment may be scheduled at 14:00, but have a MD Contour task scheduled at 13:50. Hence, if any waiting time differences between CT-Sim and any other step is smaller than 30 minutes, we adjust that difference to zero.

5 Discussion

5.1 Performance of Prediction Algorithm

The machine learning algorithm used is the Gradient Boosting Regression, which from previous works, shows best results. Key features used to create the models includes primary oncologist, doctor assigned for contouring, cancer type, priority value, start date of the appointment or task, priority due date, age, and gender of the patient. The patient history used to do the prediction takes the first appointment date for each treatment stage. As each patient inside the database may be at different stages, the patient information is not uniform. Therefore, the program must be able to adapt to each situation. The algorithm was hence written to create six different models depending on the current stage of the patient: “Medically Ready”, “CT-Sim”, “Ready for MD Contour”, “Ready for Dose Calculation”, “Prescription”, and “Ready for Physics QA”. If the patient is “Ready for Treatment”, no model is needed and the algorithm returns automatically the date scheduled.

To validate the accuracy of the prediction algorithm, we set a 4:1 ratio to separate the source database into a training and testing set, respectively. We then used sklearn's existing metric functions to do the cross-validation and find the mean errors [9]. The test is furthermore repeated 50 times to get the mean value. The formula for mean squared error where *x* is the actual value and *p* is the predicted value is:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - p_i)^2 \tag{1}$$

The formula for mean absolute error where *x* is the actual value and *p* is the predicted value is:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - p_i| \tag{2}$$

For each model, the results are in Table 2:

	Mean Squared Error	Mean Absolute Error	Standard Deviation
Medically Ready	10.9903	2.6780	0.2580
CT-Sim	9.4464	2.3345	0.2136
Ready for MD Contour	7.9789	2.1980	0.1910
Ready for Dose Calculation	5.9189	1.7291	0.1648
Prescription	2.2381	0.7196	0.1485
Ready for Physics QA	0.1338	0.1395	0.0412

Table 2. Error Indicators of Machine Learning Models from Each Stage (Days)

Generally, the mean absolute error is used to estimate the fitness of the prediction algorithm to the data set. However, since we are more interested by the possibilities of outliers, the mean squared error equation is a better indication as it squares the differences between a value and its predicted result. This implies that the presence of outliers would be more noticeable by MSE value. We observe that the presence of the outliers is significantly bigger earlier in the planning such that as at "Medically Ready" stage, the MSE is 10.99 while at "Ready for Physics QA" stage, when most stages have been cleared, the MSE decreases to 0.13. This indicates that the algorithm has high probability of error when the patient is at an early stage of planning process. This makes sense as we expect the error to get closer to zero as more information are given as input in the machine learning algorithm.

To evaluate the performance of the algorithm, we compare the raw output date from the prediction algorithm with the actual date when the patient is ready for treatment. This represents a random sample of 4180 records to test the prediction. Table 3 shows the absolute difference between predicted treatment starting date and the actual treatment starting date, with unit in days.

Minimum	0
1st Quartile	0.0236
Median	0.0458
3rd Quartile	0.0854
Maximum	38.9861
Average	0.2945
Standard Deviation	1.5773

Table 3. Absolute Difference Between Predicted and Actual Ready for Treatment Dates (Days)

Using R, we compute with 95% confidence that the absolute difference between predicted date and actual starting date of treatment are within interval [0.25,0.34] days. This shows us that the prediction algorithm has great accuracy. This makes sense because our data set includes all patients from 2012 to 2016 and hence, the majority of patients have already reached Ready for Physics QA step. This prediction model is the most accurate one with mean squared error of 0.1338 (see Table 2).

The extreme value are mostly for cases with few planning steps scheduled or has an unusually long actual waiting time. For example, as seen in Table 4 below, in the case of the maximum absolute difference of 39 days, the patient did not have a Ready for Physics QA step recorded and had an unusual waiting time of 33 days between Dose Calculation and Ready for Treatment.

Scheduled Planning Step	Scheduled Start Time
Ct-Sim	2016-04-20 15:30:00
READY FOR MD CONTOUR	2016-04-26 16:47:00
READY FOR DOSE CALCULATION	2016-04-27 12:38:00
READY FOR DOSE CALCULATION	2016-04-28 12:25:00
READY FOR DOSE CALCULATION	2016-04-28 14:31:00
READY FOR TREATMENT	2016-06-15 15:10:00

Table 4. Example of Long Waiting Time between Dose Calculation and Ready for Treatment

To complete the prediction evaluation, we also compared the waiting time computed by the prediction algorithm with the actual waiting time duration by fitting a linear regression model. Since values are estimated using the same data to construct the input matrix, the regression line is expected to be unbias. As it can be observed in Figure 1, the fitted line with a slop of 0.9938 and an intercept of -0.102 fits values nicely and shows that the prediction model is extremely accurate.

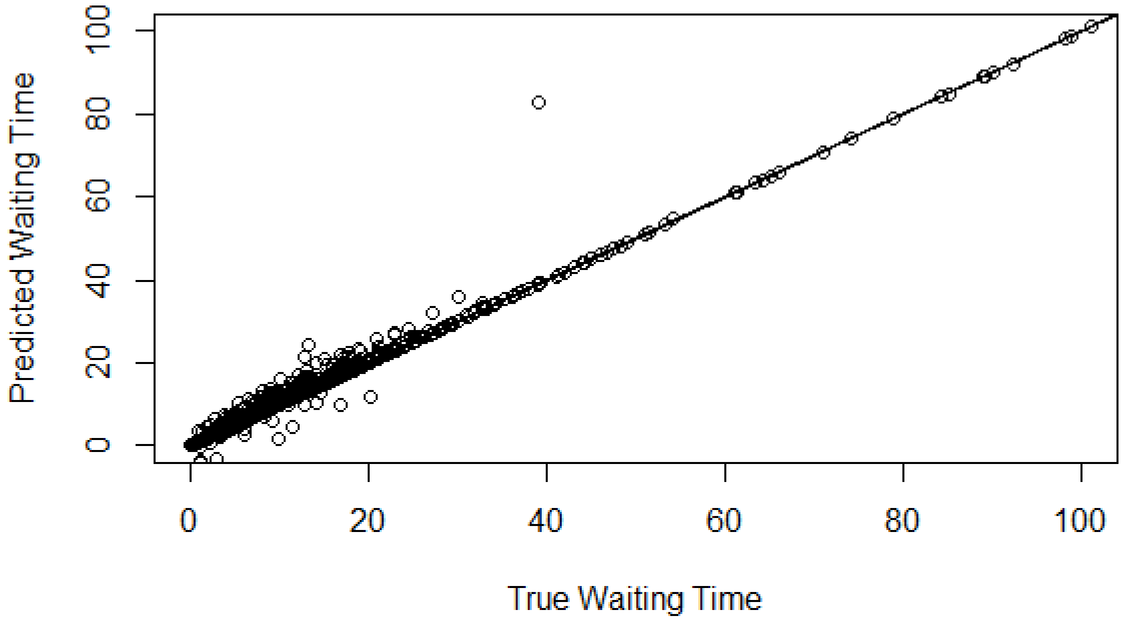

Figure 1. Regression Line Using Predicted Waiting Time as Y and Actual Waiting Time as X

In general, the possibility of having an unusual planning schedule is very small and most planning behaves like the machine learning algorithm predicts. With a large training set, the algorithm has great accuracy.

5.2 Waiting Time Analysis

The study of waiting time focuses on analyzing the duration between the first CT-Sim appointment and the treatment starting date. The work in this section includes finding a statistical distribution to explain the behavior of the waiting time, studying the waiting time between stages and looking at how variables like type of cancer and when treatment starts influence average waiting time. For each section, we discuss how these findings may help identify outliers.

5.2.1 Waiting Time Until Ready for Treatment

With 4180 patient waiting time records available, the mean waiting time from CT-Sim to Ready for Treatment is 10 working days. As shown in Table 5, by looking at the 1st and 3rd quartiles, it is observed that 50% of waiting time is within [5.02, 12.35] days, indicating that the waiting time distribution is right skewed. Note that the maximum waiting time duration of 594 days will be explained extensively in the section 5.3 ”Identifying Outliers”.

Minimum	0.0410
1st Quartile	5.0151
Median	8.1875
3rd Quartile	12.3462
Maximum	593.8486
Average	10.1891
Standard Deviation	15.4927

Table 5. Waiting Time Duration between first CT-Sim appointment to Ready for Treatment date (Days)

The best statistical distribution to fit the data hence needs to be a positive continuous right skewed model. Possible distributions are log-normal, exponential, gamma and chi-squared distributions. The Akaike Information Criterion is a measure of the relative fitness quality of a statistical model to a given set of data. With L being the maximum likelihood function of the model and k being the number of parameters in the model, the formula is:

$$AIC = 2 * \left\{k - lnL\right\} \tag{3}$$

We use the AIC to find the best statistical distribution fit by computing its value for each distribution mentioned above: 30662.544 for chi-squared distribution, 27738.94 for exponential distribution, 27161.14 for gamma distribution and is 27093.75 for log-normal distribution. The log-normal distribution, with the smallest AIC value is the best fit for our waiting time data. We plot the Quantile to Quantile (Q-Q) plot for log-normal model and are satisfied to see that the majority of values fit around a straight line. It is understandable that the distribution does not fit the values perfectly because of the many variability not taken into account by the distribution. Next, we fit the log-normal density function to the histogram of the total waiting time. As expected, the histogram is right skewed and the density function fits nicely. This allows us to see that even if the log-normal is the best statistical distribution to explain the behavior of the waiting time, it is still lacking in accuracy compared to the machine learning algorithm.

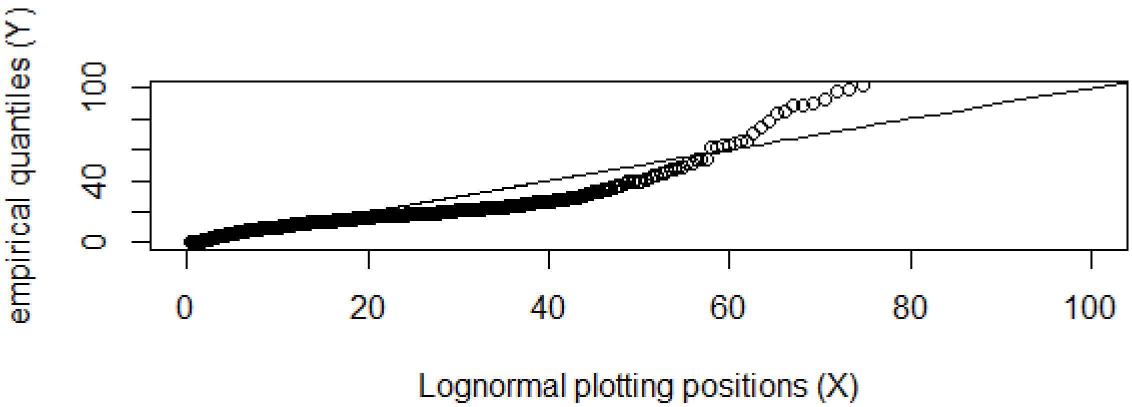

Figure 2. Q-Q Plot for Fitting Log-Normal Distribution to Total Waiting Time Duration

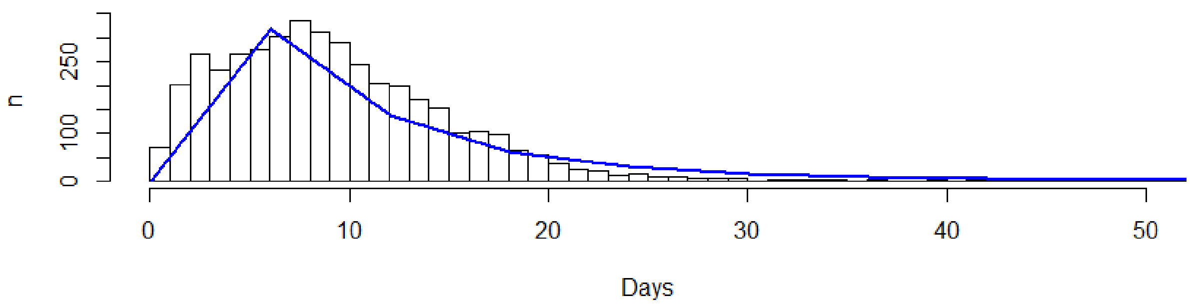

Figure 3. Histogram of Total Waiting Time Duration with Fitted Log-Normal Density Function

5.2.2 Waiting Time between stages

As previously discussed, the waiting time depends extensively on the planning steps. For each period between steps, we computed relevant statistics in Table 6 and plotted the histogram (histograms can be referred to in Appendix A).

	CT-Sim to MD Contour	MD Contour to Dose Calculation	Dose Calculation to Physics QA	Physics QA to Ready for Treatment
Minimum	-28.0035	-5.9674	0.0188	0.0007
1st Quartile	0.0819	0.6780	2.0691	0.0351
Median	0.2639	1.7576	4.5764	0.0792
3rd Quartile	0.9299	3.1448	7.8424	0.1514
Maximum	587.9917	204.8507	42.4243	7.1646
Average	2.0456	2.4051	5.5471	0.1954
Standard Deviation	14.5046	4.1950	4.4460	0.3865

Table 6. Waiting Time Duration Planning Stages (Days)

We use the median instead of the average to estimate waiting time duration as this indicator is less influenced by outliers. Additionally, the mean is typically greater than median for right skewed distribution. First, the step from CT-Sim to MD Contour is generally short (with a median at 0.2639). The MD Contour to Dose Calculation duration can be estimated to take around 1.76 days for completion. Next, the dose Calculation to Physics QA takes in median 4.58 days. Finally, the step of Physics QA to Ready for Treatment is the shortest as it is done within the day and has a small standard deviation of 0.3865.

When comparing these values, we notice the presence of both positive and negative extremes. We consider that negative time difference within one day are due to logistic error and do not represent an outlier to the actual waiting time duration. Especially, negative values are mostly due to an interchange of order between the steps CT-Sim, MD Contour and Dose Calculation. From patient history, it can be observed that multiple CT-Sim, MD Contour and Dose Calculation appointments/tasks were scheduled for the same patient, increasing possibilities of errors when the algorithm takes the latest date in the database. Dose calculation especially is a step that is repeated often since patient condition may vary as the planning goes on. This hypothesis is supported by the larger standard deviation value for each of these three stages compared to the smaller standard deviation of 0.3865 for Dose Calculation to Physics QA.

As for maximum waiting times, most outliers occur because the patient starts with a first step, takes a break and then continues with the radiotherapy planning. As the prediction algorithm extracts the patient history by the first instance of each step, the recorded first step may be a few years apart from when the planning restart. Although rare, these breaks could go up to two years, which lead our results to extremes. One noticeable trend is for patients to do a first CT-Sim, but then choosing to wait a period of time until the next appointment. This explains why even if the median waiting time between CT-Sim and MD Contour is small and most values are within 0 to 1 day (see Appendix A for histogram distribution), we still get extreme value that skews the standard deviation up to 14.5 days. We count 79 out of 4982 records with waiting time between two stages going beyond 30 days.

5.2.3 Waiting Time by Type of Cancer

Furthermore, other variables such as the type of cancer plays an important role in the patient's waiting time before treatment. In Table 9, all types of cancer had been listed in order of frequency. Breast cancer is the most common diagnosis, representing a quarter of all records and has an average of 11 days. The waiting time average is lowest for METS at 5.7 days in average and highest for Hepato-Pancreatic Biliary (HPB) cancer at 22 days.

Diagnosis	Number of Patient	Average Waiting Time
Breast	1127	11.40
Resp	487	8.60
METS	452	5.68
Prostate	383	14.70
CNS	308	8.91
HN	283	8.89
Haeme	232	9.78
LGI	232	9.32
Sarcoma	151	9.07
GU except prostate	143	11.12
Gyne	125	12.02
UGI	80	10.30
Skin	61	5.83
HPB	54	22.63
Other	27	22.54
Unknown Prim	26	5.77
Eye	9	5.72
Grand Total	4180	10.19

Table 7. Cancer Type Distribution and Average Waiting Time.

For the four most common diagnosis (breast, respiratory, METS and prostate), the histogram of the total waiting time are added to Appendix B. Breast cancer seems to be well distributed with a median of around 10 days. By comparing the histogram for the various cancer types, it is noticeable that a few diagnoses has distribution very different from our log-normal hypothesis. The waiting time for prostate diagnosis seems to be uniformly distributed within the 10 to 20 days range. A possible explanation would be that different oncologist and procedure needed for each diagnosis causes this disparity in waiting time. Since the distribution is so different by cancer type, the definition of outliers for diagnosis like prostate may be very different from the one for METS.

5.2.4 Waiting Time by Month of Planning

Finally, we also looked at the distribution of waiting time depending on the period of the year when the patient starts the treatment. The number of patients and the average time they had to wait are calculated by month in Table 8. The number of patients includes patients who started radiotherapy treatment between January 2012 to December 2015. From the computed results, it seems that the hospital is most productive at starting treatments from August to October, with a slow down during the new year period (December to January). Furthermore, an interesting observation is that high number of patients often correlates with small average waiting time. This makes sense as the number of working hours is approximately the same in each month and we would expect the hospital to balance the workload according to the demand. Overall, the average waiting time turns around 10 days for each month which is what we expected to see.

Month Start Treatment	Number of Patients	Average Waiting Time
Jan	121	9.67
Feb	261	8.04
Mar	261	8.12
Apr	254	9.14
May	255	9.13
Jun	258	9.62
Jul	265	8.11
Aug	277	9.71
Sep	259	11.22
Oct	284	9.00
Nov	253	11.09
Dec	225	11.59
Grand Total	2973	9.50

Table 8. Number of Patients and Average Waiting Time by Month

This study does not show that extreme values are dependent on specific months. However, this proves that there is a correlation between the period of the year and the total waiting time. It could hence be interesting to separate the year into two periods: busy and unbusy. This feature could then be used for improving both the prediction and outlier identification algorithms.

5.3 Identifying Outliers

In statistic, an outlier is defined to be a data point that is distant from other observation. From the research done above, we understand that the prediction algorithm can run into various different outliers:

- Patients may have an atypical planning process where stages like CT-Sim, MD Contour and Dose Calculation are not executed in order. This causes the waiting time difference between these stages to have negative values. If the absolute value of the difference is bigger than twice the standard deviation, we may identify the patient's file as an outlier. One example of our identified outlier is shown in Table 9. Here, the Dose Calculation was done four days before MD Contour which causes a negative waiting time difference between the two stages.
- Because the patient's condition is changing throughout the planning process, many stages may be needed to reassess the patient's prescription. A large number of recorded

steps implies longer planning time and higher chance of uncertainty. Especially, any record with more than 15 appointments or tasks can be considered to be an uncommon case.

- Some patient may choose to take a break in the middle of planning process. For example, a patient schedules a CT-Sim appointment, but wait for two years later to restart the planning. Because waiting time is defined to be from the first scheduled appointment until Ready for Treatment, these patients will have very long defined waiting time and can hence be considered as outliers. One example of our identified outlier is shown in Table 10. In this case, the patient's first appointment was in 2014, but it takes that person 2 years before restarting the planning.

Scheduled Planning Step	Scheduled Start Time
READY FOR DOSE CALCULATION	2015-02-05 11:42:00
READY FOR DOSE CALCULATION	2015-02-05 11:43:00
Ct-Sim	2015-02-05 12:00:00
READY FOR MD CONTOUR	2015-02-11 11:23:00
READY FOR PHYSICS QA	2015-02-13 13:22:00
READY FOR TREATMENT	2015-02-13 14:03:00

Table 9. Example of Patient with Unusual Order of Treatment Steps

Scheduled Planning Step	Scheduled Start Time
Ct-Sim	2014-05-06 15:30:00
Ct-Sim	2016-09-01 13:30:00
READY FOR MD CONTOUR	2016-09-06 15:18:00
READY FOR DOSE CALCULATION	2016-09-09 07:42:00
READY FOR DOSE CALCULATION	2016-09-09 08:22:00
READY FOR DOSE CALCULATION	2016-09-09 08:53:00
READY FOR PHYSICS QA	2016-09-14 11:26:00
READY FOR TREATMENT	2016-09-14 11:52:00

Table 10. Example of Patient with Very Long Waiting Time between First Ct-Sim and MD Contour

A new algorithm was hence written to identify these three types of outliers. The code is available on the HIG Github repository. First, the algorithm extracts the patient's history and keeps the first instance of each step scheduled. If there is more than 15 appointments/tasks recorded, this patient is considered as an outlier.

Second, the program computes the time difference between each available stages within the patient history (Ct-Sim, MD Contour, Dose Calculation, Physics QA, Ready for Treatment) as well as the total waiting time (difference between Ct-Sim and Ready for Treatment dates). The advantage of computing multiple time differences allows the program to find outliers even if the patient does not have many appointments or tasks scheduled yet. Each waiting time was classified as outlier if it is outside 2 standard deviation from the evaluated mean; we know with 95% confidence that this value is an outlier.

Third, if the waiting time is negative between Dose Calculation and Physics QA, between Physics QA and ready for Treatment or between CT-Sim and Ready for Treatment, the patient record would also be marked as an outlier. This is because we expect these steps to be proceeded in order. A negative value among these three waiting times is hence unlikely.

Finally, the program returns the result in a JSON string where values specifies the PatientAriaSer number, whether the patient is an outlier, the reason for the assumption and the total number of steps recorded for the planning process.

6 Conclusion

In conclusion, this project first looked at the machine learning algorithm performance over all available data. This showed that the prediction algorithm, by choosing to create different models, is extremely accurate, but is not immunized to outliers. We then looked at which statistical distribution best describes the waiting time behavior, the log-normal distribution and considered how different factors can influence total waiting time: waiting time between each steps until beginning of treatment, type of cancer diagnosed, and period of the year when patient starts treatment. This allowed us to define three cases of outliers: 1) irregular planning steps order causing negative values, 2) large number of appointments or tasks scheduled and 3) unusually long waiting time between steps leading to treatment. A program was then developed to consider each of these cases and output if the patient can be considered as an outlier and why. The discoveries made through this research as well as the developped outlier algorithm can be used to make the prediction algorithm more robust, analyze previous data and discover new patterns.

7 Future Works

As this initial research on outliers is done, the next step would be to improve the rules for the outlier algorithm. As the parameters used to categorize the outliers are based on the standard deviation, average and median we computed beforehand, an improvement would be to create a dynamic formula that automatically calculates the rules on whether patient data should be considered on outlier. Having such a tool would improve the prediction algorithm and help the hospital professionals to better plan treatment. Finally, although this study found interesting results on how type of diagnosis and when the treatment starts may impact waiting time, these results were not included into sorting outliers. Hence, another recommendation would be to look at how other factors can help identify outliers.

References

[1] Y. Jungkyoo and S. Marieke. Anxiety of patients in the waiting room of the emergency department. *TEI '10*, 4:279–286, 2010.

[2] M. Palaci. Muhc ml algorithm, 2016.

[3] I. El Naqa and M.J. Li, R. Murphy, editors. *Machine Learning in Radiation Oncology Theory and Applications*. Springer International Publishing Switzerland, 2015.

[4] A. Leung. Analyzing radiation oncology data for prediction of radiotherapy patient waiting time, 2014.

[5] Cancer Quality Council of Ontario. Wait time from diagnosis to radiation therapy. `http://www.csqi.on.ca/by_patient_journey/treatment/wait_time_from_diagnosis_to_radiation_therapy/`, 2015. Accessed: 2016-12-10.

[6] V. Benk, R. Przybysz, T. McGowan, and L. Paszat. Waiting times for radiation therapy in ontario journal=canadian journal of surgery. 49:16–21, 2006.

[7] Canadian Institute of Health Information. Benchmarks for treatment and wait time trending across canada. `http://waittimes.cihi.ca/QC`, 2016. Accessed: 2016-11-15.

[8] L.L. Holland, L.L. Smith, and K.E. Blick. Reducing laboratory turnaround time outliers can reduce emergency department patient length of stay. *American Journal of Clinical Pathology*, 124, 2005.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Appendix A: Histogram of Waiting Time Duration by Step

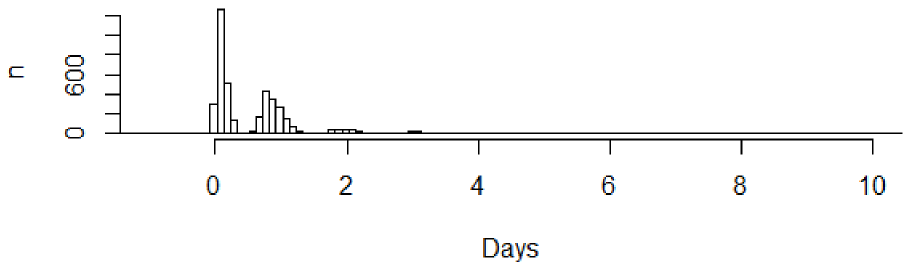

Figure A1. Histogram of Waiting Time Duration from Ct-Sim to MD Contour.

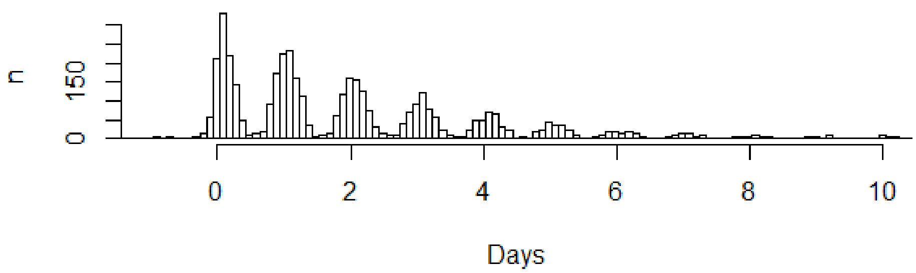

Figure A2. Histogram of Waiting Time Duration from MD Contour to Dose Calculation.

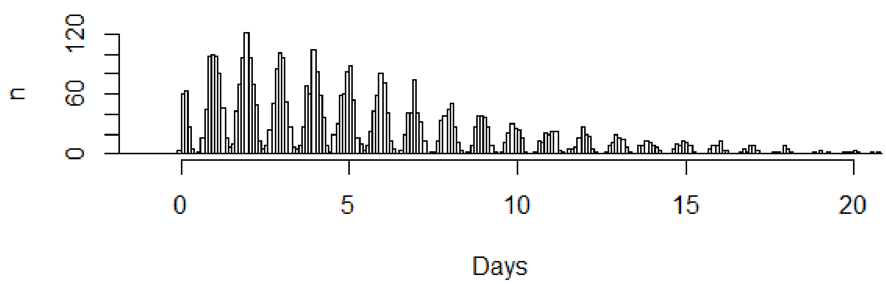

Figure A3. Histogram of Waiting Time Duration from Dose Calculation to Physics QA.

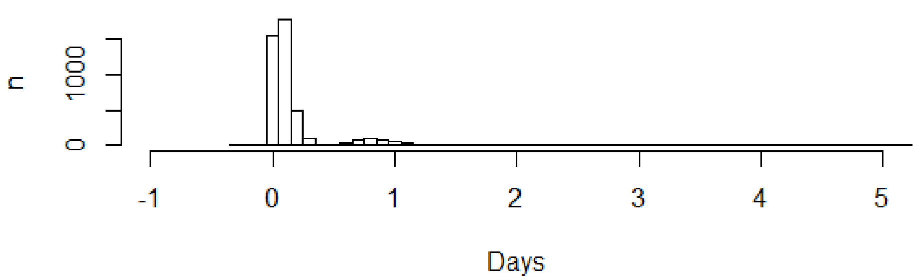

Figure A4. Histogram of Waiting Time Duration from to Physics QA to Ready for Treatment.

Appendix B: Histogram of Waiting Time Duration by Diagnosis

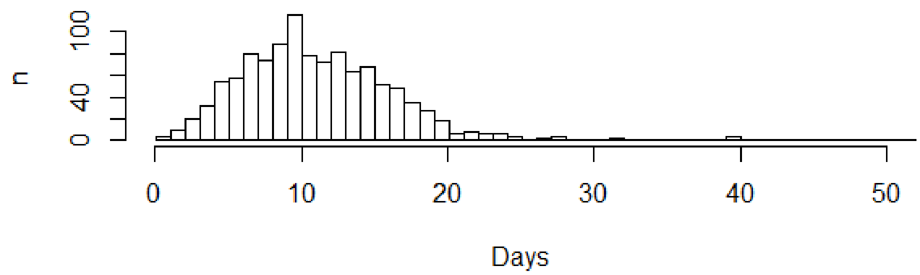

Figure B1. Histogram of Total Waiting Time Duration for Breast

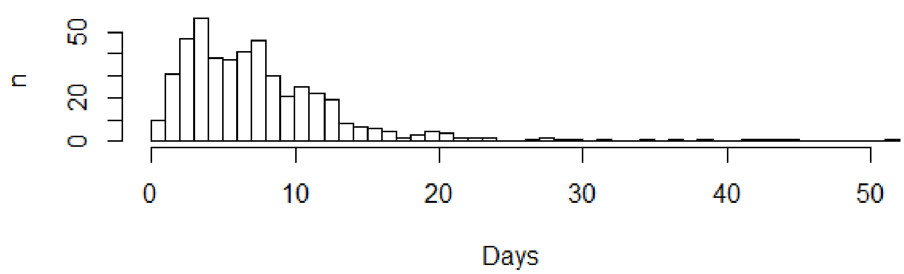

Figure B2. Histogram of Total Waiting Time Duration for Respiratory

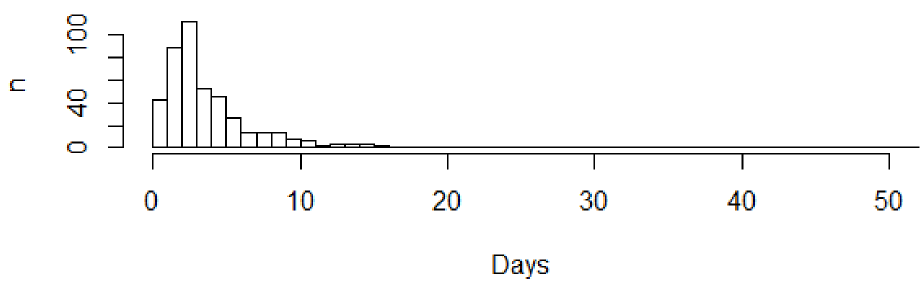

Figure B3. Histogram of Total Waiting Time Duration for METS

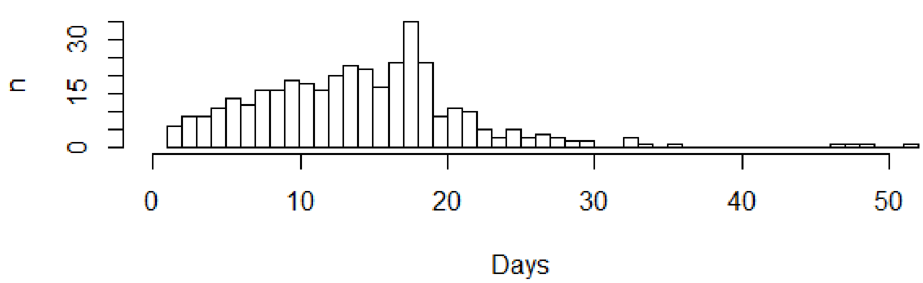

Figure B4. Histogram of Total Waiting Time Duration for Prostate