

Ethical Frameworks in Lung Cancer Survival: Predictive Modeling with Statistical and Machine Learning approaches on a Medical cohort

BSc. internship Thesis

Krijn van der Burg

14th January, 2019

Submitted in partial fulfilment of the requirements
for semester 5 of Bsc. Applied Data Science.

at

Fontys University of Applied Sciences.

STAGEVERSLAG VOOR FONTYS HOGESCHOOL ICT

Gegevens student:

Krijn R. van der Burg

Studentnummer: 3076458

Profiel / innovatiegebied: Software engineering / Applied data science

Stageperiode datum van 27-07-2018 t/m 18-01-2019 (90 werkdagen)

Gegevens bedrijf:

Radboud universitair medisch centrum

Radiotherapie

Nijmegen

René Monschouwer

Klinisch fysicus

Gegevens docentbegeleider:

Chris Geene

Gegevens verslag:

Titel stageverslag: No statistical difference found between lung cancer SBRT schedules overall survival and predicting 2-year lung cancer survival using machine learning.

Datum uitgifte stageverslag: 14-01-2018

Getekend voor gezien door bedrijfsbegeleid(st)er:

Datum:

14-1-18

De bedrijfsbegeleider,



Stagerichting:

Data Science

Gegevens student(e):

Naam student(e): Krijn van der Burg Studentnr.: 3076458Gegevens 1^o assessor:Naam: Chris Geene**A. Werkzaamheden**

Criterium	Niveau *				
	U	S	G	O	N/A
Theoretisch inzicht	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
Praktisch inzicht	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
Inzet	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
Methodiek en werkplanning	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
Onderzoekende houding	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
Breedte van de werkzaamheden	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
Diepgang van de werkzaamheden	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
Originaliteit en creativiteit	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
Werkdiscipline (schriftelijke rapportage, intern, extern, begeleiders)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input checked="" type="radio"/>
Zelfstandigheid	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
Mondelinge uitdrukkingsvaardigheid	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
Functionele samenwerking	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
Effect gedrag student op andere mensen	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
Waarde van de bereikte resultaten	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input checked="" type="radio"/>
Kostenbewust werken	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
Onderbouwing gemaakte keuzes en kritische houding	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>

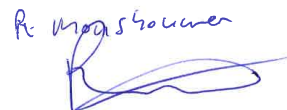
*
 U (unsatisfactory/onvoldoende)
 S (satisfactory/voldoende)
 G (good/goed)
 O (outstanding/uitmuntend)

Toelichting:

- Krijn heeft heel zelfstandig gewerkt met duidelijke focus
- Zijn eindpresentatie was op het juiste niveau voor het publiek
- Krijn heeft een goede ontwikkeling doorgeemaakt tijdens de stage.



R. Mooshouwer



Preface

This research began with Radboud Universitair Medisch Centrum’s goal of creating a lasting, innovative, and affordable healthcare system. My interest in machine learning and data analysis led me to join this cause, driven by a passion to improve lives in the non-commercial or public sector. I’m dedicated not only to discovering but also to developing tools that explore the possibilities of machine learning in medical care.

I could not have achieved my results and concluded this research without the excellent help of my supervisors, René Monshouwer and Martijn Kusters who provided advice and guidance throughout the research process.

Contents

1	Introduction	8
2	Radboudumc	8
2.1	Radiotherapy department	9
3	Patient process	9
3.1	Complaints and symptoms	9
3.2	Diagnosis	10
3.3	Treatment	10
3.4	Post-treatment follow-up	10
4	Project description	10
4.1	Project goal	11
4.2	Research question	11
4.3	Patient collected data	11
5	Available data sources	12
5.1	CASTOR, Patient’s General Treatment and Toxicity Information	12
5.2	RTHweb, patient’s basic information	12
5.3	Basisregistratie personen (BRP), patient’s survival status	12
5.4	DICOM, patient’s radiotherapy radiomics	12
6	Standardizing values tool	13
7	Data preparation	13
7.1	Multiple treatments	13
7.2	DICOM long to wide format	14
7.3	Combining data sources	14

8 Lung cancer SBRT survival research	14
8.1 Methods and materials	16
8.2 treatment planning and delivery	16
8.3 Dosimetric parameters	18
8.4 Analysis	18
9 Lung Cancer Survival Prediction Model	22
9.1 Choosing the Best Classifier	22
9.1.1 Why Random Forest Was Disregarded	23
9.1.2 Feature Analysis	23
10 Bibliography	31

Summary

Van der Burg’s study delved into the potential connection between overall survival and patient characteristics, along with radiation dose to the body in lung cancer patients undergoing various treatment schedules. The cohort, consisting of 302 patients, was categorized based on treatment schedules, and the Kaplan-Meier curve illustrated an overall survival average of 62.5% at 2 years. Log-rank and Wilcoxon methods gauged the calculated probability-value (p-value). Throughout the curve, spanning from 0 to 2500+ days, no significant survival difference emerged in either method, leading to the conclusion that different treatment schedules do not impact the overall survival of lung cancer patients at any interval.

Cox regression, principal component analysis, and Spearman’s rank-order correlation were employed to identify data features associated with cancer-related death. Univariate cox regression analysis pinpointed a significant association between cancer-related death and patient’s WHO status, gender, and dosimetric values such as PTV mean dose, PTV std dose, PTV d2pct_ingy, PTV v10, PTV v20, PTV v65, lungs maximum dose, oesophagus minimum dose, oesophagus v25, oesophagus v30, and oesophagus v35. Multivariate cox regression analysis revealed that patient’s WHO status and oesophagus minimum dosage were significantly linked to an increased cancer-related death, while female gender was associated with a decreased cancer-related death. Principal component analysis demonstrated an identical linear distribution for both classifications (deceased and alive) on a 2D plane. After selecting variables with over a 1% contribution in principal components 1 and 2, the result was nearly identical, with only event classification showing a slimmer distribution while remaining mostly under the no-event classification distribution. Spearman’s rank-order correlation indicated no correlation between dosimetric parameters and overall survival.

Six different classifiers (random forest, elastic net logistic regression, neural net, support vector machine, and LogitBoost) were applied to the entire dataset,

including follow-up toxicity-related datapoints, and a subset of datapoints available only post-treatment, excluding most toxicity features. All machine learning models underwent training and testing with 100 repetitions, 5 inner-folds, and 5 outer-folds. Ranked by AUC score, random forest (RF) and generalized linear model with elastic net (GLMNET) emerged as the two best discriminators. Random forest was not further explored due to traceability requirements. The generalized linear model was fitted by both elastic net regularization and features selected from cox regression analysis, principal component analysis, and Spearman’s rank-order correlation matrix. These features included patient’s WHO status, gender, and dosimetric values such as PTV mean dose, PTV std dose, PTV d2prct_ingy, PTV v10, PTV v20, PTV v65, lungs maximum dose, oesophagus minimum dose, oesophagus v25, oesophagus v30, oesophagus v35, and oesophagus v65. Both GLM and GLMNET models achieved AUC scores of 0.72, indicating similar importance values for features. In conclusion, patient characteristics and radiomics data exhibit predictive values for machine learning models, although insufficient to establish a fully reliable and accurate machine learning model.

Acronyms

DICOM Digital Imaging and Communications in Medicine

EPD Electronic patient's dossier

BRP (GBA) Basic person registration (Basisregistratie personen)

SBRT Stereotactic body's radiotherapy

PTV Planned target volume

Oes Oesophagus

Dose STD Dose standard deviation

Toxicity Degree to which irradiation can harm humans and cause effect.

Radiomics data Data features from medical imaging.

SBRT fractioning schema Planned schema of treatment frequency and its irradiation dosage.

Survival chance Chance of patient's survival related to health events occurring at specific time.

1 Introduction

This document provides an in-depth account of a 6-month-long research endeavor focused on cancer survival, employing machine learning and data analysis for classification purposes. While acknowledging that readers unfamiliar with these fields may find the detailed intricacies challenging, the narrative ensures clarity in presenting the problem description, initial circumstances, methodology, and ultimate conclusions. Chapter 7, titled "Lung Cancer SBRT Survival Research," stands as an independent manuscript intricately woven into this comprehensive report. Notably, this manuscript, slated for completion, is a prerequisite for author Van der Burg's pre-master study in Q3 and Q4 of 2019.

Stereotactic body radiation therapy (SBRT) takes center stage in this research, a technique deployed for treating both primary and secondary lung cancers. The benefits of hypo-fractionated radiotherapy in treating lung tumors are highlighted, emphasizing a condensed treatment course that minimizes clinic visits compared to conventional programs. Additionally, greater setup precision allows for a smaller irradiated volume. Citing studies, it is noted that patients with inoperable non-small cell lung cancer undergoing SBRT exhibit a notable 55.8% survival rate at 3 years, a marked improvement over the 20%-35% rate associated with conventional radiotherapy (Timmerman, 2010). However, potential drawbacks, such as uncertain effects of altered fractionation and the theoretical risk of altering the normal tissue-to-tumor tissue ratio with higher doses per fraction, are acknowledged. Notably, doses primarily targeting the upper heart region in lung cancer patients undergoing SBRT have been linked to non-cancer-related deaths (Stam, et al., 2016).

This study aims to evaluate the clinical outcomes of overall lung cancer patient survival under different SBRT fractioning schemas. Given the inherent difficulty in estimating cancer survival rates and times, especially considering the multifaceted influence of patient characteristics and treatments, the study explores the potential of machine learning models in predicting lung cancer survival duration. The goal is not only to benefit patients by providing predictive insights but also to offer doctors valuable information on the predictive values associated with patient characteristics and radiomics data. Beyond its research focus, this document serves as a comprehensive resource, delving into the project's intricacies, work approach, planning, communication agreements, conclusions, and evaluation.

2 Radboudumc

Radboudumc, Radboud universitair medisch centrum, is an academic hospital in Nijmegen that collaborates with Radboud university Nijmegen and is part of the Nederlandse Federatie van Universitair Medische Centra (NFU). UMC Radboud organization aims to be pioneers in shaping a sustainable, innovative and affordable health-care system for generations to come. Through a person-centred and innovative way and in close collaboration with Radboud's network.

The hospital has over a thousand beds and about ten thousand employees. More than three thousand students are trained at Radboudumc in Medicine, Biomedical studies Sciences, Dentistry, Molecular Mechanisms or Disease and Quality and Safety in the patient care. ("Over het Radboudumc", n.d.)

2.1 Radiotherapy department

The department Radiotherapy treats cancer through radiation, called radiotherapy; destroying cancer cells or inhibiting growth. Sometimes radiotherapy alone is sufficient, but often a combination of surgery, chemotherapy, or both is used. Patients who cannot be cured can have their quality of life highly improved by irradiating painful metastases. ("Over de afdeling", n.d.)

3 Patient process

Detailing the process how a lung cancer patient is diagnosed and treated and what data is collected and used in this study.

3.1 Complaints and symptoms

Lung cancer is mainly caused by smoking, responsible for more than 85% of all lung cancer cases. Tobacco has many different toxic substances which can lead to a high increase in cancer risk compared to non-smokers. Common lung cancer symptoms are:

- a cough that doesn't go away after two or three weeks
- a long-standing cough that gets worse
- persistent chest infections
- coughing up blood
- an ache or pain when breathing or coughing
- persistent breathlessness
- persistent tiredness or lack of energy
- loss of appetite or unexplained weight loss

Source: ("Lung cancer - Causes - NHS", 2015)

3.2 Diagnosis

When experiencing these symptoms, the person is examined by a general practitioner, asking about his/her general health and symptoms. A general practitioner may choose to examine further using a spirometer, measuring how much air the person breathes in and out. Diagnosis is established mostly in two ways; by X-ray or CT imaging; showing the tumour as a white-grey mass; or by tissue sample, surgically collecting a sample of the supposed tumour and analysing it in a lab. Patients treatment, survival and cure chance are mostly determined by the TNM cancer staging. Categorizing the tumour-size, metastases and lymph nodes which are all crucial for treatment and survival of the patient. ("Lung cancer", 2018)

3.3 Treatment

Patients with non-small-cell lung cancer with no other risks or complications, used to be operated to surgically remove the tumour and use chemotherapy to destroy any remaining leftover cancer cells; patients with small-cell lung cancer were always treated using radiotherapy as generally the cancer would've already spread to other body parts. ("Lung cancer", 2018) However, nowadays both cases are generally always treated using stereotactic body radiotherapy (SBRT). Which, simply put, is radiotherapy with high precision, thus not damaging healthy tissue. This, relatively new, method allows for the same effective result compared to conventual treatment methods while decreasing risks, e.g. infection from surgery. SBRT delivers various fractionated radiation dosages over multiple sessions (patients are radiated multiple times [over multiple days] to add to the total required irradiation dosage). There are 4 main treatment schedules determining the amount of treatment sessions and radiation dosage given. ("FAQs: SBRT", n.d.).

3.4 Post-treatment follow-up

The patient can have a follow up 1 week, 2 weeks, 3 weeks, 1 month, 6 months, 12 months and 24 months after the radiotherapy treatment stopped. Consisting mainly of toxicity (side effects) of the treatment but also general health the patient answers a questionnaire which the doctor fills in.

4 Project description

Practically all research at a hospital is to find new cures, quality control, improve patient's care or improve the hospital processes to provide better care. This project falls under both quality control and to improve patient's care.

4.1 Project goal

Result of this project is to conclude whether the different 4 different treatment methods, called schedules, have a statistically significant difference in survival probability and if the patient's medical data can be used to predict the patient's survival. All treatment schedules serve a slightly different purpose but can be categorized under maximizing survival chance and minimizing toxicity. By analysing the overall survival rate per schedule, the doctors get insight into the effectiveness of each schedule which the doctors can use to improve treatment or better inform their patients. Many patient's medical and biological characteristics and treatment affect survival chance and time. A machine learning model to predict lung cancer survival time would not only benefit patients but also give doctors insight into predictive values of patient's characteristics and radiomics data. A successful model also gives doctors an additional objective source to base their patient's survival time estimation on.

4.2 Research question

“Are there statistical differences in the survival rate between the different fractionating schedules of stereotactic irradiation treated lung cancer patients, and can the survival chance and toxicity of the patient be predicted using the patient's medical data?” This research question was divided into 4 distinct end products.

1. Software that can calculate, visualize and plot possible statistical differences in overall patient survival rate between the different fractioning schemas.
2. Report (dis)proving a statistical difference in overall survival between the different fractioning schemas of lung cancer patients.
3. Machine learning model predicting the survival chance of lung cancer patients.
4. Report detailing whether available data is enough to train an accurate machine learning model.

4.3 Patient collected data

Data available for this study includes; CT images, stored in DICOM format, and connected to the patient's electronic dossier; treatment planning and radiation dosage, extracted from previously mentioned DICOM files; cancer TNM-staging, detailing information about the tumour, aggression, location and more; questionnaire whether patient is experiencing side-effects, called toxicity; general medical and biological characteristics, such as age and gender. All data was stored in different formats, datasets, and online location; more dataset details in the chapter “available data sources”.

5 Available data sources

5.1 CASTOR, Patient’s General Treatment and Toxicity Information

Radboud radiotherapy has been using Castor for a few years now. A database which is tightly connected with forms and access control, aimed in this case at researches. Administrators and researchers can create forms (questionnaires) which doctors and patients can fill in, data is immediately stored in an online database. Each database has access control features allowing other researchers to read, write and edit the data as well as restrict certain datapoints.

For my research I was given read access to all pseudonymized lung cancer patients treated with SBRT. Allowing me export of all SBRT patients ever treated at Radboud while not having access to any direct personal identifying information. Somewhat identifiable information was still anonymized as much as possible, e.g. birthdates were converted to numerical age values.

This data source mainly contained data about patient’s toxicity but also treatment information and patient’s characteristics such as gender.

5.2 RTHweb, patient’s basic information

Formerly used database of the radiotherapy department accessible on the intranet by authorized employees. I did not have access to the online database web environment; but was given an export of all its data.

This data source was very minimal and most of its datapoints could also be found in the Castor dataset. However, RTHweb was critical by determining which treatment was most relevant of patient’s who underwent multiple SBRT treatments.

5.3 Basisregistratie personen (BRP), patient’s survival status

Basisregistratie personen (basic registry of persons), formerly known as GBA (gemeentelijke basis administratie). All patient’s survival status was requested at the government which returned a sheet when each patient had died or was still alive.

Castor also has a patient survival datapoint, more importantly also to what causes. However, the BRP data source was critical as 75% of the Castor survival status was unknown.

5.4 DICOM, patient’s radiotherapy radiomics

Text BoxImage result for lung cancer ct scan contouringA CT-scan is made of each patient, which is stored as a DICOM (Digital Imaging and Communications in Medicine) file, on which the tumour is visible as a white-grey mass. Doctors create contourings on each CT slice detailing the tumour and the visible organs

such as the lungs, heart and oesophagus (see fig 2. For an example). Allowing the doctor to create a better treatment plan whereas minimal healthy tissue is irradiated during the treatment.

Using inhouse software, for each patient's CT-scan the irradiation amount of each organ and the tumour received can be calculated. This data source is critical in creating a machine learning model predicting survival chance as the treatment is just as important as the tumour itself for predicting survival.

6 Standardizing values tool

Names of all contours drawn by doctors in the DICOM files are filled in by hand. Each contour drawn is given a descriptive name to identify the contour, however the doctor can type in anything, resulting in contours of the same area but given slightly different naming.

After processing DICOM files through the inhouse software, all information is extracted as numerical and textual based values and inserted in a SQL database. A table for all patients and another table with all extracted contours linked to a treatment plan and a patient. E.g. the contour for tumours is named PTV (planned target volume) but of all 400 patients, there were 360 unique names for the PTV. Difference ranging from upper- and lowercase letters, names including radiation dosage, incorrect spelling, and many more; resulting in the same contour area across different DICOM files being named different.

To tackle this problem, a Java GUI application was developed where you can select a CSV file (exporting the database to csv) and select a column. All the distinct column values are displayed, the user can select which values are equal and should have the same value.

After using the Java tool all contours of the same drawn area now have the same name and can be used in data analysis.

7 Data preparation

Four different datasets, CASTOR, RTHweb, BRP and DICOM files, all were combined into one dataset to make data analysis simpler. Unfortunately, not all datasets could easily be joined by a universal identifier. Castor dataset contains the most usable values of all datasets thus the plan was to join all other datasets to castor. However, the patient number in Castor was automatically not accessible to me due to privacy restrictions, its unique identifier was a generated auto increment number which could not be directly linked to a patient. Another dataset, a link table, was given to me which connected all Castor records to a patient number.

7.1 Multiple treatments

Each treatment record is only connected to a patient number; patients who underwent multiple treatments are only distinguishable by their values and not

an identifier. Same applies to the RTHweb dataset, treatments are linked to a patient but do not have an identifier. Even worse, not every listed treatment in both datasets existed in the other datasets, e.g. a patient might have had a SBRT treatment according to the castor dataset but not according to RTHweb dataset. For each treatment in Castor a treatment of the same patient in RTHweb was attempted to be joined by the dates of treatment. The most recent treatment is selected and all treatments within 14 days fill in missing values and aggregate multiple parameters into one. Patients who underwent multiple treatments are not only difficult to join to other data sources but also raise a statistical analysis worry; multiple treatments mean a higher total dose to the patient, influencing the survival probability. This issue was chosen to be ignored as it complicates the research too much in respect to time available.

7.2 DICOM long to wide format

After using inhouse software on all DICOM files, an SQL database is created with all its relevant values. All contouring area values of these files were exported to CSV. Result dataset was in long format, one record per contour. Other datasets had 1 record per treatment or patient thus the DICOM dataset was converted from long to wide format, turning all rows per CT- scan into columns. Now one record per treatment exist with all dosage parameters.

7.3 Combining data sources

All datasets (CASTOR, RTHweb, BRP and DICOM files) now each contain one record for every patient with the most recent or aggregate treatment information. By inner joining the various datasets one final dataset rolls out containing all patient’s information, toxicity information, treatment information and applied radiation dosages.

8 Lung cancer SBRT survival research

The primary objective of this investigation is to explore potential associations between overall survival and the dosage administered to the planned target volume (PTV), lungs, heart (sub)structures, and the esophagus. This examination focuses on patients with early-stage and locally-advanced non-small-cell lung carcinoma (NSCLC) and small-cell lung carcinoma (SCLC) who underwent various stereotactic body radiotherapy (SBRT) fractioning schemas.

The cohort comprises a combined total of 302 NSCLC and SCLC patients. Those with minimal lung damage and tumors surrounded by normal tissue received SBRT predominantly in schedules of 3 x 18 Gy and 5 x 11 Gy. Patients with either significantly damaged lungs, possibly from previous treatments, or tumors near tissues contraindicating irradiation—primarily the ribs—were subjected to SBRT schemas of 8 x 7.5 Gy or 12 x 5 Gy. All patients were anatomically registered to an average anatomy, with their planned doses adjusted ac-

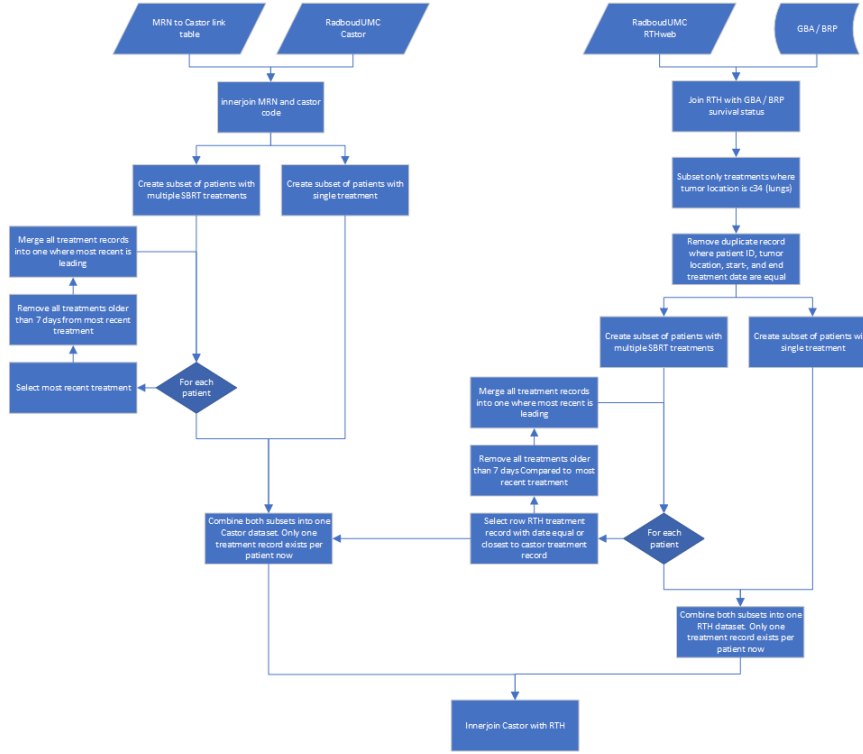


Figure 1: How data is processed and combined into one dataset.

cordingly. Subsequently, dose-volume histogram (DVH) parameters for PTV, lungs, heart (sub)structures, and the esophagus were acquired.

Various analytical approaches, including Cox regression, principal component analysis, and Spearman rank-order, were employed to pinpoint doses to the lungs, heart (sub)structures, and esophagus with negligible associations with cancer-related death.

Stereotactic body radiation therapy (SBRT) emerges as a crucial treatment modality for primary and secondary lung cancer. The advantages of hypofractionated radiotherapy in treating lung tumors include a condensed treatment course that reduces clinic visits compared to conventional programs. This is coupled with the ability to adopt a smaller irradiated volume, facilitated by enhanced setup precision [1]. Notably, patients with inoperable non-small cell lung cancer receiving SBRT demonstrated a substantial 55.8% survival rate at 3 years, a notable improvement over the 20%-35% rate associated with conventional radiotherapy [2]. However, SBRT comes with potential drawbacks, such as uncertain effects of altered fractionation and the theoretical risk of worsening the normal tissue-to-tumor tissue ratio with higher doses per fraction. For instance, doses primarily targeting the upper heart region in NSCLC patients

treated with SBRT were significantly associated with non-cancer-related deaths [3]. This study aims to evaluate the clinical outcomes of overall lung cancer patient survival under different SBRT fractioning schemas, shedding light on the nuanced impact of radiation dosage on patient outcomes.

8.1 Methods and materials

Between 2011 and 2016, Radboud University Medical Centre in Nijmegen, The Netherlands, treated 322 patients primarily diagnosed with stage T1-T2N0M0 non-small-cell lung carcinoma (NSCLC) and small-cell lung carcinoma (SCLC). Following meticulous selection criteria, 302 patients were included in the study (refer to Table 1), with exclusions made for individuals lacking SBRT plans, which couldn't be inferred from other available patient data, or if the last SBRT treatment information was unknown.

The treatment involved four distinct SBRT fractioning schedules, primarily determined by the patient's tumor location. Patients with minimal lung damage and tumors surrounded by normal tissue were administered SBRT with prevailing schedules of 18 Gy in 3 fractions and 11 Gy in 5 fractions. In contrast, patients with significant lung damage, potentially from prior treatments, or tumors in proximity to tissues posing irradiation risks—primarily the ribs—received SBRT with a schema of 7.5 Gy in 8 fractions or 5 Gy in 12 fractions.

However, the cohort with the 5 Gy in 12 fractions schema was deemed too small for statistically significant comparisons to other fractioning schedules. Consequently, all patients with the 12 x 5 Gy schedule were excluded from the study.

Follow-up (FU) encompassed a comprehensive toxicity questionnaire administered by the attending physician at intervals of 1 week, 2 weeks, 3 weeks, 1 month, 6 months, 12 months, and 24 months post the completion of radiotherapy treatment.

The survival status of all patients was diligently obtained from the Dutch Ministry of Internal Affairs Basic Register of Persons (Basisregistratie Personen [BRP]). This meticulous approach to data collection ensures a robust foundation for analyzing the outcomes and potential associations within the studied patient cohort.

8.2 treatment planning and delivery

Patients underwent a 4D-CT treatment planning scan to facilitate the tracking of tumors during radiation administration, a critical consideration given the dynamic nature of tumor movement during breathing.

The contouring process, crucial for delineating treatment areas, adhered to specific techniques and parameters, guided by the expertise of a designated professional, denoted by their employee title, utilizing a specific machine—details which can be customized as per the actual scenario. All contouring and treatment planning occurred on a CT scan performed with a particular machine,

Table 1: Patient, tumor, and treatment characteristics.

Characteristic	All Patients	N % Median (Range)
Gender		
Male	185	61.3
Female	115	38.1
Unknown	2	0.7
Diagnose Age (y)	73	(27-94)
Tumor Diameter (cm)	2	(0.3-6)
Tumor Location		
Upper Lobe	164	54.3
Middle Lobe	18	6.0
Lower Lobe	113	37.4
Unknown	7	2.3
Survival		
Alive	152	50.3
Deceased	150	49.7
Known Cancer Death	49	x
Unknown Death Cause	101	x
Treatment		
3 x 18 Gy	110	36.4
5 x 11 Gy	113	37.4
8 x 7 Gy	64	21.2
12 x 5 Gy	15	5.0
FEV1		
< 70	174	56.6
≥ 70	65	21.5
Unknown	63	20.9
T-stage		
1	144	x
2	47	x
3	18	x
Unknown	93	x
WHO Performance Status		
Asymptomatic	46	x
Symptomatic but completely ambulatory	187	x
Symptomatic	51	x
Bedbound	1	x
Unknown	17	x

incorporating advanced technologies such as contrast enhancement for precise delineation.

For the primary tumor and potentially other metastatic tumors, the GTV-PTV margins were established at x millimeters, with potential additional variables considered, such as lung breathing capacity. The documentation could optionally delve into further specifics, such as the utilization of additional machines for different aspects of the treatment process, including the types of actual radiation machines employed.

Position verification and setup correction procedures were executed using specific software, possibly coupled with the use of a plastic mask or other immobilization devices to enhance precision in treatment delivery. These meticulous steps in the treatment planning and execution process underscore the commitment to ensuring accuracy and efficacy in the administration of radiation therapy.

8.3 Dosimetric parameters

Following the exclusion of patients with missing dose data for either the PTV or one of the specified organs—lungs, heart (sub)structures, and esophagus (x patients)—the cohort was refined to include 288 patients. Subsequently, dose-volume histograms (DVHs) for lung tissue, esophagus, heart (sub)structures, and PTV were generated for all patients. This involved extracting information from archived DICOM files using in-house software.

The analysis focused on relative volume parameters, specifically the percentage of the contour volume receiving specified dosages, ranging from V5 to V75 in 5 Gy increments. This comprehensive approach to dose evaluation provides a nuanced understanding of the radiation exposure across different anatomical structures, contributing valuable insights to the study.

8.4 Analysis

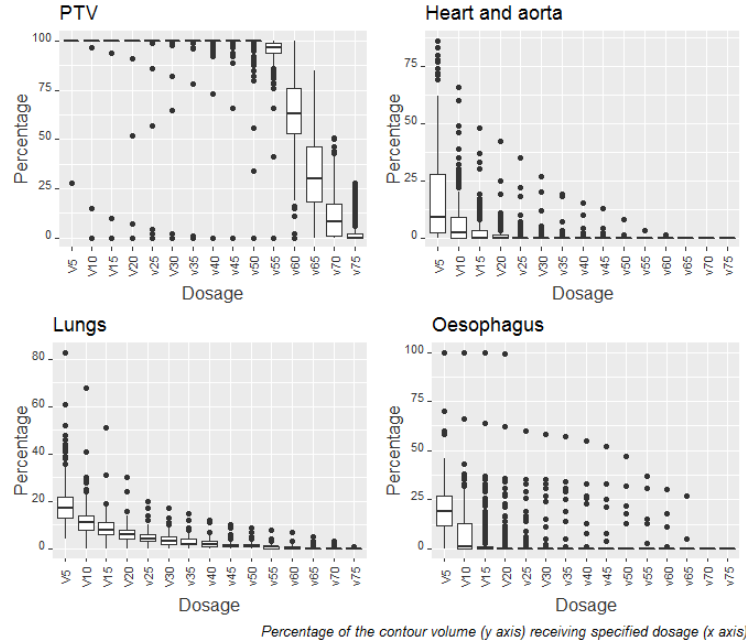
Survival status is known for each patient. Non-deceased patients were right-censored with the last date being when the respective patient’s survival status was obtained from BRP.

Kaplan-Meier curve was used to visualize the survival curve of each SBRT fractioning schema, log-rank to compare the difference between the survival curves [and possible other groupings], and Cox proportional hazard regression to describe the (in)effect of treatment and patient’s factors on survival.

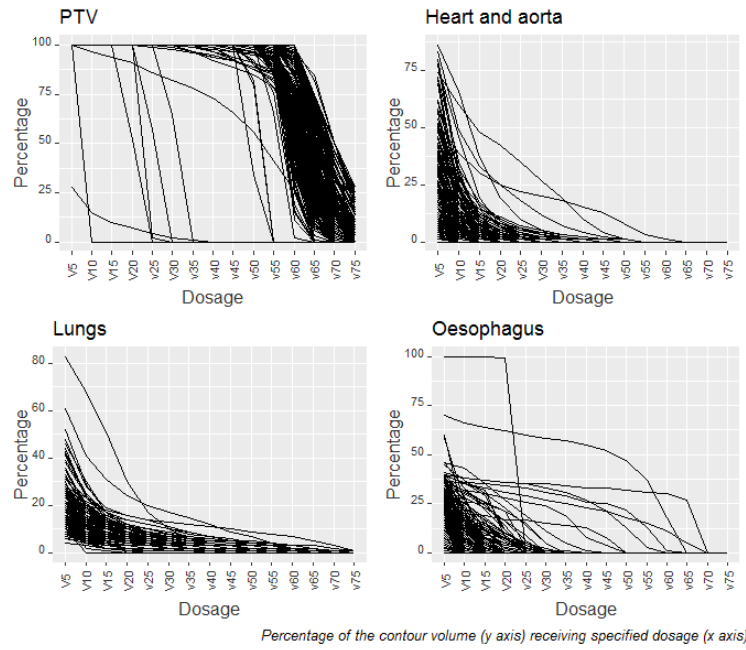
Survival probability $S(t_i)$ at time t_i is calculated using the following equation: $S(t_i) = S(t_{i-1}) \cdot (1 - \frac{d_i}{n_1})$.

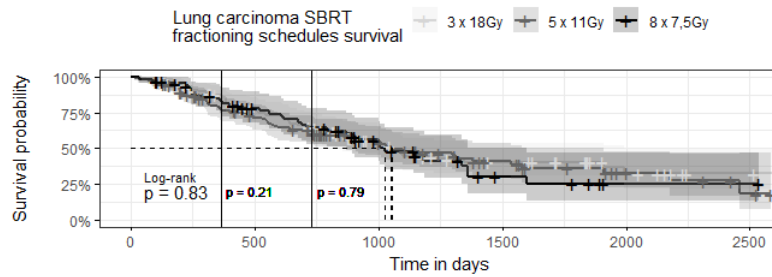
Survival time is measured as the number of days from the first treatment till event occurrence or censoring. Survival probability $S(t)$ is the probability that an individual survives from the time origin (diagnosis) to a specified future time t . Hazard probability, $h(t)$, is the probability that the event occurs to the patient who is under observation at time t .

Dosimetric parameters



Dosimetric parameters





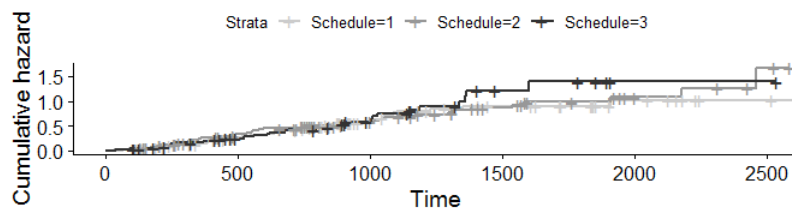
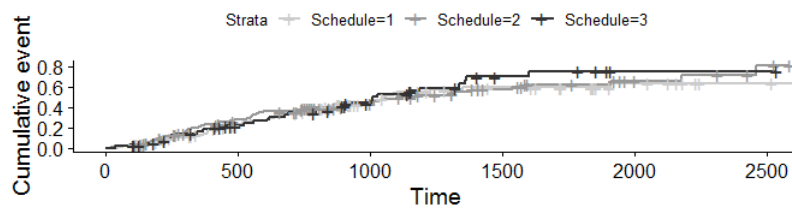
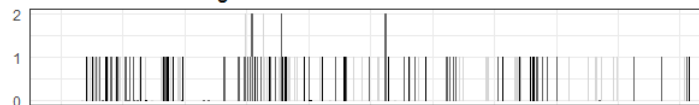
Number at risk

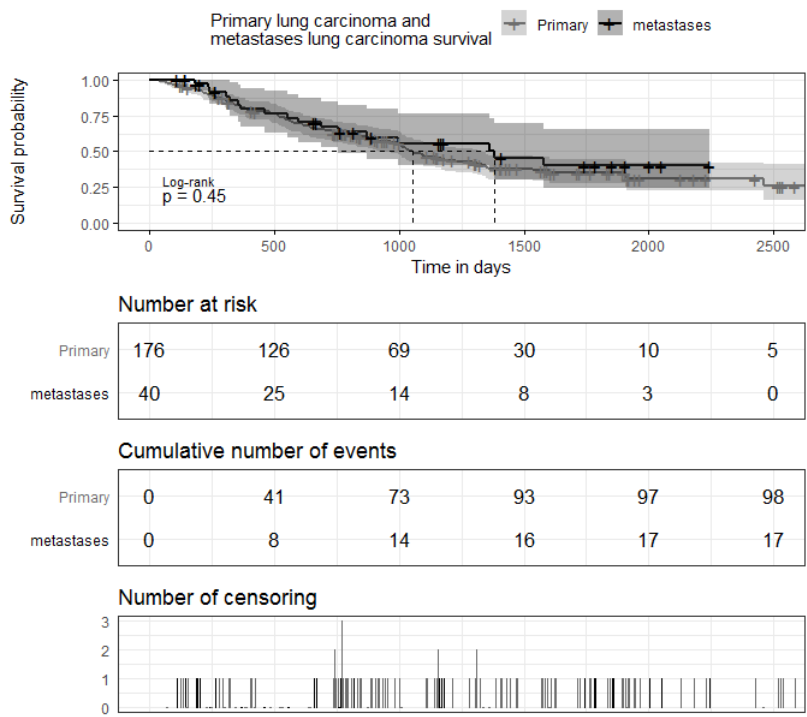
3 x 18Gy	110	67	32	19	8	2
5 x 11Gy	113	69	37	18	7	2
8 x 7.5Gy	64	40	21	6	1	1

Cumulative number of events

3 x 18Gy	0	23	42	49	51	51
5 x 11Gy	0	30	45	53	56	58
8 x 7.5Gy	0	13	24	32	33	33

Number of censoring





To assess the potential significant association between dosimetric parameters and overall survival, a meticulous series of steps was undertaken, with all tests conducted at a statistical significance level of $p < 0.05$. For the planned target volume (PTV), lungs, esophagus, and heart (sub)structures, the three dosimetric parameters exhibiting the highest likelihood were selected for analysis.

Initially, univariate analysis was applied to all dose parameters. The dosimetric parameter demonstrating the most robust and significant association with overall survival, as determined by a univariate Cox regression using the maximum likelihood estimator, was then singled out. Subsequently, the dosimetric parameters with the highest likelihood were subjected to multivariate Cox regression analysis, considering potential confounders such as diagnosis age, FEV1, pulmonary artery (PA) proven status, tumor size, and tumor location.

To uphold quality control and ensure the statistical significance of retained parameters, a second multivariate Cox regression was conducted. This aimed to explore their possible associations not only with all-cause death but also specifically with non-cancer-related death. The factors considered in this subsequent analysis included a large list of remaining parameters. This rigorous approach serves to establish the robustness and reliability of identified dosimetric parameters in predicting overall survival outcomes, while also addressing potential confounding factors and examining associations with different causes of death.

9 Lung Cancer Survival Prediction Model

Many patients' medical and biological characteristics and treatment affect survival chance and time. A machine learning model to predict lung cancer survival time would not only benefit patients but also give doctors insight into predictive values of patients' characteristics and radiomics data. A successful model also gives doctors an additional objective source to base their patients' survival time estimation on.

Identifying whether patient information and treatment radiomics-collected data have predictive values and can be modeled into a survival estimation analysis machine learning model, including finding the best classifier for said data.

All ML models in this study were trained & tested with 100 repetitions with 5 inner-folds and 5 outer-folds.

9.1 Choosing the Best Classifier

A few researchers in the data science field developed an automated process to fit various machine learning models with built-in feature selection and compare model accuracy scores (T.M. Deist, F.J.W.M. Dangers et al., 2018). Six different classifiers (random forest, elastic net logistic regression, neural net, support vector machine, and LogitBoost) were applied to the entire dataset, including follow-up toxicity-related data points, and a subset of data which was only available the moment treatment finished. Ranked by AUC score, random forest (RF) and generalized linear model with elastic net (GLMNET) scored as the

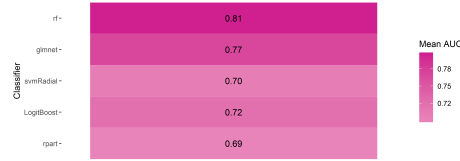


Figure 2: Classifiers AUC ranking score

two best discriminators respectively. Thus, RF and GLM(NET) models were investigated and trained in further detail.

9.1.1 Why Random Forest Was Disregarded

Random forest is an ensemble of decision trees. Trained by randomly sampling subsets of the training data, fitting a model to the subsets and aggregating prediction results. For classification using the majority vote method. An important factor for any machine learning model in the medical world is traceability, how did the model come to a classification or prediction? Random forest is a 'black box'; input variables go in and a result variable comes out; how it achieved its result is not supposed to be known. However, one can peek at some trees to get a glimpse of which variables the model found important to manually determine a crude accuracy. Result of said peek resulted in the unanimous decision not to use random forest; various trees, as second leaf node, had the question *tumour size* > 0; when answered false, the follow-up leaf node question was *tumour diameter* > 0. The first said leaf node should already not exist since nobody is treated with a tumour that has no size. Second leaf node is even more interesting since nothing can have a diameter if it doesn't have a size.

For the reasons mentioned above, traceability and questionable leaf nodes, it was unanimously decided not to use random forest.

9.1.2 Feature Analysis

Cox regression, principal component analysis and Spearman's rank-order correlation were applied to identify data features association with cancer-related death. These analysis methods measure association significance and correlation between features. Results of these analysis determine what variables are used to train the machine learning model.

Cox Regression Univariate cox regression analysis showed significant cancer-related death association with patient's WHO status, gender and dosimetric values PTV mean dose, PTV std dose, PTV d2prct_ingy, PTV v10 PTV v20 PTV v65, lungs maximum dose, esophagus minimum dose, esophagus v25, esophagus v30, and esophagus v35. Multivariate cox regression analysis show that the WHO status symptomatic and esophagus minimum dose features have a

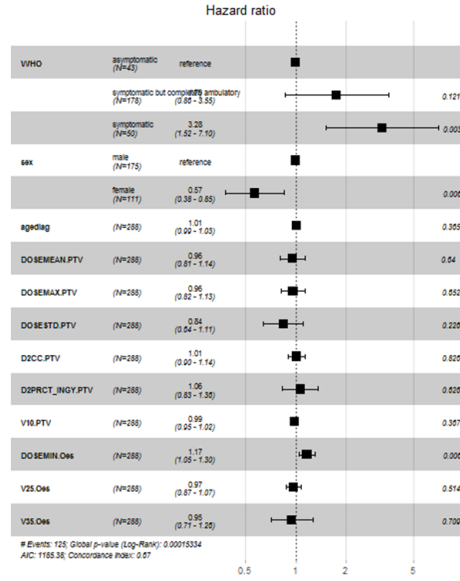


Figure 3: Multivariate cox regression

significant association with cancer-related death and the gender female feature a significant association with cancer-related survival.

Principal Component Analysis Principal component analysis shows the exact same linear distribution for both classifications (deceased and alive) on a 2D plane. After sub-selecting variables with greater than a 1% contribution in principal component (PC) 1 and PC2, the result is nearly identical; only event classification shows a difference, a slimmer distribution while remaining almost entirely under no-event classification distribution.

Spearman Rank-Order Spearman's rank-order correlation shows no correlation between all dosimetric parameters and overall survival. Generalized linear model was fit by selected features (patient's WHO status, gender and dosimetric values PTV mean dose, PTV std dose, PTV d2prct_ingy, PTV v10 PTV v20 PTV v65, lungs maximum dose, esophagus minimum dose, esophagus v25, esophagus v30, esophagus v35 and esophagus v65) from the cox regression analysis, principal component analysis and Spearman's rank-order correlation matrix.

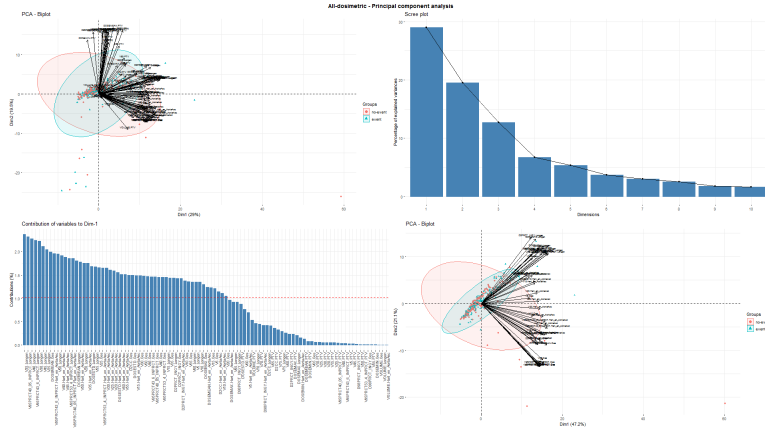


Figure 4: Principal component analysis and variable importance

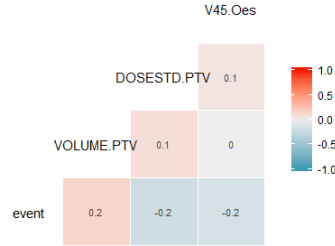


Figure 5: Spearman heatmap of the highest correlated features

Training a Generalized Linear Model Using Elastic Net Regularization

The CRAN package 'GLMNET' uses elastic net regularization (NET) as automatic feature selection to fit a linear regression model via penalized maximum likelihood. The regularization path is computed for the lasso or elastic net penalty at a grid of values for the regularization parameter lambda. Elastic net mixes two penalty algorithms; Ridge penalty, which shrinks the coefficients of correlated predictors towards each other; and Lasso, which tends to pick one of them and discards the others. If predictors are correlated in groups, an $\alpha = 0.5$ tends to select the groups in or out together. (Hastie & Qian, 2014) Generalized linear model was fit by both Elastic net regularization and selected features (patient's WHO status, gender and dosimetric values PTV mean dose, PTV std dose, PTV d2prct_ingy, PTV v10 PTV v20 PTV v65, lungs maximum dose, esophagus minimum dose, esophagus v25, esophagus v30, esophagus v35 and esophagus v65) from the cox regression analysis, principal component analysis and Spearman's rank-order correlation matrix.

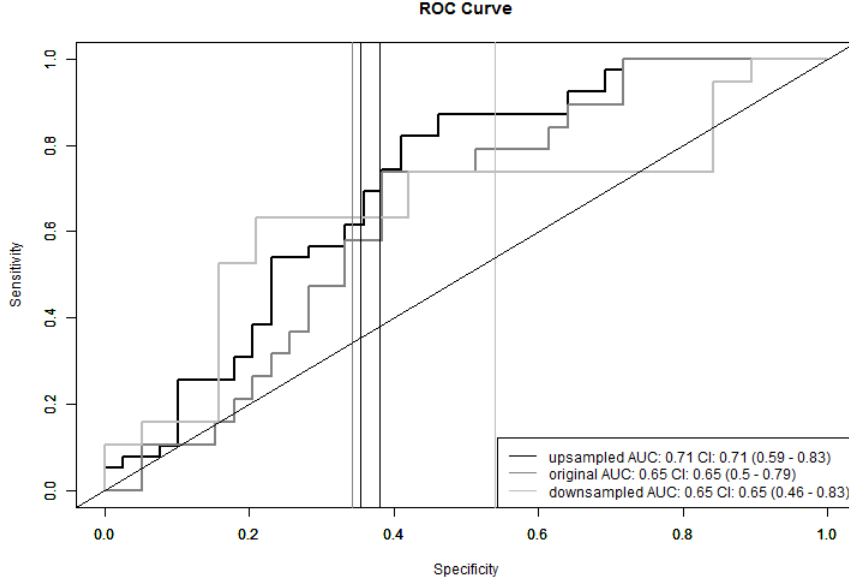


Figure 6: GLMNET model score (all variables)

Model Scores & Conclusion

AUC scores of 0.72 were achieved by both GLM and GLMNET models, showing similar features importance values. GLM with manual feature engineering achieved the same AUC score as GLMNET while using fewer features which are easier to traceback and control. However, manual feature engineering for a GLM is drastically more complex and time-consuming compared to elastic net regularization's automatic feature selection which is also easier in both programming and data analysis.

Concluding that patient characteristic and radiomics data have predictive values for machine learning models; however, not enough to create a trustworthy accurate machine learning model, but also that elastic net regularization's automatic feature selection is superior, compared to manual feature engineering, in terms of time constraints and complexity.

Ethics in Data

By default, personal information of patients such as names and addresses were not accessible to me. Some personal information was required for the research, but all irrelevant details were removed, e.g., date of birth was converted to a numerical age value.

Patient unique identifiers such as a patient number were pseudonymized;

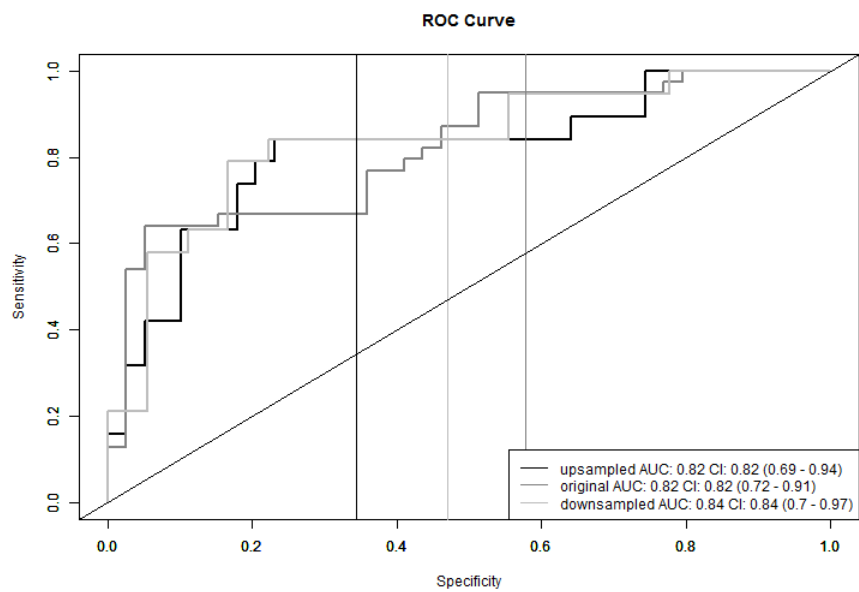


Figure 7: GLM model score (WHO, gender & PTV dose STD)

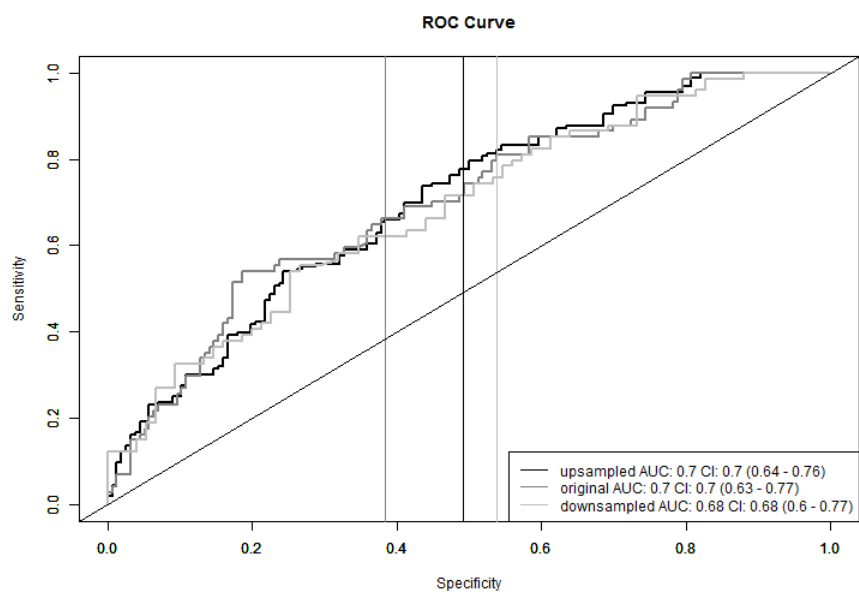


Figure 8: GLM model score (WHO, gender & 16 dosimetric features)

anonymizing or replacing the patient number was not an option so that, if the research required it, a supervisor could verify research findings or claims directly to patient's all information.

Data available for the research never left the work computer at the radio-therapy, including not storing it on cloud services or sending data over email.

Advice

Using CASTOR to record patient data is a considerable improvement compared to the RTHweb system in terms of the quantity of data points and accessibility; however, data quality and consistency are worse, which when relating to data volume is understandable and acceptable, yet more work to a data analyst. More improvements could be made to CASTOR by configuring more data constraints, e.g., only specified acceptable data types can be entered into open fields.

Various in-house tools used to obtain the data and results detailed in this study were abominably slow, so slow that about 300 DICOM files took the better half of a day to process. One slow process isn't the end of the world; however, if software repeatedly allows for lengthy coffee breaks, it negatively impacts the workflow.

A new research should be conducted to discover why there is no difference in survival between SBRT schedules. Why do schedules aimed to maximize survival changes have the same survival rate as schedules aimed to minimize toxicity?

Patient characteristic and radiomics data have predictive values for machine learning models; however, not enough to create a trustworthy accurate machine learning model. More data is required to achieve more accurate results. Radboudumc could work together with other medical universities or hospitals to acquire more data or create a joint study. More data can always change results but more importantly, machine learning models always profit from a higher quantity in terms of accuracy.

A new research should be conducted to discover more underlying predictive values from patient's information, characteristics, and treatment radiomics data. Why do females have a decreased chance of dying from lung cancer compared to men and why does a WHO symptomatic status have an increased chance of dying from lung cancer?

Research Framework

Knowledge required to conduct this study and to verify used research methods were gained and cited from published articles on pubmed.gov and found on Google Scholar similar to this study's research question. Online courses and other sources were used as self-study material to acquire the skills to understand and apply analysis and modeling methods. To learn about existing solutions, presentations and lectures regarding data analysis and machine learning applied

in the healthcare sector were attended at the Big Data Expo 2018 in Utrecht. Webinars regarding machine learning in healthcare were followed together with supervisors to remain up to date about new study results and applied techniques.

To understand how and at what stages the data I'm analyzing is gathered, my clinical physicist supervisor showed and explained the patient process of diagnosis and treatment. Weekly meetings were held with supervisors to share and discuss progress and achieved results. Presentations and expositions were attended to gain more insight into the topic and technical aspects such as PhD defenses and a Big Data Expo with topics around healthcare, data, and machine learning.

Sketches, designs, and functionality priority lists about the developed DICOM values normalizing tool were created and discussed with stakeholders. All developed software had prototypes which were discussed with stakeholders. Feedback from other co-workers was asked to improve design and functionality.

Six different classifiers were applied on the entire dataset with a repetition of 100, 5 inner, and 5 outer folds to discover the best models to investigate in the limited time available. Selected machine learning algorithms' performance results were compared and pitched to supervisors to decide on the best algorithm to build a model around.

All model results and classifier testing were conducted with a repetition of 100, and 5 inner and 5 outer folds. Leaf choices or variable importance were extracted from models to manually determine a crude accuracy. Survival research and model results were evaluated using various calculated probability values to determine statistical significance.

Survival research findings and work methodology were checked and verified by both supervisors and doctors to assure quality and reason results on a medical level of possible accuracy.

Evaluation

The first thought that comes to mind when evaluating my internship at Radboud is not just how enjoyable it was to work there but also the feeling of contributing to improving people's lives. What kept me going the most was pitching my own arguments why and especially how a subject should be analyzed, instead of simply being told what to do; resulting in a feeling of appreciation and eventually accomplishment.

Combining the various data sources turned out to be the most difficult and time-consuming task. I was attempting for too long to graciously combine the datasets. Eventually, after spending too much time on this issue, it became clear there was no proper solution. Acknowledging there isn't always the right way, a perfect way, of achieving one's goal is what allowed me to make decisions, and thus results, faster. Not just in terms of code but also making decisions that affect research results. The time spent pondering on questions, such as whether to include patients in the cohort who underwent multiple treatments; results could have been achieved for both included patients and excluded. Doing

anything is time better spent than waiting for an answer, even though eventually it might be the wrong choice.

While updating stakeholders on the progress of combining the various data sources, one constant important question was how many patients in total are in the combined dataset. Too many times I couldn't explain in detail how many patients were excluded at each process. Prompting me to develop code to log and write results obtained with all details available directly to disk. Adding traceability to the result made life at the end of the project much easier. All results delivered now contain all the original values it's based upon and what processing was applied, code is not required to read nor understand the output.

Communication during the project was excellent. Only once there was a misunderstanding, a simple graphical user interface mix-up, which was corrected the very next day. Weekly meetings were planned in the intranet scheduler with reminders, none were attended late or missed. Regularly my supervisor was unavailable, but any email sent would be answered within an hour by email or walk by my office personally to answer my inquiry. I look up to my supervisor's responsiveness and communication skills, hoping to match them one day.

From the start, it was clear that I did not have all the required knowledge to conduct the research. A planning was made in crude segments which would later be split into agile sprints set by myself; e.g. feature engineering was planned for 4 weeks, when I started this part of the project, I researched the best way to conduct feature engineering for my scenario and split the 4 weeks into sprints covering 3 feature analysis methods. I deviated from my planning multiple times but was always corrected without impeding or endangering the product's quality.

10 Bibliography

1. Nagata, Y., et al. (2005). Clinical outcomes of a phase I/II study of 48 Gy of stereotactic body radiotherapy in 4 fractions for primary lung cancer using a stereotactic body frame. *International Journal of Radiation Oncology*Biophysics*, 63(5), 1427-1431.
2. Timmerman, R. (2010). Stereotactic Body Radiation Therapy for Inoperable Early Stage Lung Cancer. *JAMA*, 303(11), 1070.
3. Stam, B., et al. (2017). Dose to heart substructures is associated with non-cancer death after SBRT in stage I–II NSCLC patients. *Radiotherapy and Oncology*, 123(3), 370-375.
4. Radboudumc - Over het Radboudumc. Retrieved December 11, 2018, from <https://www.radboudumc.nl/over-het-radboudumc>
5. Radboudumc Department - Over de afdeling. Retrieved December 11, 2018, from <https://www.radboudumc.nl/afdelingen/radiotherapie/over-de-afdeling>
6. NHS - Lung cancer - Causes. Retrieved December 11, 2018, from <https://www.nhs.uk/conditions/lung-cancer/causes/>
7. Mayo Clinic - Lung cancer. Retrieved February 11, 2018, from <https://www.mayoclinic.org/diseases-conditions/lung-cancer/diagnosis-treatment/drc-20374627>
8. FAQs: SBRT. Retrieved December 11, 2018, from <https://www.uclahealth.org/radonc/faqs-sbrt>
9. CMD Methods Pack. Retrieved December 12, 2018, from <http://cmdmethods.nl/>
10. Deist, T. M., Dankers, F. J. W. M., et al. (2018). Machine learning algorithms for outcome prediction in (chemo)radiotherapy: an empirical comparison of classifiers. *Med. Phys.* doi:10.1002/mp.12967 [status: epub ahead of print]
11. Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society: Series B*, 34, 187–220.