Big data and interpretable models for outcome prediction in radiation oncology

SUMMARY

Introduction

Machine learning models have seen considerable success, from automating conventional workflows for improving operational efficiency and performance to providing fast, personalized, and reliable recommendations. Examples of real-world applications are Netflix movie recommendations, Tesla self-driving cars, and Amazon speech recognition (Alexa). These applications are possible thanks to the (Big) Data generated and willingly donated by end-users in their daily activities. Patients and caregivers need help translating this level of success to address healthcare challenges and assist them in making personalized decisions for optimal outcomes using the available healthcare and patient (Big) Data.

Therefore, this thesis focuses on healthcare Big Data and the value of interpretable machine learning models for outcome prediction in radiation oncology and highlights the benefits of experts' involvement in the model development process. This thesis is partitioned into a theoretical and practical section, with the theoretical section consisting of two literature review chapters that discuss Big Data in radiation oncology and healthcare in general. In contrast, the practical section contains three parts (Regression, Decision tree, and Bayesian network) representing the fitted interpretable machine learning models, with each part also containing two original research chapters. Chapter 1 briefly introduces the contents of subsequent chapters and provides the blueprint of this thesis.

Part: Big Data

Data has become one of the most valuable commodities in recent years, to the point that it is being likened to crude oil. The healthcare domain is very interested, amongst others, in using (Big) Data to improve cancer care. Thus, chapter 2 gives a gentle introduction to Big Data and its main characteristics in healthcare. Chapter 3 then paints a more detailed picture of Big Data characteristics and its different sources within the confines of radiation oncology. Both chapters 2 and 3 discuss the solutions provided by Big Data for healthcare challenges, with examples where Big Data has improved operational efficiency for clinical excellence. Chapter 3 furthers the discussion on domain applications, barriers, and the future of radiation oncology Big Data.

Part: Regression

Due to the mortality rate, cancer patients are predominantly concerned about how long they have to live, more so for patients whose tumor has metastasized to other parts of the body like the spine. However, accurate prediction of complex endpoints like overall survival is challenging, even for an experienced clinician. Chapter 4 looks at the prediction of progression-free survival and overall survival for cervical cancer patients, while Chapter 5 looks at overall survival within a 1, 3, and 6 months time frame for patients with spinal bone metastases. Both chapters use a Cox proportional hazard regression model and stratify

patients into different survival risk groups to assist caregivers with patient management. Chapter 5 goes a step further and translates the model into a nomogram to provide caregivers with a tool for individualized estimates of survival probabilities for patients with spinal bone metastases.

Part: Decision tree

This part involves a tree-based algorithm, one of the most popular machine learning techniques, because of its non-parametric nature and ability to naturally classify observation into various groups. In addition, they are effortlessly understandable by a less technical audience due to their IF-THEN nature, making them valuable as clinical decision aids. Chapter 6 focuses on discriminating before the start of treatment between elderly cancer patients who will comply with their planned radiotherapy treatment and those likely not to comply using a decision tree (Compliance tree). Chapter 7 extends the application of decision trees to time-to-event data. This chapter developed a decision tree to predict overall survival in women treated in the Netherlands with radiotherapy for squamous cell cervical carcinoma FIGO-stage IIB-IVA (Survival tree) and externally validates the tree on a Korean population. Both chapters found age to be associated with the outcomes. One logical explanation of this result is immunosenescence which makes individuals susceptible to numerous diseases and morbidity, leading to a reduced chance of survival and treatment completion. Chapter 7 compared the survival curves of each risk group created from the decision tree splits of the leaf nodes on the external validation data to ascertain that the model is generalizable.

Part: Bayesian network

Prediction models can assist caregivers with their decision-making by estimating an individual's probability of developing the outcome of interest. However, the biological process which leads to a particular outcome consists of complex relationships interdependent over and within time. This complexity poses a significant challenge for statistical analysis since the likelihood of correlated features is almost inevitable. Also, clinical researchers will have difficulty determining whether or where a variable should be included for model development, making domain experts' contributions indispensable in the model-building process. Predictive models which can probabilistically reason under uncertainty such as Bayesian networks are more suitable to model such information. Chapter 8 tackles some of these problems by eliciting multiple experts' opinions from different countries on the interplay between variables to develop a Bayesian network structure capable of predicting tumor recurrence for rectal cancer patients. This chapter also compares an expert-elicited structure with an algorithmic structure. Chapter 9 builds on the knowledge of Chapter 8 to develop a Bayesian network that predicts two-year survival for lung cancer patients from a symbiotic relationship between experts' opinions and algorithmic knowledge of the data. Both chapters highlight the value of including experts in the model-building process to develop clinically valid Bayesian network structures. Chapter 8 shows the need to include

essential variables in a model based on their availability on a timeline of extraction to have a Bayesian network structure whose reasoning aligns with the clinical process. A bird's-eye view of this part shows that structures developed by multiple experts or received input from multiple experts are clinically more plausible than algorithmic structures or structures developed by a single expert.

Chapter 10 discusses some of the challenges encountered during the model-building process with possible options to tackle these challenges and emphasizes the need to include end-users in the model-building process in healthcare. Finally, this chapter discusses the importance of interpretable machine learning models in healthcare with some future directions for research.