# Big data and interpretable models for outcome prediction in radiation oncology

# RESEARCH IMPACT

## Introduction

Radiation oncology is a fertile field for applying revolutionary Big Data techniques for better cancer care. However, with data sharing across institutions restricted by administrative, political, and legal barriers, radiation oncology Big Data can not function at its full potential. The clinical data science (CDS) research group at Maastricht University, from which this thesis stems, stands on a three-legged platform to front these data barriers problems and provide the necessary assistance to caregivers and patients. The first pillar focuses on developing global FAIR data-sharing infrastructures, the second uses state-of-the-art machine learning techniques to build prediction models from these (FAIR) data, and the third applies these prediction models for better cancer care.

## Knowledge dissemination

The three CDS pillars mentioned earlier represent the group's main research areas and reflect this thesis. The theoretical section relates to the antecedent of the first pillar, which is the benefits of a functional FAIR data-sharing infrastructure since access to FAIR data stations will lead to more mature radiation oncology (Big) Data. The analysis section fits the second pillar's goal and develops different interpretable machine learning models for outcome prediction in radiation oncology. The discussion and theoretical sections do not entirely align with the third pillar but discuss issues that hamper the application of these developed models for better cancer care, with one of the root causes being the lack of model interpretability by end-users.

When the logical flow of the patterns of the clinical events leading to an outcome or endpoint of interest is captured within a prediction model such as a Bayesian network, it contributes to the clinical understanding and interpretability of the model's output. In chapter 8, the Bayesian network structure was developed to capture the sequential events on how the variables are extracted in the treatment process on a timeline, making the model's reasoning easier to understand. More so, it prevents the structure from having unrealistic dependencies such that a future variable influences a precedent variable at a particular time point, irregularities that even non-experts can identify when the process is explained.

## Cultural focus

Generally, most researchers believe there is a trade-off association between the accuracy and interpretability of a machine learning model. However, chapter 8 does not support this theory and proves this belief is nothing but a "myth" since the predictive performance of the expert Bayesian network is almost as good as the complicated algorithm-based model. The problem is that most researchers allocate the same amount of time to build interpretable and complex models, which will disfavor interpretable models' performance. Interpretable models do require a significant amount of time and effort to construct, with a lot of domain expert(s) involvement.

This thesis tackles some challenges involved in introducing these developed models to improve healthcare. It will seem end-users are more open and accepting of models that automate time-consuming tasks like auto contouring in the clinic rather than a model that supports decision-making, such as predicting the best treatment option for a patient. For automated models, end-users can visualize the model's output and decide whether to use or discard the model's solution based on how much they agree. This is tricky for prediction models because end-users can not do much with the models' output except when the model is interpretable, as then they can base their judgment on the model's reasoning. Therefore, it all boils down to end-users trust in the model. Chapter 10 provides some suggestions to help increase the end-users' trust in these models, one of which is including the end-users in the building process because if they assist in building the model. They understand how the model works then, there is a high probability of them making use of the model.

## Clinical focus

Radiation oncology is a data-rich domain and the perfect field to leverage machine learning techniques to unlock hidden potential knowledge that could assist healthcare professionals and benefit the oncology community. Therefore, this thesis focuses mainly on developing models that can serve as decision aids to caregivers for better patient management. The nomogram in chapter 5 and the decision trees of chapter 6 and 7 can be readily printed to serve as a decision support tool after they have been adequately validated.

Cancer is among the top three leading causes of death worldwide, with metastatic tumors responsible for approximately 90% of all these deaths, making the management of this type of cancer a major clinical challenge. The electronic version of the developed nomogram in chapter 5 provides personalized survival plots, which could come in handy during shared decision-making sessions.

Regression methods are the most common machine learning approach well-versed with physicians and their choice model, especially in decision-making. Unlike regression methods, Bayesian networks are better at dealing with uncertainty related to incomplete domain coverage because the variables used as inputs for the model and their relationships are direct representations of real-world features and their interplay, which is different from other models based on purely mathematical constructs like regression models. Bayesian networks' ability to probabilistically reason about any variable in the structure makes them a valuable tool in any field, more so in medical sciences since explainable visual reasoning increases interpretability. This thesis introduces caregivers to Bayesian networks to increase their knowledge and flexibility in using other machine learning models.

## Economical focus

This thesis focuses on developing prediction models which end-users can easily understand and apply to improve healthcare. In addition, chapter 4, 5 and 7 have created risk strata to group patients into subgroups of risk based on their clinical and lifestyle characteristics to assist caregivers with their patient management decisions. These risk stratification can help caregivers reduce costs and improve patient care since they will focus greater attention and resources on the high-risk group patients.

Chapter 6 developed a decision tree capable of discriminating before the start of treatment between patients who will complete their planned radiotherapy treatment and those likely to discontinue (compliance tree). This tree has a two-fold benefit for patients likely to discontinue treatment: the treatment cost and unnecessary treatment-related toxicity. For caregivers, the model saves them valuable time and helps them make better treatment decisions for patients. This chapter also provides some suggestions to help boost the compliance rate for radiotherapy treatment.

The Bayesian network in chapter 8 models the clinical process which leads to tumor recurrence, a predominant concern for most cancer survivors. The structure predicts if a rectal cancer patient will develop a tumor recurrence (True or False) within a specific time after treatment. Patients with a low probability of developing a tumor recurrence could be discharged with limited need for regular check-ups or worries from the patients if their tumor will resurface. However, those predicted to develop a tumor recurrence could be triaged to special hospital services, intensive outpatient case management, and early clinical visits post-discharge. Such applications, therefore, allow for early interventions to reduce readmissions for recurrence patients and maximize cost-effectiveness, especially for patients who are unlikely to develop a tumor recurrence.

## Technological focus

This thesis did not develop a fully functional technology per se, but two chapters contain some promising technology still under development. The first product is the nomogram in chapter 5 which is transformed into a shiny application[1]. The app provides personalized predicted survival curves for individuals. It also predicts a patient's survival probability and confidence interval around its predictions at any given time point.

The second product is a transformed version of the expert-elicited Bayesian network structure[2] of chapter 8 for the prediction of local tumor recurrence in rectal cancer patients into an interactive user interface model. The probabilistic dependencies between the variables in the structure align with the clinical process, which means one can use the

[1] https://bich.shinyapps.io/SpinalMets/

[2] https://thomas.zakbroek.com/app/network/rectalcancer

structure to reason forward, from causes to consequences, or backward, and deduce the probabilities of different causes given the consequences.

Although these applications are still under development, the first round of external validation of hopefully many to assess their clinical applicability and generalizability is underway. Their shareable link will ease their integration into the current radiotherapy workflow to assist caregivers in making better decisions for better cancer care when fully developed.