Distributed learning and prediction modelling in radiation oncology

Appendix I

Summary

This thesis discusses machine learning methods to analyze patient data in radiation oncology. Due to the increasing availability of computing power, digitalization of medical records, and the success of machine learning in other fields, increasingly sophisticated data analysis procedures are sought after also in the medical sciences. In radiation oncology, such machine learning methods might improve the prediction of radiotherapy outcomes for individual patients, which could aid physicians and patients in selecting suitable treatments. Current research efforts are dedicated to attaining a thorough understanding of the existing machine learning methods and how they can be used for treatment outcome prediction.

Furthermore, developing reliable machine learning applications requires access to large amounts of patient data. Patient data is stored by each healthcare provider and collaboration between these healthcare providers is needed to reach sufficient patient data volumes. Regulatory barriers to protect patient-privacy complicate sharing patient data across healthcare providers and therefore hamper the implementation of machine learning applications in clinical practice. Technological solutions are needed to allow machine learning across healthcare providers while meeting privacy regulations. The existing concept of *distributed learning* might pose a solution, i.e. training machine learning algorithms on patient data stored at distinct healthcare providers without patient data being exchanged.

The studies presented in this thesis form two parts:

- the development of a distributed learning infrastructure to facilitate privacy-preserving machine learning studies across healthcare providers
- the analysis of existing machine learning methods in the context of radiotherapy outcome prediction and the development of a new machine learning method.

With a small proof-of-concept study in chapter 2, the distributed learning infrastructure was described and it was shown that distributed learning across radiotherapy institutes is possible. Support vector machine models to predict post-radiotherapy dyspnea (grade 2 or higher) were trained on lung cancer patient data from five radiotherapy clinics located in three countries (Belgium, Germany, The Netherlands).

To demonstrate the infrastructure's scalability, another study in chapter 3 applied this infrastructure in eight healthcare providers across five countries (China, England, Italy, The Netherlands, Wales) to train and validate a logistic regression for predicting post-treatment two-year survival based on tumor staging data of more than 20 000 non-small cell lung cancer patients. This study was executed in four months demonstrating the distributed learning infrastructure's potential for fast-paced machine learning studies.

In a study of existing machine learning methods for (chemo)radiotherapy outcome prediction (chapter 4), the discriminative performance of six machine learning algorithms (decision tree, random forest, neural network, support vector machine, elastic net logistic regression, LogitBoost) was compared and ranked on twelve patient data sets. Random forest and elastic net logistic regression showed a small increase in discriminative performance. These findings might guide researchers in selecting appropriate machine learning methods for future studies.

Finally, a new kernelized machine learning approach was presented in chapter 5, which allows combining simulation models and machine learning methods. Both simulations and machine learning methods allow inferring predictions: simulation models use prior knowledge gained by (experimental) analysis of a system while machine learning methods derive predictions from data with statistical means. The presented results from four synthetic scenarios indicated that merging simulation models and machine learning methods might pose an advantage in scenarios where insufficient data is available to train standard machine learning algorithms. Chapter 6 discussed challenges and future prospects for distributed learning infrastructures and the use of machine learning methods in radiation oncology. Challenges for distributed learning addressed in this chapter are infrastructure sustainability and its acceptance by users. Furthermore, the risks of misusing machine learning methods and overstating results, and how reporting standards and pre-publication registration of studies can mitigate negative consequences are considered.

I