

Clinical data science in Radiotherapy

Citation for published version (APA):

van Soest, J. P. A. (2018). *Clinical data science in Radiotherapy: data extraction and quantitative analysis*. Datawyse / Universitaire Pers Maastricht. <https://doi.org/10.26481/dis.20181128js>

Document status and date:

Published: 01/01/2018

DOI:

[10.26481/dis.20181128js](https://doi.org/10.26481/dis.20181128js)

Document Version:

Publisher's PDF, also known as Version of record

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.umlib.nl/taverne-license

Take down policy

If you believe that this document breaches copyright please contact us at:

repository@maastrichtuniversity.nl

providing details and we will investigate your claim.

Summary

The amount of data recorded in clinical practice is growing. To facilitate the primary (clinical practice) and secondary use (e.g. scientific research) of this data, we need technical infrastructures which are flexible in terms of data types (e.g. images and structured databases) and scalable in terms of the amount of data (**Chapter 2**). This thesis mainly focused on the development of such an infrastructure, and the application of several components in this infrastructure for the analysis of treatment effects on outcomes in rectal cancer patients.

Like other fields of engineering, an IT infrastructure needs an architecture where the envisioned infrastructure is described. The high-level concept of such an infrastructure is presented in **Chapter 3**, where we describe the key components needed. Next to these technical components, a common method to collect data (when to record which variables? How to code these variables? By whom and how recorded?) needs to be defined and is presented in an “umbrella protocol” (**Chapter 4**). Furthermore, a detailed description of how this “umbrella protocol” should be used in combination with the proposed infrastructure is given **Chapter 5**. This includes semi-automatic data collection/extraction from clinical data sources (such as an electronic medical record), and the possible application of (distributed) machine learning algorithms.

Although the infrastructure itself is important, specific focus has been targeted towards data representation and interoperability. In this thesis, several chapters have been dedicated to development and (re)use of available data and terminologies in radiotherapy; for example, by linking multiple (institutional) data sources. We have shown that private (patient-specific) data using standardized terminologies can be linked to public knowledge: by using information about the patient’s specific treatment protocol, we were able to find scientific articles explaining the results in patients with similar treatment protocols. This specific example was defined in one question (query), by connecting multiple data sources and therefore enriching private data with public knowledge (**Chapter 6**). Next to public knowledge, different institutional data sources (e.g. medical images and image derived information) can be linked as well. Linking these additional data sources enables more advanced questions for exploratory and investigational analysis (**Chapter 7**).

Although the need for standardized terminologies is commonly recognized, in practice this only happens at the institutional level. Resulting into issues during automated (statistical) analysis of data across different institutes. To overcome this issue, we developed an application to add standardized terminological concepts to local naming conventions by storing them in an institutional repository (**Chapter 8**).

Next to handling data, several chapters use components of this infrastructure for clinical data analysis purposes. This thesis describes the validation of three models predicting local recurrence, distant metastases and overall survival in locally-advanced rectal cancer (**Chapter 9**). These models have been trained on European trials (and thus European

patients), where external validation was performed on the clinical dataset of a Chinese hospital. Hence this validation showed whether the previously developed models were applicable to a geographically different area, including a different kind of data collection (routine clinical data, instead of manually curated trial data used during training). The validation above hypothesized a difference between training and validation datasets due to geographical and population differences, although these differences are hard to identify.

To address this identification issue, and to include more background into validation results, **Chapter 10** applies a novel method for detecting (subtle) differences between cohorts. This chapter explains how to measure and interpret the results of an external validation and applied this method to previously developed models predicting for pathologic complete response in rectal cancer patients. Pathologic complete response means that there are no tumor cells found in the pathologic specimen after pre-operative radiotherapy. Prediction models could serve as an indication whether surgery is needed, or a more intensive follow-up would be sufficient and would eliminate risks of surgical complications. However, only one previously developed model showed sufficient performance during validation due to small sample sizes and related variability in cohort characteristics.

Based on the validation results in Chapter 9, we conducted a new pooled analysis of rectal cancer trial datasets, including longer follow-up time and more trials (**Chapter 11**). This chapter shows many aspects related to clinical data science, including semantic interoperability, implicit knowledge in datasets and the question what improves the statistical performance of a prediction model (more data, more variables, or better analysis techniques?). This study showed no significant improvement in model performance (e.g. better prediction of local or distant recurrence after many years) after application of strict inclusion criteria. Hence, we hypothesize that only more advanced machine learning techniques or including additional variables could improve model performance.

Although prediction models are mainly targeted towards prediction of clinical outcomes (e.g. survival, recurrence or development of toxicities), they can be applied for other purposes. For example, **Chapter 12** reports on the evaluation of an automatic contouring application of medical images, where a certain type of machine learning model is used to contour specific organs on CT scans. This study showed a reduction in time for medical personnel to contour these images. In the discussion (**Chapter 14**), we hypothesize whether prediction models not directly interfering with the decision-making process are more easily accepted in clinical practice.

The proposed architecture and infrastructure as presented in chapter 3, is reused in **Chapter 13**. This chapter describes and presents a method and application to safely combine sensitive (patient) data when strictly needed. This chapter serves as a future

perspective on the proposed infrastructure, suggesting use for specific (research) questions or needs (e.g. temporarily combining data from different institutes regarding the same subject), while safeguarding participants' privacy.

However, to combine these different data sources, we need to make sure terminologies among the different institutes do align. Based on the chapters before, we hypothesize in the general discussion (**Chapter 14**) whether the medical domain can adhere to one standard. As this level of standardization is currently not present, we suggest how to use FAIR (findable, accessible, interoperable, reusable) principles to address this issue when developing, evaluating or implementing prediction models in for clinical practice.

Samenvatting

In de klinische praktijk wordt een grote en groeiende hoeveelheid data geregistreerd en opgeslagen. Voor primair (klinische zorg) en secundair gebruik (bijv. onderzoek) van deze data, zijn technische infrastructuren nodig die flexibel omgaan met de variëteit aan data (bijv. afbeeldingen en gestructureerde databases), en mee kan groeien met de hoeveelheid data (**Hoofdstuk 2**). Deze thesis richt zich dan ook op de ontwikkeling van een dergelijke infrastructuur, inclusief de ontwikkeling en toepassing van verschillende componenten binnen deze infrastructuur. Verschillende componenten worden dan ook gebruikt bij de analyse van behandelresultaten in patiënten met endeldarmkanker.

Zoals in vele technische beroepen heeft een IT-infrastructuur ook een architectuur nodig waar de beoogde infrastructuur wordt beschreven. Deze architectuur is dan ook beschreven in **Hoofdstuk 3**, waarbij een aantal componenten worden uitgelicht. Naast deze technische componenten is er ook een algemene methodiek nodig voor het registreren van gegevens (wanneer worden specifieke gegevens opgeslagen, hoe worden deze gecodeerd, en door wie/hoe?). Dit is beschreven in een “paraplu protocol” (**Hoofdstuk 4**). Een gedetailleerde beschrijving hoe dit “paraplu protocol” de infrastructuur faciliteert, wordt gegeven in **Hoofdstuk 5**. Hierbij wordt ook aangegeven hoe specifieke informatie uit klinische databronnen (zoals onder andere een elektronisch patiëntendossier) geëxtraheerd kan worden, waarna het gebruikt kan worden voor (gedistribueerde) data-analyse.

Naast deze technische infrastructuur speelt de data zelf een grote rol: specifiek hoe deze wordt opgeslagen, en de bijbehorende definities van deze data. Een aantal hoofdstukken in deze thesis zijn dan ook gewijd aan het ontwikkelen en hergebruik van beschikbare terminologieën binnen radiotherapie, en het koppelen van verschillende intramurale (binnen bijvoorbeeld een ziekenhuis) en transmurale databronnen. In deze thesis wordt onder andere de mogelijkheid getoond om persoonlijke data te koppelen aan publieke kennis, met gebruik van gestandaardiseerde terminologieën. **Hoofdstuk 6** is hiervan een voorbeeld, waarbij de mogelijkheid wordt getoond om op basis van behandelprotocollen de bijbehorende wetenschappelijke artikelen te zoeken, voor individuele patiënten. Dit specifieke voorbeeld toont aan dat het mogelijk is om in één vraagstelling (query) meerdere databronnen te raadplegen, en hierdoor de persoonlijke data te verrijken met publiek beschikbare kennis (**Hoofdstuk 6**). Naast publieke kennis kan deze techniek worden gebruikt om verschillende intramurale databronnen te koppelen. Het koppelen van deze databronnen betekent dan ook dat meer verschillende data beschikbaar komt, en dus meer geavanceerde vragen en (exploratieve) analyses beantwoord kunnen worden (**Hoofdstuk 7**).

Hoewel iedereen het eens is dat standaard terminologieën nodig zijn, wordt dit nog niet tussen verschillende instellingen goed vormgegeven. Naast de primaire gezondheidszorg levert dit ook problemen tijdens geautomatiseerde (statistische) analyse van data tussen verschillende centra. Deze thesis beschrijft dan ook de ontwikkeling van een

applicatie die een deel van dit probleem aanpakt. Deze applicatie kan lokale terminologieën omzetten in standaard terminologieën, en hiervoor een database per instelling creëren (**Hoofdstuk 8**).

Naast de extractie en omgang met data beschrijft deze thesis ook het gebruik van een aantal infrastructuur componenten voor klinische data-analyse. Deze thesis beschrijft dan ook de validatie van drie modellen die lokale recidieven, metastases en overleving voorspellen voor patiënten met endeldarmkanker (**Hoofdstuk 9**). Deze modellen zijn geleerd op Europese klinische studies (en dus overwegend Europese patiënten), waarbij een externe validatie is uitgevoerd op een dataset uit een Chinees ziekenhuis. Deze validatie toonde dan ook aan of modellen toepasbaar zouden zijn in een ander geografisch gebied, en een andere manier van data verzamelen (data uit de klinische routine, in plaats van handmatig gecontroleerde studie data). De hypothese was dan ook dat er een verschil zou zijn tussen de Europese studie data waarop was geleerd, en de Chinese klinische routine data waarop was gevalideerd; al zijn deze verschillen lastig te identificeren. Dit wordt dan ook in **Hoofdstuk 10** aangepakt, waarbij een nieuwe methodiek wordt toegepast om (subtiele) verschillen in cohorten van patiënten te detecteren. In dit hoofdstuk wordt beschreven hoe deze verschillen gekwantificeerd worden, en hoe deze resultaten geïnterpreteerd kunnen worden. Deze methodiek wordt dan ook toegepast op het valideren van voorheen ontwikkelde modellen die pathologisch complete remissie voorspellen bij patiënten met endeldarmkanker. Pathologisch complete remissie betekent dat er, na radiotherapie, geen tumorcellen worden gevonden in het chirurgisch verwijderde endeldarm weefsel. Deze voorspellende modellen kunnen een indicatie geven of chirurgie in bepaalde groepen patiënten nodig is, of een meer intensieve controle na behandeling volstaat. Onze analyse toonde aan dat één voorspellend model een redelijke accuraatheid had. De afname in accuraatheid van de andere twee modellen werd veroorzaakt door een klein aantal patiënten geïncludeerd in deze studie, en de variabiliteit in cohort (en patiënt) karakteristieken.

Gebaseerd op de resultaten in Hoofdstuk 9 is er een nieuwe gecombineerde analyse van verschillende endeldarmkankerstudies uitgevoerd in **Hoofdstuk 11**. Deze bevat meer trials (en dus meer patiënten), en een langere opvolging van patiënten voor de studies geïncludeerd in Hoofdstuk 9. Dit hoofdstuk laat dan ook meerdere aspecten van klinische data-analyse zien; onder andere semantische interoperabiliteit, impliciete kennis over data(sets) en de vraag hoe een voorspellend model verbeterd kan worden (meer data, meer variabelen of betere analysetechnieken)? De resultaten laten zien dat er geen significante verbetering was in de accuraatheid van het model (bijvoorbeeld een betere voorspelling van lokale recidieven of metastasen 5 jaar na behandeling) na het toepassen van specifieke inclusiecriteria. Hieruit suggereren we dat verbetering in accuraatheid alleen kan worden behaald door het toepassen van meer geavanceerde “machine learning” technieken, of door het includeren van additionele patiënt informatie (variabelen).

Naast het gebruik van voorspellende modellen voor klinische uitkomsten (zoals overleving, recidieven of ontwikkeling van toxiciteiten/bijwerkingen) kunnen voorspellende modellen ook voor andere doeleinden ingezet worden. **Hoofdstuk 12** is hiervan een voorbeeld, waarbij intekeningen van organen op medische beelden door een computer automatisch worden uitgevoerd. Deze studie toonde aan dat er een tijdsreductie mogelijk is door het inzetten van deze automatische intekening modellen en software. In de discussie (**Hoofdstuk 14**) komt dit dan ook terug waarbij een hypothese wordt gevormd of voorspellende modellen die niet direct het beslissingsproces beïnvloeden makkelijker worden geaccepteerd in de klinische praktijk.

De architectuur zoals gepresenteerd in hoofdstuk 3, wordt dan ook verder gebruikt in **Hoofdstuk 13**. Dit hoofdstuk beschrijft een methode en de toepassing voor het veilig combineren van gevoelige (patiënt)data, in situaties waar het strikt noodzakelijk is. Dit hoofdstuk dient als een toekomstblik op toepassingen die op de voorgestelde infrastructuur uitgevoerd kunnen worden. Een voorbeeld hiervan is het tijdelijk combineren van data vanuit verschillende instellingen over dezelfde persoon, waarbij de privacy van deelnemers wordt gewaarborgd.

Gebaseerd op de hoofdstukken voorafgaand wordt er in **Hoofdstuk 14** (algemene discussie) de vraag gesteld of het medisch domein zich kan conformeren aan één terminologie standaard. In een infrastructuur waar we niet de directe patiënt data uitwisselen, maar juist de applicaties die patiënt data verwerken en alleen resultaten terugsturen, is dit van groot belang. In het geval het niet mogelijk is om tot één standaard te komen, geeft dit hoofdstuk een suggestie hoe FAIR (vertaald als Vindbaar, Bereikbaar, Interoperabel en herbruikbaar) principes toegepast kunnen worden om dit probleem te omzeilen. Specifiek tijdens de ontwikkeling, evaluatie en implementatie van voorspellende modellen.