Learning from routinely produced clinical data and Big Data technology in Radiation Oncology

Summary

To improve our health care systems, we need to learn from all the data available to us. Collecting, combining and learning from large amounts of data is often referred as Big Data. Dan Ariely stated: *"Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it..."*. This is represented by the many definitions available for Big Data. To provide context on the subject **Chapter 2** explains the meaning and current state of Big Data in radiation oncology. To utilize 'Big Data' in healthcare the data has to be FAIR (Findable, Accessible, Interoperable and Reusable). But the FAIR-data principles are guidelines, they do not enforce the use of certain technology. Chapter 3 and 4 demonstrate what (partially) applying these guidelines will achieve.

Chapter 3 explains the technology used in the Semantic DICOM (SeDI) project. One of the biggest sources of data in radiation oncology are the digital representations of anatomy and treatment intent, saved as DICOM images. These images consist of two main components, the information about the image (metadata) and the actual image. Traditionally the metadata is partly stored in a separate database next to the DICOM-files. When using a traditional PACS system, trying to gather all CT-scans for lung cancer patients with a tumor of a certain volume, is almost impossible. However, using the Semantic Web technology explained in Chapter 3, we were able to answer this question.

Chapter 4 builds upon the technology presented in Chapter 3, further expanding the use of FAIR data guidelines to improve data utilization. One of the remaining challenges that wasn't solved by SeDI is the free-text delineation names in the DICOM metadata fields. These names are added by the radiation oncology staff when delineating the organs and tumor(s) of a patient. Inconsistencies in these names, due to language, different guidelines, typos, etc., can be a huge problem in large scale image feature extraction. Chapter 4 provides a scalable solution to overcome these challenges using Semantic Web technology to create linked data out of the free-text source data.

Chapter 5 describes a clinical study which used an early version of the software described in Chapter 4. The goal of the study was to compare set-up and 2D EPID dosimetry data of breast cancer patients treated during voluntary moderately Deep Inspiration Breath Hold and free breathing. Identifying the correct DICOM images and processing them automatically resulted in the insight that both image sources indicate that reproducibility of radiotherapy for patients treated during free breathing and voluntary moderately Deep Inspiration Breath Hold is similar.

Chapter 6 describes the validation of a survival prediction model of larynx cancer patients using a rapid learning platform. For this study, a datamining infrastructure was set up to collect the data required for the prediction model to create a comparison to the original clinical dataset. A third dataset was included from a clinical trial performed by the RTOG. Analyzing the data available through routine clinical practice and collected a prospective clinical trial provided insights into the performance of the survival prediction model. The study showed the model performed as expected for the clinical validation data but performed poorly on the clinical trial dataset.

Chapter 7 evaluates the usefulness of automatic contouring of organs at risk (OARs) for lung cancer patients. To perform this evaluation two new auto-contouring methods were compared to the current clinical standard of contouring for lung cancer patients. The time required to perform the delineations following the standard procedures, adjusting an atlas based contour and adjusting a contour generated by a Deep Learning algorithm were recorded. The study showed that for these patients, pre-generating the contours and correcting the inconsistencies saved time over the current clinical practice while adhering to the local guidelines.

Finally, **Chapter 8** summarizes the lessons learned. A FAIR-data platform design is explained to provide future direction for data research.

Samenvatting

Om innovatie in de zorg te verbeteren is het belangrijk dat we leren om alle beschikbare gegevens nuttig in te zetten. Het verzamelen, combineren en verwerken van gegevens op grote schaal wordt vaker onder de noemer "Big Data" geschaard. Een bekend citaat van Dan Ariely is: "*Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it...*". Het komische citaat geeft aandacht aan het probleem dat er vele definities van Big Data te vinden zijn. In **hoofdstuk 2** wordt deze definitie en de huidige status van Big Data verder uitgewerkt voor radiotherapeutische oncologie. Om de zorg daadwerkelijk in het Big Data domein te krijgen is het belangrijk dat gegevens FAIR (Findable, Accessible, Interoperable, Reusable) opgeslagen worden. De Nederlandse vertaling van de term FAIR is dat de gegevens vindbaar, toegankelijk, interoperabel en herbruikbaar moeten zijn. De FAIR-richtlijnen stellen geen eisen aan de technologie die ingezet wordt om het doel te bereiken. Hoofstuk 3 en 4 beschrijven software implementaties die (deels) conformeren aan de FAIR-richtlijnen, en wat het voordeel hiervan is.

In **hoofdstuk 3** wordt de software die geschreven is voor het Semantic DICOM (SeDI) project beschreven. Een van de grootste bronnen aan gegevens binnen de radiotherapeutische oncologie zijn de vele medische beelden die gemaakt worden als onderdeel van het diagnose- en behandelproces, deze worden opgeslagen als DICOM-beelden. Dit type beeldrepresentatie is opgedeeld in twee onderdelen: het daadwerkelijke beeld (de pixels) en de gegevens over het beeld (meta-data). De huidige klinische praktijk is dat de beelden in een PACS worden opgeslagen, waarbij de meta-data voor een deel opgeslagen wordt in een relationele database. Om praktische redenen zijn niet alle meta-data beschikbaar via de relationele database en moeten de bestanden uitgelezen worden om de informatie te achterhalen. Daarnaast is er meestal geen optie om additionele gegevens over het beeld op te slaan in deze database. Hierdoor zijn sommige onderzoeksvragen moeilijk te beantwoorden. Het verzamelen van alle beelden van longkankerpatienten met een tumor van een bepaald volume is met een PACS niet mogelijk zonder alle beeldbestanden uit te lezen. Het is met de Semantic Web technologie uit hoofdstuk 3 wel mogelijk om dit snel en efficiënt te verzamelen.

Hoofdstuk 4 beschrijft de oplossing voor een van overgebleven uitdagingen van hoofdstuk 3, de naamgeving van de intekeningen. De naamgeving van de tumoren orgaanomtrek intekeningen worden opgeslagen als vrije tekst. SeDI slaat deze meta-data op inclusief typefouten, taalverschillen en definitie verschillen. Voor

automatisering kunnen deze verschillen een groot probleem zijn. Het hindert de verwerking van deze gegevens op grote schaal. De softwareoplossing uit hoofdstuk 4 beschrijft een schaalbare oplossing voor dit probleem met minimale invoer van de gebruikers. Met behulp van Semantic Web technologie worden de vrije tekstvelden gekoppeld aan een definitie die leesbaar is voor mens en machine.

Hoofstuk 5 beschrijft een klinische studie waar gebruik wordt gemaakt van een eerdere versie van de software uit hoofdstuk 4. Voor de studie zijn de set-up beelden en 2D EPID dosimetrie beelden van borstkankerpatiënten verzameld. De beelden van twee patiëntengroepen werden met elkaar vergeleken: patiënten die tijdens de behandeling vrij mochten ademhalen en patiënten waarbij gebruik werd gemaakt van een techniek waarbij de patiënt vrijwillig de adem inhoudt (voluntary moderately Deep Inspiration Breath Hold). De medische beelden zijn met de software geïdentificeerd en automatisch verwerkt om de reproduceerbaarheid van de behandelingen voor beide technieken te evalueren. Uit deze analyse is te concluderen dat de reproduceerbaarheid vergelijkbaar is.

Hoofdstuk 6 beschrijft de validatie van een voorspellingsmodel voor larynxkankerpatiënten gebruikmakend van een "rapid learning" platform. Voor deze studie werd een datamining infrastructuur opgezet om structureel klinische gegevens te verzamelen die nodig zijn voor het voorspellingsmodel. De resultaten van het voorspellingsmodel werden vergeleken met de klinische dataset waar het model op geleerd was. Daarnaast werd een derde dataset geanalyseerd van een klinisch trial uitgevoerd door de RTOG. De validatie toonde aan dat het voorspellingsmodel vergelijkbaar presteerde voor de klinische validatie dataset maar het voorspellingsmodel presteerde slecht voor de klinische trial dataset.

Hoofdstuk 7 evalueert de klinische inzetbaarheid van automatische intekensoftware voor risico-organen bij longkankerpatiënten. Hierbij werden twee automatische inteken methodes vergeleken met de huidige klinische handmatige standaardmethode. De automatische intekeningen werden gemaakt met behulp van atlas gebaseerde intekeningen en "Deep Learning" algoritme. Om de klinische inzetbaarheid te meten werd de benodigde tijd om de organen in te tekenen gemeten voor de standaardmethode en de benodigde tijd om de automatische intekeningen aan te passen tot ze aan de klinische standaard voldeden. De studie toonde aan dat het inzetten van automatische intekensoftware met een menselijke validatie tijd bespaart in de kliniek terwijl de lokale richtlijnen blijven gehandhaafd.

Tot slot, vat **hoofdstuk 8** de inzichten verkregen door de verschillende studies samen. Daarnaast wordt een FAIR-data platform beschreven om richting te geven aan vervolgonderzoek.