Text mining real-world data to evaluate systemic anti-cancer therapy

Laar, S.A. van

Chapter 1

General introduction

General introduction

In the Netherlands, every year, approximately 120,000 people are diagnosed with cancer, 600,000 live with cancer, and approximately 45,000 patients die from cancer making it the most common cause of death. Also, both the incidence and absolute cancer related mortality are still rising, since the aging population and higher obesity prevalence are putting people at significant risk for cancer. Though overall survival is improving [1]. Furthermore, the costs related to cancer care in the Netherlands are high and rising, and currently estimated at 6.5 billion euro's yearly [2]. Making the development and use of (cost-)effective treatments a continuously pressing topic.

A significant proportion of all cancer patients receives systemic cancer therapy with anti-cancer drugs either as curative or palliative treatment, however exact numbers are unknown. Since 1940, chemotherapeutic antitumor drugs have become available and are known for their toxicity to non-cancerous tissue and the occurrence of drug resistance. Targeted therapies are designed to target mutations in the pathways of specific tumor cell types and include for example, tyrosine kinase inhibitors (TKIs) and monoclonal antibodies (MABs) [3]. Approval of the first MAB, rituximab, was in 1997, and the first TKI, imatinib, followed in 2001 [3, 4]. Onwards, the MABs were directed towards T-cell protein receptors and tumor antigens downregulating the immune response, creating the immune checkpoint inhibitors (ICI), of which ipilimumab was the first in 2011 [3]. All these innovations in cancer treatments have resulted in 196 approved anticancer drugs by the European Medicines Agency (EMA) in November 2022 [5].

For decades, the randomized controlled trial (RCT) has been the basis of the marketing access of anticancer agents. RCTs can be characterized by their strict patient in- and exclusion criteria, and highly protocolized treatment regimens and rigid statistical analysis plans. As the blinding and randomization further reduce the chance of bias, this method has become the golden standard to investigate treatment efficacy and safety before granting marketing access [6-8]. However, RCTs have limited external validity, as the in- and exclusion criteria are strict, and therefore may result in a highly selected patient population, not representative for the patient population in clinical practice [9]. Furthermore, as the unmet medical need of patients with cancer remains, new treatments often receive faster marketing approval through the conditional marketing approval pathway of the EMA. These approvals are often initially based on surrogate

end-points as progression-free survival, recurrence-free survival or response rates, all not always good predictors for overall survival [10, 11]. Regarding treatment safety, rare or long-term safety outcomes may be undetected during the trial period. Furthermore, it is not uncommon that new anti-cancer drugs show moderate improvements of overall survival only, that disappear if patients in clinical practice differ from those included in RCTs. More concrete, 40% of oncologic drugs with an orphan designation authorized between 2000 and 2017 did not show any clinically relevant gain in overall survival [12]. And, even though the RCT is the golden standard, there is a trend that initial registration trials only have a single-arm and a non-randomized cohort design [13]. Therefore, results from the registration trials are not always representative for the effectiveness of anticancer treatments in clinical practice, thus a knowledge gap remains.

Real-world data (RWD) are defined as all observatory healthcare data not collected in conventional trials, and in contrary to trials, RWD are generated in daily practice [14, 15]. Even though RWD are not new, the interest in RWD to generate real-world evidence, increased significantly in recent years, since they comprehend the valuable information needed to complement trial results [16]. RWD have the potential to gain insights in patients, their characteristics, treatments and outcomes in clinical practice. All of these are important variables since both individual patient characteristics and specific treatment details can influence the overall outcomes. This is extremely relevant in the field of cancer treatments, as the treatments usually enter the market with limited evidence from clinical studies, but also since the number of treatment options are rapidly expanding and treatments come with high costs [17]. Therefore, being able to determine which patients might or might not benefit (most) from a specific treatment, and if treatments are really cost-effective in clinical practice is of high importance [18]. Therefore, continuously generated RWD have the potential to be used for the evaluation of new therapies and to establish whether efficacy depends on certain patient characteristics.

The electronic health record (EHR) is the hospitals' digital version of the once paper patient files. In the EHR a vast amount of RWD is stored. Initially, EHRs were used purely for billing purposes, but the function and functionalities have widely expanded and now it is used, e.g., to keep track of patients' health information, test results, decision supports, and communication with healthcare partners [19]. As a result, EHRs contain longitudinal data on patients, their diseases and treatments collected during

routine care, especially on cancer patients, which frequently visit the hospital [20]. One part of the data is stored structurally in the EHR, including the demographics, laboratory data, medication orders and vital signs of patients. However, 80% of the information is stored in unstructured data – free-text notes, including visit summaries and correspondence – and these notes contain valuable details and nuances on a patients' treatment trajectory [21, 22]. As these notes are unstructured, manual chart review is still the standard method of data extraction, but this is time-consuming, laborious, and error- prone [21, 23, 24]. Since, this method cannot be implemented on a large scale to systematically review cancer treatments, a more advanced data collection method is warranted.

Text mining is a general term for a wide range of techniques that are used for information extraction of data from files with unstructured text. Natural language processing is the most important one, since it can be implemented to parse, segment, extract, or analyze text data. The specific algorithms used can be divided into 1) rule-based algorithms that are manually comprised, based on knowledge of the field and the expected used terminology, and 2) statistical-based algorithms, where the algorithm is trained based on large datasets with machine learning [24-27]. Text mining has been implemented in many other fields to capture data from text, however, until now the use of text mining of EHRs has been limited [28, 29]. We hypothesize that text mining has a high potential to improve the efficiency of data capture from EHR, and therefore the collection of RWD on cancer treatments.

Aim and outline of this thesis

The general aim of this thesis is to investigate if a text-mining tool is suitable for the collection of real-world data from EHRs to evaluate cancer treatments in clinical practice. More specific, we aim to investigate, first, if it is possible to capture data on cancer treatments with the same accuracy as manual review by text mining the EHR. And second, if text mining can be applied to evaluate a variety of aspects of cancer treatments in clinical practice, including treatment patterns, effectiveness, adverse events and guideline adherence.

Part I: Methods for real-world data collection

First, in **Chapter 2**, we will provide an overview of the current state of the use of real-world data. In this chapter all commonly used real-world data sources will be summarized, and we will reflect on the strengths and limitations of each source. Furthermore, we summarize the different aspects of oncologic treatments that can be reviewed with these data as the real-world effectiveness, adverse events, cost-effectiveness and treatment utilization patterns. In **Chapter 3** we will validate the text-mining tool CTcue for data collection from EHRs by performing a study in a population of renal cell carcinoma patients and compare these data with results of a manual data collection.

Part 2: Real-world treatment patterns and effectiveness

The focus of part 2 of this thesis will be on the application of text mining EHR to study treatment effectiveness and treatment patterns. In **Chapter 4** we will apply the validated tool from chapter 3 in an additional hospital and summarize the treatment patterns and the first-line effectiveness results of a metastatic renal cell carcinoma population. Next, based on the rule-based queries from chapter 4, the treatment patterns and effectiveness of population of patients with metastatic hepatocellular carcinoma will be studied in **Chapter 5.** Since hepatocellular carcinoma has many known potential prognostic factors, these are added in this study, to further investigate which patients are most likely to benefit from treatment.

Part 3: Real-world treatment safety

Part 3 includes three chapters regarding the safety of treatments in clinical practice. Compared to palliative treatments, the acceptance of adverse events is lower for adjuvant treatments in the curative setting. Therefore, we will investigate in **Chapter 6** the tolerability, safety, and preliminary efficacy of adjuvant melanoma treatments. Furthermore, some adverse events are severe but can be prevented with prophylactic treatment. An example is the occurrence of febrile neutropenia during certain types of breast cancer treatments. This can be prevented by use of granulocyte colony-stimulating factors (G-CSFs), which is indicated for a high-risk and intermediate risk group. In **Chapter 7** we will evaluate with text mining whether G-CSFs are applied according to the guideline. At last, we use the data-extraction method to fast discover the severity of adverse events of a potential treatment of Covid-19. In **Chapter 8** we react on a signal that the treatment of Covid-19 with remdesivir could result in renal- and liver damage and investigate the changes in renal- and liver function after treatment initiation.

Part IV: General discussion, summaries and appendix

This thesis will conclude with a general discussion on future perspectives in **Chapter 9**, and summaries in English and Dutch are provided in **Chapter 10**. Table 1.1 gives an overview of the investigated topics per chapter illustrating overlap and differences between studies.

Table 1.1. Overview of topics per chapter

Chapter	2	3	4	5	6	7	8
Type of study							
Review	●						
Original study		●	●	●	●	●	●
Study topics							
Method validation		●					
Treatment effectiveness		●	●	●	●		
Treatment safety		●			●	●	●
Treatment patterns			●	●			
Treatment tolerability					●		
Guideline adherence						●	
Prognostic factors				●		●	
Disease							
Renal cell carcinoma		●	●				
Hepatocellular carcinoma				●			
Melanoma					●		
Breast cancer						●	
Covid-19							●
Treatments							
Immune checkpoint inhibitors			●	●	●		
Tyrosine kinase inhibitors			●	●	●		
Chemotherapy						●	
Granulocyte-colony stimulating factors						●	
Antiviral treatment							●

References

1. *NKR cijfers*. 10 December 2022; Available from: https://iknl.nl/nkr-cijfers.
2. J. Praagman, E.S., L. van Disseldorp, V. Lemmens, *Kanker in Nederland - trends & prognoses tot en met 2032*. 2022.
3. Falzone, L., S. Salomone, and M. Libra, *Evolution of Cancer Pharmacological Treatments at the Turn of the Third Millennium.* Frontiers in Pharmacology, 2018. **9**.
4. Cohen, P., D. Cross, and P.A. Jänne, *Kinase drug discovery 20 years after imatinib: progress and future directions.* Nature Reviews Drug Discovery, 2021. **20**(7): p. 551-569.
5. Pantziarka, P., et al., *An Open Access Database of Licensed Cancer Drugs.* Frontiers in Pharmacology, 2021. **12**.
6. Franklin, J.M. and S. Schneeweiss, *When and How Can Real World Data Analyses Substitute for Randomized Controlled Trials?* Clin Pharmacol Ther, 2017. **102**(6): p. 924-933.
7. Bothwell, L.E. and S.H. Podolsky, *The Emergence of the Randomized, Controlled Trial.* 2016. **375**(6): p. 501-504.
8. Verweij, J., et al., *Innovation in oncology clinical trial design.* Cancer Treatment Reviews, 2019. **74**: p. 15-20.
9. Lakdawalla, D.N., et al., *Predicting Real-World Effectiveness of Cancer Therapies Using Overall Survival and Progression-Free Survival from Clinical Trials: Empirical Evidence for the ASCO Value Framework.* Value in Health, 2017. **20**(7): p. 866-875.
10. Pasalic, D., et al., *Progression-free survival is a suboptimal predictor for overall survival among metastatic solid tumour clinical trials.* Eur J Cancer, 2020. **136**: p. 176-185.
11. Davis, C., et al., *Availability of evidence of benefits on overall survival and quality of life of cancer drugs approved by European Medicines Agency: retrospective cohort study of drug approvals 2009-13.* BMJ, 2017. **359**: p. j4530.
12. Schuller, Y., et al., *Oncologic orphan drugs approved in the EU – do clinical trial data correspond with real-world effectiveness?* Orphanet Journal of Rare Diseases, 2018. **13**(1): p. 214.
13. Chen, E.Y., V. Raghunathan, and V. Prasad, *An Overview of Cancer Drugs Approved by the US Food and Drug Administration Based on the Surrogate End Point of Response Rate.* JAMA Internal Medicine, 2019. **179**(7): p. 915-921.
14. Makady, A., et al., *What Is Real-World Data? A Review of Definitions Based on Literature and Stakeholder Interviews.* Value in Health, 2017. **20**(7): p. 858-865.
15. Skovlund, E., H.G.M. Leufkens, and J.F. Smyth, *The use of real-world data in cancer drug development.* Eur J Cancer, 2018. **101**: p. 69-76.
16. Ramamoorthy, A. and S.-M. Huang, *What Does It Take to Transform Real-World Data Into Real-World Evidence?* Clinical Pharmacology & Therapeutics, 2019. **106**(1): p. 10-18.
17. Vokinger, K.N., et al., *Prices and clinical benefit of cancer drugs in the USA and Europe: a cost–benefit analysis.* The Lancet Oncology, 2020. **21**(5): p. 664-670.
18. Wild, C.P., et al., *Cancer Prevention Europe.* Molecular Oncology, 2019. **13**(3): p. 528-534.
19. Kim, E., et al., *The Evolving Use of Electronic Health Records (EHR) for Research.* Seminars in Radiation Oncology, 2019. **29**(4): p. 354-361.
20. Cowie, M.R., et al., *Electronic health records to facilitate clinical research.* Clin Res Cardiol, 2017. **106**(1): p. 1-9.
21. Casey, J.A., et al., *Using Electronic Health Records for Population Health Research: A Review of Methods and Applications.* Annu Rev Public Health, 2016. **37**: p. 61-81.
22. Fessele, K.L., *The Rise of Big Data in Oncology.* Semin Oncol Nurs, 2018. **34**(2): p. 168-176.

23. Haerian, K., et al., *Detection of pharmacovigilance-related adverse events using electronic health records and automated methods.* Clin Pharmacol Ther, 2012. **92**(2): p. 228-34.
24. Assale, M., et al., *The Revival of the Notes Field: Leveraging the Unstructured Content in Electronic Health Records.* Front Med (Lausanne), 2019. **6**: p. 66.
25. Ford, E., et al., *Extracting information from the text of electronic medical records to improve case detection: a systematic review.* Journal of the American Medical Informatics Association : JAMIA, 2016. **23**(5): p. 1007-1015.
26. Wong, A., et al., *Natural Language Processing and Its Implications for the Future of Medication Safety: A Narrative Review of Recent Advances and Challenges.* Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy, 2018. **38**(8): p. 822-841.
27. Sun, W., et al., *Data Processing and Text Mining Technologies on Electronic Medical Records: A Review.* Journal of Healthcare Engineering, 2018. **2018**: p. 4302425.
28. Wang, Y., et al., *Clinical information extraction applications: A literature review.* J Biomed Inform, 2018. **77**: p. 34-49.
29. Sohn, S., et al., *Drug side effect extraction from clinical narratives of psychiatry and psychology patients.* J Am Med Inform Assoc, 2011. **18 Suppl 1**: p. i144-9.