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

Laar, S.A. van

Part I

Methods for real-world data collection

Chapter 2

Use of real-world data sources to evaluate cancer treatments: current practice and future perspectives

Sylvia A. van Laar, Kim B. Gombert-Handoko,
Henk-Jan Guchelaar, Juliëtte Zwaveling

Manuscript in preparation

Abstract

Over the last two decades the interest in real-world data (RWD), specifically on cancer treatments, has significantly increased since they can bridge the knowledge gap in evidence not addressed in randomized clinical trials. All sources of RWD used in studies have their own strengths and limitations and therefore can be used to investigate various aspects of oncologic drug treatments in daily practice. This review first gives an overview of the currently available real-world sources used to evaluate cancer treatments in clinical practice. It includes sources as the case report form, electronic health records, administrative claims data, patient reported outcomes, cancer registries, mobile applications, wearable devices, and social media data and their strengths and limitations. Next, it elaborates on the potential applications of RWD, and summarizes how data from these sources can, in varying degrees, be used to estimate real-world effectiveness, monitor pharmacovigilance, calculate cost-effectiveness, describe general treatment utilization patterns, and support personalized treatment choice of cancer treatments. And, it gives a perspective of how RWD use can be optimized as the field of RWD on oncologic treatments is rapidly growing. The improvements include, but are not limited to, more harmonized and complete RWD sources, more connected or merged datasets, and more use of artificial intelligence in data processing. Furthermore, improved privacy legislation can even further enhance the insights generated with RWD. Concluding, a variety of RWD sources can be used to generate significant insights into cancer treatments in clinical practice. With further improvements RWD use can reach its full potential.

1. Introduction

Randomized controlled trials (RCTs) are the golden standard in determining the efficacy and safety of medical drug treatments due to its proven and rigorous study design, which makes RCTs crucial in market authorization for drug treatments [1]. However, RCTs also have potential weaknesses: by design, they are time and resource-intensive, and only include a selected patient group; they are often not designed for follow-up of long-term toxicity, frequently use surrogate parameters, and they have limited external validity [2, 3].

In general, all data regarding the effects of health interventions in patients not collected in RCTs are defined as real-world data (RWD) [4, 5]. RWD includes data primarily generated for studies in clinical practice, and the reuse or secondary use of data already generated for other (healthcare) purposes. As health-related information is increasingly stored digitally, these routinely generated data, e.g., electronic health records, administrative claims- and patient-derived data, can also be used as RWD sources [4, 6, 7]. Evidence derived from RWD, real-world evidence (RWE), in general, has greater external validity and generalizability RCTs. For example, by including larger populations, RWD can gain insight in treatment effectiveness and long-term and rare toxicity. Also, topics not included in trials can be studied, such as treatment patterns, effectivity in rare cancer tumours, and the effects in specific subgroups can be analysed [2]. In this way, RWE can add valuable information to the evidence which is obtained from RCTs [8].

In the last two decades an exponential growth is seen in publications with RWD and RWE on PubMed. It increased from 12 articles in 2002 to 3,310 articles in 2021, including opinion and perspective papers, original studies, and (systematic) reviews [9]. Furthermore, in 2021 around a quarter of the publications included the topic "cancer" (Figure 2.1). All sources of RWD used in studies have their own strengths and limitations and can be used to investigate various aspects of oncologic drug treatments in daily practice [10]. In addition, the increase in the number of studies results in a rapidly changing field. Therefore, the aim of this review is to give an overview of the current options for real-world sources used to evaluate cancer treatments in clinical practice, and how it can be optimized in the future.

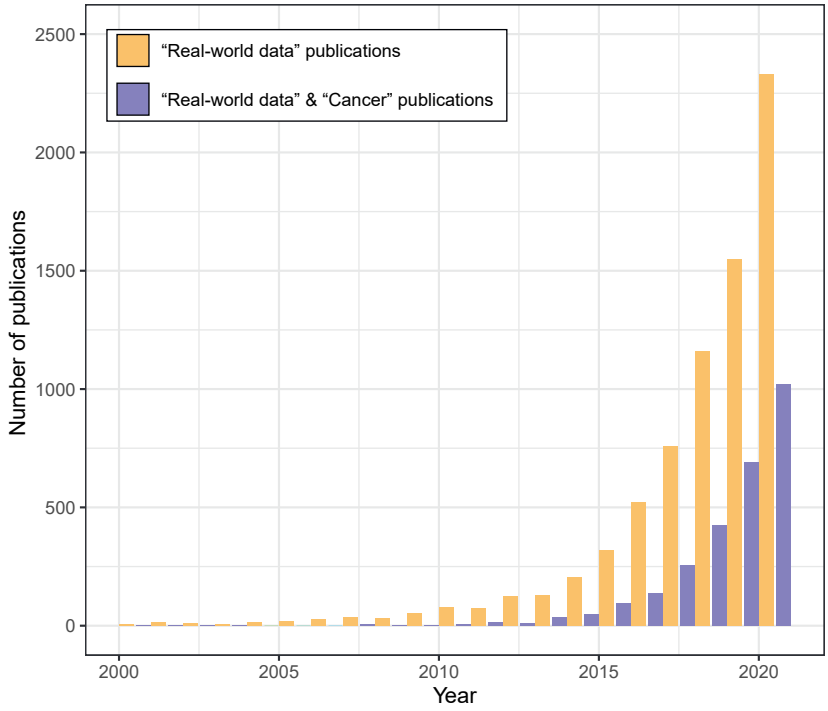

Figure 2.1. Number of publications on "Real-world data" and "Real-world data" combined with "cancer" per year on PubMed.

2. Sources

Multiple sources contain RWD on cancer treatments, and each source has its advantages and challenges regarding the time and effort it takes to collect the data, the quality of the data, the coverage of information and therefore the potential risk of bias. Table 2.1 shows a summary of basic characteristics and the strengths and limitations per source.

2.1 Case report form

The standard method to generate data prospectively, and primarily for study purposes, is by using case report forms (CRFs). Traditionally CRFs were paper forms, but they are now mostly digital (electronic CRFs [eCRFs]). Based on a study protocol, they are used to collect all study results, which is mandatory by the Good Clinical Practice guidelines for clinical trials [11, 12]. However, CRFs can also be used to collect data from clinical practice. The process of data collection, recording, monitoring, and auditing is done by staff, which makes the use of CRFs to collect data a time intensive and costly process [11].

Table 2.1. Strengths and limitations of real-world data sources

	Real-world data source	Strengths	Limitations
	Case report forms	- All essential data can be collected - High data quality	- Expensive - Laborious - Data is not readily available
	Electronic health records	- Extensive and nuanced longitudinal patient, disease, and treatment data	- Often limited to one healthcare center - Complex data source - Risk of missing or incorrect data
	Administrative claims data	- Longitudinal data across healthcare facilities - Insights in costs	- High variety in sensitivity - Limited to reimbursable healthcare
	Patient reported outcomes	- Direct insight in patients' perception of treatments and quality of life - Multiple validated questionnaires available	- Use of PROMS is not yet standard of care - Patients' willingness and capacity to participate can bias outcomes
	Registries	- Large, structured databases with patient, disease, and treatment data - Reflects treatments on population level	- Laborious - Data may lag current practice
	Wearable devices	- High density longitudinal biometric data	- Data generation is limited to users - Still in development
	Social media data	- Insight in patients' perception on treatments and disease	- Limited to social media users - Missing data on patient demographics

2.2 Electronic medical records

A vast amount of RWD is collected from electronic health records (EHRs). Primarily, EHRs were implemented for billing purposes, but has expanded to many core functions including keeping track of patients' health information, test results, decision support, and improvement of communication with care partners [13]. 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 [14]. It includes data on primary care-, hospital and emergency department visits, demographics, vital signs, laboratory data, medication orders, procedures, imaging, health behaviour, and free-text notes, e.g., summarizing encounters and correspondence [15]. These free-text notes can make up to 80% of the healthcare information, and contain valuable details and nuances on a patients' treatment trajectory [16]. EHR data has limitations when used for research purposes. Since accurate and correct documentation may be difficult in clinical practice, data may be missing, erroneous, uninterpretable, or inconsistent, in particular in free-text notes. Also, EHRs of a single center may not represent the complete medical history of a patient and a patient population [17, 18]. Furthermore, high variety in expressions in terminology – the linguistic complexity – and the level of detail or specificity used to describe a specific case can limit the accuracy of automatic data extraction [17-19].

2.3 Administrative claims data

Administrative claims data are generated for reimbursement purposes. However, since the billing codes from claims for insurance providers include coded information on diagnoses and healthcare services provided, they can also give insight in provided cancer treatments [20]. What makes claims data unique, is that they capture longitudinal data from individuals across healthcare facilities, including in- and outpatient visits, and dispensing data of prescription medication from outpatient or community pharmacies [20, 21]. Therefore, these data can be used to study specific elements on healthcare delivery such as the hospital volumes, healthcare providers, and prescribing patterns [20]. Claims data only contains reimbursable healthcare, therefore excludes, e.g., over-the-counter medicines. Furthermore, they are more likely to present more severe clinical conditions, i.e., conditions that require surgery, or persistent healthcare increase the accuracy of correct coding. Therefore, the sensitivity of specific outcomes in claims data varies [20, 21]. Furthermore, compared to other sources, claims data often lack clinical, laboratory and patient's behavioral data [20].

2.4 Patient reported outcome measures

Patient reported outcome measures (PROMs) are standardized questionnaires through which patients themselves directly share their experience and perception of a specific disease or treatment by reflecting e.g., on health-related quality of life, symptoms, and

treatment adherence and satisfaction. These questionnaires can be generic for quality of life, e.g., the EQ-5D [22], or disease specific, like the questionnaire for patients with cancer: EORTC-QLQ-C30 [23, 24]. Primarily, they were developed to assess the effectiveness of treatment in clinical trials [25, 26]. However, the applications have further broadened, by including them in clinical practice, improving the communication regarding and assessment and management of patients' symptoms [25]. Furthermore, these PROMs create the possibility to retrospectively review all patient reported experiences. However, due to the lack of time and knowledge and inadequate digital infrastructures clinicians have difficulty with successfully integrate, interpret and act upon the PROMs data, which must be taken into account when using PROMs as RWD source [27]. Furthermore, it must be taken into account that the willingness and capacity of patients to continuously participate with the questionnaires can bias results [28].

2.5 Cancer registries

In cancer registries data from all patients with a certain type of cancer and/or with a specific treatment in clinical practice is systematically and continuously stored without specific prospectively defined outcomes [20, 29-31]. They are often population-based on a regional or national level, but can also be health-care setting based (e.g., one hospital) [31]. In registries, in general, a wealth of information on patients' sociodemographic and clinical characteristics is collected, making them a unique source of data on all patients in a certain region and therefore ideal to retrospectively investigate, e.g., cancer incidence and treatment effectiveness in a specific geographical area, social group or time period [31, 32]. For the collection of data, registries are mostly dependent on EHR data from hospitals, which are usually manually entered [33], but also data from independent (pathology) labs, physician offices, administrative claims data, biomarker data, pharmacy claims, patient reported outcomes and general demographic data can be included [20]. Even though registries are large and structured databases, they lack uniformity. Therefore, not every registry will contain the same variables [33]. Also, information on disease progression or recurrence and adverse drug reactions is not always structurally collected [20]. And, as data collection from various sources is time-consuming, the results from registries can lag 1–2 years behind current practice [20, 34].

2.6 Mobile applications and wearable devices

Wearable devices are items worn around the body (e.g., around the wrist, upper arm, waist, or hip) that can monitor and objectively collect biometric data (e.g., heart rate, respiratory rate, oxygen saturation, physical activity, sleep patterns) [35, 36]. These devices also can be used for the evaluation of cancer treatments. Multiple recent review articles mention that these tools have the potential to investigate adverse events, treatment adherence, symptoms, practice of physical activity, and overall, the quality of life [35-37]. Limitations in the use of these wearable items include, firstly, that users must have a certain level of digital literacy, which may be lacking in the elder generations. Furthermore, the devices themselves must be worn and function without technical problems, in order to generate data. Therefore the consumer-grade tools have to be validated and monitored [36]. Also, consensus still must be found on the terminology regarding wearable devices, in order to increase uniformity of the outcome measures in studies [37].

2.7 Social media data

Next to all data collected in and around the healthcare setting, social media use has significantly increased. This concurrent expansion of generated digital information has the potential to be used for cancer treatment evaluation. A review of Kalf et al. [38] showed that data collection from social media sources as forum topics and discussions, Facebook pages of patient communities or support groups and Twitter conversations can be a potential source to study relative effectiveness, adverse events, symptoms, quality of life, and adherence behaviour of patients with cancer and their treatments. Social media listening can also be used to give insight in patients experiences with (new) treatments [39]. Social media data can be biased due to the specific user profile, e.g. younger and female patients, and post may be duplicated, topics can be recurrent, and reporting on social media can also induce the reporting from patients with similar events [38]. Also, the amount of demographic or clinical information of patients on social media is limited, e.g., regarding patients' age, sex, or specific disease characteristics [39].

3. Applications of RWD in the evaluation of drug treatments

All sources of RWD can be used for specific research purposes in various phases of the life cycle of a drug – from drug discovery to post-registration. Multiple RWD applications have been defined and categorized.

3.1 Effectiveness

In real-world studies the treatment outcomes, and more specifically the effectiveness of a treatment in a specific population, are studied frequently for new cancer treatments post-marketing. Thereby the effectiveness of treatments under different circumstances than the trial and trial population can be studied, e.g., elderly, patients with comorbidities or limited kidney-, liver-, or cardiac function. But also, for the investigation of long-term efficacy outcomes, RWD are relevant, since in RCTs the outcome of a treatment is mostly studied during a shorter period [40]. For an accurate determination of the treatment effectiveness in clinical practice, insight in patient characteristics, previous treatments a patient received and the accompanying outcomes as recurrence, progression, or mortality are necessary. Electronic health records data, especially when merged into a registry or other larger databases, are suitable for this purpose, when the caveats as inaccurate or incomplete data are considered [17, 41]. Also claims data are used for effectiveness studies, mostly regarding the overall survival after treatment initiation [42]. The growing quantity of effectiveness studies even enables systematically reviewing specific outcomes [43].

3.2 Pharmacovigilance

The monitoring of adverse events is continued after marketing authorization, which is an important part of pharmacovigilance. Indeed, not all adverse events (AEs) are identified in RCTs; especially rare and late AEs do often not surface during this phase [44]. Especially for oncologic treatments pharmacovigilance is important, as almost all treatments are accompanied with serious toxicity. New oncologic treatments frequently enter the market before their safety profile is complete, and more often fast track routes are applied to accelerate market entry [45]. Furthermore, new targeted treatments and immunotherapies come with their own unique, complex safety profiles [46]. AEs can arise as a signal from multiple sources indicating a potential causal association between a treatment and the specific AE. Spontaneous reporting systems, in which a reporter, e.g., physician, reports their suspicion of an adverse event, have been the cornerstone in the pharmacovigilance process. However, the disadvantages of this system are, in general, a combination of underreporting, overreporting of publicized events, biased reporting, incomplete patient information, and missing of a denominator to estimate the frequency of this AE [44, 46, 47]. Therefore, in the field of pharmacovigilance, there is increasing interest in various sources of RWD. Reviews of Crestan et al. [46] and Lavertu et al. [47] emphasize the potential use of big data collections as EHR,

claims, and social media data. And both groups underline the potential application of innovative methods for the synthesis and analysis of data, as in general the use of artificial intelligence, including but not limited to, data mining, cognitive computing, and natural language processing. Furthermore, for more insights in the patients' perspective on adverse events, PROMs are a relevant RWD source, as observed in the study of Stormoen et al., who investigated the burden of adverse events in a population of prostate cancer patients receiving medical treatment [48].

3.3 Cost-effectiveness

Demonstrating cost-effectiveness is especially relevant for cancer treatments, since new treatments often come with high prices even though the uncertainty on the added clinical value [49, 50]. RWD already has already extensively be used in the field of cancer treatments for cost-effectiveness. For example, almost all (96%) single technology appraisals of oncologic drug treatment performed by the National Institute for Health and Care Excellence of the United Kingdom contained some RWD in the cost-effectiveness analysis [51]. Moreover, insight into the clinical value of treatments with RWD can substantiate more accurate economic evaluations, and therefore influence reimbursement decisions and price negotiations [52]. Key components of the estimation are survival outcomes (discussed in "3.1 effectiveness"), quality of life, and costs [53]. The quality of life can best be evaluated by using patient reported outcomes, and increasing the use of PROMs will improve these estimations, as currently the quality of life is not often systematically recorded [54]. Furthermore, administrative claims data are the prime source for costs data [55]. Furthermore, detailed information as actual number of used vials of a specific treatment can be extracted from EHR, as is done in the real-world cost-effectiveness analysis of Van Kampen et al. [56].

3.4 Indication expansion

Often indications are expanded later in the drug treatment life-cycle [45]. Real-world evidence on the off-label use reflecting on the safety and efficacy of a treatment can be useful in this process. To illustrate, between 1 January 2018 until 31 December 2019, 12 of the 78 applications for the extensions of indications regarding antineoplastic and immunomodulating agents submitted at the European Medicines agency included RWE, mainly derived from registries and EHR data [57]. One example is the expansion of the application of palbociclib in men with hormone receptor-positive/human epidermal growth factor receptor 2-negative metastatic breast cancer in addition to the

registration for women. This was based on the study of Kraus et al. [58], who showed by reviewing EHRs, medical claims and prescription data, that palbociclib use in men was safe and consistent with the treatment of women.

3.5 General treatment prescribing and utilization patterns

How medication is used in clinical practice, can only be studied by analysing RWD. Treatment-related aspects include the specific choice of treatment, the dose, and the accompanying order in lines of treatment, reflected in treatment pattern studies in certain populations and moments [40]. These studies give insight in (shifts in) preference, and the adherence to treatment guidelines. Also, the time on treatment, treatment compliance and dose (reductions) over time add important understandings of how treatments are used and tolerated in clinical practice. For example, the review of Waser et al. [59] summarized twenty treatment pattern studies in non-metastatic non-small-cell lung cancer patients, to conclude that the patterns of increasing use of (neo)adjuvant treatment with increasing stage were, in general, in line with guidelines, and these studies mainly used data from registries, claims and medical records. Furthermore, Gan et al. [60] showed maintained effectiveness of cabozantinib in first- to fourth line treatment use, and associated toxicity-related dose reductions with improved outcomes based on a database with hospital and pharmacy records data. These studies also show that for comprehensive insights in patterns, larger datasets are needed, and therefore claims data, or larger registries containing EHR data are suitable sources.

3.6 Personalized treatment choice

Next to drug utilization in the general population, the characterisation of specific patient populations receiving treatments is relevant to interpret use and effectiveness of treatments of real-world cancer patients. It is well-known that trial populations are not always representative due to strict eligibility criteria and underrepresentation of certain demographics, e.g., related to race or age [61, 62]. Besides, disease related factors including biomarkers and non-related factors as a patients' physical state, genetics, comorbidities, comedication can influence outcomes [1, 63]. Detailed RWD, especially from health records, can be used to generate better insight in the patient characteristics and the accompanying outcomes can potentially help in identifying what treatment a specific patient might benefit the most from. Ideally, RWD is used to estimate the influence of all these individual factors on specific treatments and outcomes for clinical decision support in therapeutic decision making [64]. Machine learning algorithms

can potentially be the key for the pattern recognition in the growing amount of data, however due to the complexity of the data, this is not yet practice [65].

4. Future perspectives

The amount of available RWD of patients with cancer is constantly growing and can be found in several data sources. These RWD have the potential to replace or add data to the evidence generated by conventional research methods by being able to generate datasets that are quicker, cheaper, or larger, and generate data that were previously not available. This manuscript highlights the most common sources used to generate insights in oncologic treatment effects after registration, and the aspects of a treatment of which insight can be improved. It is shown that not every data source is suitable for all research questions, but a combination of sources can generate a wide range of data of oncologic patients and their treatments in clinical practice. However, improvements are needed before the use of RWD will reach its full potential.

4.1 Optimising source data

In the basis, the quality of the data is essential for collecting meaningful RWD which is suitable for analysis. However, a repeating disadvantage of all sources is that the primary purpose of the data capture was not research, and therefore the data quality can be limited. Maissenhaelter et al. [2] summarized the challenges in data quality in data completeness, data structure, data accuracy and the challenge of novel types of data. Part of the solution can be improving and harmonizing the data capture by generating more user-friendly, but clear front-ends of the used software, for example of EHRs. However, it will remain impossible to capture all patient data, therefore, minimum standards have to be established and harmonized over all used software programs [66]. Also, data entry can in the future potentially be replaced with new techniques as automatic speech-recognition based clinical documentation. However, these techniques are still in development [67]. Also, with more uniform data sources, combination of data and the implementation of application of smart analysis (4.3) will be made easier, as currently most data is siloed in their own systems [65]. However, only improving the data input will not solve all challenges with RWD.

4.2 Integration of sources and systems

The next challenge lies in the better integration of the sources and systems. This problem can be split into two parts, namely that that data from the same source type can be easily merged, and that different source types can be combined. If data sources are more harmonized, they can more easily be combined in joint data warehouses, and thereby the pool of RWD will be enlarged [2]. Subsequently, these large pools of RWD enable larger population studies. Registries are already a form of merged datasets, however, by harmonizing their sources, data collection can potentially be further automated. Furthermore, for example, integration of data from wearable devices into the EHR can add extra perspective, not only for research purposes, but also for healthcare practice. Already some devices have been piloted or implemented in EHR systems. Until now, systems are not always interoperable or cannot connect with wearable devices. In addition, EHR systems have to be ready to store the amount of data that wearables generate [68].

4.3 Artificial intelligence for smart analysis

Next to the efforts to optimizing and combining sources, the method of handling data also can be improved, especially with artificial intelligence. Data language processing techniques have the potential to process and extract information from large amounts of unstructured data, such as captured in the EHR and social media. To transfer EHR data into a form fit for retrospective analysis, manual extraction has been the standard method [69]. However, as this method is time-consuming, it is not very scalable [70]. With use of text mining, including natural language processing (NLP) strategies, the automatization of data extraction from EHR is currently in development [71, 72]. This technique can be divided into rule-based NLP, a system based on a coded set of rules for data extraction, and machine-learning-based NLP, a system based on a statistical learning algorithm that identifies the relevant context based on annotated training data [72]. Automated techniques, including machine learning algorithms, natural language processing software with sentiment analysis, are already used to analyze large social media datasets [38, 39, 73]. And with this, social media studies may develop into a fast method to collect patient experiences and add information to PROs [38].The challenge in the field of NLP is generating an algorithm that is specific and precise enough to fit the heterogenous sources by rule-based or machine learning [74]. Furthermore, when datasets have sufficient quality and quantity, ultimately, AI can be used for pattern recognition in the rich and complex data to evaluate specific cancer treatments [65].

4.4 Privacy

With these constantly growing sources of patient data, it remains fundamental to protect the privacy of the patients. In the European Union (EU) patient data is protected by the General Data Protection Regulation (GDPR). And, in line, performing research with RWD should always be accompanied with the appropriate consents, anonymization or pseudonymisation of the data and data handling should only be performed by trained and qualified personnel [2]. However, the current privacy- and data regulations also hinder optimal reuse of these valuable real-world data, due to, e.g., hurdles on gaining access to data [75]. Therefore, the individual right of data privacy competes with the populations' right on (cancer) treatment improvements that could be enabled with these RWD [2]. New types of consent from patients, for example, the proposal on data altruism, in the proposed Data Governance Act of the EU, in which patients voluntarily make their data available for scientific and public purposes, could be an outcome for this barrier [75, 76].

5. Conclusion

The amount of RWD available on oncologic patients and their disease and treatment trajectory has significantly grown. This manuscript showed a non-exhaustive overview of the current practice and summarized RWD sources as traditional claims data, EHRs and more recently emerging fields as social media and wearable devices. Data from these sources are already used to gain insight into a wide field of oncologic treatment aspects in daily practice, including, effectiveness, pharmacovigilance, and prescribing patterns and can add relevant data for indication expansion, personalized treatment choice and cost-effectiveness. We conclude that the field of RWD is rapidly developing, and more harmonized and complete sources, smarter connections between data sources, use of artificial intelligence on several levels, and improvement of patient privacy legislation can even further enhance the insights generated with RWD.

References

1. Phillips, C.M., et al., *Assessing the efficacy-effectiveness gap for cancer therapies: A comparison of overall survival and toxicity between clinical trial and population-based, real-world data for contemporary parenteral cancer therapeutics.* Cancer, 2020. **126**(8): p. 1717-1726.
2. Maissenhaelter, B.E., A.L. Woolmore, and P.M. Schlag, *Real-world evidence research based on big data: Motivation-challenges-success factors.* Onkologe (Berl), 2018. **24**(Suppl 2): p. 91-98.
3. 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 Intern Med, 2019. **179**(7): p. 915-921.
4. Makady, A., et al., *What Is Real-World Data? A Review of Definitions Based on Literature and Stakeholder Interviews.* Value Health, 2017. **20**(7): p. 858-865.
5. Ramamoorthy, A. and S.M. Huang, *What Does It Take to Transform Real-World Data Into Real-World Evidence?* Clin Pharmacol Ther, 2019. **106**(1): p. 10-18.
6. Pisaniello, H.L. and W.G. Dixon, *What does digitalization hold for the creation of real-world evidence?* Rheumatology (Oxford), 2020. **59**(1): p. 39-45.
7. de Lusignan, S. and L. Crawford, *Creating and using real-world evidence to answer questions about clinical effectiveness.* J Innov Health Inform., 2015. **22**(3): p. 368-373.
8. Feinberg, B.A., et al., *Use of Real-World Evidence to Support FDA Approval of Oncology Drugs.* Value in Health, 2020. **23**(10): p. 1358-1365.
9. *PubMed.gov*. 2 February 2022]; Available from: https://pubmed.ncbi.nlm.nih.gov/?term=real-world+evidence&filter=pubt.review.
10. Simon, G.E., et al., *When Can We Trust Real-World Data To Evaluate New Medical Treatments?* Clin Pharmacol Ther, 2022. **111**(1): p. 24-29.
11. O'Leary, E., et al., *Data collection in cancer clinical trials: Too much of a good thing?* Clin Trials, 2013. **10**(4): p. 624-32.
12. Bellary, S., B. Krishnankutty, and M.S. Latha, *Basics of case report form designing in clinical research.* Perspect Clin Res, 2014. **5**(4): p. 159-66.
13. Kim, E., et al., *The Evolving Use of Electronic Health Records (EHR) for Research.* Seminars in Radiation Oncology, 2019. **29**(4): p. 354-361.
14. Cowie, M.R., et al., *Electronic health records to facilitate clinical research.* Clin Res Cardiol, 2017. **106**(1): p. 1-9.
15. 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.
16. Fessele, K.L., *The Rise of Big Data in Oncology.* Semin Oncol Nurs, 2018. **34**(2): p. 168-176.
17. Hersh, W.R., et al., *Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research.* Medical Care, 2013. **51**.
18. Gianfrancesco, M.A. and N.D. Goldstein, *A narrative review on the validity of electronic health record-based research in epidemiology.* BMC Medical Research Methodology, 2021. **21**(1): p. 234.
19. Hanauer, D.A., et al., *Complexities, variations, and errors of numbering within clinical notes: the potential impact on information extraction and cohort-identification.* BMC Medical Informatics and Decision Making, 2019. **19**(3): p. 75.
20. Penberthy, L.T., et al., *An overview of real-world data sources for oncology and considerations for research.* CA Cancer J Clin, 2022. **72**(3): p. 287-300.

21. Gross, M.D., B. Al Hussein Al Awamlh, and J.C. Hu, *Assessing Treatment-Related Toxicity Using Administrative Data, Patient-Reported Outcomes, or Physician-Graded Toxicity: Where Is the Truth?* Seminars in Radiation Oncology, 2019. **29**(4): p. 333-337.
22. Herdman, M., et al., *Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L).* Quality of Life Research, 2011. **20**(10): p. 1727-1736.
23. Aaronson, N.K., et al., *The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology.* J Natl Cancer Inst, 1993. **85**(5): p. 365-76.
24. Nolte, S., et al., *General population normative data for the EORTC QLQ-C30 health-related quality of life questionnaire based on 15,386 persons across 13 European countries, Canada and the Unites States.* Eur J Cancer, 2019. **107**: p. 153-163.
25. Pérez-Alfonso, K.E. and V. Sánchez-Martínez. *Electronic patient-reported outcome measures evaluating cancer symptoms: A systematic review*. in *Seminars in Oncology Nursing*. 2021. Elsevier.
26. Churruca, K., et al., *Patient-reported outcome measures (PROMs): A review of generic and condition-specific measures and a discussion of trends and issues.* Health Expectations, 2021. **24**(4): p. 1015-1024.
27. Nguyen, H., et al., *A review of the barriers to using Patient-Reported Outcomes (PROs) and Patient-Reported Outcome Measures (PROMs) in routine cancer care.* Journal of Medical Radiation Sciences, 2021. **68**(2): p. 186-195.
28. Zini, M.L.L. and G. Banfi, *A Narrative Literature Review of Bias in Collecting Patient Reported Outcomes Measures (PROMs).* International Journal of Environmental Research and Public Health, 2021. **18**(23): p. 12445.
29. Garrison, L.P., et al., *Using Real-World Data for Coverage and Payment Decisions: The ISPOR Real-World Data Task Force Report.* Value in Health, 2007. **10**(5): p. 326-335.
30. Skovlund, E., H.G.M. Leufkens, and J.F. Smyth, *The use of real-world data in cancer drug development.* European Journal of Cancer, 2018. **101**: p. 69-76.
31. Thong, M.S.Y., et al., *Population-based cancer registries for quality-of-life research.* Cancer, 2013. **119**(S11): p. 2109-2123.
32. Forsea, A.M., *Cancer registries in Europe-going forward is the only option.* Ecancermedicalscience, 2016. **10**: p. 641.
33. Hoeijmakers, F., et al., *National quality registries: how to improve the quality of data?* J Thorac Dis, 2018. **10**(Suppl 29): p. S3490-s3499.
34. Parums, D.V., *Editorial: Registries and Population Databases in Clinical Research and Practice.* Med Sci Monit, 2021. **27**: p. e933554.
35. Albino de Queiroz, D., et al., *Internet of Things in active cancer Treatment: A systematic review.* J Biomed Inform, 2021. **118**: p. 103814.
36. Low, C.A., *Harnessing consumer smartphone and wearable sensors for clinical cancer research.* NPJ Digit Med, 2020. **3**: p. 140.
37. Beauchamp, U.L., H. Pappot, and C. Holländer-Mieritz, *The Use of Wearables in Clinical Trials During Cancer Treatment: Systematic Review.* JMIR Mhealth Uhealth, 2020. **8**(11): p. e22006.
38. Kalf, R., et al., *Use of Social Media in the Assessment of Relative Effectiveness: Explorative Review With Examples From Oncology.* JMIR Cancer, 2018. **4**: p. e11.
39. Mazza, M., et al., *Social Media Listening to Understand the Lived Experience of Individuals in Europe With Metastatic Breast Cancer: A Systematic Search and Content Analysis Study.* Frontiers in Oncology, 2022. **12**.

40. Pulini, A.A., et al., *Impact of Real-World Data on Market Authorization, Reimbursement Decision & Price Negotiation.* Ther Innov Regul Sci, 2021. **55**(1): p. 228-238.
41. Kumar, A., et al., *Evaluation of the Use of Cancer Registry Data for Comparative Effectiveness Research.* JAMA Network Open, 2020. **3**(7): p. e2011985-e2011985.
42. Luyendijk, M., et al., *Assessment of Studies Evaluating Incremental Costs, Effectiveness, or Cost-Effectiveness of Systemic Therapies in Breast Cancer Based on Claims Data: A Systematic Review.* Value in Health, 2020. **23**(11): p. 1497-1508.
43. Johal, S., et al., *Real-world treatment patterns and outcomes in small-cell lung cancer: a systematic literature review.* J Thorac Dis, 2021. **13**(6): p. 3692-3707.
44. Patadia, V.K., et al., *Using real-world healthcare data for pharmacovigilance signal detection – the experience of the EU-ADR project.* Expert Review of Clinical Pharmacology, 2015. **8**(1): p. 95-102.
45. Michaeli, D.T., et al., *Initial and supplementary indication approval of new targeted cancer drugs by the FDA, EMA, Health Canada, and TGA.* Invest New Drugs, 2022. **40**(4): p. 798-809.
46. Crestan, D., et al., *Pharmacovigilance of anti-cancer medicines: opportunities and challenges.* Expert Opin Drug Saf, 2020. **19**(7): p. 849-860.
47. Lavertu, A., et al., *A New Era in Pharmacovigilance: Toward Real-World Data and Digital Monitoring.* Clinical Pharmacology & Therapeutics, 2021. **109**(5): p. 1197-1202.
48. Stormoen, D.R., et al., *Patient reported outcomes interfering with daily activities in prostate cancer patients receiving antineoplastic treatment.* Acta Oncologica, 2021. **60**(4): p. 419-425.
49. Shin, G., H.Y. Kwon, and S. Bae, *For Whom the Price Escalates: High Price and Uncertain Value of Cancer Drugs.* Int J Environ Res Public Health, 2022. **19**(7).
50. Vokinger, K.N., et al., *Prices and clinical benefit of cancer drugs in the USA and Europe: a cost-benefit analysis.* Lancet Oncol, 2020. **21**(5): p. 664-670.
51. Bullement, A., et al., *Real-world evidence use in assessments of cancer drugs by NICE.* International Journal of Technology Assessment in Health Care, 2020. **36**(4): p. 388-394.
52. Guggenbickler, A.M., et al. *Rapid Review of Real-World Cost-Effectiveness Analyses of Cancer Interventions in Canada*. Current Oncology, 2022. **29**, 7285-7304 DOI: 10.3390/curroncol29100574.
53. Hall, P.S., *Real-world data for efficient health technology assessment.* European Journal of Cancer, 2017. **79**: p. 235-237.
54. Parody-Rúa, E., et al., *Economic Evaluations Informed Exclusively by Real World Data: A Systematic Review.* International Journal of Environmental Research and Public Health, 2020. **17**(4): p. 1171.
55. Shih, Y.T. and L. Liu, *Use of Claims Data for Cost and Cost-Effectiveness Research.* Semin Radiat Oncol, 2019. **29**(4): p. 348-353.
56. van Kampen, R.J.W., et al., *Real-world and trial-based cost-effectiveness analysis of bevacizumab in HER2-negative metastatic breast cancer patients: a study of the Southeast Netherlands Breast Cancer Consortium.* European Journal of Cancer, 2017. **79**: p. 238-246.
57. Flynn, R., et al., *Marketing Authorization Applications Made to the European Medicines Agency in 2018–2019: What was the Contribution of Real-World Evidence?* Clinical Pharmacology & Therapeutics, 2022. **111**(1): p. 90-97.
58. Kraus, A.L., et al., *Real-World Data of Palbociclib in Combination With Endocrine Therapy for the Treatment of Metastatic Breast Cancer in Men.* Clinical Pharmacology & Therapeutics, 2022. **111**(1): p. 302-309.
59. Waser, N., et al., *Real-world treatment patterns in resectable (stages I–III) non-small-cell lung cancer: a systematic literature review.* Future Oncology, 2022. **18**(12): p. 1519-1530.

60. Gan, C.L., et al., *Cabozantinib real-world effectiveness in the first-through fourth-line settings for the treatment of metastatic renal cell carcinoma: Results from the International Metastatic Renal Cell Carcinoma Database Consortium.* Cancer Med, 2021. **10**(4): p. 1212-1221.
61. Liu, R., et al., *Evaluating eligibility criteria of oncology trials using real-world data and AI.* Nature, 2021. **592**(7855): p. 629-633.
62. Jayakrishnan, T., et al., *Landmark Cancer Clinical Trials and Real-World Patient Populations: Examining Race and Age Reporting.* Cancers, 2021. **13**(22): p. 5770.
63. Sarfati, D., B. Koczwara, and C. Jackson, *The impact of comorbidity on cancer and its treatment.* CA: A Cancer Journal for Clinicians, 2016. **66**(4): p. 337-350.
64. Beauchemin, M., et al., *Clinical decision support for therapeutic decision-making in cancer: A systematic review.* Int J Med Inform, 2019. **130**: p. 103940.
65. Kann, B.H., A. Hosny, and H.J.W.L. Aerts, *Artificial intelligence for clinical oncology.* Cancer Cell, 2021. **39**(7): p. 916-927.
66. Rolland, B., et al., *Toward Rigorous Data Harmonization in Cancer Epidemiology Research: One Approach.* Am J Epidemiol, 2015. **182**(12): p. 1033-8.
67. Tran, B.D., et al., *How does medical scribes' work inform development of speech-based clinical documentation technologies? A systematic review.* J Am Med Inform Assoc, 2020. **27**(5): p. 808-817.
68. Dinh-Le, C., et al., *Wearable Health Technology and Electronic Health Record Integration: Scoping Review and Future Directions.* JMIR Mhealth Uhealth, 2019. **7**(9): p. e12861.
69. Post, A.R., Z. Burningham, and A.S. Halwani, *Electronic Health Record Data in Cancer Learning Health Systems: Challenges and Opportunities.* JCO Clinical Cancer Informatics, 2022(6): p. e2100158.
70. Chen, Y., et al., *Automated medical chart review for breast cancer outcomes research: a novel natural language processing extraction system.* BMC Medical Research Methodology, 2022. **22**(1): p. 136.
71. Harpaz, R., et al., *Text Mining for Adverse Drug Events: the Promise, Challenges, and State of the Art.* Drug Safety, 2014. **37**(10): p. 777-790.
72. Percha, B., *Modern Clinical Text Mining: A Guide and Review.* Annu Rev Biomed Data Sci, 2021. **4**: p. 165-187.
73. Zhang, L., M. Hall, and D. Bastola, *Utilizing Twitter data for analysis of chemotherapy.* Int J Med Inform, 2018. **120**: p. 92-100.
74. Sheikhalishahi, S., et al., *Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review.* JMIR Med Inform, 2019. **7**(2): p. e12239.
75. Shabani, M., *Will the European Health Data Space change data sharing rules?* Science, 2022. **375**(6587): p. 1357-1359.
76. Horgan, D., et al. *European Health Data Space—An Opportunity Now to Grasp the Future of Data-Driven Healthcare*. Healthcare, 2022. **10**, DOI: 10.3390/healthcare10091629.