



Review

<https://doi.org/10.1631/jzus.B2400285>

Real-world data and evidence: pioneering frontiers in precision oncology

Jingxin JIANG^{1,2,3*}, Weiwei PAN^{1,2,3*}, Liyang SUN^{4*}, Liwei PANG^{1,2,3}, Hailang CHEN⁵, Jian HUANG^{1,2,3}✉, Wuzhen CHEN^{1,2,5}✉

¹Department of Breast Surgery (Surgical Oncology), Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou 310009, China

²Key Laboratory of Tumor Microenvironment and Immune Therapy of Zhejiang Province, Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou 310009, China

³Cancer Institute, Zhejiang University, Hangzhou 310009, China

⁴Department of Neurosurgery, Lanxi People's Hospital, Jinhua 413045, China

⁵Department of Oncology, Lanxi People's Hospital, Jinhua 413045, China

Abstract: Real-world studies (RWSs) have emerged as a transformative force in oncology research, complementing traditional randomized controlled trials (RCTs) by providing comprehensive insights into cancer care within routine clinical settings. This review examines the evolving landscape of RWSs in oncology, focusing on their implementation, methodological considerations, and impact on precision medicine. We systematically analyze how RWSs leverage diverse data sources - including electronic health records (EHRs), insurance claims, and patient registries - to generate evidence that bridges the gap between controlled clinical trials and real-world clinical practice. The review underscores the key contributions of RWSs, including capturing therapeutic outcomes in traditionally underrepresented populations, expanding drug indications, and evaluating long-term safety and effectiveness in routine clinical settings. While acknowledging significant challenges, including data-quality variability and privacy concerns, we discuss how emerging technologies like artificial intelligence are helping to address these limitations. Integration of RWSs with traditional clinical research is revolutionizing the paradigm of precision oncology and enabling more personalized treatment approaches based on real-world evidence.

Key words: Real-world study (RWS); Precision oncology; Real-world data (RWD); Study design; Data characterization

1 Introduction

The landscape of oncology research has undergone a paradigm shift with the emergence of real-world studies (RWSs), which have become an indispensable complement to traditional randomized controlled trials (RCTs). This evolution reflects the growing recognition that while RCTs remain the gold standard for establishing treatment efficacy, they may not fully capture the complexities and heterogeneity of cancer care in routine clinical practice (Berger et al., 2017).

Real-world data (RWD) encompasses a broad spectrum of health-related information generated beyond routine clinical trials, including electronic health records (EHRs), medical and pharmacy claims, disease

✉ Wuzhen CHEN, hjys@zju.edu.cn

✉ Jian HUANG, chenwuzhen@zju.edu.cn

* These authors contributed equally to this work

Jingxin JIANG, <https://orcid.org/0000-0001-7415-158X>

Wuzhen CHEN <https://orcid.org/0000-0003-0210-6645>

Jian HUANG, <https://orcid.org/0000-0003-3340-5007>

registries, and patient-generated data from wearable devices and social media platforms. The integration of these diverse data sources, particularly when combined with multi-omics data, offers unprecedented opportunities to understand cancer biology, treatment responses, and patient outcomes in real-world settings. This comprehensive approach has proven invaluable in supporting clinical drug development, regulatory decision-making, and the advancement of precision oncology (Gross et al., 2023; Verkerk and Voest, 2024). High-quality, purpose-built RWSs have emerged as crucial tools for expanding drug and medical-device indications, monitoring post-market safety, and validating clinical trial outcomes in broader patient populations.

Although there are inherent challenges in implementing RWSs - including data quality assurance, methodological rigor, and analytical complexity - this review provides a comprehensive analysis of their application in oncology by examining the characteristics and quality requirements of data sources, essential design elements, current challenges, and emerging solutions. We aim to illustrate how the integration of RWSs with traditional clinical research is transforming our understanding of cancer biology and treatment, while providing a framework for future developments in precision oncology.

2 Real-world studies in oncology

2.1 Insights into disease epidemiology and burden

RWSs that utilize comprehensive registry systems have revolutionized our understanding of cancer epidemiology and disease burden (Global Burden of Disease Cancer Collaboration et al., 2019; Global Burden of Disease 2019 Cancer Collaboration et al., 2022). Leading administrative cancer registries, notably the Surveillance, Epidemiology, and End Results (SEER) program, systematically collect and analyze data on cancer incidence, treatment patterns, and outcomes across diverse populations. These registries are particularly valuable for studying demographic groups traditionally underrepresented in RCTs, including elderly patients, those with multiple comorbidities, and vulnerable populations, enabling more targeted and effective public health interventions.

The integration of population-based cancer registries, exemplified by the United States (U.S.) National Cancer Database, has further enhanced our understanding by linking epidemiological data with detailed individual-level information (Roshandel et al., 2023). These registries capture comprehensive patient profiles, including sociodemographic factors, lifestyle behaviors, and genomic characteristics, while connecting them to treatment outcomes and insurance claims. This integration provides invaluable insights into disease burden, prognosis, and survival patterns. A landmark example of the impact of RWD is the identification of hormone-replacement therapy as a breast-cancer risk factor, a finding initially observed through observational studies and subsequently validated by RCTs (Collaborative Group on Hormonal Factors in Breast C, 1997; Chlebowski et al., 2003). This illustrates the complementary relationship between real-world observation and controlled clinical research.

The impact of registry-based studies extends beyond clinical insights to inform healthcare policy and resource allocation. Global Burden of Disease (GBD) studies have been particularly influential in shaping public-health strategies, especially in cancer screening and prevention. Notable successes include the implementation of cost-effective human papillomavirus (HPV) vaccination programs and systematic screening protocols, which have made cervical cancer elimination a realistic goal in many regions (Global Burden of Disease Cancer Collaboration et al., 2019; Voelker, 2023). Similarly, evidence-based mammography screening programs have significantly reduced breast-cancer mortality in developed nations (Farkas and Nattinger, 2023). These achievements have catalyzed the establishment of national and regional cancer-control initiatives which extend to low-income countries (Newman, 2022), demonstrating how real-world evidence (RWE) can drive meaningful improvements in global cancer care.

2.2 Assessment of treatment modalities and outcomes through real-world studies

RWSs provide critical insights into the effectiveness and accessibility of oncological treatments in routine clinical practice, often revealing significant variations from outcomes observed in RCTs (Moore et al., 2019; Jazieh et al., 2021; Chakiryan et al., 2021; Lin et al., 2021; Samlowski et al., 2023; Yan et al., 2023; Martin et al., 2023). These studies are particularly valuable in evaluating treatment efficacy, safety profiles, and patient-reported outcomes (PROs) among populations typically underrepresented in clinical trials, including patients with poor performance status, advanced age, multiple malignancies, or rare cancers. A landmark example is the expansion of palbociclib indications to include men with hormone-receptor (HR)-positive, human epidermal growth factor receptor 2 (HER2)-negative metastatic breast cancer, approved based on RWD from electronic health records, insurance claims, and global safety databases (Wedam et al., 2020).

RWSs offer unique advantages in scenarios where traditional prospective trials may be impractical or ethically challenging. The "trial effect" or "Hawthorne effect" (McCarney et al., 2007) observed in RCTs often yields outcomes that differ from real-world clinical experience. By leveraging diverse data sources, RWSs provide deeper insights into tumor pathology and treatment impacts on patient quality of life (Penberthy et al., 2022; Tang et al., 2023).

Comparative analyses of RWSs and RCTs findings have revealed important disparities that inform clinical practice improvements. For example, RWSs have shown that 40-50% of younger women with early-stage breast cancer experience interruptions in endocrine therapy, including ovarian suppression - a rate substantially higher than that reported in RCTs (Cluze et al., 2012). Similarly, investigations into trastuzumab emtansine (T-DM1) effectiveness in stage IV HER2-positive breast cancer have demonstrated reduced benefits in second-line treatment compared to pertuzumab-naïve patients (Ethier et al., 2021). The study found that only 50.1% of patients received T-DM1 as second-line therapy, with median overall survival of 12-19 months, considerably shorter than the 30 months reported in RCTs (Diéras et al., 2017). These findings underscore the limitations of generalizing RCTs results to broader clinical settings.

Recent advances in artificial intelligence (AI) have enhanced RWSs capabilities through the integration and analysis of complex multi-omics datasets, including genomics, pathomics, and radiomics (Stein-O'Brien et al., 2023). These AI-driven approaches efficiently process censored data and complex interactions, enabling the identification of novel prognostic markers (Prelaj et al., 2024). However, while promising, these findings require validation through well-designed prospective trials before implementation in clinical practice, highlighting the complementary relationship between RWSs and traditional clinical research.

2.3 Economic evaluation of cancer treatments through real-world studies

RWSs have emerged as essential tools for evaluating the economic impact of cancer treatments because they offer more accurate assessments than traditional economic forecasting models based on idealized assumptions. By analyzing comprehensive data from actual clinical settings, RWSs provide crucial insights into treatment costs, patient economic burden, and comparative cost-effectiveness across different therapeutic strategies (Dai et al., 2022; Pollard et al., 2022; Sinha et al., 2022; Soliman et al., 2023; van den Puttelaar et al., 2023). These evaluations are particularly valuable in understanding the financial implications of complex treatment decisions, including drug sequencing and dosing protocols, which enables healthcare systems to optimize resource allocation while maintaining quality care.

A compelling example of the economic impact of RWSs is seen in the evaluation of immune-checkpoint inhibitors - breakthrough therapies that have revolutionized cancer treatment but pose significant financial challenges to healthcare systems worldwide. Initial RWE suggested that lower doses of anti-programmed cell death protein 1 (PD-1) therapies (nivolumab and pembrolizumab) could achieve comparable efficacy to standard dosing (Malmberg et al., 2022). This observation led to a pivotal randomized trial examining the combination of low-dose nivolumab with triple metronomic chemotherapy in head and neck cancer patients (Patil et al., 2023). The positive results from this study demonstrated that strategic dose optimization could

maintain therapeutic efficacy while substantially reducing financial toxicity, thus establishing an alternative standard of care for patients with limited access to full-dose treatments. This exemplifies how RWSs-driven insights can inform cost-effective treatment strategies without compromising clinical outcomes.

3 Real-world data in oncology

3.1 Characteristics of RWD in oncology

The Food and Drug Administration (FDA) defines RWD as "data consistently accumulated from diverse sources pertinent to a patient's health status and/or the provision of medical services" (United States Food and Drug Administration, 2018). While RCTs generate data in highly controlled environments with selected populations, RWD encompasses information from routine clinical practice, including preventive care, diagnostics, treatment outcomes, health examinations, insurance claims, and patient registries (Booth et al., 2019; Jarow et al., 2017; Sherman et al., 2016). This fundamental difference enables RWD to complement RCTs findings by providing comprehensive insights into treatment effectiveness and safety across diverse patient populations (Fig. 1), ultimately enhancing clinical and regulatory decision-making (Gross et al., 2023; Merola et al., 2023). To better understand the complementary relationship between RWSs and RCTs, a comprehensive comparison of their respective strengths and limitations is essential (Table 1). This framework highlights how RWSs offer advantages in terms of generalizability and real-world applicability, while RCTs excel in internal validity and controlled assessment, underscoring the value of integrating both approaches in oncological research.

RCTs often exclude key patient populations encountered in routine practice, such as elderly patients, those with poor performance status, or individuals with significant comorbidities. RWD is particularly valuable for studying rare cancers, specific molecular subtypes, and complex patient groups typically excluded from traditional research settings. Post-marketing studies utilizing RWD are therefore essential for validating trial findings in real-world settings and understanding therapeutic effectiveness across diverse patient populations (Black, 1996; Kennedy-Martin et al., 2015; Skovlund et al., 2018). Despite its advantages, RWD faces significant challenges in completeness and accuracy (Cook and Collins, 2015). The integration and management of data from multiple sources requires extensive standardization and verification processes. Methodological limitations include population representativeness bias, incomplete data in registries, and reliability concerns with certain data sources (Booth et al., 2019; Tsai et al., 2019; Collins et al., 2020; Velmovitsky et al., 2021). Additionally, establishing causal relationships and conducting safety evaluations present unique analytical challenges (Flynn et al., 2022; Zhu et al., 2023). These limitations suggest that RWD should complement rather than replace RCTs in clinical decision-making, particularly in safety-critical areas.

3.2 Applying RWD in real-world precision oncology research

3.2.1 Enhancing screening programs

RWD has revolutionized cancer-screening programs by enabling comprehensive population-level analyses that integrate diverse data sources, including EHRs, cancer registries, and multidimensional clinical-genetic features. For example, the CanPredict model for lung-cancer risk assessment (Liao et al., 2023) was developed using multiple data dimensions: sociodemographic factors (age, sex, ethnicity, etc.), lifestyle characteristics (body mass index (BMI), smoking status, alcohol consumption, etc.), and comprehensive medical history (comorbidities, previous medical history, and family history). By analyzing data from 19.67 million individuals across two English primary care databases, this model has emerged as a robust tool for identifying high-risk individuals who would benefit most from targeted lung-cancer screening.

The impact of RWD on screening programs extends beyond risk stratification. A landmark study by Trentham-Dietz et al. (Trentham-Dietz et al., 2024) demonstrated this potential by integrating six Cancer Intervention and Surveillance Modeling Network (CISNET) models with national breast-cancer data. Their

analysis, which showed significant reductions in breast-cancer mortality with mammography screening, provided compelling evidence to support the initiation of biennial screening at age 40. RWD has also proven valuable in uncovering disparities in screening adherence, as demonstrated by a U.S. national lung-cancer screening program analysis (Núñez et al., 2021), which highlighted significant gaps - particularly among low-income individuals, Black populations, and those with mental health conditions - thereby informing targeted interventions to improve screening accessibility and utilization.

3.2.2 Advancing diagnostic accuracy and treatment optimization

RWD has transformed diagnostic practices by providing comprehensive insights into disease presentation and progression across diverse populations, transcending the traditional "symptom-imaging-pathology" paradigm. AI-assisted tools have transformed radiological and pathological image interpretation by offering remarkable accuracy in tasks such as image classification, reconstruction, detection, segmentation, registration, and synthesis (Kleppe et al., 2021; Chen et al., 2022; Jiang et al., 2023). Platforms like QuPath (Bankhead et al., 2017) exemplify this advancement by integrating deep learning and machine-learning capabilities; this enables automated tumor detection and classification while supporting customized models for specific cellular identification and biomarker quantification.

The integration of big data has catalyzed a paradigm shift in oncological treatment strategies, from uniform histology-based approaches toward genomic-guided precision oncology. This evolution demands robust methodologies for evaluating novel patient subgroups, assessing real-world drug efficacy, and validating new biomarkers (Terranova and Venkatakrishnan, 2024; Verkerk and Voest, 2024). The discrepancy between clinical trial outcomes and real-world treatment patterns highlights the importance of RWD in optimizing therapeutic strategies. A compelling example is the treatment of older women (aged >75) with ductal carcinoma in situ (DCIS), a population traditionally excluded from clinical trials. Analysis of real-world treatment patterns revealed that while these patients are not necessarily frail, they often receive de-escalated interventions, potentially increasing the risk of disease progression, particularly in estrogen receptor-negative DCIS (Karakatsanis and Markopoulos, 2020). This finding emphasizes the need for individualized treatment approaches that consider not only tumor characteristics but also patient-specific factors such as life expectancy, performance status, and personal preferences, with the goal of optimizing oncological outcomes while preserving quality of life.

3.2.3 Accelerating drug development

The integration of RWD into drug-approval processes has significantly enhanced regulatory decision-making in oncology. RWE provides crucial insights into post-approval drug performance and safety while enabling expedited effectiveness assessments, particularly for treatments that address urgent health needs or rare diseases (Arondekar et al., 2022). Major regulatory agencies, including the FDA, European Medicines Agency (EMA), and China's National Medical Products Administration (NMPA), have formally recognized the value of RWD in drug-approval frameworks (Arlett et al., 2022; Xu et al., 2024). Blinatumomab's accelerated approval in 2014 exemplifies the successful integration of RWE in drug development. Its approval for treating adults with poor-prognosis Philadelphia chromosome-negative relapsed/refractory B-cell precursor acute lymphoblastic leukemia was based on a single-arm Phase II study of 189 patients (Przepiorka et al. 2015), supported by RWE from a historical control group of 694 patients (Topp et al., 2015). A subsequent Phase III RCT confirmed blinatumomab's survival benefit over standard chemotherapy (Kantarjian et al., 2017), validating the initial RWE-supported approval.

3.2.4 Applying RWD/RWSs in breast-cancer research

RWD has especially advanced breast-cancer research across multiple domains, including screening and diagnosis patterns (Duggento et al., 2021; Zhang et al., 2023; Gennaro et al., 2024; Trentham-Dietz et al., 2024), quality-of-life assessments (Kirkham et al., 2019; Timmins et al., 2024), treatment outcomes (Agiro et al., 2018;

Veitch et al., 2019; Anwar et al., 2021; Lin et al., 2021; Dai et al., 2022; Wang et al., 2022; Wu et al., 2023), prognostic predictions (Stabellini et al., 2023), and cost-effectiveness analyses (Dinan et al., 2019). The Canadian Rethinking Clinical Trials (REaCT) program demonstrates the efficiency of RWE in addressing recruitment challenges; it results in faster enrollment and improved cost-effectiveness compared to traditional RCTs (Basulaiman et al., 2019). A recent example is the expedited approval of goserelin (10.8 mg) for breast-cancer treatment, based on the Ezreal study presented at the 2024 European Society for Medical Oncology (ESMO) Congress (Wang et al., 2024). This nationwide multicenter observational case-control study screened 15,629 patients across 16 Chinese hospitals, enrolling 1,060 eligible participants. After propensity-score matching, analysis of 590 patients (295 per group) compared the estradiol-suppressing effects of 10.8 mg versus 3.6 mg goserelin in pre- and perimenopausal HR-positive breast-cancer patients, demonstrating non-inferiority of the higher dose formulation.

4 Design essentials for real-world studies

The exponential growth of RWD applications in oncology research, marked by a seven-fold increase in publications over the past decade (Malone et al., 2018), highlights the critical need for robust methodological frameworks in conducting RWSs. While RWD offers unprecedented opportunities to bridge the gap between scientific discovery and clinical practice, the variable quality of studies underscores the importance of standardized approaches. Success in RWSs depends fundamentally on data accessibility, transparency, and methodological rigor. Essential elements include protocol pre-registration, adherence to data-privacy regulations, and maintenance of high research standards. Given the inherent complexities of RWD, generating high-quality RWE requires well-defined research questions, appropriate data selection, and robust study design. This systematic approach is crucial for identifying disparities between clinical evidence and practice, and ultimately advancing cancer care and improving patient outcomes (Franklin and Schneeweiss, 2017; Khozin et al., 2017; Miksad and Abernethy, 2018; Tang et al., 2023).

4.1 Articulating research questions using the PICOTS framework

When a term appears for the first time in the Abstract or in the formal text, it should be expanded, with its abbreviation in parentheses following immediately: RWSs serve three primary functions: evaluating intervention impacts, characterizing disease patterns in specific populations, and developing predictive models. The foundation of effective RWSs lies in well-defined research questions structured through the PICOTS framework: Population, Intervention, Comparison, Outcomes, Time, and Setting (U.S. Food and Drug Administration, 2023). This systematic approach ensures comprehensive consideration of all critical study elements, which facilitates proper study design and execution (Fig. 2).

Research questions in RWSs must be both scientifically meaningful and pragmatically answerable using available RWD. The focus should be on questions suited to observational studies rather than those requiring randomized controlled trials, such as establishment of initial treatment efficacy (Baumfeld Andre et al., 2020). Successful RWSs design demands clear definition of the target population and setting, precise specification of interventions or exposures, selection of appropriate comparison groups, and establishment of well-defined, measurable outcomes within relevant time frames. Careful consideration must be given to the availability and quality of data sources that can support these research objectives.

4.2 Identifying the appropriate data source

Selection of optimal data sources is fundamental to the success of RWS. This process requires comprehensive understanding of various data-source characteristics, including population representativeness, variable scope, and data architecture (Table 2). Researchers must ensure that selected datasets not only align with their research questions but also adhere to ethical guidelines, while providing sufficient coverage of

target populations and necessary variables (Miksad and Abernethy, 2018). Different research objectives demand different data sources. For instance, cancer-burden assessments should primarily utilize comprehensive national or global epidemiological registries rather than EHRs. However, EHRs play a crucial role in providing detailed clinical information, including patient demographics, diagnoses, laboratory results, and treatment data. The integration of such longitudinal data enables thorough investigation of disease characteristics, treatment patterns, and outcomes in routine clinical practice (Penberthy et al., 2022).

The complexity of RWSs often necessitates integration of multiple data sources to achieve comprehensive analysis. Successful examples include drug-safety research, in which combining EHRs with national adverse-event monitoring systems provides complementary insights into both adverse-event patterns and their contributing factors (Booth et al., 2019; Kim, 2024). Similarly, oncology survival analyses benefit from supplementing traditional follow-up data with household registration records to minimize patient attrition. However, this integration process presents significant challenges, particularly in maintaining patient privacy while ensuring reliable cross-dataset identification. While AI-based solutions offer promising approaches for data integration, their implementation requires careful evaluation of accuracy and completeness (Kovačević et al., 2024), and balancing data integration with privacy protection remains a central consideration in RWSs methodology.

4.3 Ensuring data-collection quality

RWSs require rigorous standardization of data extraction and processing protocols to maintain scientific credibility. Adherence to established reporting guidelines and transparent documentation of methodology are essential for ensuring result traceability and reproducibility (Wang et al., 2021). This transparency extends to acknowledging data-handling limitations and their potential impact on study outcomes, thereby minimizing the risk of the results being misinterpreted. Quality assurance in RWSs demands comprehensive understanding of data-generation processes. This includes systematic assessment of variable completeness, identification of missing or miscoded data, and regular quality verification through sampling. The implementation of standardized data-collection methods helps minimize quality discrepancies at the source. Notable initiatives such as the Observational Health Data Sciences and Informatics (OHDSI) Common Data Model (Hripcsak et al., 2015) and the Fast Healthcare Interoperability Resources (FHIR) standard (Vorisek et al., 2022) have emerged to address these challenges by facilitating data harmonization across disparate sources, though careful consideration must be given to maintaining the integrity of underlying pathophysiological relationships.

4.4 Confounding factors: identification and control

RWSs face inherent challenges in terms of controlling confounding variables. To enhance study validity, researchers must implement robust methodological approaches, including propensity-score matching and careful study design. Cross-validation with independent data sources, such as clinical trials or national registries, further strengthens the credibility of findings and ensures appropriate comparison between treatment groups.

Pragmatic clinical trials (PCTs) represent an innovative solution that bridges the gap between traditional RCTs and RWSs (Derksen et al., 2019; Franklin and Schneeweiss, 2017). In addition to maintaining randomization, PCTs accommodate broader patient populations and implement streamlined processes, including simplified consent procedures and flexible data-collection methods. This approach not only reduces participant burden but also facilitates long-term follow-up data collection. The above-mentioned REaCT program exemplifies successful PCT implementation in oncology (Basulaiman et al., 2019). By employing simplified eligibility criteria, streamlined consent processes, and efficient data collection methods, REaCT achieved remarkable success, enrolling over 2,100 patients across 11 cancer centers in 4 years.

5 Application of AI in RWD analysis

AI has emerged as a transformative tool in oncology research, enabling sophisticated pattern recognition and prediction from complex real-world datasets (Perez-Lopez et al., 2024). The integration of AI with RWD encompasses multiple complementary approaches, each addressing specific analytical challenges in cancer research.

5.1 Machine-learning algorithms

Traditional machine-learning methods, such as random forests and gradient boosting machines (GBM), provide robust frameworks for predictive modeling and patient classification (Poirion et al., 2021). Deep learning approaches have shown promise, with convolutional neural networks (CNNs) excelling in medical image analysis and pattern recognition, while recurrent neural networks (RNNs) effectively handle sequential and temporal data. For example, CNNs have demonstrated superior performance in analyzing medical imaging data, achieving results comparable to or exceeding expert radiologists in certain diagnostic tasks (Li et al., 2019).

5.2 Natural language processing (NLP)

NLP has revolutionized the extraction of insights from unstructured clinical data (Yim et al., 2016). Through algorithms such as Latent Dirichlet Allocation and TF-IDF, NLP enables automated analysis of clinical notes, pathology reports, and patient feedback. This capability extends to analyzing PROs and social media data, providing valuable insights into treatment satisfaction and patient experiences (Lu et al., 2021; Sim et al., 2023).

5.3 Reinforcement learning (RL)

RL represents a promising frontier in treatment optimization (Xu et al., 2023). By learning from patient interactions and outcomes, RL algorithms can simulate and identify optimal treatment pathways. This approach has successfully informed personalized cancer-screening strategies and diagnosis (Qaiser and Rajpoot, 2019; Yala et al., 2022), treatment-regimen design (Lu et al., 2024; Tortora et al., 2021), and clinical trial optimization (Zhao et al., 2009).

5.4 Challenges and considerations

Despite the potential benefits, the effectiveness of AI applications depends critically on data quality and representativeness. The "black box" nature of many AI models presents challenges for clinical interpretation, driving ongoing development of more transparent approaches, such as the SHapley Additive exPlanations (SHAP) analysis (Hsu et al., 2023). These challenges underscore the importance of combining AI capabilities with robust clinical validation and interpretability frameworks.

6 Challenges and advancements of real-world studies in oncology

At the revision stage, authors who have created their files using a drawing or painting program such as Visio, Origin, Excel, AutoCAD, While RWD offers unprecedented advantages in terms of scale, efficiency, and cost-effectiveness, it also faces significant challenges with regard to data quality, heterogeneity, and inherent biases. The diverse nature of RWD, which is collected from routine clinical practice, presents complex challenges including missing values, coding inconsistencies, and the integration of disparate "data islands". Standardization of data-collection, management, and analysis processes is therefore key for oncological research, and this will require sophisticated data-cleansing algorithms and robust research guidelines.

The management of sensitive health information poses additional ethical challenges, demanding stringent

protocols for data security, access control, and privacy protection. These considerations necessitate the development of standardized ethical review procedures and strict data-anonymization protocols to safeguard patient confidentiality while maintaining data utility for research purposes.

To address these challenges, the oncology research community is developing innovative solutions and collaborative frameworks. The Cancer Core Europe (CCE) initiative, comprising seven leading European cancer centers, exemplifies this approach through its establishment of a virtual data hub linking molecular profiles with clinical outcomes while addressing legal and privacy concerns (Eggermont et al., 2019). Such initiatives highlight the importance of multi-institutional collaboration in establishing unified data standards and enhancing the representativeness of research findings.

7 Conclusions

RWSs are bringing about a paradigm shift in oncological research by providing detailed information about patient demographics, long-term outcomes, and treatment safety. The potential of RWSs is only becoming more substantial as regulatory frameworks change and technology advances. To fully realize this potential, it is important to have standardized protocols for data collection, analysis, and interpretation. Strict research standards and ethical guidelines will ensure that the medical community can use RWSs to advance precision oncology, thus improving patient outcomes.

Data availability statement

No new data was created or analyzed in this study.

Acknowledgments

This work is supported by the Natural Science Foundation of Zhejiang Province (No. ZCLY24H1601), the Natural Science Foundation of China (No. 82403697), and the China's National Key R&D Program (NO. 2022YFC2505100). We thank Dr. Chao NI, Dr. Xuan SHAO, Dr. Zhigang CHEN, Dr. Ke WANG, Dr. Jun YE, and Dr. Pin WU (All from the Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China) for their invaluable advice and insights during the preparation of this manuscript. Their expertise significantly contributed to the completion of this work.

Author contributions

Jiangxin JIANG led the design and conceptualization of the review, drafted significant sections of the manuscript, and was responsible for the creation of the figures to visually represent key concepts. Wuzhen CHEN assisted in the conceptualization and contributed to drafting and refining the manuscript. Weiwei PAN, Liyang SUN, Liwei PANG and Hailang CHEN were instrumental in collecting and analyzing the literature, and they participated actively in revising the manuscript. Jian HUANG supervised the overall project and focused on critically revising the manuscript for important intellectual content. All authors reviewed, edited, and approved the final manuscript, ensuring the accuracy and integrity of the work.

Compliance with ethics guidelines

Jingxin JIANG, Weiwei PAN, Liyang SUN, Liwei PANG, Hailang CHEN, Jian HUANG and Wuzhen CHEN declare that they have no conflict of interest. Given that this research did not engage human participants, animal subjects, or sensitive data, it was exempt from the ethical approval process.

Table 1 Characteristics of Real-World Studies (RWSs) and Randomized Controlled Trials (RCTs)

Characteristic	RWSs	RCTs
Study Design	Observational studies using routine clinical data; Multiple design options	Controlled experiments with randomization; Standardized protocols
Population	Broad, diverse patient populations	Selected participants meeting strict criteria; Excluding comorbidities
Data Quality	Variable quality; Potential bias and confounding	High quality; Minimized bias through randomization
Outcomes	Long-term effectiveness; Patient-reported outcomes; Real-world safety	Specific endpoints; Efficacy under ideal conditions; Short-term focus
Implementation	Cost-effective; Faster recruitment; Applying existing data	Resource-intensive; Lengthy recruitment; Complex logistics
Strengths	Real-world applicability; Population representativeness; Long-term insights	Strong causal inference; Internal validity; Standardized evidence
Limitations	Data quality concerns; Bias control challenges; Privacy issues	Limited generalizability; High cost; Artificial setting
Regulatory Role	Complementary evidence; Growing acceptance	Primary evidence; Established standard

Table 2 Characterization of Different Data Sources in Real-World Studies

Data Source	Volume & Format	Accessibility & Quality	Primary Applications	Major Limitations
Clinical Sources				
Electronic Health Records	Large; Unstructured	Medium; Variable Quality	<ul style="list-style-type: none"> Clinical outcomes Treatment patterns Patient history 	<ul style="list-style-type: none"> Data inconsistency Integration challenges
Disease Registries	Medium; Semi-structured	Medium; High Quality	<ul style="list-style-type: none"> Disease-specific outcomes Population studies Long-term follow-up 	<ul style="list-style-type: none"> Limited scope Selection bias
Administrative Sources				
Insurance Claims	Large; Structured	Low; Standardized	<ul style="list-style-type: none"> Resource utilization Cost analysis Treatment patterns 	<ul style="list-style-type: none"> Limited clinical detail Privacy restrictions
Patient-Generated Data				
Wearable Devices & Apps	Small; Continuous	High; Variable	<ul style="list-style-type: none"> Real-time monitoring Patient behavior Quality of life Patient experience 	<ul style="list-style-type: none"> Small samples Validation needs
Social Media	Medium; Unstructured	High; Low Quality	<ul style="list-style-type: none"> Treatment satisfaction Side effects 	<ul style="list-style-type: none"> Reliability issues Verification challenges
Regulatory Data				
Safety Monitoring Systems	Medium; Structured	Low; High Quality	<ul style="list-style-type: none"> Adverse event Safety signals Post-market surveillance 	<ul style="list-style-type: none"> Reporting delays Incomplete capture

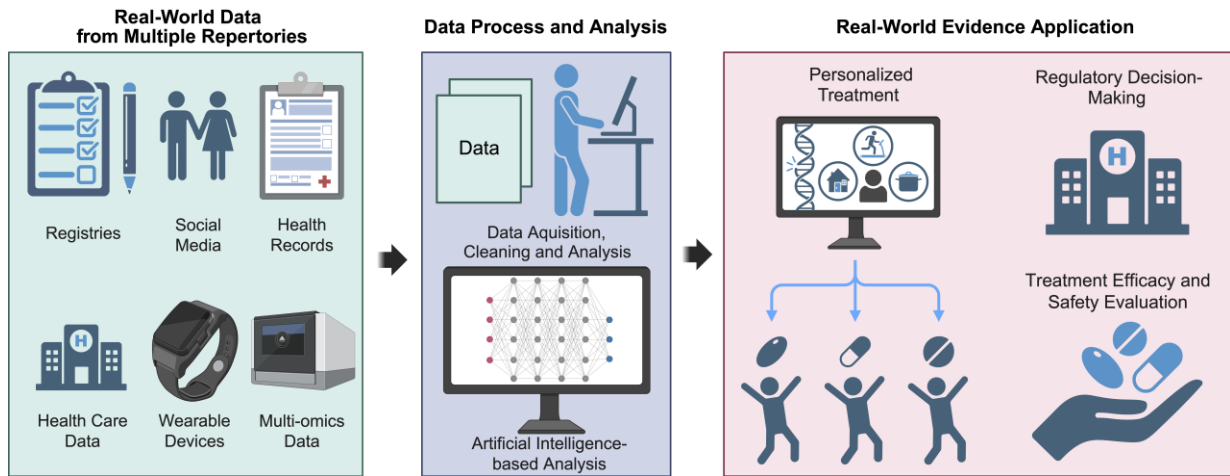


Fig. 1 Process of Real-World Data to Real-World Evidence (Created with Biorender.com)

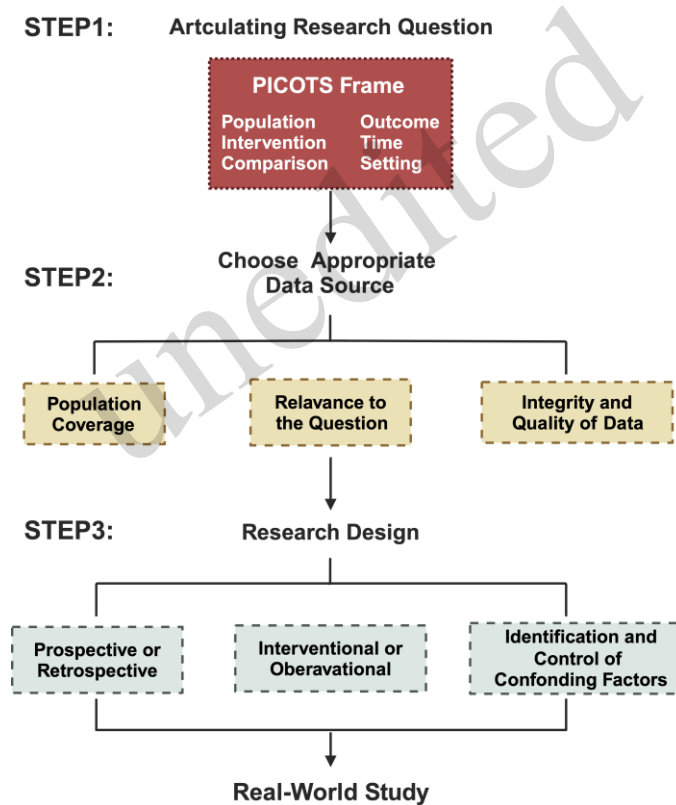


Fig. 2 Design Essentials for Real-World Study (Created with Biorender.com)

References

Berger M, Daniel G, Frank K, et al., 2017. A Framework for Regulatory Use of Real-World Evidence. Available from <https://www.researchgate.net/publication/319650416>. [Accessed on Nov. 25, 2024]
 Gross AJ, Pisano CE, Khunsriraksakul C, et al., 2023. Real-World Data: Applications and Relevance to Cancer Clinical Trials. *Semin Radiat Oncol*, 33(4):374-385.
<http://doi.org/10.1016/j.semradonc.2023.06.003>

- Verkerk K, Voest EE, 2024. Generating and using real-world data: A worthwhile uphill battle. *Cell*, 187(7):1636-1650.
<http://doi.org/10.1016/j.cell.2024.02.012>
- Global Burden of Disease Cancer Collaboration, 2019. Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-Years for 29 Cancer Groups, 1990 to 2017: A Systematic Analysis for the Global Burden of Disease Study. *JAMA Oncol*, 5(12):1749-1768.
<http://doi.org/10.1001/jamaoncol.2019.2996>
- Global Burden of Disease Cancer Collaboration, 2022. Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life Years for 29 Cancer Groups From 2010 to 2019: A Systematic Analysis for the Global Burden of Disease Study 2019. *JAMA Oncol*, 8(3):420-444.
<http://doi.org/10.1001/jamaoncol.2021.6987>
- Roshandel G, Badar F, Barchuk A, et al., 2023. REPCAN: Guideline for REporting Population-based CANcer Registry Data. *Asian Pac J Cancer Prev APJCP*, 24(9):3297-3303.
<http://doi.org/10.31557/APJCP.2023.24.9.3297>
- Chlebowski RT, Hendrix SL, Langer RD, et al., 2003. Influence of estrogen plus progestin on breast cancer and mammography in healthy postmenopausal women: the Women's Health Initiative Randomized Trial. *JAMA*, 289(24):3243-3253.
<http://doi.org/10.1001/jama.289.24.3243>
- Collaborative Group on Hormonal Factors in Breast Cancer, 1997. Breast cancer and hormone replacement therapy: collaborative reanalysis of data from 51 epidemiological studies of 52,705 women with breast cancer and 108,411 women without breast cancer. *Lancet Lond Engl*, 350(9084):1047-1059.
- Voelker RA, 2023. Cervical Cancer Screening. *JAMA*, 330(20):2030.
<http://doi.org/10.1001/jama.2023.21987>
- Farkas AH, Nattinger AB, 2023. Breast Cancer Screening and Prevention. *Ann Intern Med*, 176(11):ITC161-ITC176.
<http://doi.org/10.7326/AITC202311210>
- Newman LA, 2022. Breast cancer screening in low and middle-income countries. *Best Pract Res Clin Obstet Gynaecol*, 83:15-23.
<http://doi.org/10.1016/j.bpobgyn.2022.03.018>
- Moore S, Leung B, Wu J, Ho C, 2019. Real-World Treatment of Stage III NSCLC: The Role of Trimodality Treatment in the Era of Immunotherapy. *J Thorac Oncol*, 14(8):1430-1439.
<http://doi.org/10.1016/j.jtho.2019.04.005>
- Jazieh AR, Onal HC, Tan DSW, et al., 2021. Real-World Treatment Patterns and Clinical Outcomes in Patients With Stage III NSCLC: Results of KINDLE, a Multicountry Observational Study. *J Thorac Oncol*, 16(10):1733-1744.
<http://doi.org/10.1016/j.jtho.2021.05.003>
- Chakiryan NH, Jiang DD, Gillis KA, et al., 2021. Real-World Survival Outcomes Associated With First-Line Immunotherapy, Targeted Therapy, and Combination Therapy for Metastatic Clear Cell Renal Cell Carcinoma. *JAMA Netw Open*, 4(5):e2111329.
<http://doi.org/10.1001/jamanetworkopen.2021.11329>
- Lin AP, Huang TW, Tam KW, 2021. Treatment of male breast cancer: meta-analysis of real-world evidence. *Br J Surg*, 108(9):1034-1042.
<http://doi.org/10.1093/bjs/znab279>
- Samlowski W, Robert NJ, Chen L, et al., 2023. Real-World nivolumab dosing patterns and safety outcomes in patients receiving adjuvant therapy for melanoma. *Cancer Med*, 12(3):2378-2388.
<http://doi.org/10.1002/cam4.5061>
- Yan Y, Lv R, Wang T, et al., 2023. Real-world treatment patterns, discontinuation and clinical outcomes in patients with B-cell lymphoproliferative diseases treated with BTK inhibitors in China. *Front Immunol*, 14:1184395.
<http://doi.org/10.3389/fimmu.2023.1184395>
- Martin P, Cohen JB, Wang M, et al., 2023. Treatment Outcomes and Roles of Transplantation and Maintenance Rituximab in Patients With Previously Untreated Mantle Cell Lymphoma: Results From Large Real-World Cohorts. *J Clin Oncol*, 41(3):541-554.
<http://doi.org/10.1200/JCO.21.02698>
- Wedam S, Fashoyin-Aje L, Bloomquist E, et al., 2020. FDA Approval Summary: Palbociclib for Male Patients with Metastatic Breast Cancer. *Clin Cancer Res*, 26(6):1208-1212.
<http://doi.org/10.1158/1078-0432.CCR-19-2580>
- McCarney R, Warner J, Iliffe S, et al., 2007. The Hawthorne Effect: a randomised, controlled trial. *BMC Med Res Methodol*, 7:30.
<http://doi.org/10.1186/1471-2288-7-30>
- Penberthy LT, Rivera DR, Lund JL, et al., 2022. An overview of real-world data sources for oncology and considerations for

- research. *CA Cancer J Clin*, 72(3):287-300.
<http://doi.org/10.3322/caac.21714>
- Tang M, Pearson SA, Simes RJ, Chua BH, 2023. Harnessing Real-World Evidence to Advance Cancer Research. *Curr Oncol*, 30(2):1844-1859.
<http://doi.org/10.3390/curroncol30020143>
- Cluze C, Rey D, Huiart L, et al., 2012. Adjuvant endocrine therapy with tamoxifen in young women with breast cancer: determinants of interruptions vary over time. *Ann Oncol*, 23(4):882-890.
<http://doi.org/10.1093/annonc/mdr330>
- Ethier JL, Desautels D, Robinson A, et al., 2021. Practice Patterns and Outcomes of Novel Targeted Agents for the Treatment of ERBB2-Positive Metastatic Breast Cancer. *JAMA Oncol*, 7(9):e212140.
<http://doi.org/10.1001/jamaoncol.2021.2140>
- Diéras V, Miles D, Verma S, et al., 2017. Trastuzumab emtansine versus capecitabine plus lapatinib in patients with previously treated HER2-positive advanced breast cancer (EMILIA): a descriptive analysis of final overall survival results from a randomised, open-label, phase 3 trial. *Lancet Oncol*, 18(6):732-742.
[http://doi.org/10.1016/S1470-2045\(17\)30312-1](http://doi.org/10.1016/S1470-2045(17)30312-1)
- Stein-O'Brien GL, Le DT, Jaffee EM, Fertig EJ, Zaidi N, 2023. Converging on a Cure: The Roads to Predictive Immunotherapy. *Cancer Discov*, 13(5):1053-1057.
<http://doi.org/10.1158/2159-8290.CD-23-0277>
- Prelaj A, Miskovic V, Zanitti M, et al., 2024. Artificial intelligence for predictive biomarker discovery in immuno-oncology: a systematic review. *Ann Oncol*, 35(1):29-65.
<http://doi.org/10.1016/j.annonc.2023.10.125>
- Dai WF, Beca JM, Nagamuthu C, et al., 2022. Cost-effectiveness Analysis of Pertuzumab With Trastuzumab in Patients With Metastatic Breast Cancer. *JAMA Oncol*, 8(4):597-606.
<http://doi.org/10.1001/jamaoncol.2021.8049>
- Soliman R, Oke J, Sidhom I, et al., 2023. Cost-effectiveness of childhood cancer treatment in Egypt: Lessons to promote high-value care in a resource-limited setting based on real-world evidence. *EClinicalMedicine*, 55:101729.
<http://doi.org/10.1016/j.eclinm.2022.101729>
- van den Puttelaar R, Meester RGS, Peterse EFP, et al., 2023. Risk-Stratified Screening for Colorectal Cancer Using Genetic and Environmental Risk Factors: A Cost-Effectiveness Analysis Based on Real-World Data. *Clin Gastroenterol Hepatol*, 21(13):3415-3423.e29.
<http://doi.org/10.1016/j.cgh.2023.03.003>
- Pollard S, Weymann D, Chan B, et al., 2022. Defining a Core Data Set for the Economic Evaluation of Precision Oncology. *Value Health*, 25(8):1371-1380.
<http://doi.org/10.1016/j.jval.2022.01.005>
- Sinha S, Laskar SG, Wadasadawala T, et al., 2022. Adopting Health Economic Research in Radiation Oncology: A Perspective From Low- or Middle-Income Countries. *JCO Glob Oncol*, 8:e2100374.
<http://doi.org/10.1200/GO.21.00374>
- Malmberg R, Zietse M, Dumoulin DW, et al., 2022. Alternative dosing strategies for immune checkpoint inhibitors to improve cost-effectiveness: a special focus on nivolumab and pembrolizumab. *Lancet Oncol*, 23(12):e552-e561.
[http://doi.org/10.1016/S1470-2045\(22\)00554-X](http://doi.org/10.1016/S1470-2045(22)00554-X)
- Patil VM, Noronha V, Menon N, et al., 2023. Low-Dose Immunotherapy in Head and Neck Cancer: A Randomized Study. *J Clin Oncol*, 41(2):222-232. <http://doi.org/10.1200/JCO.22.01015>
- United States Food and Drug Administration, 2018. Framework for FDA's Real-World Evidence Program. Available from <https://www.fda.gov/media/123160/download>. [Accessed on Nov. 25, 2024]
- Sherman RE, Anderson SA, Dal Pan GJ, et al., 2016. Real-World Evidence - What Is It and What Can It Tell Us? *N Engl J Med*, 375(23):2293-2297.
<http://doi.org/10.1056/NEJMs1609216>
- Jarow JP, LaVange L, Woodcock J, 2017. Multidimensional Evidence Generation and FDA Regulatory Decision Making: Defining and Using "Real-World" Data. *JAMA*, 318(8):703-704.
<http://doi.org/10.1001/jama.2017.9991>
- Booth CM, Karim S, Mackillop WJ, 2019. Real-world data: towards achieving the achievable in cancer care. *Nat Rev Clin Oncol*, 16(5):312-325.
<http://doi.org/10.1038/s41571-019-0167-7>
- Merola D, Campbell U, Gautam N, et al., 2023. The Action Coalition to Advance Real-World Evidence through Randomized Controlled Trial Emulation Initiative: Oncology. *Clin Pharmacol Ther*, 113(6):1217-1222.
<http://doi.org/10.1002/cpt.2800>

- Black N, 1996. Why we need observational studies to evaluate the effectiveness of health care. *BMJ*, 312(7040):1215-1218.
<http://doi.org/10.1136/bmj.312.7040.1215>
- Kennedy-Martin T, Curtis S, Faries D, et al., 2015. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. *Trials*, 16:495.
<http://doi.org/10.1186/s13063-015-1023-4>
- Skovlund E, Leufkens HGM, Smyth JF, 2018. The use of real-world data in cancer drug development. *Eur J Cancer*, 101:69-76.
<http://doi.org/10.1016/j.ejca.2018.06.036>
- Cook JA, Collins GS, 2015. The rise of big clinical databases. *Br J Surg*, 102(2):e93-e101.
<http://doi.org/10.1002/bjs.9723>
- Tsai CJ, Riaz N, Gomez SL, 2019. Big Data in Cancer Research: Real-World Resources for Precision Oncology to Improve Cancer Care Delivery. *Semin Radiat Oncol*, 29(4):306-310.
<http://doi.org/10.1016/j.semradonc.2019.05.002>
- Collins R, Bowman L, Landray M, Peto R, 2020. The Magic of Randomization versus the Myth of Real-World Evidence. *N Engl J Med*, 382(7):674-678. <http://doi.org/10.1056/NEJMs1901642>
- Velmovitsky PE, Bevilacqua T, Alencar P, et al., 2021. Convergence of Precision Medicine and Public Health Into Precision Public Health: Toward a Big Data Perspective. *Front Public Health*, 9:561873.
<http://doi.org/10.3389/fpubh.2021.561873>
- Flynn R, Plueschke K, Quinten C, et al., 2022. Marketing Authorization Applications Made to the European Medicines Agency in 2018–2019: What was the Contribution of Real-World Evidence? *Clin Pharmacol Ther*, 111(1):90-97.
<http://doi.org/10.1002/cpt.2461>
- Zhu R, Vora B, Menon S, et al., 2023. Clinical Pharmacology Applications of Real-World Data and Real-World Evidence in Drug Development and Approval—An Industry Perspective. *Clin Pharmacol Ther*, 114(4):751-767.
<http://doi.org/10.1002/cpt.2988>
- Liao W, Coupland CAC, Burchardt J, et al., 2023. Predicting the future risk of lung cancer: development, and internal and external validation of the CanPredict (lung) model in 19-67 million people and evaluation of model performance against seven other risk prediction models. *Lancet Respir Med*, 11(8):685-697.
[http://doi.org/10.1016/S2213-2600\(23\)00050-4](http://doi.org/10.1016/S2213-2600(23)00050-4)
- Trentham-Dietz A, Chapman CH, Jayasekera J, et al., 2024. Collaborative Modeling to Compare Different Breast Cancer Screening Strategies: A Decision Analysis for the US Preventive Services Task Force. *JAMA*, 331(22):1947-1960.
<http://doi.org/10.1001/jama.2023.24766>
- Núñez ER, Caverly TJ, Zhang S, et al., 2021. Adherence to Follow-up Testing Recommendations in US Veterans Screened for Lung Cancer, 2015-2019. *JAMA Netw Open*, 4(7):e2116233.
<http://doi.org/10.1001/jamanetworkopen.2021.16233>
- Chen X, Wang X, Zhang K, et al., 2022. Recent advances and clinical applications of deep learning in medical image analysis. *Med Image Anal*, 79:102444.
<http://doi.org/10.1016/j.media.2022.102444>
- Jiang X, Hu Z, Wang S, Zhang Y, 2023. Deep Learning for Medical Image-Based Cancer Diagnosis. *Cancers*, 15(14):3608.
<http://doi.org/10.3390/cancers15143608>
- Kleppe A, Skrede OJ, De Raedt S, et al., 2021. Designing deep learning studies in cancer diagnostics. *Nat Rev Cancer*, 21(3):199-211.
<http://doi.org/10.1038/s41568-020-00327-9>
- Bankhead P, Loughrey MB, Fernández JA, et al., 2017. QuPath: Open source software for digital pathology image analysis. *Sci Rep*, 7(1):16878.
<http://doi.org/10.1038/s41598-017-17204-5>
- Terranova N, Venkatakrishnan K, 2024. Machine Learning in Modeling Disease Trajectory and Treatment Outcomes: An Emerging Enabler for Model-Informed Precision Medicine. *Clin Pharmacol Ther*, 115(4):720-726.
<http://doi.org/10.1002/cpt.3153>
- Karakatsanis A, Markopoulos C, 2020. The challenge of avoiding over- and under-treatment in older women with ductal cancer in situ: A scoping review of existing knowledge gaps and a meta-analysis of real-world practice patterns. *J Geriatr Oncol*, 11(6):917-925.
<http://doi.org/10.1016/j.jgo.2020.02.005>
- Arondekar B, Duh MS, Bhak RH, et al., 2022. Real-World Evidence in Support of Oncology Product Registration: A Systematic Review of New Drug Application and Biologics License Application Approvals from 2015-2020. *Clin Cancer Res*, 28(1):27-35.
<http://doi.org/10.1158/1078-0432.CCR-21-2639>
- Xu J, Wu W, Zhang X, et al., 2024. The Use of Real-World Evidence for Regulatory Decisions in China. *Clin Pharmacol Ther*,

- 116(1):82-95.
<http://doi.org/10.1002/cpt.3257>
- Arlett P, Kjaer J, Broich K, et al., 2022. Real-World Evidence in EU Medicines Regulation: Enabling Use and Establishing Value. *Clin Pharmacol Ther*, 111(1):21-23.
<http://doi.org/10.1002/cpt.2479>
- Przepiorka D, Ko CW, Deisseroth A, et al., 2015. FDA Approval: Blinatumomab. *Clin Cancer Res*, 21(18):4035-4039.
<http://doi.org/10.1158/1078-0432.CCR-15-0612>
- Topp MS, Gökbuget N, Stein AS, et al., 2015. Safety and activity of blinatumomab for adult patients with relapsed or refractory B-precursor acute lymphoblastic leukaemia: a multicentre, single-arm, phase 2 study. *Lancet Oncol*, 16(1):57-66.
[http://doi.org/10.1016/S1470-2045\(14\)71170-2](http://doi.org/10.1016/S1470-2045(14)71170-2)
- Kantarjian H, Stein A, Gökbuget N, et al., 2017. Blinatumomab versus Chemotherapy for Advanced Acute Lymphoblastic Leukemia. *N Engl J Med*, 376(9):836-847.
<http://doi.org/10.1056/NEJMoa1609783>
- Gennaro G, Bucchi L, Ravaioli A, et al., 2024. The risk-based breast screening (RIBBS) study protocol: a personalized screening model for young women. *Radiol Med (Torino)*, 129(5):727-736.
<http://doi.org/10.1007/s11547-024-01797-9>
- Zhang J, Wu J, Zhou XS, Shi F, Shen D, 2023. Recent advancements in artificial intelligence for breast cancer: Image augmentation, segmentation, diagnosis, and prognosis approaches. *Semin Cancer Biol*, 96:11-25.
<http://doi.org/10.1016/j.semcancer.2023.09.001>
- Duggento A, Conti A, Mauriello A, Guerrisi M, Toschi N, 2021. Deep computational pathology in breast cancer. *Semin Cancer Biol*, 72:226-237.
<http://doi.org/10.1016/j.semcancer.2020.08.006>
- Timmins IR, Jones ME, O'Brien KM, et al., 2024. International Pooled Analysis of Leisure-Time Physical Activity and Premenopausal Breast Cancer in Women From 19 Cohorts. *J Clin Oncol*, 42(8):927-939.
<http://doi.org/10.1200/JCO.23.01101>
- Kirkham AA, Bland KA, Wollmann H, et al., 2019. Maintenance of Fitness and Quality-of-Life Benefits From Supervised Exercise Offered as Supportive Care for Breast Cancer. *J Natl Compr Cancer Netw*, 17(6):695-702.
<http://doi.org/10.6004/jnccn.2018.7276>
- Agiro A, DeVries A, Malin J, et al., 2018. Real-World Impact of a Decision Support Tool on Colony-Stimulating Factor Use and Chemotherapy-Induced Febrile Neutropenia Among Patients With Breast Cancer. *J Natl Compr Cancer Netw*, 16(2):162-169.
<http://doi.org/10.6004/jnccn.2017.7033>
- Wang X, Feng K, Wang W, et al., 2022. Long-term outcomes of intraoperative radiotherapy for early-stage breast cancer in China: a multicenter real-world study. *Cancer Commun Lond Engl*, 42(3):277-280.
<http://doi.org/10.1002/cac2.12258>
- Veitch Z, Khan OF, Tilley D, et al., 2019. Real-World Outcomes of Adjuvant Chemotherapy for Node-Negative and Node-Positive HER2-Positive Breast Cancer. *J Natl Compr Cancer Netw*, 17(1):47-56.
<http://doi.org/10.6004/jnccn.2018.7066>
- Anwar M, Chen Q, Ouyang D, et al., 2021. Pyrotinib Treatment in Patients With HER2-positive Metastatic Breast Cancer and Brain Metastasis: Exploratory Final Analysis of Real-World, Multicenter Data. *Clin Cancer Res*, 27(16):4634-4641.
<http://doi.org/10.1158/1078-0432.CCR-21-0474>
- Wu SY, Li JW, Wang YJ, et al., 2023. Clinical feasibility and oncological safety of non-radioactive targeted axillary dissection after neoadjuvant chemotherapy in biopsy-proven node-positive breast cancer: a prospective diagnostic and prognostic study. *Int J Surg Lond Engl*, 109(7):1863-1870.
<http://doi.org/10.1097/JS9.0000000000000331>
- Dai WF, Beca JM, Nagamuthu C, et al., 2022. Comparative Effectiveness and Safety of Pertuzumab and Trastuzumab Plus Chemotherapy vs Trastuzumab Plus Chemotherapy for Treatment of Metastatic Breast Cancer. *JAMA Netw Open*, 5(2):e2145460.
<http://doi.org/10.1001/jamanetworkopen.2021.45460>
- Stabellini N, Cao L, Towe CW, et al., 2023. Validation of the PREDICT Prognostication Tool in US Patients With Breast Cancer. *J Natl Compr Cancer Netw*, 21(10):1011-1019.e6.
<http://doi.org/10.6004/jnccn.2023.7048>
- Dinan MA, Wilson LE, Reed SD, 2019. Chemotherapy Costs and 21-Gene Recurrence Score Genomic Testing Among Medicare Beneficiaries With Early-Stage Breast Cancer, 2005 to 2011. *J Natl Compr Cancer Netw*, 17(3):245-254.
<http://doi.org/10.6004/jnccn.2018.7097>
- Basulaiman B, Awan AA, Fergusson D, et al., 2019. Creating a pragmatic trials program for breast cancer patients: Rethinking

- Clinical Trials (REaCT). *Breast Cancer Res Treat*, 177(1):93-101.
<http://doi.org/10.1007/s10549-019-05274-0>
- Wang YS, 2024. 123P - Goserelin 10.8 mg and 3.6 mg depots in breast cancer (BC): A large real-world noninferiority study. *Ann Oncol*, 9(suppl_4):1-25.
<http://doi.org/10.1016/esmooop/esmooop103096>
- Malone DC, Brown M, Hurwitz JT, et al., 2018. Real-World Evidence: Useful in the Real World of US Payer Decision Making? How? When? And What Studies? *Value Health J Int Soc Pharmacoeconomics Outcomes Res*, 21(3):326-333.
<http://doi.org/10.1016/j.jval.2017.08.3013>
- Franklin JM, Schneeweiss S, 2017. When and How Can Real World Data Analyses Substitute for Randomized Controlled Trials? *Clin Pharmacol Ther*, 102(6):924-933.
<http://doi.org/10.1002/cpt.857>
- Khozin S, Blumenthal GM, Pazdur R, 2017. Real-world Data for Clinical Evidence Generation in Oncology. *J Natl Cancer Inst*, 109(11).
<http://doi.org/10.1093/jnci/djx187>
- Miksad RA, Abernethy AP, 2018. Harnessing the Power of Real-World Evidence (RWE): A Checklist to Ensure Regulatory-Grade Data Quality. *Clin Pharmacol Ther*, 103(2):202-205.
<http://doi.org/10.1002/cpt.946>
- US Food and Drug Administration, 2023. Using the PICOTS framework to strengthen evidence gathered in clinical trials—guidance from the AHRQ’s evidence-based practice centers program.
 Available from <https://www.fda.gov/media/109448/download>. [Accessed on Nov. 25, 2024]
- Baumfeld Andre E, Reynolds R, Caubel P, et al., 2020. Trial designs using real-world data: The changing landscape of the regulatory approval process. *Pharmacoepidemiol Drug Saf*, 29(10):1201-1212.
<http://doi.org/10.1002/pds.4932>
- Kim HS, 2024. Dark Data in Real-World Evidence: Challenges, Implications, and the Imperative of Data Literacy in Medical Research. *J Korean Med Sci*, 39(9):e92.
<http://doi.org/10.3346/jkms.2024.39.e92>
- Kovačević A, Bašaragin B, Milošević N, et al., 2024. De-identification of clinical free text using natural language processing: A systematic review of current approaches. *Artif Intell Med*, 151:102845.
<http://doi.org/10.1016/j.artmed.2024.102845>
- Wang SV, Pinheiro S, Hua W, et al., 2021. STaRT-RWE: structured template for planning and reporting on the implementation of real world evidence studies. *BMJ*, 372:m4856.
<http://doi.org/10.1136/bmj.m4856>
- Hripcsak G, Duke JD, Shah NH, et al., 2015. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud Health Technol Inform*, 216:574-578.
- Vorisek CN, Lehne M, Klopfenstein SAI, et al., 2022. Fast Healthcare Interoperability Resources (FHIR) for Interoperability in Health Research: Systematic Review. *JMIR Med Inform*, 10(7):e35724.
<http://doi.org/10.2196/35724>
- Derksen JWG, May AM, Koopman M, 2019. The era of alternative designs to connect randomized clinical trials and real-world data. *Nat Rev Clin Oncol*, 16(9):589.
<http://doi.org/10.1038/s41571-019-0250-0>
- Perez-Lopez R, Ghaffari Laleh N, Mahmood F, Kather JN, 2024. A guide to artificial intelligence for cancer researchers. *Nat Rev Cancer*. 24(6):427-441.
<http://doi.org/10.1038/s41568-024-00694-7>
- Poirion OB, Jing Z, Chaudhary K, et al., 2021. DeepProg: an ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data. *Genome Med*, 13(1):112.
<http://doi.org/10.1186/s13073-021-00930-x>
- Li X, Zhang S, Zhang Q, et al., 2019. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. *Lancet Oncol*, 20(2):193-201.
[http://doi.org/10.1016/S1470-2045\(18\)30762-9](http://doi.org/10.1016/S1470-2045(18)30762-9)
- Yim WW, Yetisgen M, Harris WP, et al., 2016. Natural Language Processing in Oncology: A Review. *JAMA Oncol*, 2(6):797-804.
<http://doi.org/10.1001/jamaoncol.2016.0213>
- Sim JA, Huang X, Horan MR, et al., 2023. Natural language processing with machine learning methods to analyze unstructured patient-reported outcomes derived from electronic health records: A systematic review. *Artif Intell Med*, 146:102701.
<http://doi.org/10.1016/j.artmed.2023.102701>
- Lu Z, Sim JA, Wang JX, et al., 2021. Natural Language Processing and Machine Learning Methods to Characterize

- Unstructured Patient-Reported Outcomes: Validation Study. *J Med Internet Res*, 23(11):e26777.
<http://doi.org/10.2196/26777>
- Xu C, Song Y, Zhang D, et al., 2023. Spatiotemporal knowledge teacher-student reinforcement learning to detect liver tumors without contrast agents. *Med Image Anal*, 90:102980.
<http://doi.org/10.1016/j.media.2023.102980>
- Yala A, Mikhael PG, Lehman C, et al., 2022. Optimizing risk-based breast cancer screening policies with reinforcement learning. *Nat Med*, 28(1):136-143.
<http://doi.org/10.1038/s41591-021-01599-w>
- Kaiser T, Rajpoot NM, 2019. Learning Where to See: A Novel Attention Model for Automated Immunohistochemical Scoring. *IEEE Trans Med Imaging*, 38(11):2620-2631.
<http://doi.org/10.1109/TMI.2019.2907049>
- Lu Y, Chu Q, Li Z, et al., 2024. Deep reinforcement learning identifies personalized intermittent androgen deprivation therapy for prostate cancer. *Brief Bioinform*, 25(2):bbae071.
<http://doi.org/10.1093/bib/bbae071>
- Tortora M, Cordelli E, Sicilia R, et al., 2021. Deep Reinforcement Learning for Fractionated Radiotherapy in Non-Small Cell Lung Carcinoma. *Artif Intell Med*, 119:102137.
<http://doi.org/10.1016/j.artmed.2021.102137>
- Zhao Y, Kosorok MR, Zeng D, 2009. Reinforcement learning design for cancer clinical trials. *Stat Med*, 28(26):3294-3315.
<http://doi.org/10.1002/sim.3720>
- Hsu WH, Ko AT, Weng CS, et al., 2023. Explainable machine learning model for predicting skeletal muscle loss during surgery and adjuvant chemotherapy in ovarian cancer. *J Cachexia Sarcopenia Muscle*, 14(5):2044-2053.
<http://doi.org/10.1002/jcsm.13282>
- Eggermont AMM, Apolone G, Baumann M, et al., 2019. Cancer Core Europe: A translational research infrastructure for a European mission on cancer. *Mol Oncol*, 13(3):521-527.
<http://doi.org/10.1002/1878-0261.12447>

Supplementary information:

No supplementary data was provided by this study.