



Аналитические методы в работе с данными реальной клинической практики и доказательствами реальной клинической практики: критический обзор

Идриса К. ¹, Драгунс В. ², Булкрэн М.С. ¹

¹ ФГАОУ ВО «Национальный исследовательский университет ИТМО», Санкт-Петербург, Российская Федерация

² ООО «Экс Севен Ресеч», Санкт-Петербург, Российская Федерация

Аннотация

Актуальность. Формирование медицинских доказательств все меньше опирается исключительно на рандомизированные контролируемые исследования (РКИ) и всё больше — на данные реальной клинической практики (RWD). Собираемые в различных условиях рутинной клинической практики, данные RWD служат основой для получения доказательств, основанных на данных реальной клинической практики (RWE), предоставляя понимание исходов пациентов, эффективности лечения и организации здравоохранения. Хотя RWE повышает обобщаемость результатов и экономическую эффективность по сравнению с РКИ, присущая ей сложность, такие как проблемы с качеством данных, пропущенные значения, смешивающие факторы (конфаундинг) и ошибки отбора, требует применения строгих аналитических подходов.

Методы. В данном критическом обзоре обобщается спектр аналитических методологий, применяемых к RWD, начиная с традиционных статистических методов и подходов (фреймворков) к каузальному анализу и заканчивая машинным обучением и продвинутыми методами, включая обработку естественного языка (NLP), байесовское моделирование и сетевой анализ. Каждый подход оценивается с точки зрения его сильных сторон, ограничений и пригодности для решения уникальных задач, возникающих при работе с RWD.

Результаты. Традиционные статистические модели обеспечивают интерпретируемость и контроль наблюдаемых смешивающих факторов, но ограничены строгими допущениями и уязвимы перед неучтёнными смешивающими факторами. Методы каузального анализа, такие как инструментальные переменные и имитация целевого исследования, усиливают каузальную интерпретацию, но требуют строгих допущений и специализированных знаний. Подходы машинного обучения превосходны в прогнозировании и анализе многомерных данных, но сталкиваются с проблемами интерпретируемости и обобщаемости. Продвинутые методы, включая NLP и байесовские модели, расширяют аналитические возможности, но требуют значительного опыта и вычислительных ресурсов. Во всех категориях триангуляция (использование нескольких методов) повышает надёжность и достоверность результатов.

Выводы. RWE играет критическую роль в разработке лекарств, пострегистрационном наблюдении, сравнительных исследованиях эффективности, исследованиях в области экономики здравоохранения и оценки исходов (HEOR), а также в персонализированной медицине. Будущий прогресс зависит от новых инноваций, таких как федеративное обучение, генерация синтетических данных и объяснимый ИИ (XAI), наряду с достижениями в области интеграции данных, гармонизации и воспроизводимости результатов. Методологическая прозрачность, строгая валидация и междисциплинарное сотрудничество необходимы для обеспечения того, чтобы RWE предоставляла надёжные, этичные и клинически значимые выводы, способные служить основой для регуляторных решений и оптимизации здравоохранения во всём мире.

Ключевые слова: доказательства реальной клинической практики; RWE; данные реальной клинической практики; RWD; каузальный анализ; машинное обучение; аналитика в здравоохранении

Для цитирования: Идриса К., Драгунс В., Булкрэн М.С. Аналитические методы в работе с данными реальной клинической практики и доказательствами реальной клинической практики: критический обзор. *Реальная клиническая практика: данные и доказательства*. 2026;6(2):44-55. <https://doi.org/10.37489/2782-3784-myrd-103>. EDN: NKOOPM.

Поступила: 16.03.2026. В доработанном виде: 20.04.2026. Принята к печати: 06.05.2026. Опубликовано: 30.06.2026.

Analytical methods in real-world data/evidence (RWD/E): a critical review

Kiryowa Idrisa¹, Vitaly Draguns², Mohamed Said Boulkrane¹

¹ITMO University, St. Petersburg, Russian Federation

²X7 Research LLC, St. Petersburg, Russian Federation

Abstract

Background. Healthcare evidence generation is shifting beyond randomized controlled trials (RCTs) to embrace real-world data (RWD). Collected from diverse routine care settings, RWD underpins real-world evidence (RWE), offering insights into patient outcomes, treatment effectiveness, and healthcare delivery. While RWE enhances generalizability and cost-effectiveness compared with RCTs, its inherent complexities such as data quality issues, missingness, confounding, and selection bias demand rigorous analytical approaches.

Methods. This critical review synthesizes the spectrum of analytical methodologies applied to RWD, ranging from traditional statistical techniques and causal inference frameworks to machine learning and advanced methods including natural language processing (NLP), Bayesian modeling, and network analysis. Each approach is appraised in terms of strengths, limitations, and suitability for addressing the unique challenges of RWD.

Results. Traditional statistical models provide interpretability and control for observed confounding but are limited by strong assumptions and vulnerability to unmeasured confounders. Causal inference methods, such as instrumental variables and target trial emulation, strengthen causal interpretation but require strong assumptions and specialized expertise. Machine learning approaches excel in prediction and high-dimensional data analysis but face interpretability and generalizability challenges. Advanced methods, including NLP and Bayesian models, extend analytical capacity but demand significant expertise and computational resources. Across categories, triangulation of multiple methods enhances robustness and credibility.

Conclusions. RWE plays a critical role in drug development, post-market surveillance, comparative effectiveness research, health economics and outcomes research, and personalized medicine. Future progress hinges on emerging innovations such as federated learning, synthetic data generation, and explainable AI (XAI), alongside advances in data integration, harmonization, and reproducibility. Methodological transparency, rigorous validation, and interdisciplinary collaboration are essential to ensure that RWE delivers trustworthy, ethical, and clinically actionable insights capable of informing regulatory decisions and optimizing healthcare globally.

Keywords: real-world evidence; RWE; real-world data; RWD; causal inference; machine learning; healthcare analytics

For citation: Idrisa K, Draguns VYu, Boulkrane MS. Analytical methods in real-world data/evidence (RWD/E): a critical review. *Real-World Data & Evidence*. 2026;6(2):44-55. <https://doi.org/10.37489/2782-3784-myrd-103>. EDN: NKOOPM.

Received: 16.03.2026. **Revision received:** 20.04.2026. **Accepted:** 06.05.2026. **Published:** 30.06.2026.

Introduction

The development and application of clinical evidence has resulted in a profound paradigm change in the healthcare environment [1]. Randomized controlled trials (RCTs) have long been regarded as the gold standard for assessing the efficacy and safety of medical products due to their ability to reduce bias through randomization [2]. However, RCTs frequently include highly selected patient populations, take place in controlled surroundings, and can be prohibitively expensive and time-consuming [3, 4]. These constraints frequently limit the generalizability of their findings to the numerous patient populations and clinical practice settings encountered in the real world [5, 6]. In response to these challenges, real-world data (RWD) has emerged as a crucial complementary source of information [7]. RWD refers to data about patient health status and/or healthcare delivery that is routinely obtained from a range of sources other than standard randomized controlled trials [8–10]. RWD is defined by the US Food and Drug Administration

(FDA) as data produced from sources such as electronic health records (EHRs), medical claims data, product and disease registries, and data collected from mobile devices and other digital health technologies [11]. The examination of RWD yields real-world evidence (RWE), which is clinical evidence about the use and possible benefits or dangers of a medical product [11, 12]. This increased emphasis on RWD and RWE represents a fundamental shift in evidence creation, motivated by the awareness that RWD provides higher generalizability to various patient populations and real-world clinical settings, typically at a lower cost than traditional trials [13]. This trend suggests that regulatory agencies, payers, clinicians, and the pharmaceutical industry are increasingly valuing external validity and cost-efficiency alongside the internal validity previously prioritized by RCTs, necessitating a strong foundation for RWE creation [14, 15].

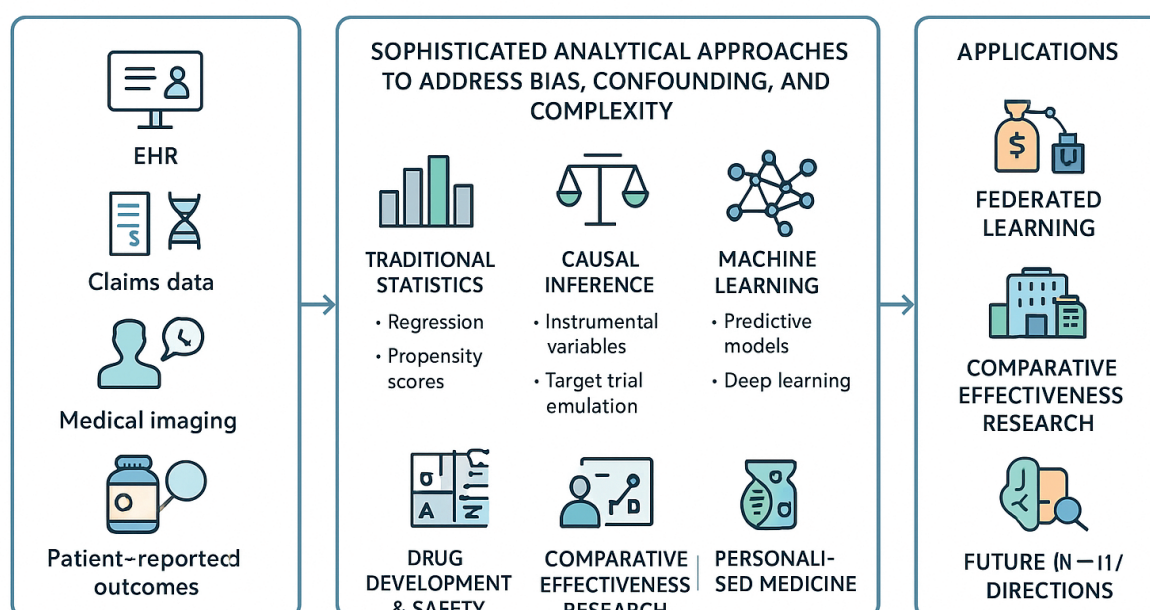
While RWD has enormous potential to provide a broader, more representative view of patient

populations and clinical practice, its inherent complexities present significant analytical challenges [15]. Because RWD is not collected for research purposes, it leads to issues such as data quality inconsistencies, missing information, confounding factors, and various types of selection bias [12, 16]. These issues are common across various RWD sources and, if not addressed, can lead to misleading or inaccurate evidence. The principle “flawed input leads to flawed output” is especially important in the context of RWE [17]. Inadequate analytical approaches for RWD can produce unreliable data while also undermining trust in the broader RWD/RWE paradigm. If the data produced is insufficient, it can lead to bad decisions in crucial areas such as medication development, regulatory approvals, and clinical practice. Repeated instances of unreliable RWE may weaken trust among regulatory agencies, such as the FDA, and the broader medical community, hampering the very progress that RWD and RWE want to achieve. This circumstance highlights the important necessity for methodological transparency, thorough validation, and the use of sophisticated analytical approaches to ensure that the RWE generated is credible and actionable [18]. This review seeks to assess the range of analytical approaches used in RWD/RWE research, which include but not limited to statistical approaches, advanced machine learning algorithms, and specialized causal inference techniques to assess their respective strengths, limitations, and applicability in addressing the unique challenges of RWD/RWE analytics [8, 19]. Furthermore, it identifies present methodological limitations and suggests future options for developing and applying these analytical tools thereby

aiming to be a significant resource for researchers, physicians, and policymakers seeking to understand and implement rigorous analytical approaches for producing high-quality and dependable RWE.

Foundations of Real-World Data and Evidence

RWD includes information about patient health status and/or healthcare delivery that is frequently acquired from sources other than RCTs [5]; hence, understanding the fundamentals of RWD and RWE is critical for grasping their analytical journey from raw data to meaningful insights. This data comes from a variety of sources, including electronic health records (EHRs), medical claims data, product and disease registries, and, increasingly, data collected from mobile devices and other digital health technologies [20]. Raw data (RWD) is essentially the raw material (observations from routine clinical practice or administrative processes) for real-world evidence (RWE) [11]. RWE is the clinical evidence regarding the use and potential benefits or risks of a medical product derived from RWD analysis [12]; thus, it represents the refined product (the actionable insight that emerges after rigorous analytical processing of RWD [14]), and it is increasingly being used for critical functions such as regulatory decision-making, informing drug development strategies, and conducting post-market surveillance of medical products [5]. The quality and dependability of RWE are completely dependent on the quality of the underlying RWD and, more importantly, the analytical techniques used on it since such data maybe incomplete and lack end-point due to the fact that its initial purpose is not research [21].



Toward reproducibility, transparency, and ethical RWE for global healthcare impact

Fig. 1. From Real-World Data to Reliable Real-World Evidence

Common Sources and Types of RWD

RWD comes from a variety of sources, each providing unique insights and posing distinct analytical issues given that most of the collected data is unstructured, voluminous, and dynamic [22].

Electronic health records (EHRs) are comprehensive digital records of patient health information that include diagnoses, procedures, prescriptions, laboratory data, clinical notes, and more [18]. They provide extensive, longitudinal clinical detail, resulting in a thorough comprehension of patient journeys. However, because EHR data are primarily used for clinical treatment rather than research, they may contain structured and unstructured information (e. g. images, clinical notes etc.), missing data, coding variations, and errors [23, 24]. These require careful and intensive pre-processing. This type of RWD had created unprecedented opportunities for data-driven approaches in healthcare (pattern recognition, preoperative planning and others) [23, 25, 26]. This data has recently been reported to be more useful with the use of new analytical techniques such as Machine Learning and Neural networks to improve predictions in selected outcomes those linked to administrative and data usage hence allowing validation and replication of findings from RCT [27, 28].

Medical Claims Data are administrative datasets primarily generated for billing and reimbursement purposes, these provide a valuable insights into healthcare utilization and expenditures for large populations [29].

This type of data is a rich source for tracking diagnoses and procedures but has significant limitations. A notable drawback is the lack of detailed clinical information, such as specific laboratory results, disease severity, or lifestyle factors [26]. Furthermore, these datasets often only reflect services for which a claim was successfully processed, potentially excluding services that were not billed or reimbursed [23].

Patient registrations are organized systems that collect standardized data on individuals who have a given disease, condition, or are undergoing a specific intervention [26]. They frequently have greater data quality for specific outcomes and exposures important to their objective, but their generalizability may be limited to the group enrolled.

Patient-Generated Health Data (PGHD) is gathered through wearables and mobile technologies, is a rapidly expanding domain of real-world data [26]. These sources, which include data from wearable sensors and mobile applications, offer continuous and granular insights into patient behavior, physiological metrics (heart rate and sleep patterns), and symptoms in real time [30]. However, the use of PGHD presents notable challenges, including data heterogeneity, validity concerns, and significant privacy issues.

The integration of these diverse Real-World Data (RWD) sources such as Electronic Health Records (EHRs), claims data, patient registries, genomics, and PGHD results in complex, multi-modal datasets. While these datasets provide an unprecedented



Fig. 2. RWD Types and Sources (source: [Swift et al., 2018])

opportunity for a holistic understanding of patient health, they also amplify the complexities related to data harmonization, interoperability, and the selection of appropriate analytical methodologies [31].

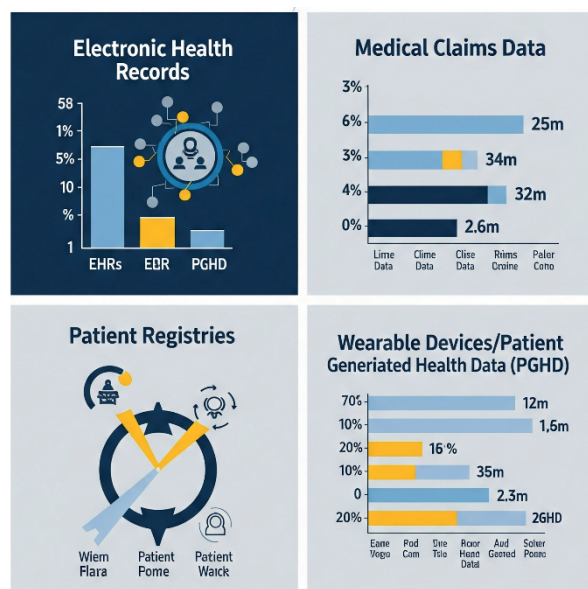


Fig. 3. RWD sources and their respective data effectiveness

Challenges in RWD analytics

The utility of real-world data (RWD) is inherently dependent on both its quality and the robustness of the analytical methodologies employed. Since RWD is frequently collected for purposes unrelated to research, issues such as inconsistencies, coding variability, inaccurate entries, and substantial amounts of missing data are common, potentially undermining study validity [32]. Moreover, unlike randomized controlled trials (RCTs), the observational nature of RWD precludes, and random allocation of treatment lead to several potential biases. These include confounding by indication, in which treatment decisions are correlated with clinical outcomes; selection bias, arising from systematic differences between exposed and unexposed populations; and information bias, resulting from systematic errors in data collection [33, 34]. Heterogeneity in RWD is another key limitation, reflecting variability in patient demographics, clinical practice patterns, and data capture methodologies across healthcare systems, which complicates the synthesis and generalization of findings [35]. Furthermore, the non-experimental nature of RWD poses substantial challenges for causal inference, necessitating careful methodological adjustments to mitigate residual confounding and bias [36].

Table 1. Sources, data, application and limitations in RWD analytics

Source Type & Data	Strengths (Analytical)	Limitations (Analytical)	References
<i>Electronic Health Records (EHRs).</i> Diagnoses, Procedures, Medications, Lab Results, Clinical Notes, Demographics.	Rich clinical detail, Longitudinal patient history, Reflects real-world practice	Unstructured data, Missingness, Coding variability, Data entry errors	25
<i>Medical Claims Data.</i> Billing Codes (ICD, CPT), Utilization, Costs, Dates of Service, Provider Information.	Large populations, Cost-effectiveness data, Healthcare utilization patterns	Lack of clinical depth (e. g., lab results, severity), No direct clinical outcomes	33
<i>Patient Registries.</i> Disease-specific outcomes, Treatment details, Demographics, Quality of life.	High data quality for specific outcomes, Targeted populations, Follow-up data	Limited generalizability, Potential selection bias (volunteer bias)	26
<i>Wearable Devices / Patient-Generated Health Data (PGHD).</i> Activity, Heart Rate, Sleep Patterns, Self-reported symptoms, Glucose levels.	Continuous, Granular, Behavioral insights, Data from outside clinical settings	Data validity and reliability issues, Privacy concerns, Heterogeneity of devices/formats	32

Analytical Methodologies for RWD/RWE

The unique characteristics of RWD necessitate a diverse toolkit of analytical methodologies, while foundational statistical methods remain relevant, the complexities of RWD often demand more advanced causal inference techniques and machine learning approaches to generate robust and reliable evidence [8].

Real-world evidence (RWE) studies frequently employ traditional statistical methods as foundational tools for descriptive analyses and preliminary hypothesis testing. Regression models, including linear

regression for continuous outcomes, logistic regression for binary outcomes, and Cox proportional hazards regression for time-to-event outcomes, are commonly used to model associations and predict outcomes [22]. These models necessitate careful covariate adjustment to address confounding variables [37]. Propensity score methods are critical for mitigating confounding in observational studies by balancing observed covariates between treatment and control groups. Approaches include matching pairing treated and untreated individuals with similar propensity scores; weighting using

Table 2. Comprehensive Overview of Analytical Methods in Real-World Data Studies

Category	Methodology	Summative description	Reference
Traditional Statistics	Linear Regression	Models work with continuous outcomes based on explanatory variables. Predicting changes in blood pressure. Simple, interpretable, widely used. Sensitive to confounding; assumes linearity.	22, 37
	Logistic Regression	Models operate on binary outcomes using predictor variables. Predicting disease presence/absence. Handle binary data effectively. Requires correct model specification.	
	Cox Proportional Hazards Regression	Models use time-to-event data. Survival analysis in oncology. Efficient with censored data. Assumes proportional hazards.	
Propensity Score Methods	Matching	Pairs treated and untreated individuals with similar propensity scores. Reducing bias in treatment effect estimation. Balances observed covariates. Limited by measured variables only.	38
	Weighting (IPTW)	Assigns weights to create balanced pseudo-populations. Comparative effectiveness research. Mimics randomization. Sensitive to extreme weights.	
	Stratification	Divides participants into strata by propensity scores. Analysis within strata to improve comparability. Simple to implement. Reduces sample size per stratum.	
Causal Inference	Instrumental Variables (IV)	Uses external variables affecting treatment but not the outcome. Evaluating policy impacts on treatment. Addresses the unmeasured confounding. Requires strong instrument validity.	12, 18
	Difference-in-Differences (DiD)	Compares outcome changes over time between groups. Policy changes evaluation. Controls for time-invariant confounding. Assumes parallel trends.	
	Regression Discontinuity Design (RDD)	Leverages treatment assignment thresholds. Threshold-based intervention analysis. Strong internal validity. Limited external generalizability.	
	Target Trial Emulation	Designs observational studies to mimic RCTs. Comparative drug safety/effectiveness studies. Increases transparency, reduces bias. Requires detailed, high-quality data.	
Machine Learning	Predictive Modeling (RF, GBM)	Learn patterns from data to predict outcomes. Adverse event risk prediction. Handles high-dimensional datasets. May lack interpretability.	43
	Unsupervised Learning	Finds hidden structure in unlabeled data. Patient phenotyping. Identifies novel subgroups. Require large datasets.	
	Deep Learning & Image Analysis	Learning hierarchical features from raw data. Automated medical image interpretation. Excels at unstructured data processing. Computationally intensive.	
Advanced Methods	Natural Language Processing (NLP)	Extracts structured data from unstructured clinical text. Identifying adverse events from clinical notes. Unlocks rich clinical details. Requires large, annotated corpora.	40–42
	Bayesian Methods	Incorporates prior knowledge into statistical inference. Rare disease modeling. Handles uncertainty; integrates expert knowledge. Sensitive to choice of priors.	
	Network Analysis	Examine relationships between healthcare entities. Drug – drug interaction mapping. Captures complex interdependencies. Interpretation can be challenging.	

inverse probability of treatment weighting (IPTW) to generate a balanced synthetic population; and stratification dividing participants into strata by propensity scores [38]. Causal inference methods aim to establish cause effect relationships in the presence of confounding and selection bias [18]. Instrumental variables (IV) exploit exogenous variation that influences treatment assignment but not the outcome, allowing estimation of causal effects even with unmeasured confounding. Difference-in-differences (DiD) compare pre- and post-intervention changes between treatment and control groups, while regression discontinuity designs (RDD) leverage threshold-based treatment assignment [12]. Target trial emulation explicitly structures observational studies to mimic randomized controlled trials, enhancing transparency in assumptions [39].

Beyond core analytical categories, advanced methodologies provide specialized capabilities for real-world data (RWD) analysis. Natural Language Processing (NLP) is particularly valuable for extracting structured information from unstructured text sources such as clinical notes, discharge summaries, and pathology reports within electronic health records (EHRs). By leveraging NLP, researchers can access nuanced clinical details often overlooked by structured data alone [40]. Bayesian methods offer a flexible statistical framework that incorporates prior knowledge or expert opinion into the analytical process. This approach is especially advantageous when addressing uncertainty or analyzing limited datasets, such as those in rare disease research [41]. Network analysis enables the exploration of complex interconnections among entities such as drugs, diseases, genes, or patients. Applying network theory to RWD can reveal important patterns, including drug-drug interactions, disease progression pathways, and patient similarity networks, thereby supporting a more integrated understanding of clinical and biological systems [42].

Critical Appraisal of Analytical Methods

The generation of robust real-world evidence (RWE) depends on the careful analysis of real-world data (RWD) hence keen evaluation of analytical methods used must be done, particularly their strengths, limitations, and suitability for specific data and research contexts. By this, it allows us to understand the best methodology and also to predict the expected results.

Traditional statistical methods, such as regression models and propensity score techniques, remain foundational due to their interpretability and widespread use in controlling observed confounders [22, 37]. These approaches are well suited for hypothesis testing with structured data but are highly vulnerable to unmeasured confounding, restrictive model assumptions, and difficulties handling high-dimensional data [44, 45]. Causal inference methods, such as instrumental variables, regression discontinuity, and

target trial emulation, explicitly aim to address confounding and selection bias [12, 18]. These approaches strengthen causal interpretation in observational settings and improve the credibility of findings for regulatory purposes [38, 46]. However, they rely on strong, often untestable assumptions such as valid instruments or no unmeasured confounding which make results sensitive to be specific and necessitate substantial methodological expertise [47].

Machine learning approaches excel at pattern discovery, prediction, and modeling non-linear interactions in high-dimensional RWD [43]. They automate feature extraction, support risk stratification, and enable novel insights into patient phenotypes. However, they are often criticized for their “black box” nature, raising interpretability and transparency challenges in clinical and regulatory contexts [43, 48]. Moreover, generalizability remains a concern, as models trained on one dataset may fail in another, underscoring the importance of rigorous validation [44]. Advanced methodologies such as natural language processing (NLP), Bayesian methods, and network analysis extend analytical capacity [40–42]. NLP enables extraction of rich information from unstructured clinical notes [49], Bayesian methods allow incorporation of prior knowledge to model uncertainty in sparse datasets [41], and network analysis reveals structural relationships among diseases, drugs, and patients [42]. Despite these strengths, their implementation requires specialized expertise, substantial computational resources, and careful handling of model assumptions [50].

Considerations for Bias, Confounding, and Generalizability

Bias mitigation and confounding control are critical in RWE research. While tools such as propensity scores and instrumental variables reduce confounding, no single method fully addresses unmeasured confounders [33, 35]. Sensitivity analysis and triangulation comparing results across multiple approaches therefore it is indispensable for robustness [22, 46]. Even although RWD generally enhances external validity compared to RCTs, machine learning models may still lack portability across populations, thus highlighting the necessity of external validation [51].

Applications and Future Directions of RWD/E

The analytical rigor applied to real-world data (RWD) translates into substantial impacts across multiple domains of healthcare, underscoring the transformative potential of real-world evidence (RWE).

Drug Development and Post-Market Surveillance. RWE is increasingly integrated into the lifecycle of medical products, from early development to post-market monitoring. It accelerates drug development

Table 3. Critical Appraisal of Analytical Methods in Real-World Evidence (RWE)

Method Category	Strengths	Limitations	Bias & Confounding Considerations	Regulatory & Ethical Implications	References
Traditional Statistics (Regression, Propensity Scores)	Widely understood; transparent; effective for hypothesis testing; good for controlling observed confounders.	Vulnerable to unmeasured confounding; relies on strong assumptions (e. g., linearity, proportional hazards); limited handling of high-dimensional/unstructured data.	Can address observed confounders but fails with unmeasured ones.	Well-accepted by regulators due to interpretability; limited use for complex RWD.	44, 45
Causal Inference Methods (IV, RDD, Target Trial Emulation)	Designed to address confounding and selection bias; enables causal inference; enhances credibility of RWE for decision-making.	Relies on untestable assumptions (e. g., valid instruments, no unmeasured confounding); requires specialized expertise; sensitive to model specification.	Stronger control of confounding compared to traditional statistics, but assumptions remain critical.	Increasingly valued in regulatory settings for causal interpretation; requires transparency in assumptions.	38, 46
Machine Learning Approaches (Predictive Models, Phenotyping)	Handles high-dimensional and complex data; captures non-linear relationships; automates feature engineering; strong for prediction.	Interpretability issues ("black box"); poor generalizability across datasets; requires very large, high-quality data.	Can reduce bias in feature selection but not inherently designed for causal inference.	Regulatory concerns about transparency and explainability; risks of algorithmic bias.	22, 43, 48
Advanced Methods (NLP, Bayesian, Network Analysis)	NLP extracts insights from unstructured text; Bayesian methods integrate prior knowledge & handle uncertainty; network analysis captures complex interrelationships.	Requires specialized expertise; computationally intensive; results sensitive to model assumptions.	Helps uncover hidden biases and patterns in complex systems; Bayesian methods useful for rare diseases.	Ethical implications in text mining and network inference; computational demands may limit regulatory adoption.	40–42

by informing trial design, optimizing eligibility criteria, and identifying patient subgroups more likely to benefit from specific therapies [52]. Post-approval, RWE plays a crucial role in pharmacovigilance by detecting rare adverse events, evaluating long-term outcomes, and complementing randomized controlled trials (RCTs) with evidence from diverse populations in routine care [52, 53].

Comparative Effectiveness Research (CER). RWD supports comparative effectiveness research (CER) by enabling head-to-head comparisons of interventions in real-world clinical settings [12]. This extends beyond the controlled environment of RCTs to assess treatments as they are applied in practice, offering insights for treatment guidelines, payer decision-making, and policy development [54]. The observational design of CER makes it particularly valuable when RCTs are infeasible due to ethical or logistical constraints [6].

Health Economics and Outcomes Research (HEOR). HEOR relies heavily on RWE to evaluate

the economic and societal value of interventions [55]. RWD enables cost-effectiveness assessments, burden-of-illness studies, and evaluations of real-world resource utilization, informing reimbursement decisions and value-based care models [55, 56].

Personalized Medicine and Precision Health. Personalized medicine depends on the integration of RWD with genomic, imaging, and phenotypic data to tailor treatments to individual characteristics [57]. These insights enable stratification of patients based on genetic or clinical markers, enhancing prediction of therapeutic responses and risks. The analytical shift is from population-level inference to individualized risk prediction, requiring advanced machine learning and statistical approaches to handle high-dimensional, multi-modal datasets [57, 58].

Emerging Analytical Techniques. Innovations in data science are critical for maximizing the value of RWD:

- **Federated Learning** allows distributed model training across institutions without centralizing sensitive patient data, addressing privacy concerns and enabling large-scale, collaborative research [54].
- **Synthetic Data Generation** uses generative AI to create realistic, privacy-preserving data for research and method development [55].
- **Explainable AI (XAI)** seeks to improve interpretability of machine learning predictions, fostering trust in clinical practice and regulatory contexts [53].

Integration of Diverse Data Sources. Future progress depends on integrating heterogeneous RWD sources such as EHRs, claims, genomics, imaging, and social determinants of health into comprehensive patient profiles. Data harmonization and multi-modal analysis are necessary to derive robust insights [54].

Standardization and Reproducibility. A major challenge is the lack of standardization in data formats, terminologies, and analytic pipelines. Adoption of FAIR (Findable, Accessible, Interoperable, Reusable) principles and frameworks like the OMOP Common Data Model is essential to improve reproducibility and regulatory acceptance [53].

Role of AI and Advanced Computing. The scale and complexity of RWD necessitate the use of AI, high-performance computing, and cloud infrastructure for processing large datasets, training models, and conducting causal inference analyses. These tools are indispensable for unlocking the full potential of RWE [58].

Recommendations for Best Practices. Future directions emphasize methodological transparency, rigorous validation, interdisciplinary collaboration, and researcher training in advanced data science and ethics. These steps are crucial for ensuring the credibility, reproducibility, and ethical integrity of RWE [52].

Conclusion

Real-world data (RWD) and real-world evidence (RWE) have emerged as transformative pillars in contemporary healthcare, offering the ability to generate clinically meaningful insights that extend beyond the constraints of randomized controlled trials. By reflecting routine clinical practice, RWD enables the assessment of treatment effectiveness, safety, and patient-centered outcomes across diverse populations. However, the inherent complexities of RWD including but not limited to issues of data quality, missingness, heterogeneity, and confounding demand analytical rigor and innovation. This review has underscored the breadth of analytical methodologies available for RWD analysis, spanning traditional statistical models, causal inference frameworks, machine learning techniques, and advanced approaches such as natural language processing, Bayesian modeling, and network analysis. Each method presents distinct advantages and limitations, but no single strategy offers a comprehensive solution. A triangulated, multi-method approach that leverages complementary strengths is essential for enhancing validity, mitigating bias, and ensuring the robustness of findings. Future directions highlight the importance of emerging methods such as federated learning, synthetic data generation, and explainable artificial intelligence (XAI), which collectively address key concerns around data privacy, scalability, and interpretability. Equally critical are efforts to integrate heterogeneous data sources, promote standardization through common data models, and ensure reproducibility of findings across contexts. Methodological transparency, rigorous validation, and interdisciplinary collaboration will be indispensable in advancing the credibility and acceptance of RWE in clinical, regulatory, and policy decision-making. Ultimately, the promise of RWD and RWE lies not merely in methodological sophistication but in their ethical and responsible application. By aligning innovation with principles of fairness, privacy, and equity, the healthcare community can fully harness the potential of RWE to advance personalized medicine, inform evidence-based practice, and optimize healthcare delivery worldwide.

ДОПОЛНИТЕЛЬНАЯ ИНФОРМАЦИЯ

Конфликт интересов

Авторы заявляют об отсутствии конфликта интересов в связи с данной публикацией.

Участие авторов

All authors contributed equally to the preparation of the publication.

ADDITIONAL INFORMATION

Conflict of interests

The authors declare that there is no conflict of interest in connection with this publication.

Authors' participation

Все авторы сделали эквивалентный вклад в подготовку публикации.

Финансирование

Исследование выполнено без целевого внешнего финансирования.

СВЕДЕНИЯ ОБ АВТОРАХ

Идриса Кирёва — Магистрант, ФГАОУ ВО «Национальный исследовательский университет ИТМО», Санкт-Петербург, Российская Федерация
e-mail: 408079@edu.itmo.ru
ORCID ID: 0009-0007-3178-0424

Драгунс Виталий — зам. генерального директора ООО «Экс Севен Ресеч», Санкт-Петербург, Российская Федерация
e-mail: vdraguns@x7cpr.com
ORCID ID: 0009-0000-4454-2782

Булкран Мохамед Саид — доцент, ФГАОУ ВО «Национальный исследовательский университет ИТМО», Санкт-Петербург, Российская Федерация
Автор, ответственный за переписку
e-mail: mboulkrane@itmo.ru
ORCID ID: 0000-0002-1204-5041

Financing

The study was carried out without targeted external funding.

ABOUT THE AUTHORS

Kiryowa Idrisa — MSc, Student, ITMO University, St. Petersburg, Russian Federation
e-mail: 408079@edu.itmo.ru
ORCID ID: 0009-0007-3178-0424

Vitaly Draguns — deputy general director, X7 Research LLC, St. Petersburg, Russian Federation
e-mail: vdraguns@x7cpr.com
ORCID ID: 0009-0000-4454-2782

Mohamed Said Boulkrane — Associate Professor, ITMO University, St. Petersburg, Russian Federation
Corresponding author
e-mail: mboulkrane@itmo.ru
ORCID ID: 0000-0002-1204-5041

References

1. Alemayehu D, Ali R, Alvir JMaJ, et al. Examination of Data, Analytical Issues and Proposed Methods for Conducting Comparative Effectiveness Research Using “Real-World Data.” *J Manag Care Pharm.* 2011;17(9 Supp A):1-37. doi:10.18553/jmcp.2011.17.s9-a.1
2. Price G, Mackay R, Aznar M, et al. Learning healthcare systems and rapid learning in radiation oncology: Where are we and where are we going? *Radiother Oncol.* 2021;164:183-195. doi:10.1016/j.radonc.2021.09.030
3. Chodankar D. Introduction to real-world evidence studies. *Perspect Clin Res.* 2021;12(3):171-174. doi:10.4103/picr.picr_62_21
4. Freemantle N, Strack T. Real-world effectiveness of new medicines should be evaluated by appropriately designed clinical trials. *J Clin Epidemiol.* 2010;63(10):1053-1058. doi:10.1016/j.jclinepi.2009.07.013
5. Berger ML, Lipset C, Gutteridge A, Axelsen K, Subedi P, Madigan D. Optimizing the Leveraging of Real-World Data to Improve the Development and Use of Medicines. *Value Health.* 2015; 18(1):127-130. doi:10.1016/j.jval.2014.10.009
6. Franklin JM, Schneeweiss S. When and How Can Real World Data Analyses Substitute for Randomized Controlled Trials? *Clin Pharmacol Ther.* 2017;102(6):924-933. doi:10.1002/cpt.857
7. Friedman C, Rubin J, Brown J, et al. Toward a science of learning systems: a research agenda for the high-functioning Learning Health System. *J Am Med Inform Assoc.* 2015;22(1):43-50. doi:10.1136/amiajnl-2014-002977
8. Craddock M, Dempsey C, Abdulwahid D, et al. Challenges and opportunities for real-world evidence in clinical oncology—a view from the UK: proceedings of a national workshop. *ESMO Real World Data Digit Oncol.* 2024;6:100089. doi:10.1016/j.esmorw.2024.100089
9. Guérin J, Nahid A, Tassy L, et al. Consore: A Powerful Federated Data Mining Tool Driving a French Research Network to Accelerate Cancer Research. *Int J Environ Res Public Health.* 2024;21(2):189. doi:10.3390/ijerph21020189
10. Makady A, De Boer A, Hillege H, Klungel O, Goettsch W. What Is Real-World Data? A Review of Definitions Based on Literature and Stakeholder Interviews. *Value Health.* 2017;20(7):858-865. doi:10.1016/j.jval.2017.03.008
11. US Food & Drug Administration. Real world evidence. June 9, 2025. Accessed August 7, 2025. <https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence>
12. Dang A. Real-World Evidence: A Primer. *Pharm Med.* 2023;37(1):25-36. doi:10.1007/s40290-022-00456-6
13. Roche N, Reddel HK, Agusti A, et al. Integrating real-life studies in the global therapeutic research framework. *Lancet Respir Med.* 2013;1(10):e29-e30. doi:10.1016/S2213-2600(13)70199-1

14. Kim HS, Lee S, Kim JH. Real-world Evidence versus Randomized Controlled Trial: Clinical Research Based on Electronic Medical Records. *JKorean Med Sci*. 2018;33(34):e213. doi:10.3346/jkms.2018.33.e213
15. Rudrapatna VA, Butte AJ. Opportunities and challenges in using real-world data for health care. *J Clin Invest*. 2020;130(2):565-574. doi:10.1172/JCI129197
16. Hegde H, Shimpi N, Panny A, Glurich I, Christie P, Acharya A. MICE vs PPCA: Missing data imputation in healthcare. *Inform Med Unlocked*. 2019;17:100275. doi:10.1016/j.imu.2019.100275
17. Blonde L, Khunti K, Harris SB, Meizinger C, Skolnik NS. Interpretation and Impact of Real-World Clinical Data for the Practicing Clinician. *Adv Ther*. 2018;35(11):1763-1774. doi:10.1007/s12325-018-0805-y
18. Murray EJ, Swanson SA, Hernán MA. Guidelines for estimating causal effects in pragmatic randomized trials. *arXiv*. Preprint posted online November 19, 2019;arXiv:1911.06030. doi:10.48550/arXiv.1911.06030
19. Eichler H, Koenig F, Arlett P, et al. Are Novel, Nonrandomized Analytic Methods Fit for Decision Making? The Need for Prospective, Controlled, and Transparent Validation. *Clin Pharmacol Ther*. 2020;107(4):773-779. doi:10.1002/cpt.1638
20. Association of the British Pharmaceutical Industry. Demonstrating value with real world data: a practical guide. In: 2016. <https://www.abpi.org.uk/media/wugogbxq/2011-06-13-abpi-guidance-de-monstrating-value-with-real-world-data.pdf>
21. Makady A, Goettsch W. Review of Policies And Perspectives on Real-World Data for Drug Development and Assessment (Imi-Getreal Deliverable). *Value Health*. 2015;18(7):A567. doi:10.1016/j.jval.2015.09.1863
22. Liu F, Panagiotakos D. Real-world data: a brief review of the methods, applications, challenges and opportunities. *BMC Med Res Methodol*. 2022;22(1):287. doi:10.1186/s12874-022-01768-6
23. Chishtie J, Sapiro N, Wiebe N, et al. Use of Epic Electronic Health Record System for Health Care Research: Scoping Review. *J Med Internet Res*. 2023;25:e51003. doi:10.2196/51003
24. Sun W, Cai Z, Li Y, Liu F, Fang S, Wang G. Data Processing and Text Mining Technologies on Electronic Medical Records: A Review. *J Healthc Eng*. 2018;2018:1-9. doi:10.1155/2018/4302425
25. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE J Biomed Health Inform*. 2018;22(5):1589-1604. doi:10.1109/JBHI.2017.2767063
26. Subrahmanya SVG, Shetty DK, Patil V, et al. The role of data science in healthcare advancements: applications, benefits, and future prospects. *Ir J Med Sci* 1971 -. 2022;191(4):1473-1483. doi:10.1007/s11845-021-02730-z
27. Bartlett VL, Dhruva SS, Shah ND, Ryan P, Ross JS. Feasibility of Using Real-World Data to Replicate Clinical Trial Evidence. *JAMA Netw Open*. 2019;2(10):e1912869. doi:10.1001/jamanetworkopen.2019.12869
28. Veturi Y, Lucas A, Bradford Y, et al. A unified framework identifies new links between plasma lipids and diseases from electronic medical records across large-scale cohorts. *Nat Genet*. 2021; 53(7):972-981. doi:10.1038/s41588-021-00879-y
29. Javaid M, Haleem A, Singh RP. Health informatics to enhance the healthcare industry's culture: An extensive analysis of its features, contributions, applications and limitations. *Inform Health*. 2024;1(2):123-148. doi:10.1016/j.infoh.2024.05.001
30. Bazo R, Da Costa CA, Seewald LA, et al. A Survey About Real-Time Location Systems in Healthcare Environments. *J Med Syst*. 2021;45(3):35. doi:10.1007/s10916-021-01710-1
31. Simon BD, Ozyoruk KB, Gelikman DG, Harmon SA, Türkbeý B. The future of multimodal artificial intelligence models for integrating imaging and clinical metadata: a narrative review. *Diagn Interv Radiol*. Published online October 2, 2024. doi:10.4274/dir.2024.242631
32. Curtis LH, Sola-Morales O, Heidt J, et al. Regulatory and HTA Considerations for Development of REAL-WORLD Data Derived External Controls. *ClinPharmacolTher*. 2023;114(2):303-315. doi:10.1002/cpt.2913
33. Al-Sahab B, Leviton A, Loddenkemper T, Paneth N, Zhang B. Biases in Electronic Health Records Data for Generating Real-World Evidence: An Overview. *J Healthc Inform Res*. 2024; 8(1):121-139. doi:10.1007/s41666-023-00153-2
34. Segal JB, Varadhan R, Groenwold RHH, et al. Assessing Heterogeneity of Treatment Effect in Real-World Data. *Ann Intern Med*. 2023;176(4):536-544. doi:10.7326/M22-1510
35. Moler Zapata S. *Methods to Address Confounding and Heterogeneity in Cost-Effectiveness Analyses Using Real-World Data*. London School of Hygiene & Tropical Medicine; 2023. doi:10.17037/PUBS.04671315
36. Monti S, Grosso V, Todoerti M, Caporali R. Randomized controlled trials and real-world data: differences and similarities to untangle literature data. *Rheumatology*. 2018;57(Supplement_7):vii54-vii58. doi:10.1093/rheumatology/key109
37. Linden A, Yarnold PR. Combining machine learning and propensity score weighting to estimate

- causal effects in multivalued treatments. *J Eval Clin Pract.* 2016;22(6):875-885. doi:10.1111/jep.12610
38. Prosperi M, Guo Y, Sperrin M, et al. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nat Mach Intell.* 2020;2(7):369-375. doi:10.1038/s42256-020-0197-y
 39. Cui P, Shen Z, Li S, et al. Causal Inference Meets Machine Learning. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* ACM; 2020:3527-3528. doi:10.1145/3394486.3406460
 40. Tsai WC, Tsai YC, Kuo KC, et al. Natural language processing and network analysis in patients withdrawing from life-sustaining treatments: a retrospective cohort study. *BMC Palliat Care.* 2022;21(1):225. doi:10.1186/s12904-022-01119-8
 41. Arora P, Boyne D, Slater JJ, Gupta A, Brenner DR, Druzdel MJ. Bayesian Networks for Risk Prediction Using Real-World Data: A Tool for Precision Medicine. *Value Health.* 2019;22(4):439-445. doi:10.1016/j.jval.2019.01.006
 42. Ray J, Johnny O, Trovati M, Sotiriadis S, Bessis N. The Rise of Big Data Science: A Survey of Techniques, Methods and Approaches in the Field of Natural Language Processing and Network Theory. *Big Data and Cognitive Computing.* 2018; 2(3):22. <https://doi.org/10.3390/bdcc2030022>
 43. Liu F. Data Science Methods for Real-World Evidence Generation in Real-World Data. *Annu Rev Biomed Data Sci.* 2024;7(1):201-224. doi:10.1146/annurev-biodatasci-102423-113220
 44. Leist AK, Klee M, Kim JH, et al. Mapping of machine learning approaches for description, prediction, and causal inference in the social and health sciences. *Sci Adv.* 2022;8(42):eabk1942. doi:10.1126/sciadv.abk1942
 45. Weberpals J, Becker T, Davies J, et al. Deep Learning-based Propensity Scores for Confounding Control in Comparative Effectiveness Research: A Large-scale, Real-world Data Study. *Epidemiology.* 2021;32(3):378-388. doi:10.1097/EDE.0000000000001338
 46. Crown WH. Real-World Evidence, Causal Inference, and Machine Learning. *Value Health.* 2019;22(5):587-592. doi:10.1016/j.jval.2019.03.001
 47. Keith KA, Jensen D, O'Connor B. Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates. Published online 2020. doi:10.48550/ARXIV.2005.00649
 48. Zhao Y, Yu Y, Wang H, et al. Correction to: Machine Learning in Causal Inference: Application in Pharmacovigilance. *Drug Saf.* 2022;45(8):927-927. doi:10.1007/s40264-022-01199-8
 49. Feder A, Keith KA, Manzoor E, et al. Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond. *Trans Assoc Comput Linguist.* 2022;10:1138-1158. doi:10.1162/tacl_a_00511
 50. Kammoun A, Slama R, Tabia H, Ouni T, Abid M. Generative Adversarial Networks for face generation: A survey. *ACM Comput Surv.* Published online March 31, 2022:1122445.1122456. doi:10.1145/1122445.1122456
 51. Wood-Doughty Z. BALANCING THE ASSUMPTIONS OF CAUSAL INFERENCE AND NATURAL LANGUAGE PROCESSING. Published online August 2021. <https://jscholarship.library.jhu.edu/server/api/core/bitstreams/1ff558b0-700c-4c93-8914-70dbc90241fc/content>
 52. Zou KH, Salem LA, Ray A, eds. *Real-World Evidence in a Patient-Centric Digital Era.* First edition. CRC Press; 2023. doi:10.1201/9781003017523
 53. Fleurence R, Wang X, Bian J, et al. Generative AI in Health Economics and Outcomes Research: A Taxonomy of Key Definitions and Emerging Applications, an ISPOR Working Group Report. *arXiv.* Preprint posted online 2024. doi:10.48550/ARXIV.2410.20204
 54. Chahal CAA, Alahdab F, Asatryan B, et al. Data Interoperability and Harmonization in Cardiovascular Genomic and Precision Medicine. *Circ Genomic Precis Med.* 2025;18(3). doi:10.1161/CIRCGEN.124.004624
 55. Chirikov V, Kroep S. RWD197 Literature Review on the Use of Synthetic Data and AI Advances for Patient-Centered, Sustainable HEOR. *Value Health.* 2024;27(12):S612. doi:10.1016/j.jval.2024.10.3754
 56. Jafar R. Artificial Intelligence Applications in HEOR. Cytel. August 6, 2024. <https://cytel.com/perspectives/artificial-intelligence-applications-in-heor/>
 57. Zou KH, Li JZ. Enhanced Patient-Centricity: How the Biopharmaceutical Industry Is Optimizing Patient Care through AI/ML/DL. *Healthcare.* 2022;10(10):1997. doi:10.3390/healthcare10101997
 58. Vargas-Santiago M, León-Velasco DA, Maldonado-Sifuentes CE, Chanona-Hernandez L. A State-of-the-Art Review of Artificial Intelligence (AI) Applications in Healthcare: Advances in Diabetes, Cancer, Epidemiology, and Mortality Prediction. *Computers.* 2025;14(4):143. doi:10.3390/computers14040143