REVIEW

Open Access

Real-world data: a brief review of the methods, applications, challenges and opportunities

Fang Liu[1]* and Demosthenes Panagiotakos[2]

Abstract

Background The increased adoption of the internet, social media, wearable devices, e-health services, and other technology-driven services in medicine and healthcare has led to the rapid generation of various types of digital data, providing a valuable data source beyond the confines of traditional clinical trials, epidemiological studies, and lab-based experiments.

Methods We provide a brief overview on the type and sources of real-world data and the common models and approaches to utilize and analyze real-world data. We discuss the challenges and opportunities of using real-world data for evidence-based decision making This review does not aim to be comprehensive or cover all aspects of the intriguing topic on RWD (from both the research and practical perspectives) but serves as a primer and provides useful sources for readers who interested in this topic.

Results and Conclusions Real-world hold great potential for generating real-world evidence for designing and conducting confirmatory trials and answering questions that may not be addressed otherwise. The voluminosity and complexity of real-world data also call for development of more appropriate, sophisticated, and innovative data processing and analysis techniques while maintaining scientific rigor in research findings, and attentions to data ethics to harness the power of real-world data.

Keywords Real-world data (RWD), Real-world evidence (RWE), Electronic health records, Machine learning, Artificial intelligence, Causal inference

Introduction

Per the definition by the US FDA, real-world data (RWD) in the medical and healthcare field "are the data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources"[1]. The wide usage of the internet, social media, wearable devices and mobile devices, claims and billing activities, (disease) registries, electronic health records (EHRs), product and disease registries, e-health services, and other technology-driven services, together with increased capacity in data storage, have led to the rapid generation and availability of digital RWD [2].

The increasing accessibility of RWD and the fast development of artificial intelligence (AI) and machine learning (ML) techniques, together with rising costs and recognized limitations of traditional trials, has spurred great interest in the use of RWD to enhance the efficiency of clinical research and discoveries and bridge the evidence gap between clinical research and practice. For example, during the COVID-19 pandemic, RWD are used to generate or aid the generation of real-world evidence (RWE) on the effectiveness of COVID-19

*Correspondence:
Fang Liu
fliu2@nd.edu
[1] Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, 46530 Notre Dame, IN, USA
[2] School of Health Sciences and Education, Harokopio University, Athens, Greece
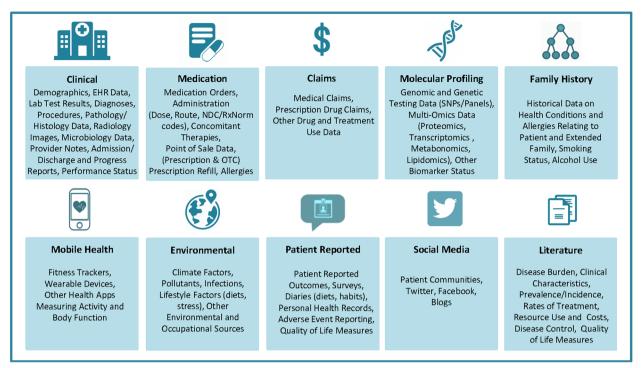
Fig. 1 RWD Types and Sources (source: Fig. 1 in [16] with written permission by Dr. Brandon Swift to use the figure)

vaccination [3–5], to model localized COVID-19 control strategies [6], to characterize COVID-19 and flu using data from smartphones and wearables [7], to study behavioral and mental health changes in relation to the lockdown of public life [8], and to assist in decision and policy making, among others.

In what follows, we provide a brief review on the type and sources of RWD (Section 2) and the common models and approaches to utilize and analyze RWD (Section 3) , and discuss the challenges and opportunities of using RWD for evidence-based decision making (Section 4). This review does not aim to be comprehensive or cover all aspects of the intriguing topic on RWD (from both the research and practical perspectives) but serves as a primer and provides useful sources for readers who interested in this topic.

Characteristics, types and applications of RWD
RWD have several characteristics as compared to data collected from randomized trials in controlled settings. First, RWD are observational as opposed to data gathered in a controlled setting. Second, many types of RWD are unstructured (e.g., texts, imaging, networks) and at times inconsistent due to entry variations across providers and health systems. Third, RWD may be generated in a high-frequency manner (e.g., measurements at the millisecond level from wearables), resulting in

voluminous and dynamic data. Fourth, RWD may be incomplete and lack key endpoints for an analysis given that the original collection is not for such a purpose. For example, claims data usually do not have clinical endpoints; registry data have limited follow-ups. Fifth, RWD may be subject to bias and measurement errors (random and non-random). For example, data generated from the internet, mobile devices, and wearables can be subject to selection bias; a RWD dataset is a unrepresentative sample of the underlying population that a study intends to understand; claims data are known to contain fraudulent values. In summary, RWD are messy, incomplete, heterogeneous, and subject to different types of measurement errors and biases. A systematic scoping review of the literature suggests data quality of RWD is not consistent, and as a result quality assessments are challenging due to the complex and heterogeneous nature of these data. The sub-optimal data quality of RWD is well recognized [9–12]; how to improve it (e.g. regulatory-grade) is work in progress [13–15].

There are many different types of RWD. Figure 1 [16] provides a list of the RWD types and sources in medicine. We also refer readers to [11] for a comprehensive overview of the RWD data types. Here we use a few common RWD types, i.e., EHRs, registry data, claims data, patient-reported outcome (PRO) data, and data collected

from wearables, as examples to demonstrate the variety of RWD and how they can be used for what purposes.

EHRs are collected as part of routine care across clinics, hospitals, and healthcare institutions. EHR data are typical RWD – noisy, heterogeneous, structured, and unstructured (e.g., text, imaging), and dynamic and require careful and intensive efforts pre-processing [17]. EHRs have created unprecedented opportunities for data-driven approaches to learn patterns, make new discoveries, assist preoperative planning, diagnostics, clinical prognostication, among others [18–27], improve predictions in selected outcomes especially if linked with administrative and claim data and usage of proper machine learning techniques [27–30], and validate and replicate findings from clinical trials [31].

Registry data have various types. For example, product registries include patients who have been exposed to a biopharmaceutical product or a medical device; health services registries consist of patients who have had a common procedure or hospitalization; and disease registries contains information about people diagnosed with a specific type of disease. Registries data enable identification and sharing best clinical practices, improve accuracy of estimates, provide valuable data for supporting regulatory decision-making [32–35]. Especially for rare diseases where clinical trials are often of small size and data are subject to high variability, registries provide a valuable data source to help understand the course of a disease, and provide critical information for confirmatory clinical trial design and translational research to develop treatments and improve patient care [34, 36, 37]. Reader may refer to [38] for a comprehensive overview on registry data and how they help understanding of patient outcomes.

Claims data refer to data generated during processing healthcare claims in health insurance plans or from practice management systems. Despite that claims data are collected and stored primarily for payment purposes originally, they have been used in healthcare to understand patients' and prescribes' behavior and how they interact, to estimate disease prevalence, to learn disease progression, disease diagnosis, medication usage, and drug-drug interactions, and validate and replicate findings from clinical trials [31, 39–46]. A known pitfall of claim data is fraud, on top of some of the common data characteristics of RMD, such as upcoding[1][47]. The data fraud problem can be mitigated with detailed audits and

adoption of modern statistical, data mining and ML techniques for fraud detection [48–51].

PRO data refer to data reported directly by patients on their health status. PRO data have been used to provide RWE on effectiveness of interventions, symptoms monitoring, relationships between exposure and outcomes, among others [52–55]. PRO data are subject to recall bias and large inter-individual variability.

Wearable devices generate continuous streams of data. When combined with contextual data (e.g., location data, social media), they provide an opportunity to conduct expansive research studies that are large in scale and scope [56] that would be otherwise infeasible in controlled trials. Examples of using wearable RWD to generate RWE include applications in neuroscience and environmental health [57–60]. The wearables generate huge amounts of data. Advances in data storage, real-time processing capabilities and efficient battery technology would be essential for the full utilization of wearable data.

Using and analyzing RWD

A wide range of research methods are available to make use of RWD. In what follows, we outline a few approaches, including pragmatic clinical trials, target trial emulation, and applications of ML and AI techniques.

Pragmatic clinical trials are trials designed to test the effectiveness of an intervention in the real-world clinical setting. Pragmatic trials leverage the increasingly integrated healthcare system and may use data from EHR, claims, patient reminder systems, telephone-based care, etc. Due to the data characteristics of RWD, new guidelines and methodologies are developed to mitigate bias in RWE generated by RWD for decision making and causal inference, especially for per-protocol analysis [61, 62]. The research question under investigation in pragmatic trials is whether an intervention works in real life and trials are designed to maximize the applicability and generalizability of the intervention. Various types of outcomes can be measured in these trials, but mostly patient-centered, instead of typical measurable symptoms or markers in explanatory trials. For example, ADAPTABLE trial [63, 64] is a high-profile pragmatic trial and is the first large-scale, EHR-enabled clinical trial conducted within the U.S. It used EHR data to identify around 450,000 patients with established atherosclerotic cardiovascular disease (CVD) for recruitment and eventually enrolled about 15,000 individuals at 40 clinical centers that were randomized to two aspirin dose arms. Electronic patient follow-up for patient-reported outcomes was completed every 3 to 6 months, with a median follow-up was 26.2 months to determine the optimal dosage of aspirin in CVD patients, with the primary endpoint being the

---

[1] Upcoding refers to instances in which a medical service provider obtains additional reimbursement from insurance by coding a service it provided as a more expensive service than what was actually performed

composite of all-cause mortality, hospitalization for non-fatal myocardial infarction, or hospitalization for a non-fatal stroke. The cost of ADATABLE is estimated to be only 1/5 to 1/2 of a traditional RCT of that scale.

Target trial emulation is the application of trial design and analysis principles from (target) randomized trials to the analysis of observational data [65]. By precisely specifying the target trial's inclusion/exclusion criteria, treatment strategies, treatment assignment, causal contrast, outcomes, follow-up period, and statistical analysis, one may draw valid causal inferences about an intervention from RWD. Target trial emulation can be an important tool especially when comparative evaluation is not yet available or feasible in randomized trials. For example, [66] employs target trial emulation to evaluate real-world COVID-19 vaccine effectiveness, measured by protection against COVID-19 infection or related death, in racially and ethnically diverse, elderly populations by comparing newly vaccinated persons with matched unvaccinated controls using data from the US Department of Veterans Affairs health care system. The simulated trial was conducted with clearly defined inclusion/exclusion criteria, identification of matched controls, including matching based on propensity scores with careful selection of model covariates. Target trial emulation has also been used to evaluate the effect of colon cancer screening on cancer incidence over eight years of follow up [67], and the risk of urinary tract infection among diabetic patients [68].

RWD can also be used as historical controls and reference groups for controlled trials, with assessment of the quality and appropriateness of the RWD and employment of proper statistical approaches for analyzing the data [69]. Controlling for selection bias and confounding is key to the validity of this approach because of the lack of randomization and potentially unrecognized baseline differences, and the control group needs to be comparable with the treated group. RWD also provide a great opportunity to study rare events given the data voluminousness [70–72]. These studies also highlight the need for improving the RWD data quality, developing surrogate endpoints, and standardizing data collection for outcome measures in registries.

In terms of analysis of RWD, statistical models and inferential approaches are necessary for making sense of RWD, obtaining causal relationships, testing/validating hypotheses, and generating regulatory-grade RWE to inform policymakers and regulators in decision making – just as in the controlled trial settings. In fact, the motivation for and the design and analysis principles in pragmatic trials and target trial emulation are to obtain causal inference, with more innovative methods beyond the traditional statistical methods to adjust for potential confounders and improve the capabilities of RWD for causal inference [73–76].

ML techniques are getting increasingly popular and are powerful tools for predictive modeling. One reason for their popularity is that the modern ML techniques are very capable of dealing with voluminous, messy, multi-modal, and various unstructured data types without strong assumptions about the distribution of data. For example, deep learning can learn abstract representations of large, complex, and unstructured data; natural language processing (NLP) and embedding methods can be used to process texts and clinical notes in EHRs and transform them to real-valued vectors for downstream learning tasks. Secondly, new and more powerful ML techniques are being developed rapidly, due to the high demand and the large group of researchers in the field attracted by the hot topic. Thirdly, there are also many open source codes (e.g., on Github) and software libraries (e.g., TensorFlow, Pytorch, Keras) out there to facilitate the implementation of these techniques. Indeed, ML has enjoyed a rapid surge in the last decade or so for a wide range of applications in RWD, outperforming more conventional approaches [77–85]. For example, ML is widely applied in in health informatics to generate RWE and formulate personalized healthcare [86–90] and was successfully employed on RWD collected during the COVID-19 pandemic to help understand the disease and evaluate its prevention and treatment strategies [91–95]. It should be noted that the ML techniques are largely used for predictions and classification (e.g., disease diagnosis), variable selections (e.g, biomarker screening), data visualization, etc, rather than generating regulatory-level RWE; but this may change soon as regulatory agencies are aggressively evaluating ML/AI for generating RWE and engaging stakeholders on the topic [96–99].

It would be more effective and powerful to combine the expertise from statistical inference and ML when it comes to generating RWE and learning causal relationships. One of the recent methodological developments is indeed in that direction – leveraging the advances in semi-parametric and empirical process theory and incorporating the benefits of ML into comparative effectiveness using RWD. A well-known framework is targeted learning [100–102] that has been successfully applied in causal inference for dynamic treatment rules using EHR data [103] and efficacy of COVID-19 treatments [104], among others.

Regardless of which area a RWD project focuses on – causal inference or prediction and classification, representativeness of RWD of the population where the conclusions from the RWD project will be generalized to is critical. Otherwise, estimation or prediction can be misleading or even harmful. The information in RWD

might not be adequate to validate the appropriateness of the data for generalization; in that case, the investigators should resist the temptation to generalize to groups that they are unsure about.

Challenges and opportunities

Various challenges – from data gathering to data quality control to decision making – still exist in all stages of a RWD life cycle despite all the excitement around their transformative potentials. We list some of the challenges below, where plenty of opportunities for improvement exist and greater efforts are needed to harness the power of RWD.

Data quality: RWD are now often used for other purposes than what they are originally collected for and thus may lack information for critical endpoints and not always be positioned for generating regulatory-grade evidence. On top of that, RWD are messy, heterogeneous, and subject to various measurement errors, all of which contribute to the lower quality of RWD compared to data from controlled trials. As a result, accuracy and precision of results based on RWD are negatively impacted and misleading results or false conclusions can be generated. While these do not preclude the use of RWD in evidence generation and decision making, data quality issues need to be consistently documented and addressed as much as possible through data cleaning and pre-processing (e.g., imputation to fill in missing values, over-sampling for imbalanced data, denoising, combining disparate pieces of information across databases, etc). If an issue can be addressed during the pre-processing stage, efforts should be made to correct it during data analysis or caution should be used when interpreting the results. Early engagement of key stakeholders (e.g., regulatory agencies if needed, research institutes, industries etc.) are encouraged to establish data quality standards and reduce unforeseen risks and issues.

Efficient and practical ML and statistical procedures: Fast growth of digital medical data and the fact that workforce and investment flood into the field also drive the rapid development and adoption of modern statistical procedures and ML algorithms to analyze the data. The availability of open-source platforms and software greatly facilitate the application of the procedures in practice. On the other hand, noisiness, heterogeneity, incompleteness, and unbalancedness of RWD may cause considerable under-performance of the existing statistical and ML procedures and demand new procedures that target specifically at RWD and can be effectively deployed in the real world. Further, the availability of the open-source platform and software and the accompanied convenience, while offered with good intentions, also increases the chance of practitioners misusing the procedures, if

not equipped with proper training or understanding the principles of the techniques before applying them to real-world situations. In addition, to maintain scientific rigor during the RWE generation process from RWD, results from statistical and ML procedures would require medical validation either using expert knowledge or conducting reproducibility and replicability studies before they are being used for decision making in the real world [105].

Explainability and interpretability: Modern ML approaches are often employed in a black-box fashion and there a lack of understanding of the relationships between input and output and causal effects. Model selection, parameter initialization, and hyper-parameter tuning are also often conducted in a trial-and-error manner, without domain expert input. This is in contrast to the medical and healthcare field where interpretability is critical to building patient/user trust, and doctors are unlikely to use technology that they don't understand. Promising and encouraging research work on this topic has already started [106–111], but more research is warranted.

Reproducibility and replicability: Reproducibility and replicability[2] are major principles in scientific research, RWD included. If an analytical procedure is not robust and its output is not reproducible or replicable, the public would call into questions the scientific rigor of the work and doubt the conclusion from a RWD-based study [113–115]. Result validation, reproducibility, and replicability can be challenging given their messiness, incompleteness, unstructured data, but need to be established especially considering that the generated evidence could be used towards regulatory decisions and affect the lives of millions of people. Irreproducibility can be mitigated by sharing raw and processed data and codes, assuming no privacy is compromised in this process. For replicability, given that RWD are not generated from controlled trials and every data set may has its own unique data characteristics, complete replicability can be difficult or even infeasible. Nevertheless, detailed documentation of data characteristics and pre-processing, pre-registration of analysis procedures, and adherence to open science principles (e.g., code repositories [116]) are critical for replicating findings on different RWD datasets, assuming they come from the same underlying population. Readers may refer to [117–119] for more suggestions and discussions on this topic.

---

[2] Reproducibility refers to "instances in which the original researcher's data and computer codes are used to regenerate the results" and replicability refers to "instances in which a researcher collects new data to arrive at the same scientific findings as a previous study." [112]
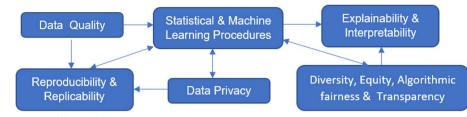
Fig. 2 Challenges in RWD and Their Relations

Privacy: Ethical issues exist when an RWD project is implemented, among which, privacy is a commonly discussed topic. Information in RWD is often sensitive, such as medical histories, disease status, financial situations, and social behaviors, among others. Privacy risk can increase dramatically when different databases (e.g., EHR, wearables, claims) are linked together, a common practice in the analysis of RWD. Data users and policymakers should make every effort to ensure that RWD collection, storage, sharing, and analysis follow established data privacy principles (i.e., lawfulness, fairness, purpose limitation, and data minimization). In addition, privacy-enhancing technology and privacy-preserving data sharing and analysis can be deployed, where there already exist plenty effective and well-accepted state-of-the-art concepts and approaches, such as differential privacy[3][120] and federated learning[4][121, 122]. Investigators and policymakers may consider integrating these concepts and technology when collecting and analyzing RWD and disseminating the results and RWE from the RWD.

Diversity, Equity, Algorithmic fairness, and Transparency (DEAT): DEAT is another important ethical issue to consider in an RWD project. RWD may contain information from various demographic groups, which can be used to generate RWE with improved generalizability compared to data collected in controlled settings. On the other hand, certain types of RWD may be heavily biased and unbalanced toward a certain group, not as diverse or inclusive, and in some cases, even exacerbate disparity (e.g., wearables and access to facilities and treatment may be limited to certain demographic groups). Greater effort will be needed to gain access to RWD from underrepresented groups and to effectively take into account the heterogeneity in RWD while being mindful of the limitation for diversity/equity. This topic also relates to

algorithmic fairness, which aims at understanding and preventing bias in ML models. Algorithmic fairness is an increasingly popular research topic in literature [123–127]. Incorrect and misleading conclusions may be drawn if the trained models systematically disadvantage a certain group (e.g., a trained algorithm might be less likely to detect cancer in black patients than white patients or in men than women). Transparency means that information and communication concerning the processing of personal data must be easily accessible and easy to understand. Transparency ensures that data contributors are aware of how their data are being used and for what purposes and decision-makers can evaluate the quality of the methods and the applicability of the generated RWE [128–131]. Being transparent when working with RWD is critical for building trust among the key stakeholders during an RWD life cycle (individuals who supply the data, those who collect and manage the data, data curators who design studies and analyze the data, and decision and policy makers).

The above challenges are not isolated but rather connected as depicted in Fig. 2. Data quality affects the performance of statistical and ML procedures; data sources and the cleaning and pre-processing process relate to result reproducibility and replicability. How data are analyzed and which statistical and ML procedures to use have an impact on reproducibility and replicability, whether privacy-preserving procedures are used during data collected and analysis and how information is shared and released relate to data privacy, DEAT, and explainability and interpretability, which can in turns affect which ML procedures to apply and development of new ML techniques.

Conclusions

RWD provide a valuable and rich data source beyond the confines of traditional epidemiological studies, clinical trials, and lab-based experiments, with lower cost in data collection compared to the latter. If used and analyzed appropriately, RWD have the potential to generate valid and unbiased RWE with savings in both cost and time, compared to controlled trials, and to enhance the efficiency of medical and health-related research and

---

[3] Differential privacy provides a mathematically rigorous framework in which randomized procedures are used to guarantee individual privacy when releasing information.

[4] Federated learning enables local devices to collaboratively learn a shared model while keeping all training data on the local devices without sharing, mitigating privacy risks.

decision-making. Procedures that improve the quality of the data and overcome the limitation of RWD to make the best of them have been and will continue to be developed. With the enthusiasm, commitment, and investment in RWD from all key stakeholders, we hope that the day that RWD unleashes its full potential will come soon.

Abbreviations

AI	artificial intelligence
CVD	cardiovascular disease
COVID	coronavirus disease
DEAT	diversity, equity, algorithmic fairness, and transparency
EHR	electronic health records
ML	machine learning
NLP	natural language processing
PRO	patient-reported outcome
RWD	real-world data
RWE	real-world evidence

References

1. US Food and Drug Administration, et al. Real-World Evidence. 2022. https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence. Accessed 1 Sep 2022.
2. Wikipedia. Real world data. 2022. https://en.wikipedia.org/wiki/Real_world_data. Accessed 19 Mar 2022.
3. Powell AA, Power L, Westrop S, McOwat K, Campbell H, Simmons R, et al. Real-world data shows increased reactogenicity in adults after heterologous compared to homologous prime-boost COVID-19 vaccination, March- June 2021, England. Eurosurveillance. 2021;26(28):2100634.
4. Hunter PR, Brainard JS. Estimating the effectiveness of the Pfizer COVID-19 BNT162b2 vaccine after a single dose. A reanalysis of a study of 'real-world' vaccination outcomes from Israel. medRxiv. 2021.02.01.21250957. https://doi.org/10.1101/2021.02.01.21250957.
5. Henry DA, Jones MA, Stehlik P, Glasziou PP. Effectiveness of COVID-19 vaccines: findings from real world studies. Med J Aust. 2021;215(4):149.
6. Firth JA, Hellewell J, Klepac P, Kissler S, Kucharski AJ, Spurgin LG. Using a real-world network to model localized COVID-19 control strategies. Nat Med. 2020;26(10):1616–22.
7. Shapiro A, Marinsek N, Clay I, Bradshaw B, Ramirez E, Min J, et al. Characterizing COVID-19 and influenza illnesses in the real world via person-generated health data. Patterns. 2021;2(1):100188.
8. Ahrens KF, Neumann RJ, Kollmann B, Plichta MM, Lieb K, Tüscher O, et al. Differential impact of COVID-related lockdown on mental health in Germany. World Psychiatr. 2021;20(1):140.
9. Hernández MA, Stolfo SJ. Real-world data is dirty: Data cleansing and the merge/purge problem. Data Min Knowl Disc. 1998;2(1):9–37.
10. Corrigan-Curay J, Sacks L, Woodcock J. Real-world evidence and real-world data for evaluating drug safety and effectiveness. Jama. 2018;320(9):867–8.
11. Makady A, de Boer A, Hillege H, Klungel O, Goettsch W, et al. What is real-world data? A review of definitions based on literature and stakeholder interviews. Value Health. 2017;20(7):858–65.
12. Franklin JM, Schneeweiss S. When and how can real world data analyses substitute for randomized controlled trials? Clin Pharmacol Ther. 2017;102(6):924–33.
13. Miksad RA, Abernethy AP. Harnessing the power of real-world evidence (RWE): a checklist to ensure regulatory-grade data quality. Clin Pharmacol Ther. 2018;103(2):202–5.
14. Curtis MD, Griffith SD, Tucker M, Taylor MD, Capra WB, Carrigan G, et al. Development and validation of a high-quality composite real-world mortality endpoint. Health Serv Res. 2018;53(6):4460–76.
15. Booth CM, Karim S, Mackillop WJ. Real-world data: towards achieving the achievable in cancer care. Nat Rev Clin Oncol. 2019;16(5):312–25.
16. Swift B, Jain L, White C, Chandrasekaran V, Bhandari A, Hughes DA, et al. Innovation at the intersection of clinical trials and real-world data science to advance patient care. Clin Transl Sci. 2018;11(5):450–60.
17. 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;2018:4302425. https://doi.org/10.1155/2018/4302425.
18. Wu J, Roy J, Stewart WF. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Med Care. 2010;48(6 Suppl):S106-13. https://doi.org/10.1097/MLR.0b013e3181de9e17, https://pubmed.ncbi.nlm.nih.gov/20473190/.
19. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary use of EHR: data quality issues and informatics opportunities. Summit Transl Bioinforma. 2010;2010:1.
20. Kawaler E, Cobian A, Peissig P, Cross D, Yale S, Craven M. Learning to predict post-hospitalization VTE risk from EHR data. In: AMIA annual symposium proceedings. vol. 2012. p. 436. American Medical Informatics Association Country United States.
21. 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. 2017;22(5):1589–604.
22. Poirier C, Hswen Y, Bouzillé G, Cuggia M, Lavenu A, Brownstein JS, et al. Influenza forecasting for French regions combining EHR, web and climatic data sources with a machine learning ensemble approach. PloS ONE. 2021;16(5):e0250890.
23. Zheng T, Xie W, Xu L, He X, Zhang Y, You M, et al. A machine learning-based framework to identify type 2 diabetes through electronic health records. Int J Med Inform. 2017;97:120–7.
24. Pivovarov R, Perotte AJ, Grave E, Angiolillo J, Wiggins CH, Elhadad N. Learning probabilistic phenotypes from heterogeneous EHR data. J Biomed Inform. 2015;58:156–65.
25. Zhao D, Weng C. Combining PubMed knowledge and EHR data to develop a weighted bayesian network for pancreatic cancer prediction. J Biomed Informa. 2011;44(5):859–68.
26. Veturi Y, Lucas A, Bradford Y, Hui D, Dudek S, Theusch E, 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):972–81.
27. Kwon BC, Choi MJ, Kim JT, Choi E, Kim YB, Kwon S, et al. Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records. IEEE Trans Vis Comput Graph. 2018;25(1):299–309.

28. Mahmoudi E, Kamdar N, Kim N, Gonzales G, Singh K, Waljee AK. Use of electronic medical records in development and validation of risk prediction models of hospital readmission: systematic review. BMJ. 2020;369:m958.

29. Desai RJ, Wang SV, Vaduganathan M, Evers T, Schneeweiss S. Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. JAMA Netw Open. 2020;3(1):e1918962.

30. Huang L, Shea AL, Qian H, Masurkar A, Deng H, Liu D. Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. J Biomed Inform. 2019;99:103291.

31. 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.

32. Dreyer NA, Garner S. Registries for robust evidence. Jama. 2009;302(7):790–1.

33. Larsson S, Lawyer P, Garellick G, Lindahl B, Lundström M. Use of 13 disease registries in 5 countries demonstrates the potential to use outcome data to improve health care's value. Health Affairs. 2012;31(1):220–7.

34. McGettigan P, Alonso Olmo C, Plueschke K, Castillon M, Nogueras Zondag D, Bahri P, et al. Patient registries: an underused resource for medicines evaluation. Drug Saf. 2019;42(11):1343–51.

35. Izmirly PM, Parton H, Wang L, McCune WJ, Lim SS, Drenkard C, et al. Prevalence of systemic lupus erythematosus in the United States: estimates from a meta-analysis of the Centers for Disease Control and Prevention National Lupus Registries. Arthritis Rheumatol. 2021;73(6):991–6.

36. Jansen-Van Der Weide MC, Gaasterland CM, Roes KC, Pontes C, Vives R, Sancho A, et al. Rare disease registries: potential applications towards impact on development of new drug treatments. Orphanet J Rare Dis. 2018;13(1):1–11.

37. Lacaze P, Millis N, Fookes M, Zurynski Y, Jaffe A, Bellgard M, et al. Rare disease registries: a call to action. Intern Med J. 2017;47(9):1075–9.

38. Gliklich RE, Dreyer NA, Leavy MB, editors. Registries for Evaluating Patient Outcomes: A User's Guide. 3rd ed. Rockville (MD): Agency for Healthcare Research and Quality (US); 2014 Apr. Report No.: 13(14)-EHC111. PMID: 24945055.

39. Svarstad BL, Shireman TI, Sweeney J. Using drug claims data to assess the relationship of medication adherence with hospitalization and costs. Psychiatr Serv. 2001;52(6):805–11.

40. Izurieta HS, Wu X, Lu Y, Chillarige Y, Wernecke M, Lindaas A, et al. Zostavax vaccine effectiveness among US elderly using real-world evidence: Addressing unmeasured confounders by using multiple imputation after linking beneficiary surveys with Medicare claims. Pharmacoepidemiol Drug Saf. 2019;28(7):993–1001.

41. Allen AM, Van Houten HK, Sangaralingham LR, Talwalkar JA, McCoy RG. Healthcare cost and utilization in nonalcoholic fatty liver disease: real-world data from a large US claims database. Hepatology. 2018;68(6):2230–8.

42. Sruamsiri R, Iwasaki K, Tang W, Mahlich J. Persistence rates and medical costs of biological therapies for psoriasis treatment in Japan: a real-world data study using a claims database. BMC Dermatol. 2018;18(1):1–11.

43. Quock TP, Yan T, Chang E, Guthrie S, Broder MS. Epidemiology of AL amyloidosis: a real-world study using US claims data. Blood Adv. 2018;2(10):1046–53.

44. Herland M, Bauder RA, Khoshgoftaar TM. Medical provider specialty predictions for the detection of anomalous medicare insurance claims. In: 2017 IEEE international conference on information reuse and integration (IRI). New York City: IEEE; 2017. p. 579–88.

45. Momo K, Kobayashi H, Sugiura Y, Yasu T, Koinuma M, Kuroda SI. Prevalence of drug–drug interaction in atrial fibrillation patients based on a large claims data. PLoS ONE. 2019;14(12):e0225297.

46. Ghiani M, Maywald U, Wilke T, Heeg B. RW1 Bridging The Gap Between Clinical Trials And Real World Data: Evidence On Replicability Of Efficacy Results Using German Claims Data. Value Health. 2020;23:S757–8.

47. Silverman E, Skinner J. Medicare upcoding and hospital ownership. J Health Econ. 2004;23(2):369–89.

48. Kirlidog M, Asuk C. A fraud detection approach with data mining in health insurance. Procedia-Soc Behav Sci. 2012;62:989–94.

49. Li J, Huang KY, Jin J, Shi J. A survey on statistical methods for health care fraud detection. Health Care Manag Sci. 2008;11(3):275–87.

50. Viaene S, Dedene G, Derrig RA. Auto claim fraud detection using Bayesian learning neural networks. Expert Syst Appl. 2005;29(3):653–66.

51. Phua C, Lee V, Smith K, Gayler R. A comprehensive survey of data mining-based fraud detection research. arXiv preprint arXiv:1009.6119. 2010.

52. Roche N, Small M, Broomfield S, Higgins V, Pollard R. Real world COPD: association of morning symptoms with clinical and patient reported outcomes. COPD J Chronic Obstructive Pulm Dis. 2013;10(6):679–86.

53. Small M, Anderson P, Vickers A, Kay S, Fermer S. Importance of inhaler-device satisfaction in asthma treatment: real-world observations of physician-observed compliance and clinical/patient-reported outcomes. Adv Ther. 2011;28(3):202–12.

54. Pinsker JE, Müller L, Constantin A, Leas S, Manning M, McElwee Malloy M, et al. Real-world patient-reported outcomes and glycemic results with initiation of control-IQ technology. Diabetes Technol Ther. 2021;23(2):120–7.

55. Touma Z, Hoskin B, Atkinson C, Bell D, Massey O, Lofland JH, Berry P, Karyekar CS, Costenbader KH. Systemic lupus erythematosus symptom clusters and their association with Patient-Reported outcomes and treatment: analysis of Real-World data. Arthritis Care & Research. 2022;74(7):1079-88.

56. Martinez GJ, Mattingly SM, Mirjafari S, Nepal SK, Campbell AT, Dey AK, et al. On the quality of real-world wearable data in a longitudinal study of information workers. In: 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops). New York City: IEEE; 2020. p. 1–6.

57. Christensen JH, Saunders GH, Porsbo M, Pontoppidan NH. The everyday acoustic environment and its association with human heart rate: evidence from real-world data logging with hearing aids and wearables. Royal Soc Open Sci. 2021;8(2):201345.

58. Johnson KT, Picard RW. Advancing neuroscience through wearable devices. Neuron. 2020;108(1):8–12.

59. Pickham D, Berte N, Pihulic M, Valdez A, Mayer B, Desai M. Effect of a wearable patient sensor on care delivery for preventing pressure injuries in acutely ill adults: A pragmatic randomized clinical trial (LS-HAPI study). Int J Nurs Stud. 2018;80:12–9.

60. Adams JL, Dinesh K, Snyder CW, Xiong M, Tarolli CG, Sharma S, et al. A real-world study of wearable sensors in Parkinson's disease. NPJ Park Dis. 2021;7(1):1–8.

61. Hernán MA, Robins JM, et al. Per-protocol analyses of pragmatic trials. N Engl J Med. 2017;377(14):1391–8.

62. Murray EJ, Swanson SA, Hernán MA. Guidelines for estimating causal effects in pragmatic randomized trials. arXiv preprint arXiv:1911.06030. 2019.

63. Hernandez AF, Fleurence RL, Rothman RL. The ADAPTABLE Trial and PCORnet: shining light on a new research paradigm. Ann Intern Med. 2015;163(8):635-6.

64. Baigent C. Pragmatic trials-need for ADAPTABLE design. N Engl J Med. 2021;384(21).

65. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. Am J Epidemiol. 2016;183(8):758–64.

66. Ioannou GN, Locke ER, O'Hare AM, Bohnert AS, Boyko EJ, Hynes DM, et al. COVID-19 vaccination effectiveness against infection or death in a National US Health Care system: a target trial emulation study. Ann Intern Med. 2022;175(3):352–61.

67. García-Albéniz X, Hsu J, Hernán MA. The value of explicitly emulating a target trial when using real world evidence: an application to colorectal cancer screening. Eur J Epidemiol. 2017;32(6):495–500.

68. Takeuchi Y, Kumamaru H, Hagiwara Y, Matsui H, Yasunaga H, Miyata H, et al. Sodium-glucose cotransporter-2 inhibitors and the risk of urinary tract infection among diabetic patients in Japan: Target trial emulation using a nationwide administrative claims database. Diabetes Obes Metab. 2021;23(6):1379–88.

69. Jen EY, Xu Q, Schetter A, Przepiorka D, Shen YL, Roscoe D, et al. FDA approval: blinatumomab for patients with B-cell precursor acute lymphoblastic leukemia in morphologic remission with minimal residual disease. Clin Cancer Res. 2019;25(2):473–7.

70. Gross AM. Using real world data to support regulatory approval of drugs in rare diseases: A review of opportunities, limitations & a case example. Curr Probl Cancer. 2021;45(4):100769.

71. Wu J, Wang C, Toh S, Pisa FE, Bauer L. Use of real-world evidence in regulatory decisions for rare diseases in the United States—Current status and future directions. Pharmacoepidemiol Drug Saf. 2020;29(10):1213–8.

72. Hayeems RZ, Michaels-Igbokwe C, Venkataramanan V, Hartley T, Acker M, Gillespie M, et al. The complexity of diagnosing rare disease: An organizing framework for outcomes research and health economics based on real-world evidence. Genet Med. 2022;24(3):694–702.

73. Hernán MA, Robins JM. Causal inference. Boca Raton: CRC; 2010.

74. Ho M, van der Laan M, Lee H, Chen J, Lee K, Fang Y, et al. The current landscape in biostatistics of real-world data and evidence: Causal inference frameworks for study design and analysis. Stat Biopharm Res. 2021. https://www.tandfonline.com/doi/abs/10.1080/19466315.2021. 1883475.

75. Crown WH. Real-world evidence, causal inference, and machine learning. Value Health. 2019;22(5):587–92.

76. Cui P, Shen Z, Li S, Yao L, Li Y, Chu Z, et al. Causal inference meets machine learning. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: Association for Computing Machinery; 2020. p. 3527–3528.

77. Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RK, Hua Y, Gueroussov S, Najafabadi HS, Hughes TR, Morris Q. The human splicing code reveals new insights into the genetic determinants of disease. Science. 2015;347(6218):1254806.

78. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. 2015;31(5):761–3.

79. Anthimopoulos M, Christodoulidis S, Ebner L, Christe A, Mougiakakou S. Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. IEEE Trans Med Imaging. 2016;35(5):1207–16.

80. Van Grinsven MJ, van Ginneken B, Hoyng CB, Theelen T, Sánchez CI. Fast convolutional neural network training using selective data sampling: Application to hemorrhage detection in color fundus images. IEEE Trans Med Imaging. 2016;35(5):1273–84.

81. Kleesiek J, Urban G, Hubert A, Schwarz D, Maier-Hein K, Bendszus M, et al. Deep MRI brain extraction: A 3D convolutional neural network for skull stripping. NeuroImage. 2016;129:460–9.

82. Gibson E, Li W, Sudre C, Fidon L, Shakir DI, Wang G, et al. NiftyNet: a deep-learning platform for medical imaging. Comput Methods Prog Biomed. 2018;158:113–22.

83. Coccia M. Deep learning technology for improving cancer care in society: New directions in cancer imaging driven by artificial intelligence. Technol Soc. 2020;60:101198.

84. Bien N, Rajpurkar P, Ball RL, Irvin J, Park A, Jones E, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. PLoS Med. 2018;15(11):e1002699.

85. Johansson FD, Collins JE, Yau V, Guan H, Kim SC, Losina E, et al. Predicting response to tocilizumab monotherapy in rheumatoid arthritis: a real-world data analysis using machine learning. J Rheumatol. 2021;48(9):1364–70.

86. Ravì D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, et al. Deep learning for health informatics. IEEE J Biomed Health Informa. 2016;21(1):4–21.

87. Suzuki K. Overview of deep learning in medical imaging. Radiol Phys Technol. 2017;10(3):257–73.

88. Shen D, Wu G, Suk HI. Deep learning in medical image analysis. Annu Rev Biomed Eng. 2017;19:221–48.

89. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88.

90. Lee JG, Jun S, Cho YW, Lee H, Kim GB, Seo JB, et al. Deep learning in medical imaging: general overview. Korean J Radiol. 2017;18(4):570–84.

91. Amyar A, Modzelewski R, Li H, Ruan S. Multi-task deep learning based CT imaging analysis for COVID-19 pneumonia: Classification and segmentation. Comput Biol Med. 2020;126:104037.

92. Oh Y, Park S, Ye JC. Deep learning covid-19 features on cxr using limited training data sets. IEEE Trans Med Imaging. 2020;39(8):2688–700.

93. Hemdan EED, Shouman MA, Karar ME. Covidx-net: A framework of deep learning classifiers to diagnose covid-19 in x-ray images. arXiv preprint arXiv:2003.11055. 2020.

94. Wang S, Zha Y, Li W, Wu Q, Li X, Niu M, et al. A fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis. Eur Respir J. 2020;56(2).

95. Ardakani AA, Kanafi AR, Acharya UR, Khadem N, Mohammadi A. Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: Results of 10 convolutional neural networks. Comput Biol Med. 2020;121:103795.

96. Food U, Administration D. Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device(SaMD) - Discussion Paper and Request for Feedback. 2019. https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf. Accessed 24 Mar 2022.

97. Food U, Administration D. Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. 2021. https://www.fda.gov/media/145022/download. Accessed 24 March 2022.

98. of Medicines Regulatory Authorities IC. Informal Innovation Network Horizon Scanning Assessment Report - Artificial Intelligence. 2021. https://www.icmra.info/drupal/sites/default/files/2021-08/horizon_scanning_report_artificial_intelligence.pdf. Accessed 24 March 2022.

99. Agency EM. Artificial intelligence in medicine regulation. 2021. https://www.ema.europa.eu/en/news/artificial-intelligence-medicine-regulation. Accessed 24 Mar 2022.

100. Van der Laan MJ, Rose S. Targeted learning: causal inference for observational and experimental data. 2011. Springer-Verlag New York Inc., United States.

101. Van der Laan MJ, Rose S. Targeted learning in data science. Causal Inference for Complex Longitudinal Studies 2018. Cham: Springer.

102. van der Laan MJ, Luedtke AR. Targeted learning of the mean outcome under an optimal dynamic treatment rule. J Causal Infer. 2015;3(1):61–95.

103. Sofrygin O, Zhu Z, Schmittdiel JA, Adams AS, Grant RW, van der Laan MJ, et al. Targeted learning with daily EHR data. Stat Med. 2019;38(16):3073–90.

104. Chakravarti P, Wilson A, Krikov S, Shao N, van der Laan M. PIN68 Estimating Effects in Observational Real-World Data, From Target Trials to Targeted Learning: Example of Treating COVID-Hospitalized Patients. Value Health. 2021;24:S118.

105. Eichler HG, Koenig F, Arlett P, Enzmann H, Humphreys A, Pétavy F, 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):773–9.

106. Chakraborty S, Tomsett R, Raghavendra R, Harborne D, Alzantot M, Cerutti F, et al. Interpretability of deep learning models: A survey of results. In: 2017 IEEE smartworld, ubiquitous intelligence & computing, advanced & trusted computed, scalable computing & communications, cloud & big data computing, Internet of people and smart city innovation. New York City: IEEE; 2017. p. 1–6.

107. Zhang Q, Zhu SC. Visual interpretability for deep learning: a survey. arXiv preprint arXiv:1802.00614. 2018.

108. Hohman F, Park H, Robinson C, Chau DHP. Summit: Scaling deep learning interpretability by visualizing activation and attribution summarizations. IEEE Trans Vis Comput Graph. 2019;26(1):1096–106.

109. Ghoshal B, Tucker A. Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection. arXiv preprint arXiv:2003.10769. 2020.

110. Raghu M, Gilmer J, Yosinski J, Sohl-Dickstein J. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. 2017; 31st Conference on Neural Information Processing Systems (NIPS 2017). Long Beach: NEURAL INFO PROCESS SYS F, LA JOLLA; 2017. ISBN: 9781510860964.

111. Cruz-Roa AA, Ovalle JEA, Madabhushi A, Osorio FAG. A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection. In: International

Conference on Medical Image Computing and Computer-Assisted Intervention. Cham; Springer; 2013. p. 403–410.

112. Barba LA. Terminologies for reproducible research. arXiv preprint arXiv:1802.03311. 2018.

113. Stupple A, Singerman D, Celi LA. The reproducibility crisis in the age of digital medicine. NPJ Digit Med. 2019;2(1):1–3.

114. Carter RE, Attia ZI, Lopez-Jimenez F, Friedman PA. Pragmatic considerations for fostering reproducible research in artificial intelligence. NPJ Digit Med. 2019;2(1):1–3.

115. Liu C, Gao C, Xia X, Lo D, Grundy J, Yang X. On the replicability and reproducibility of deep learning in software engineering. ACM Transactions on Software Engineering and Methodology. 2021;31(1):1–46.

116. Springate DA, Kontopantelis E, Ashcroft DM, Olier I, Parisi R, Chamapiwa E, et al. ClinicalCodes: an online clinical codes repository to improve the validity and reproducibility of research using electronic medical records. PloS ONE. 2014;9(6):e99825.

117. Wang SV, Schneeweiss S, Berger ML, Brown J, de Vries F, Douglas I, et al. Reporting to improve reproducibility and facilitate validity assessment for healthcare database studies V1. 0. Value health. 2017;20(8):1009–22.

118. Panagiotou OA, Heller R. Inferential challenges for real-world evidence in the era of routinely collected health data: many researchers, many more hypotheses, a single database. JAMA Oncol. 2021;7(11):1605–7.

119. Belbasis L, Panagiotou OA. Reproducibility of prediction models in health services research. BMC Res Notes. 2022;15(1):1–5.

120. Dwork C, McSherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis. In: Theory of cryptography conference. Springer; 2006. p. 265–284.

121. Konečnỳ J, McMahan B, Ramage D. Federated optimization: Distributed optimization beyond the datacenter. arXiv preprint arXiv:1511.03575. 2015.

122. Konečnỳ J, McMahan HB, Yu FX, Richtárik P, Suresh AT, Bacon D. Federated learning: Strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492. 2016.

123. McCradden MD, Joshi S, Mazwi M, Anderson JA. Ethical limitations of algorithmic fairness solutions in health care machine learning. Lancet Digit Health. 2020;2(5):e221–3.

124. Mitchell S, Potash E, Barocas S, D'Amour A, Lum K. Algorithmic fairness: Choices, assumptions, and definitions. Ann Rev Stat Appl. 2021;8:141–63.

125. Mhasawade V, Zhao Y, Chunara R. Machine learning and algorithmic fairness in public and population health. Nat Mach Intell. 2021;3(8):659–66.

126. Wong PH. Democratizing algorithmic fairness. Philos Technol. 2020;33(2):225–44.

127. Paulus JK, Kent DM. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. NPJ Digit Med. 2020;3(1):1–8.

128. Orsini LS, Berger M, Crown W, Daniel G, Eichler HG, Goettsch W, et al. Improving transparency to build trust in real-world secondary data studies for hypothesis testing—why, what, and how: recommendations and a road map from the real-world evidence transparency initiative. Value Health. 2020;23(9):1128–36.

129. Patorno E, Schneeweiss S, Wang SV. Transparency in real-world evidence (RWE) studies to build confidence for decision-making: reporting RWE research in diabetes. Diabetes Obes Metab. 2020;22:45–59.

130. White R. Building trust in real-world evidence and comparative effectiveness research: the need for transparency. Future Med. 2017;6(1):5–7.

131. Rodriguez-Villa E, Torous J. Regulating digital health technologies with transparency: the case for dynamic and multi-stakeholder evaluation. BMC Med. 2019;17(1):1–5.

Publisher's Note