REVIEW

Harnessing Real-World Data for Regulatory Use and Applying Innovative Applications

Kelly H Zou [1]
Jim Z Li [2]
Joseph Imperato [3]
Chandrashekhar N Potkar [4]
Nikuj Sethi [5]
Jon Edwards [6]
Amrit Ray [7]

[1]Research, Development and Medical, Upjohn Division, Pfizer Inc, New York, NY 10017, USA; [2]Research, Development and Medical, Upjohn Division, Pfizer Inc, San Diego, CA 92121, USA; [3]Business Technology, Upjohn Division, Pfizer Inc, New York, NY 10017, USA; [4]Research, Development and Medical, Pfizer Gulf FZ LLC, Dubai Media City, United Arab Emirates; [5]Business Technology, Upjohn Division, Pfizer Inc, Collegeville, PA 19426, USA; [6]Envision Pharma Group, Envision House, Horsham RH12 1XQ, UK; [7]Research, Development and Medical, Upjohn Division, Pfizer Inc, Collegeville, PA 19426, USA

Correspondence: Kelly H Zou
Vice President and Head of Medical, Analytics and Insights, Research, Development and Medical, Upjohn Division, Pfizer Inc, 235 East 42nd Street, MS 235-9-1, New York, NY 10017, USA
Email KellyZou@pfizer.com

Abstract: A vast quantity of real-world data (RWD) are available to healthcare researchers. Such data come from diverse sources such as electronic health records, insurance claims and billing activity, product and disease registries, medical devices used in the home, and applications on mobile devices. The analysis of RWD produces real-world evidence (RWE), which is clinical evidence that provides information about usage and potential benefits or risks of a drug. This review defines and explains RWD, and it also details how regulatory authorities are using RWD and RWE. The main challenges in harnessing RWD include collating and analyzing numerous disparate types or categories of available information including both structured (eg, field entries) and unstructured (eg, doctor notes, discharge summaries, social media posts) data. Although the use of artificial intelligence to capture, amalgamate, standardize, and analyze RWD is still evolving, it has the potential to support the increased use of RWE to improve global health and healthcare.
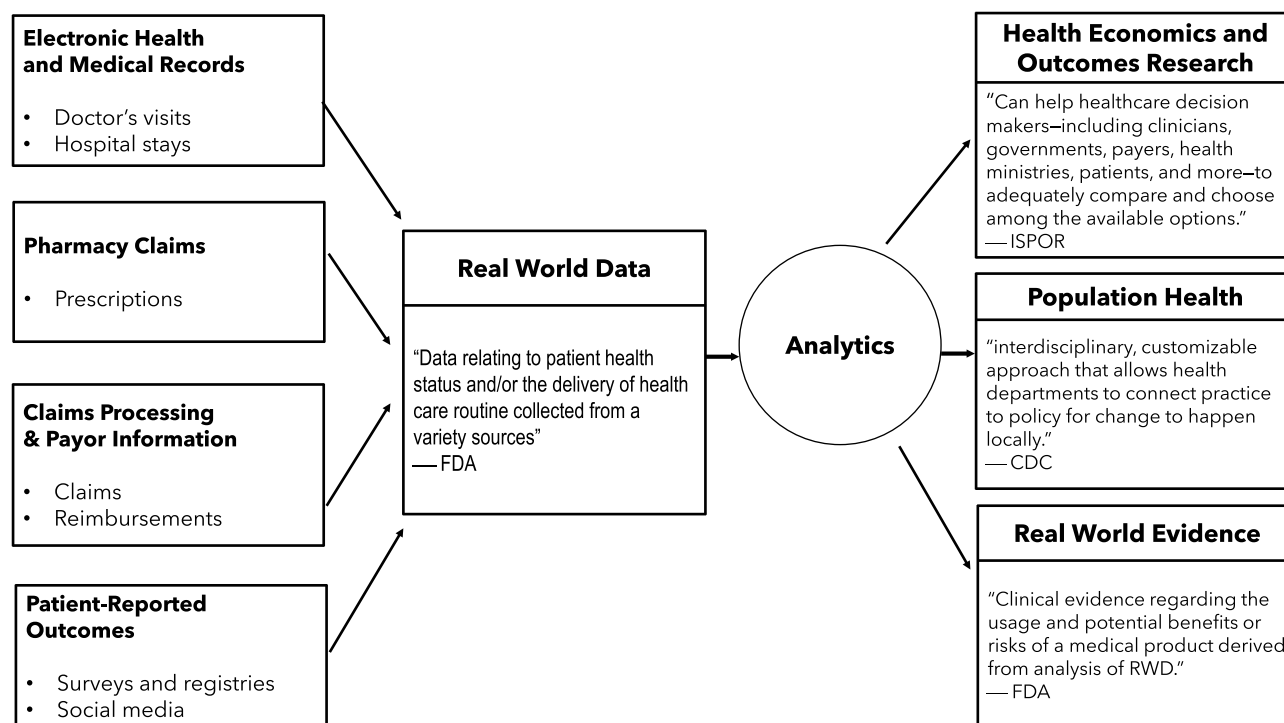
Keywords: real-world data, real-world evidence, regulatory, artificial intelligence, robotic process automation

Introduction

In the healthcare industry, real-world data (RWD) comprise data regarding patient health and/or healthcare delivery, obtained outside of randomized, controlled trials (RCTs).[1,2] RWD may be derived from electronic health records (EHRs), insurance claims and billing activity, product and disease registries, medical devices used in the home, and applications (apps) on mobile devices (Figure 1). The nature and sources of RWD provide the potential for large volumes of data to be collected, often in near-real time (eg, data streamed from mobile devices), that can provide information at granular or large-scale level. For example, the MyHeart Counts Cardiovascular Health Study, which collected data from 48,968 individuals, demonstrated that a study using smartphones to measure physical activity and cardiovascular health is feasible and provides a means of collecting large amounts of data in real time.[3] The analysis of RWD produces real-world evidence (RWE), which is clinical evidence that provides information about the usage and potential benefits or risks of a drug product.[4]

RWE may be generated from different study designs or analyses, including randomized (non-controlled) trials, large simple trials, pragmatic trials (ie, trials that incorporate elements of routine clinical practice), and prospective or retrospective observational studies.[5] RWE can be derived from a variety of sources. Therefore, its use provides an opportunity to combine diverse datasets to provide a broader understanding than might be available through RCTs.[2]

Figure 1 From real-world data to real-world evidence, key capabilities such as health economics outcomes research, real world evidence and population health.
Abbreviations: FDA, US Food and Drug Administration; WHO, World Health Organization.

In order to obtain reliable and useful evidence, RWD must be collated and analyzed using accurate and robust algorithms tailored to the specific contexts. The responsible use and application of RWD requires addressing issues around data standardization, integration, quality control, access, privacy, and security. For example, RWD may be available from multiple repositories and registries, with each source storing data in separate clouds (ie, siloed) and/or in different languages.

Data from multiple sources may include disparate types or categories of information and may include both structured (eg, field entries) and unstructured (eg, doctor notes, discharge summaries, social media posts) information.[6] Data aggregation and translation are, therefore, necessary for analyses to be performed on such data, and therefore database or repository systems must be interoperable and varied data formats must be translated and integrated. Data analysis must also take into account the level of data accuracy (eg, correct identification of all relevant patients, absence of data entry errors), miscoding of claims, and variations across health systems in data coding practices.[7] To leverage data and gain actionable insights, artificial intelligence (AI) capabilities such as machine learning, deep learning, and other algorithms can be useful.

This review article serves as an introduction to readers who would like to learn more about RWE and the potential for AI to aid the capture and analysis of RWD. Additional information and details on how RWE can provide key information on "the safety and effectiveness of a medication in large, heterogeneous populations" is provided in an earlier review.[8]

RWE for Regulatory Purposes

Across the globe, regulatory agencies and pharmaceutical companies increasingly embrace RWE and AI for its ability to support a more comprehensive picture of health.[9] In the United States, the Food and Drug Administration (FDA) 21st Century Cures Act, passed in 2016,[4] and the Prescription Drug User Fee Act VI (PDUFA VI)[10] have both expanded the opportunities for the use of RWE in the healthcare and pharmaceutical industries.

Although data from RCTs and the synthesis of these data in systematic reviews and meta-analyses remain the gold standard for evidence, due to the restrictions of trial design and patient inclusion criteria, the data produced in such trials are not necessarily applicable to real-world settings.[2] Also, RWE has the potential to help identify rare adverse events.[11] Therefore, the FDA does consider

RWD in its evaluation of medical product safety, and occasionally to inform decisions about efficacy.[12] In one case, RWD were also used in Japan to support regulatory decision-making around lorazepam when a small clinical trial failed to meet the primary endpoint due to low enrollment, yet the secondary endpoint and real-world usage supported use of the drug in the target population.[13] However, concerns remain about the use of RWE in studies in which the effectiveness (or safety signals) of a treatment are small or negligible.[11]

There has been a relatively long historical precedent for using RWD for safety purposes,[14] while the FDA has more recently established a framework for the use of RWD for generating evidence of effectiveness. A set of guidance documents has been developed by the agency, listed chronologically below:

- Use of RWE to support regulatory decision-making for medical devices (August 2017)[15]
- Use of EHR data in clinical investigations: Guidance for Industry (July 2018)[16]
- Framework for FDA's real-world evidence program (December 2018)[17]
- Submitting documents using RWD and RWE to FDA for drugs and biologics: Guidance for Industry (May 2019).[18]

Additional perspectives outside of the United States include:

- Regulatory perspective on RWE in scientific advice (European Medicines Agency; April 2018)[19]
- Key considerations in using RWE to support drug development (China Center for Drug Evaluation; May 2019).[20]

From a regulatory perspective, RWE may now be used to support label expansion (including new indications, dosages, and patient sub-populations), to assess the feasibility of Phase IV studies, to provide evidence for products in an expedited approval pathway, to provide a historical control arm for clinical studies, and in pragmatic trials.[5] Indeed, there is a growing list of drug approvals and labelling changes that have utilized RWE and data from non-traditional study designs in their regulatory submissions.[21] For example, an expansion in the indication of palbociclib was supported by RWD from EHR and insurance claims.[21,22] Such changes are the result of ever-greater

amounts of health-related data being stored by computers, mobile and wearable devices, biosensors, and medical devices.[1] A survey conducted in 2017 among life science companies showed that more than half (54%) of respondents were providing significant investment to increase their RWE programs,[23] thus showing how the use of RWE will likely grow in the near future. Subsequent surveys have demonstrated that 90% of respondents from pharma companies are looking to build their RWE capabilities across all of the product life cycle, through preclinical, clinical, approval, and post-approval stages. However, less than half (45%) of pharma companies felt that they had these sort of end-to-end RWE capabilities currently.[24] The most recent survey (conducted in January and February 2020) revealed that 94% of respondents believe that the use of RWE in research and development will be important or very important to their organization in the next three years. The importance of RWE was thought likely to move from post-approval value demonstration to research and development, notably in supporting regulatory filings and augmenting clinical trials.[25] The opportunity to harness RWD for regulatory purposes will further grow as AI experts refine the necessary software and statisticians develop new algorithms and analytical skills that refine the methods of data analysis.

For RWE to be efficiently used for regulatory purposes, certain conditions must be in place and standards met. For example, standards must be established regarding acceptable data quality, sources, and analytical methods. A clearer understanding is also needed around the ways in which RWE can be used to inform about alternative dosing schedules, benefits for patient subpopulations, and targeted/precision medicine strategies. From a practical perspective, basic infrastructure must be in place to collect RWD in any countries that would use RWE. This is particularly important in developing countries, in which EHRs or claims information may not be widely used or where patient registries collect disparate information. Similarly, trials conducted in real-world settings require sufficient patient participation to generate meaningful RWD. Where RWD are available, advanced, validated analytical methodologies are also needed to generate accurate and informative RWE. The use of standardized methods for both collecting and analyzing data will improve reproducibility and comparability of real-world studies. The drive for more standardization has been led by organizations such as the Observational Health Data Sciences and Informatics group (OHDSI).[26] The OHDSI has

produced a common-data model, which is used to encode and store clinical data in a standardized manner. Use of common-data models allows the same question to be addressed consistently across different databases and countries, thus making large-scale international observational research feasible.[27] Finally, the findings of RWE analyses and related information must be communicated in a credible and comprehensible way to patients, the medical community, and any other stakeholders (eg, policymakers).

How Can Artificial Intelligence Be Used to Generate RWE?

To obtain high-quality RWE that is acceptable for regulatory use and other decision-making applications, RWD must be collected in a large quantity, collated into an analyzable format, and a certain level of data accuracy and reliability ensured. To support these efforts, AI capabilities are increasingly being applied to the analysis of RWD.[24,28] Furthermore, in a 2018 survey, 60% of pharma industry respondents were using AI in their RWE programs and 95% anticipated utilizing AI for this purpose in the coming years.[24] AI constitutes a combination of self-learning capabilities that mimic the way the human brain works, with the intent of replicating human decision-making and interactions.

Natural language processing (NLP), machine learning, and robotic process automation (RPA) are three AI capabilities that are particularly applicable to the generation of RWE for the healthcare industry.[6] Natural language processing—the computerized interpretation and organization of human language—includes text classification, recognition of syntax, interpretation of word meaning based on location in a sentence, and language translation.[9] Machine learning, which includes a variety of predictive statistical and mathematical modeling techniques (Table 1),[9,29] is often layered onto NLP to reinterpret and correct initial assumptions following repeated usage.[30] RPA involves software that automates repeated tasks, helping to speed the processes at the same time as reducing human error and oversight.[31,32] RPA software could also be used to capture, interpret, and process patient-level data, as well as trigger responses and communicate with other digital systems.[33]

In the context of analyzing RWD to generate RWE, RPA can be used to extract large amounts of data from EHRs, insurance claims data, apps, and social media. However, much RWD is stored in unstructured formats, from which the relevant data must be interpreted and stored in a new, structured format.[6] The potential for AI tools to extract and translate large amounts of unstructured RWD for RWE generation is huge. To achieve this, RPA can be combined with natural language processing and machine learning to produce more uniform datasets containing only the information relevant to the analyses. For example, RPA might be employed to access clinical narratives in EHRs. However, because clinical narratives contain unstructured data, an NLP approach could be used to extract relevant information and translate the text (including physician shorthand) into structured field entries.[34] Machine learning algorithms and RPA might then be used to rapidly analyze the data from multiple perspectives. Applying a similar approach across a variety of data sources and therapeutic areas, previously unseen trends may emerge and new insights may be gained into aspects of patient health, drug product usage and safety, and healthcare resource utilization.

We are not aware of scenarios or studies in which all three elements (RPA, NLP, and machine learning algorithms) have been used to analyze RWE. However, the individual elements are already being incorporated in research settings. For example, a recent study of myopia in Chinese school-aged children provides an example of machine learning applied to RWD.[35] In this study, RWD were collected from electronic medical records across eight ophthalmic centers. The random forest method of machine learning was used to develop an algorithm that would predict myopia among children as young as 3 years of age. The algorithm was refined to improve accuracy, weeding out unnecessary data and resolving heterogeneous data, and then internally and externally validated. A separate study used hospital episode statistics from England's National Health Service to develop a predictive machine learning algorithm to screen patients at risk of idiopathic pulmonary arterial hypertension.[36] In that study, a multidisciplinary team that included both clinicians and AI experts collaborated to develop the algorithm throughout numerous testing and validation stages, ultimately achieving 99.99% specificity and 14.10% sensitivity. In examples such as these, the use of machine learning allows, in the investigators' words, the combination of "enormous numbers of predictors in a non-linear and highly interactive way".[35] However, before machine learning approaches can be relied upon to produce accurate RWE, the algorithms must be tested, refined, and retested in real-world situations to ensure high accuracy.[37]

Table 1 Ten Machine Learning Algorithm Techniques[29]

Algorithm	Description
Supervised Learning Algorithms	**These techniques use training data to learn mapping functions that convert input data to output data. Permits new (non-training) output information to be generated from new inputs.**
1. Linear Regression	Determines the potential relationship between input and output variables used for continuous, as opposed to categorical, variables.
2. Logistic Regression	Similar to linear regression but analyzes categorical variables (ie, dichotomous information).
3. Classification and Regression Trees (CART)	Utilizes a decision tree to make predictions for categorical variables, starting with a root node (the first decision), working through internal nodes, and ending with a leaf node.
4. Naïve Bayes	Calculates the probability of an event occurring given that another event has already happened. Bayes' theorem is utilized. Naïve refers to the assumption that variables are not associated with or independent to each other, which is often not the case in the real world, hence this technique is described as naïve.
5. K-Nearest Neighbors (KNN)	Uses the entire dataset as the training set rather than having training and test data sets. When an outcome for a new data instance is required in this method, the entire data set is evaluated for the closest match to the new data point (nearest neighbor). This is then used to generate new data.
Unsupervised Learning Algorithms	**These techniques are used when only input variables are available. There are no output variables. Unlabeled training data are used to model the structure of the results.**
6. Apriori algorithm	Generates the rules of association in data sets, eg, data items that frequently occur together in the data set. For example, if a person purchases gift wrapping paper they will often purchase sellotape or sticky tape.
7. K-means	Groups the data into clusters, with centroids calculated for K-clusters. The data points assigned to clusters have the smallest distance between the centroid and all of the other data points.
8. Principal Component Analysis (PCA)	Reduces the number of variables to make the data more straightforward to evaluate. The factor (component) that contributes to greatest degree of variation in the data set is the principal component. The second component captures all of the remaining variance in the data set, which is not correlated with the first component.
Reinforcement Learning Algorithms	**These techniques take actions that will maximize reward. Reinforcement algorithms often learn optimal actions through trial and error.**
9. Bagging with Random Forests	Constructs multiple models using a bootstrap sampling method in which random subsamples of the original data set are selected to create training data sets. These samples are the same size as the original data, but some values repeat, while other values are excluded. A bagging procedure is then used in which multiple models are created using the same algorithm on the training data sets. The next element is the random forest process in which a random selection of factors is used to find the best factors of where to split.
10. Adaptive Boosting	Builds a series of models independently to correct for the misclassification of data from the prior model.

Challenges with Using AI and RWD

Although AI shows promise for application in the healthcare industry, the diverse, complex, and observational nature of RWD presents challenges for data analysis.[9] For example, claims data are typically generated for insurance billing purposes, not adjudicated in terms of data quality, and medical errors can exist. Refinement, or training, of AI algorithms is essential to improving the accuracy of RWD analysis, and both studies described above focused on this aspect of the methodology. However, researchers must first ensure that the data obtained are complete and relevant to the condition, patient population, and treatment analyzed.[28] For example, unstructured data may contain relevant information for only certain sub-populations or information may be entered for some patients but not others. Even structured data pose challenges in the application of AI to RWD, as field-entry data may be entered using inconsistent terms, may be formatted differently between sources, or may be incomplete or contain errors.[7] Any of these situations might lead to inaccuracy in the analyses and/or data being rejected by the algorithms.

Low data recall (incorrect inclusion or exclusion of data from an analysis) also appears to pose a great challenge to

the use of AI in RWD analysis. Algorithms that are not optimized might identify "false-positive" and "false-negative" data, for example by including or excluding patients in analysis who do not actually meet the intended criteria.[28] A recent retrospective study of EHR data for patients with cardiovascular disease illustrates the difficulty that analysts encounter in algorithm development. The study results suggested that, despite the researchers taking various approaches toward generating the algorithms, the final model did not accurately analyze structured data and medication information was not sufficiently captured from unstructured data.[28] In addition, analysts must consider that data and relevant variables may change over time, and their algorithms may also need to evolve to take this into account. To circumvent these challenges, AI algorithms must be continually refined and their limitations understood by those who interpret the resulting RWE. Initial attempts to elevate the quality of analysis might include combining multiple data sources, the consideration of doctor shorthand in language processing, and systematic comparisons of machine learning algorithms within or across therapeutic areas.[34,35,37,38]

In some countries, data aggregation may pose an equally significant challenge. Legal barriers around data privacy, eg, European Union General Data Protection Regulation (GDPR),[39] practical barriers related to data storage across multiple organizations (ie, data silos), and economic barriers involving lack of incentives for organizations to collaborate and share data, all affect the availability of RWD to which AI tools can be applied.[9] Additional challenges are likely to be encountered in this fast-moving field, as new methods, applications, and learning algorithms are utilized to capture and evaluate RWD. If challenges like these are successfully met and large-volume RWD are shown to provide sufficiently accurate and comprehensive RWE, the application of AI to RWD has the potential to shorten the timeline for clinical trial design and regulatory approval, and to uncover patterns in large sets of data that would otherwise not be observed.[40]

Collaborations and Partnerships to Harness RWD

With the growing understanding of the utility of RWE, pharmaceutical researchers have begun to partner with academia and government with the intent of advancing the use of RWD in addressing major healthcare issues.

For example, the Statistical Partnerships Among Academe, Industry, and Government Award of the American Statistical Association was established in 2002 to recognize outstanding partnerships between academia, industry, and government organizations, as well as to promote new partnerships among these organizations.[41] The award recognizes outstanding collaborations between organizations, while identifying key individual contributors. A number of companies and universities have formed fruitful public-private partnerships in statistics, analytics, and data science.[41] In addition, 80% of respondents from the pharmaceutical industry who completed a survey in early 2020, reported that they are entering into strategic partnerships to utilize new sources of RWD.[25] Furthermore, some of the studies outlined in this review have involved partnerships between pharmaceutical and academic researchers.[28,36]

In order for these partnerships to be successful, all parties must view the collaboration as a long-term relationship and work together to evolve ideas from the stage of problem identification to that of solution implementation. In the context of applying AI to RWE generation, the process might involve developing innovative statistical methodology to solve the novel statistical problems that emerge in each unique context. For example, data-driven partnerships can play an important role in improving the management of diseases in patients across diverse geographical settings.[42,43]

AI capabilities such as RPA, NLP, and machine learning will necessarily be employed to facilitate the analyses of data obtained from any studies conducted through these partnerships, with careful attention paid to the technologic and statistical methods used.

The Future of AI for RWE

Although RWE has long been used to aid the understanding of patients, health conditions, and healthcare resource usage, its use in a regulatory capacity is in its infancy. Both those who generate RWE and those who interpret and use it in a practical sense must keep in mind the limitations of the source data and analytical approaches used.[2] For this reason —similar to data obtained from RCTs—transparency of methodology and the use of methodological best practices will be essential.[9] All parties will benefit from a clear understanding of the analytical methodology and to what extent conclusions can be drawn. Because of the ready availability (and relatively low cost) of RWD relative to data from RCTs, RWE may be increasingly relied upon by both the pharmaceutical industry and regulatory agencies to inform
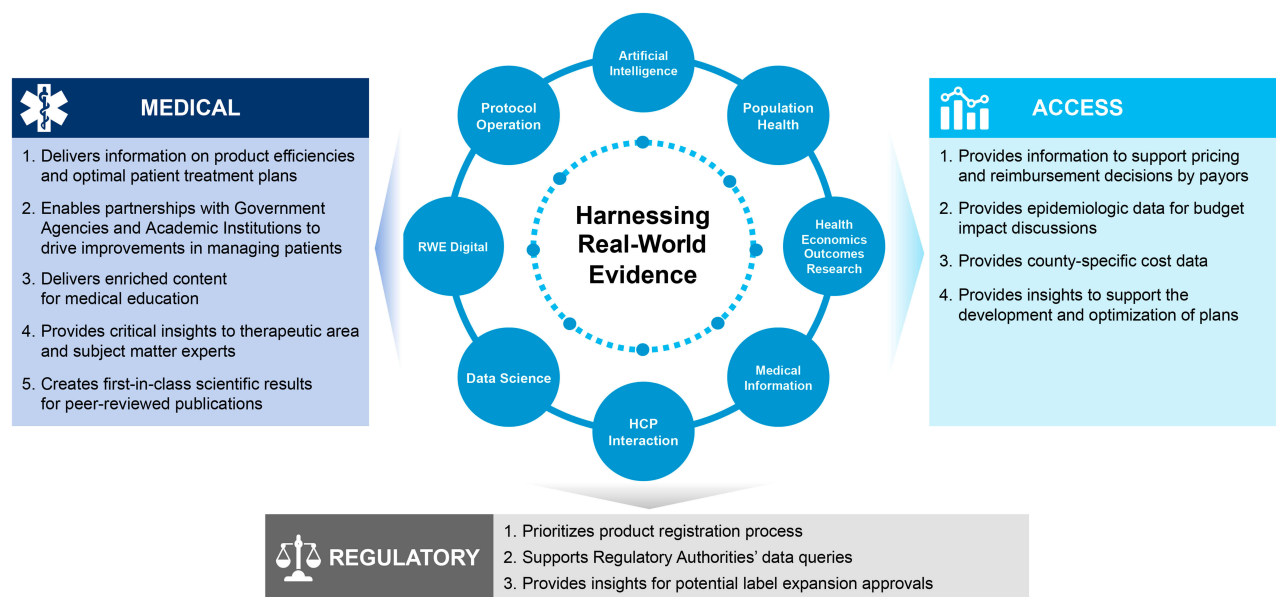
Figure 2 Benefits achieved with real-world evidence and AI.
Abbreviation: AI, artificial intelligence.

decision-makers on the clinical value and benefit of interventions.[2] However, data access, privacy, and security issues will be key needs and must be addressed with each new RWD analysis. Data quality and bias will also remain key issues for interpreting the generated RWE, and potential confounding factors must be communicated.[12,44] For regulatory purposes, early engagement with regulators will support subsequent efforts to obtain and analyze RWD. Finally, in an era of digital innovation, AI will enable extensive collection, aggregation, analyses, and interpretations to generate RWE.[24,45]

Conclusions

In practical terms, the use of AI in an era of "Big Data" and RWD is still evolving but has great potential to support the increased use of RWE to improve global health and healthcare (Figure 2). Again, the limitations of AI software and methodology will need to be considered while continual efforts are made to improve the capabilities of the AI tools. Cross-disciplinary expertise will also be necessary to ensure that the software and subsequent refinements are tailored to each analytical approach. A focus must be given to the careful application and evolution of AI technology in the context of RWD. If this is done, researchers, regulators, and other stakeholders may benefit by more clearly understanding patterns in patient treatment and behavior, disease progression, and resource usage. This will support the ultimate goal of influencing positive patient outcomes in real time.

Funding

Disclosure

Kelly H Zou, Jim Z Li, Joseph Imperato, Chandrashekhar N Potkar, Nikuj Sethi, and Amrit Ray are employees of the Upjohn Division, Pfizer Inc. Jon Edwards is an employee of Envision Pharma Group. Envision Pharma Group has a consultancy agreement with Pfizer Inc. The views expressed are their own and do not necessarily represent those of their employers. The authors report no other conflicts of interest in this work.

References

1. US Food and Drug Administration. Real-world evidence. Available from: https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence. Accessed July 7, 2020.
2. Sherman RE, Anderson SA, Dal Pan GJ, et al. Real-world evidence - what is it and what can it tell us? *N Engl J Med*. 2016;375 (23):2293–2297. doi:10.1056/NEJMsb1609216
3. McConnell MV, Shcherbina A, Pavlovic A, et al. Feasibility of obtaining measures of lifestyle from a smartphone app: the Myheart Counts Cardiovascular Health study. *JAMA Cardiol*. 2017;2(1):67–76. doi:10.1001/jamacardio.2016.4395
4. US Congress. 21st Century Cures Act. H.R. 34. Available from: https://www.congress.gov/114/bills/hr34/BILLS-114hr34enr.pdf. Accessed July 7, 2020.
5. Baumfeld Andre E, Reynolds R, Caubel P, Azoulay L, Dreyer NA. Trial designs using real-world data: the changing landscape of the regulatory approval process. *Pharmacoepidemiol Drug Saf*. 2019. doi:10.1002/pds.4932

6. Cleary M. Artificial intelligence: the key to unlocking novel real-world data? *Value Outcomes Spotlight*. 2019;5(2):16–19.

7. Khera R, Krumholz HM. With great power comes great responsibility: big data research from the national inpatient sample. *Circ Cardiovasc Qual Outcomes*. 2017;10:7. doi:10.1161/CIRCOUTCOMES.117.003846

8. Katkade VB, Sanders KN, Zou KH. Real world data: an opportunity to supplement existing evidence for the use of long-established medicines in health care decision making. *J Multidiscip Healthc*. 2018;11:295–304. doi:10.2147/JMDH.S160029

9. Matheny M, Israni ST, Ahmed M, Whicher D, editors. *Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril*. Washington, DC: National Academy of Medicine; 2019.

10. US Food and Drug Administration. PDUFA VI: fiscal years 2018-2022. Available from: https://www.fda.gov/industry/prescription-drug-user-fee-amendments/pdufa-vi-fiscal-years-20182022. Accessed July 7, 2020.

11. Collins R, Bowman L, Landray M, Peto R. The magic of randomization versus the myth of real-world evidence. *N Engl J Med*. 2020;382 (7):674–678. doi:10.1056/NEJMsb1901642

12. Alemayehu D, Cappelleri JC, Emir B, Zou KH, editors. *Statistical Topics in Health Economics and Outcomes Research*. Boca Raton, FL Chapman and Hall/CRC Press; 2017.

13. Chambers RB, Zou KH, Nakazuru Y, et al. RWE for lorazepam IV regulatory approval in Japan. Presented at Joint Statistical Meeting, Denver, CO; 2019.

14. Gottlieb S. Harnessing real world evidence for safety and innovation. https://www.fda.gov/news-events/speeches-fda-officials/harnessing-real-world-evidence-safety-and-innovation-11192018. Accessed July 7, 2020.

15. US Food and Drug Administration. Use of real-world evidence to support regulatory decision-making for medical devices: guidance for industry and Food and Drug Administration staff. Docket no. FDA-2016-D-2153. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/use-real-world-evidence-support-regulatory-decision-making-medical-devices. Accessed July 7, 2020.

16. US Food and Drug Administration. Use of electronic health record data in clinical investigations: guidance for industry. Docket no. FDA-2016-D-1224. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/use-electronic-health-record-data-clinical-investigations-guidance-industry. Accessed July 7, 2020.

17. US Food and Drug Administration. Framework for FDA's real-world evidence program. Available from: https://www.fda.gov/media/120060/download. Accessed July 7, 2020.

18. US Food and Drug Administration. Submitting documents using real-world data and real-world evidence to FDA for drugs and biologics: guidance for industry. Docket no. 2019-09529. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/submitting-documents-using-real-world-data-and-real-world-evidence-fda-drugs-and-biologics-guidance. Accessed July 7, 2020.

19. European Medicines Agency. Regulatory perspective on real world evidence (RWE) in scientific advice. EMA human scientific committees' working parties with patients' and consumers' organisations (PCWP) and healthcare professionals' organisations (HCPWP). Available from: https://www.ema.europa.eu/en/documents/presentation/presentation-regulatory-perspective-real-world-evidence-rwe-scientific-advice-emas-pcwp-hcpwp-joint_en.pdf. Accessed July 7, 2020.

20. Center for Drug Evaluation (China). *Key Considerations in Using Real-World Evidence to Support Drug Development*. NMPA; 2019.

21. Adding real-world evidence to a totality of evidence approach for evaluating marketed product effectiveness. Available from: https://healthpolicy.duke.edu/sites/default/files/u31/totality_of_evidence_appendix.pdf. Accessed Jly 7, 2020.

22. Wedam S, Fashoyin-Aje L, Bloomquist E, et al. FDA approval summary: palbociclib for male patients with metastatic breast cancer. *Clin Cancer Res*. 2020;26(6):1208–1212. doi:10.1158/1078-0432.CCR-19-2580

23. Deloitte. Getting real with real-world evidence (RWE) 2017 RWE benchmark survey. Available from: https://www2.deloitte.com/us/en/pages/life-sciences-and-health-care/articles/real-world-evidence-benchmarking-survey.html. Accessed July 7, 2020.

24. Davis B, Morgan J, Shah S. The future of real-world evidence Biopharma companies focus on end-to-end, AI-driven, internally developed solutions. Available from: https://www2.deloitte.com/us/en/insights/industry/life-sciences/2018-real-world-evidence-benchmarking.html?id=gx:2sm:3tw:4RWE_LSHC18::6Life_Sciences_and_Healthcare:20180711095200:Global&linkId=54041116. Accessed July 7, 2020.

25. Morgan J, Feghali K, Shah S, Miranda W. RWE focus is shifting to R&D, early investments begin to pay off. Available from: https://www2.deloitte.com/us/en/insights/industry/health-care/real-world-evidence-study.html?id=us:2sm:3li:4di_gl:5eng:6di. Accessed July 7, 2020.

26. Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform*. 2015;216:574–578.

27. Hripcsak G, Ryan PB, Duke JD, et al. Characterizing treatment pathways at scale using the OHDSI network. *Proc Natl Acad Sci U S A*. 2016;113(27):7329–7336. doi:10.1073/pnas.1510502113

28. Hernandez-Boussard T, Monda KL, Crespo BC, Riskin D. Real world evidence in cardiovascular medicine: ensuring data validity in electronic health record-based studies. *J Am Med Inform Assoc*. 2019;26 (11):1189–1194. doi:10.1093/jamia/ocz119

29. Shaw R. The 10 best machine learning algorithms for data science beginners. Available from: https://www.dataquest.io/blog/top-10-machine-learning-algorithms-for-beginners/. Accessed July 7, 2020.

30. Shah P, Kendall F, Khozin S, et al. Artificial intelligence and machine learning in clinical development: a translational perspective. *NPJ Digit Med*. 2019;2:69. doi:10.1038/s41746-019-0148-3

31. Xie H. Robotic process automation in healthcare will lead to better patient outcomes. In: *Information Age*. Bonhill Group Plc; 2019.

32. Dietz R, Raimondo C, Kremer J, Reynolds J, Saxena R. *Robotics in Claims: Transformation Through Intelligent Automation*. Ernst & Young LLP; 2018.

33. Association for Intelligent Information Management. Intelligent information management glossary: what is robotic process automation? Available from: https://www.aiim.org/What-is-Robotic-Process-Automation#. Accessed July 7, 2020.

34. Juhn Y, Liu H. Natural language processing to advance EHR-based clinical research in allergy, asthma, and immunology. *J Allergy Clin Immunol*. 2020;145(2):463–469. doi:10.1016/j.jaci.2019.12.897

35. Lin H, Long E, Ding X, et al. Prediction of myopia development among Chinese school-aged children using refraction data from electronic medical records: A retrospective, multicentre machine learning study. *PLoS Med*. 2018;15(11):e1002674. doi:10.1371/journal.pmed.1002674

36. Kiely DG, Doyle O, Drage E, et al. Utilising artificial intelligence to determine patients at risk of a rare disease: idiopathic pulmonary arterial hypertension. *Pulm Circ*. 2019;9(4):2045894019890549. doi:10.1177/2045894019890549

37. Blaivas M, Blaivas L. Are all deep learning architectures alike for point-of-care ultrasound?: evidence from a cardiac image classification model suggests otherwise. *J Ultrasound Med*. 2020;39 (6):1187–1194. doi:10.1002/jum.15206

38. Mahajan V, Venugopal VK, Murugavel M, Mahajan H. The algorithmic audit: working with vendors to validate radiology-ai algorithms-how we do it. *Acad Radiol*. 2020;27(1):132–135. doi:10.1016/j.acra.2019.09.009

39. European Commission. *Communication from the Commission to the European Parliament and the Council: Data Protection Rules as a Trust-Enabler in the EU and Beyond Taking Stock.* Brussels, Belgium. 2019.

40. Hu K, LaFontaine P, Mandic A, et al. An assessment of real-world data, artificial intelligence, and their use in drug development. Presented at Drug Information Association; February 57, 2019; Vienna, Austria.

41. American Statistical Association. Statistical Partnerships Among Academe, Industry, and Government (SPAIG) award. Available from: https://www.amstat.org/ASA/Your-Career/Awards/Statistical-Partnerships-Among-Academe-Industry-and-Government-Award.aspx. Accessed July 7, 2020.

42. Rizvi S, Majumdar M, Zou KH, Ray A. Partnerships in action: advancing healthcare through collaborative science in emerging markets. Available from: https://www.cioapplications.com/cxoinsights/partnerships-in-action-advancing-healthcare-through-collaborative-science-in-emerging-markets-nid-5717.html. Accessed July 7, 2020.

43. Ali R, El Shahawy O, Zou KH, Sherman S, Weitzman M, Ray A. NYU Abu Dhabi and Pfizer Inc. collaboration drives progress. Available from: https://magazine.amstat.org/blog/2020/04/01/nyu-abu-dhabi-pfizer/. Accessed July 7, 2020.

44. Berger ML, Dreyer N, Anderson F, Towse A, Sedrakyan A, Normand SL. Prospective observational studies to assess comparative effectiveness: the ISPOR Good Research Practices Task Force Report. *Value Health.* 2012;15:217–230. doi:10.1016/j.jval.2011.12.010

45. Zou KH, Imperato J, Ovalle JC, Li JZ, Sethi N, Ray A. Enhanced patient-centricity: optimizing patient care through AI/ML/DL. *PharmaVoice.* 2020;20(6):52–54.