Journal of the American Medical Informatics Association, 2025, **32(11)**, 1746–1755
https://doi.org/10.1093/jamia/ocaf133
Advance access publication 26 September 2025
Case Report

Case Report

Towards responsible artificial intelligence in healthcare—getting real about real-world data and evidence

Eileen Koski, MPhil[1], Amar Das, MD, PhD[2], Pei-Yun Sabrina Hsueh, PhD[3], Anthony Solomonides, PhD[4], Amanda L. Joseph, BCom, MSc[5,6,7], Gyana Srivastava, BA[7], Carl Erwin Johnson, MD, EdM, MSc[8], Joseph Kannry, MD[9], Bilikis Oladimeji, MBBS, MMCi[10], Amy Price, DPhil[11,12], Steven Labkoff, MD[7,13], Gnana Bharathy, PhD[14], Baihan Lin, PhD[15,16,17], Douglas Fridsma, MD, PhD[18], Lee A. Fleisher, MD[19], Monica Lopez-Gonzalez, PhD[20], Reva Singh, JD, MA[21], Mark G. Weiner, MD[22], Robert Stolper, AB[23], Russell Baris, MS[24], Suzanne Sincavage, PhD[25,26], Tristan Naumann, PhD[27], Tayler Williams, BA[21], Tien Thi Thuy Bui, PharmD[7,28], Yuri Quintana, PhD*[,6,7,29]

[1]IBM Research, Yorktown Heights, NY 10598, United States, [2]Real World Evidence, Guardant Health, Palo Alto, CA 94304, United States, [3]Department of Ethical AI and External Innovation, Pfizer Inc., New York, NY 10001, United States, [4]Research Institute, Endeavor Health, Evanston, IL 60201, United States, [5]School of Health Information Science, University of Victoria, Victoria, BC V8P 5C2, Canada, [6]Homewood Research Institute, Guelph, ON N1E 6K9, Canada, [7]Division of Clinical Informatics, Beth Israel Deaconess Medical Center, Boston, MA 02215, United States, [8]Outcomes Research, Merck & Co., Inc., North Wales, PA 19454, United States, [9]Division of General Internal Medicine, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029, United States, [10]Clinical Technology Strategy, UnitedHealth Group, Eden Prairie, MN 55344, United States, [11]The Dartmouth Institute for Health Policy and Clinical Practice, Geisel School of Medicine at Dartmouth, Hanover, NH 03755, United States, [12]BMJ, London, WC1H 9JP, United Kingdom, [13]Quantori, Cambridge, MA 02142, United States, [14]Faculty of Engineering and IT, University of Technology Sydney, Sydney, NSW 2007, Australia, [15]Department of AI, Hasso Plattner Institute for Digital Health, Icahn School of Medicine at Mount Sinai, New York, NY 10029-5674, United States, [16]Department of Psychiatry, Mount Sinai, New York, NY 10029, United States, [17]Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, NY 10029, United States, [18]Health Universe, San Francisco, CA 94114, United States, [19]Anesthesiology and Critical Care, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, United States, [20]Cognitive Insights for Artificial Intelligence, Baltimore, MD 21202, United States, [21]Public Policy, American Medical Informatics Association, Washington, DC 20011, United States, [22]Department of Population Health Sciences, Weill Cornell Medicine, New York, NY 10065, United States, [23]Strategy Consulting, IQVIA, Boston, MA 02210, United States, [24]eLumindata, Westport, CT, United States, [25]Foundation for Biodefense Research, Fallbrook, CA, United States, [26]IDIQ Inc., Fallbrook, CA 92028, United States, [27]Microsoft New England Research and Development Center, Microsoft Research, Cambridge, MA 02142, United States, [28]Massachusetts College of Pharmacy and Health Sciences, Boston, MA, United States, [29]Department of Medicine, Harvard Medical School, Boston, MA 02115, United States

*Corresponding author: Yuri Quintana, PhD, Division of Clinical Informatics, Beth Israel Deaconess Medical Center, 133 Brookline Avenue, HVMA Annex, Suite 2200, Boston, MA 02215, United States (yquintan@bidmc.harvard.edu)

Abstract

Background: The use of real-world data (RWD) in artificial intelligence (AI) applications for healthcare offers unique opportunities but also poses complex challenges related to interpretability, transparency, safety, efficacy, bias, equity, privacy, ethics, accountability, and stakeholder engagement.

Methods: A multi-stakeholder expert panel comprising healthcare professionals, AI developers, policymakers, and other stakeholders was assembled. Their task was to identify critical issues and formulate consensus recommendations, focusing on the responsible use of RWD in healthcare AI. The panel's work involved an in-person conference and workshop and extensive deliberations over several months.

Results: The panel's findings revealed several critical challenges, including the necessity for data literacy and documentation, the identification and mitigation of bias, privacy and ethics considerations, and the absence of an accountability structure for stakeholder management. To address these, the panel proposed a series of recommendations, such as the adoption of metadata standards for RWD sources, the development of transparency frameworks and instructional labels likened to "nutrition labels" for AI applications, the provision of cross-disciplinary training materials, the implementation of bias detection and mitigation strategies, and the establishment of ongoing monitoring and update processes.

Received: November 9, 2024; **Revised:** March 11, 2025; **Editorial Decision:** April 5, 2025; **Accepted:** July 25, 2025

Conclusion: Guidelines and resources focused on the responsible use of RWD in healthcare AI are essential for developing safe, effective, equitable, and trustworthy applications. The proposed recommendations provide a foundation for a comprehensive framework addressing the entire lifecycle of healthcare AI, emphasizing the importance of documentation, training, transparency, accountability, and multi-stakeholder engagement.

Key words: real-world data; real-world evidence; artificial intelligence; responsible AI; AI accountability; AI transparency.

Introduction

The use of Artificial Intelligence (AI) to generate real-world evidence (RWE) from available real-world data (RWD) offers unique promise for accelerating healthcare advances by enabling the analysis and interpretation of data at an unprecedented scale. However, the use of RWD to develop AI applications in healthcare also poses complex challenges that must be understood and addressed, including but not limited to interpretability, transparency, safety, efficacy, bias, equity, privacy, ethics, accountability, and stakeholder engagement.[1]

The term "RWD" refers to any data gathered or documented during routine healthcare interactions, as opposed to data collected in the context of a research protocol.[2] RWD encompasses a wide variety of information, including electronic health records (EHRs), insurance claims, laboratory test results, patient-generated data from wearable or implanted devices, and surveys and digital health applications. Analysis of RWD can provide real-world evidence (RWE) to advance research, support healthcare decision-making and innovation, and inform public policy. The terms RWD and RWE are sometimes used interchangeably. However, it is important to recognize their fundamental differences, as described in the US Food and Drug Administration (FDA) framework for their Real-World Evidence program.[3] RWD represents data collected on patient health and care delivery, while RWE reflects clinical evidence about the use and potential benefits of an intervention based on the analysis of RWD.

Some studies have illustrated key challenges related to bias and privacy when using RWD in healthcare AI applications. A study by Obermeyer et al[4] demonstrates one type of risk in using RWD: an inappropriate choice of proxy for health status led to disparities in healthcare access and utilization. The authors found that a widely used commercial algorithm systematically underestimated the health needs of Black patients compared to White patients with similar health status due to the algorithm's reliance on healthcare costs as a proxy for health needs. This example highlights the importance of carefully examining RWD for potential biases and developing strategies to mitigate these biases in AI model development and deployment, as well as the need for transparency in how patient data is used to ensure that AI models do not inadvertently lead to disparate impact in certain patient populations.

In a similar vein, a study by Esteva et al[5] found that deep learning models trained on smartphone photos to detect skin cancer performed significantly worse on images of darker skin tones, likely due to the underrepresentation of these skin tones in the training data.

Another example is the widely used Framingham Heart Score, a scoring system used to predict the risk of cardiovascular events in practice. The score was based on a longitudinal study conducted in Framingham, Massachusetts primarily because of the engagement of local area physicians and its proximity to cardiologists at the Massachusetts General Hospital and Harvard Medical School.[6] Framingham was a middle-class community that was predominantly White of European descent and the score was later shown not to perform as well among Black patients as it does among White patients, exhibiting both over and underestimations of risk.[7]

These examples underscore how biases in RWD can be introduced and lead to inequitable model performance and should help drive home the importance of proactively identifying and mitigating bias and protecting patient privacy when leveraging RWD for AI applications.

Principles for all phases of the lifecycle of AI development and deployment in healthcare have already been enumerated in cross-organizational alliances and professional societies such as the American Medical Informatics Association (AMIA).[7] Some activities, such as the AMIA AI Evaluation Showcase,[8] have been conducted to help identify best practices in healthcare AI development and deployment.

This paper aims to identify challenges specifically related to the use of RWD and RWE in healthcare AI and propose guidelines and other recommendations to promote its responsible use for research, diagnosis, treatment recommendations, and predictive analytics. The recommendations aim to ensure that AI innovations adhere to robust assurance and privacy standards and are backed by a clear accountability structure to guarantee safe, effective, equitable, and trustworthy (SEET) use while remaining aligned with multi-stakeholder priorities.

Methods

A multi-stakeholder expert consensus process was conducted to develop recommendations on the ethical deployment of healthcare AI applications focused on identifying major challenges and potential strategies to ensure adherence to the SEET standard for AI solutions across three specific domains: the use of real-world data and evidence (RWD/RWE), clinical decision support, and consumer health applications. The consensus process encompassed educational webinars, an in-person consensus conference entitled "Blueprints for Trust," and ongoing deliberations over several months.[9] The conference convened 50 experts, including clinicians, informatics researchers, AI scientists, solution architects, legal and policy specialists, patients, patient and consumer health advocates, and industry leaders. A planning committee from AMIA and the DCI Network, an initiative of the Division of Clinical Informatics at Beth Israel Deaconess Medical Center, identified participants.

The process involved multiple phases, beginning with a series of educational webinars to introduce foundational concepts and key challenges. This was followed by the in-person conference, which included panel discussions, breakout sessions, and a full-day workshop to collaboratively develop an initial framework. Meeting transcripts were analyzed using natural language processing (NLP) to extract key themes that were brought to the group for discussion and development recommendations to address challenges in the use of RWD and RWE in AI applications. Although participant consent

for the use of transcripts was limited to summarization, key points were synthesized into this manuscript to reflect agreed-upon recommendations. The recommendations underwent iterative refinement through regular online group meetings and structured deliberative processes over several months. Feedback was sought from additional stakeholders with expertise in healthcare, AI ethics, and governance, ensuring comprehensive insights into practical, clinical, and ethical objectives. External peers invited to the meeting also helped refine the framework through their feedback. Discrepancies were addressed during facilitated discussions, where participants presented their viewpoints and rationale. A neutral moderator ensured balanced representation and guided the group toward consensus.

Ethical concerns surrounding financial incentives, particularly in AI monetization, were deliberated and all members disclosed any potential conflicts or interests and included them in this paper. The work was not funded by any source and authors represented their own views not of their employers. Ensuring transparency and mitigating conflicts of interest are paramount to safeguarding patient rights.

This four-month deliberative process resulted in a collaboratively developed, carefully vetted framework, revealing three distinct governance models tailored to the requirements of RWE, CDS, and CH domains. Each governance model reflects the unique needs and challenges of its respective domain, providing a structured foundation for ethical AI deployment in healthcare. The recommendations for AI in CDS[10] Consumer health[11] appear in other papers. This paper reports on the recommendations for the use of RWD and RWE in healthcare AI, with specific emphasis on RWD since the authors feel that many serious challenges arise during the selection, evaluation and validation of the appropriateness of the data to be used in developing AI applications.

Results

Panel deliberations highlighted gaps in RWD documentation, emphasizing the importance of Findable, Accessible, Interoperable, and Reusable (FAIR) principles.[12] These insights informed specific recommendations to address data biases and stakeholder accountability. The group identified the following as the most critical issues that need to be addressed when using RWD in conjunction with AI in healthcare.

1) **Data literacy and documentation**
   The characteristics and context of the vast majority of RWD are not always sufficiently documented for different types of users to understand its precise meaning and appropriate use. Healthcare AI application developers and users may need more background or training to understand the limits of applicability of any given dataset to a specific question in each healthcare context, which encompasses issues of data source, type of care, population, location—both geographic and setting of care.
2) **Bias**
   RWD reflects the societal issues and systemic biases present in healthcare, which affects why, how, where, and when people interact with the healthcare system, the type and level of care they may receive, and their outcomes.[13,14] If recognized and addressed, these biases in RWD may contribute to inaccurate and inequitable applications, both when an AI application is first trained and when such an application evolves or is ported to a new environment after its initial implementation. Critical issues were identified based on expert discussions and validated against existing literature.[4,15–17]
3) **Privacy and ethics**
   By definition, RWD represents data generated on or about individuals interacting with the healthcare system or addressing health related issues in any way that produces data, such as the use of wellness apps. While the use of patient data in AI applications may be covered during the process of consenting to care, it is not clear that all RWD used in this manner is subject to any privacy or ethical oversight. In addition, as more patients and consumers interact with AI directly, there will be a greater need for safeguards regarding the applications available and the use of their data.
4) **Lack of accountability structure for stakeholder management**: These problems apply to various stakeholders, from RWD producers, curators, and platform providers to AI application developers, deployers, and distributors. Finding an accountability structure that can delineate end-users' responsibilities is, therefore, essential to the fast-evolving regulatory and guideline development pathways.

The expert consensus group recognized the need for an overall responsible AI framework that adheres to accepted principles[8] encompassing the entire lifecycle of design, development, validation, implementation, and monitoring and that addresses the needs of a broad array of stakeholders concerning interpretability, transparency, safety, efficacy, bias, equity, privacy, ethics, accountability, and stakeholder engagement. This framework must include processes to maintain this broad stakeholder engagement continuously.

As one component of such a framework, in Table 1 we have put forth the specific guidelines to address issues we consider critical concerning the use of RWD in healthcare AI.

Discussion

This paper provides recommendations for the responsible use of RWD and RWE in healthcare AI and builds on previous work from this consortium on AI in CDS[11] and AI in Consumer Health.[21] While SEET standards can be applied to all three domains (RWD/RWE, CDS, and consumer health) future work should evaluate the cross-domain implementation of these standards. The stakes are much higher for the use of RWD in "black-box" tools, such as deep learning and generative AI (Gen-A) however, because the risk of not fully understanding the data being used is amplified beyond the experience using such data and evidence as the basis for conclusions using standard statistical and analytic methods, or even newer data mining and machine-learning methodologies. There are two primary reasons for this: (1) the methodology by which an AI algorithm is created and operates cannot always be examined or understood the way a more traditional analytic technique can; and (2) the sheer scale and complexity of the data being used.

While offering advantages of availability and scale, RWD reflects the inherent variability, complexity, and biases in how different populations access, utilize, and receive

Table 1. Challenges and recommendations for use of RWD in healthcare AI.

Data quality and documentation	
Challenge	Available data sets are not always sufficiently documented for use by end-users who may not understand the complexities of the data, the context in which it was generated, or any relevant limitations to representativeness with respect to either patient characteristics (geographic, temporal, socio-demographic representativeness), or data source (e.g, EHR, out-patient clinic, laboratory, pharmacy, etc.)
Recommendations **Establish metadata standards for RWD sources** • Metadata standards should be established from existing standards under the same Findable, Accessible, Interoperable, and Reusable (FAIR) principles[12] recommended to characterize RWD sources, including provenance, context of creation, collection methods, coverage (ie, geographic, temporal, socio-demographic representativeness), and formats. • Adherence to such standards would allow stakeholders to better evaluate fitness for purpose and mitigate risks in using the data. **Develop transparency frameworks and "Data Characteristics Labels" for AI systems, similar to commonly used "nutrition labels"**[18,19] • These frameworks and labels should document data sources, algorithms, evaluation results, and known limitations, recommended use cases, and any known limitations, particularly concerning potential applicability in differing contexts of use or populations such as in Sendak[20]	**Stakeholder roles** **Standards Development Organizations (SDOs):** • Develop robust metadata standards for documentation of data set components and characteristics, including the context of use, detailed description of how data was compiled, source population characteristics, etc. • Develop training materials targeted to both developers and end-users providing examples of common issues and mitigation strategies **Data providers:** • Implement appropriate metadata standards in their documentation • Provide relevant examples of unique issues or limitations to their data **AI developers:** • Evaluate potential data sets with respect to their fitness for the intended use • Implement metadata standards and transparency frameworks. **Clinicians and researchers:** • Review available training materials on RWD characteristics, biases, focusing on applicability and limitations with respect to their untended use cases.
Design and development	
Challenge	The appropriate guidelines and requirements must be understood from project inception, to assure that all of the appropriate steps are taken to comply with requirements for scientific rigor and transparency at every step of the project.
Recommendations • Require pre-specifications for use of RWD and RWE to ensure appropriate rigor and transparency related to the development, testing, validation and procedures for implementation of the tools.	**Stakeholder roles** **Standards Development Organizations (SDOs):** • Support standards for relevant pre-specification. **AI developers:** • Assure that pre-specification requirements are fully met at all stages of project design, planning, execution and evaluation **AI users:** • Require validation that the pre-specification requirements have been met.

(continued)

Table 1. (continued)

Bias detection and mitigation	
Challenge	Real-world data (RWD) often reflects population and contextual biases, such as geography or socio-demographics, which can distort insights and lead AI models to perform poorly on under-represented groups. Identifying and addressing these biases is critical for equitable AI outcomes.
Recommendations • All RWD considered for use in developing AI applications should be carefully evaluated in terms of both known and potential biases, particularly with respect to the intended context of use. • AI developers should Implement bias detection and mitigation strategies during model development and validation. • AI users should evaluate the model in their population and context of use before implementation. • AI users should carefully re-evaluate the model on an appropriate population and context of use before any implementation in a different setting. • AI users should regularly monitor and update models to detect evidence of drift in performance over time, as well as across different populations and settings.	**Stakeholder roles** **Standards Development Organizations (SDOs):** • Support continued development of standards for collection and evaluation of socio-demographic patient and population characteristics. **Data providers:** • Document any known limitations to the representativeness of their data—e.g, geographic, temporal or socio-demographic. **AI developers:** • Implement bias detection and mitigation strategies during data selection and evaluation, model development and validation • Regularly monitor and update models to ensure performance is maintained over time • Re-assess the model when applied across different populations or settings **Healthcare organizations:** • Assess AI model performance in the context of local patient populations and use cases before deployment • Regularly monitor and update models to ensure performance is maintained over time
Privacy and consent	
Challenge	Although patients typically give consent for medical interactions and interventions, the consent process may only address secondary use of their data in a very limited fashion and is unlikely to address potential use of their data in the development of AI algorithms. Patients also have a right to know if AI is being used in their care. In order to engender trust, appropriate processes need to be developed in order to inform patients, caregivers and the general public about the use of their data in AI applications.
Recommendations **Transparency** • Develop transparent consent processes to inform patients about the use of their data in AI applications. • Develop educational materials to help patients understand how use of AI in their care may benefit them, as well as any potential risk to them from its use. • Include explanations of risk-based data collection strategies, highlighting how only the minimum necessary data will be collected and used based on risk assessments. • Explicitly communicate the potential risks of data breaches, including the possibility of re-identification or misuse of sensitive health information.	**Stakeholder roles** **AI developers and healthcare organizations:** • Develop transparent consent processes that inform patients about the use of their data in AI applications • Implement robust data privacy and security safeguards in compliance with relevant regulations (e.g, HIPAA, GDPR) • Develop and implement transparent, tiered consent processes that explain data use, risk levels, and AI integration in patient care. **Policymakers and regulatory bodies:** • Guide privacy and consent requirements for the use of RWD in AI development and deployment • Engage diverse stakeholders, including patients and advocacy groups, in developing these guidelines
Stakeholder Education and Engagement	
Challenge	There is a wide range of stakeholders from the individual to the government that have a stake in understanding how and when to best deploy AI in support of healthcare and medical research, and as importantly, how and when not to deploy it. Given the technical nature of development and application of AI, there is a compelling need for accessible educational materials and ongoing engagement of the broadest possible stakeholder community.

(continued)

Table 1. (continued)

Recommendations
• Develop accessible educational materials on the potential benefits, risks, and limitations of RWD-driven AI in healthcare.
• Training materials need to:
  • Be geared to a wide range of stakeholders
  • Educate stakeholders about the problems that can arise from failing to understand how a dataset relates to the intended context of use.
  • Include case studies, best practices for bias detection, and guidance on metadata documentation to enhance stakeholder understanding and collaboration.
  • Provide a clear description of accountable responsibilities from each stakeholder group's perspective.
• Foster ongoing multi-stakeholder dialogue to align AI development with societal values and patient needs.
	Stakeholder roles
	Data Providers:
	• Develop descriptions of how data is used and how privacy and confidentiality are protected when sharing data
	AI Developers:
	• Develop accessible educational materials on RWD-driven AI's benefits, risks, and limitations in healthcare.
	AI users:
	• Foster ongoing multi-stakeholder dialogue to align AI development and deployment with societal values and patient needs
	Patients, caregivers and the general public:
	• Provide resources to help individuals understand how their health data may be used in AI applications and their rights regarding data privacy and consent.
	All Stakeholders:
	• Engage in ongoing multi-stakeholder dialogue to align AI development and deployment with societal values and patient needs.

Monitoring and updates	
Challenge	AI systems must be carefully monitored over time to assure that their assessments and recommendations remain in sync as data evolves over time, which can occur due to population changes, new treatments or practice patterns. Once implemented, such systems need to be monitored carefully over time to detect, and address, any drift or changes observed. In addition, a tool developed and implemented in one setting or context of use, may not work the same in a different setting and must be evaluated appropriately in the new setting or application prior to implementation.
Recommendations • A monitoring program must be implemented to detect and assess the validity of a tool's output or validity given RWD's ongoing and changing input, based on the metadata standards and transparency framework proposed above. • All software systems in healthcare need to be monitored and updated, but AI tools particularly adaptive ones, pose specific challenges in that the output may change over time as RWD accrues and the output of the tools is refined • New applications of an existing tool to a different population or clinical context must be re-trained or re-validated against appropriate RWD for the new application.	**Stakeholder roles** **AI developers:** • Design and implement monitoring systems for continuous assessment of tool outputs. • Integrate diverse datasets during initial development and validation phases to enhance tool generalizability. • Regularly retrain and refine AI models to ensure outputs remain accurate and unbiased. • Conduct re-training and validation processes for tools deployed in new populations or clinical settings. **Healthcare organizations:** • Establish protocols for real-time validation and feedback on AI tool performance. • Allocate resources for ongoing tool evaluation and updates to maintain reliability. • Identify and flag discrepancies in tool performance across populations and initiate re-validation requests. • Perform localized testing to ensure contextual applicability of AI tools. **Researchers** • Conduct independent evaluations of tool performance in various settings to identify limitations and recommend improvements. **Regulatory bodies:** • Oversee adherence to metadata and transparency standards in monitoring practices. • Define guidelines for periodic audits and updates of adaptive AI systems. • Require documentation and approval of re-validation efforts to maintain compliance. **Policymakers:** • Engage diverse stakeholders, including patients and advocacy groups, in guideline development and implementation

healthcare.[22] Lacking the familiarity of experimental research data collection (e.g, clinical trials), the context and meaning of RWD elements are often not as formally or fully documented and, thus, may need additional support or information to be adequately understood. For example, laboratory tests may be performed with different methodologies or be reported in different units, particularly over time. If this is not well documented, understood, and accounted for in an analysis, a user may assume that the values are comparable or have been normalized, leading to erroneous conclusions. Such straightforward issues can be detected by preliminary descriptive statistical analyses of data distributions based on selected characteristics. Other issues, such as differential population representation or incomplete or inaccurate sociodemographics, may have profound implications for the applicability of outcomes; however, these issues may be more difficult to detect and require special effort and methodologies to mitigate.[23]

Understanding the characteristics and complexities of the source data can help to ensure the validity, safety, and efficacy of AI applications that utilize such data for training, predictions, and recommendations. A full understanding of the data being used is a fundamental prerequisite for any method of knowledge generation based on RWD. Efforts to use RWD and RWE in AI will also benefit from the use of Common Data Models (CDMs), such as the OMOP Common Data Model,[24] that are critical for harmonizing data across different sources. Such data models also enable standardized queries and interoperability between datasets in the US and Europe.

There have already been numerous well-documented instances of harm caused by algorithms that were developed for one purpose or on one population but had an unintended or opposite effect through a misinterpretation of what the data meant. The seminal work done by Obermeyer et al[4] showed the impact of selection of an inappropriate proxy measure, and the issue with the Framingham Risk Score[6,7] showed how under-representation could distort the outcome. These are just two examples of ways in which problems arise due to lack of understanding or representation and the impact that can have. There may be ways to address these biases however, as demonstrated by Zink et al[14] in their use of a race-adjusted algorithm. These examples underscore the critical need for demographic subgroup validation as well as the opportunity to use tailored strategies, such as race adjustments, to address inequities in data representation and predictive performance.

Similar errors have occurred even in cases where those involved had significant knowledge and training in data analysis and the healthcare domain. As AI applications become more readily available and widely deployed, it will become increasingly important to ensure that users have a solid grounding in fundamental scientific concepts, such as differentiating causation from association.

The inadequate capture of contextual details on RWD sources can lead to AI applications producing incorrect, unreliable, irreproducible, biased, or unsafe outputs. While there has been progress in sourcing, aggregating, and harmonizing RWD for secondary use, clear metadata standards and transparency frameworks do not yet exist for appropriate documentation that sufficiently characterizes RWD to enable appropriate use in different AI models and scenarios by people who are not already familiar with the complexities of the data being used. In December 2023, the FDA published a transparency framework to evaluate the potential use of RWD in RWE that helps support the approval of new indications or post-approval study requirements as specified in the RWE program.[25] A more general version of such a framework and RWD metadata standards for responsible AI[26] would be essential to warrant the responsible use of RWD for RWE incorporated in AI applications. This is particularly important since application developers may not fully understand the nature of the data they are working with, while clinical users may not understand the AI applications, leading to inappropriate design assumptions and inadequate communication. Patients who interact with AI applications may not have any way to judge whether or not to trust the information provided.

The framework's reliance on expert opinions presents a limitation, as it may not fully address all challenges associated with RWD and RWE in AI applications. Empirical validation across diverse settings is essential to assess its generalizability and applicability. Regular updates will also be required to align with evolving governmental perspectives, such as the FDA's RWE framework, which underscores the importance of addressing misconceptions about RWD and emphasizes rigor in defining data provenance and contextual use.[27] Future research should prioritize the empirical testing of these guidelines and their iterative refinement based on insights gained from real-world implementation.

There are also tools and processes that can help assure appropriate scientific rigor throughout the entire lifecycle of an AI application that is based on or utilizes RWD and/or RWE. For example, pre-specifications, as requested by regulatory bodies such as the FDA and highlighted by professional organizations, can play a crucial role in ensuring the rigor and transparency of RWE studies through the use of predefined protocols that establish clear methodologies and criteria for data collection, analysis, and interpretation, which are essential for maintaining reproducibility and reliability.

There must also be careful consideration of issues related to specific uses of patient data, such as digital phenotyping, which is defined as the collection and analysis of data from personal digital devices, such as smartphones and wearables. Such data may offer valuable insights into patient behavior, activity, and health patterns and can be leveraged for AI in Real-World Evidence (RWE) by providing granular, real-time data to improve understanding of patient outcomes, adherence to treatment, and early detection of health risks. However, the use of digital phenotyping must prioritize transparency and patient consent to prevent privacy violations. It is essential to ensure that patients are fully informed about the nature of the data being collected, how it will be used, and who will have access to it.

These issues ultimately affect every stakeholder and every step involved in the resulting tool or model, encompassing design, development, testing, validation, regulation, implementation, usage, monitoring, and maintenance of such applications. Similarly, the stakeholders cover a broad range of scientists, developers, clinicians, administrators, and patients. In particular, it is essential to ensure patient privacy in these implementations by requiring clear consent protocols and robust anonymization techniques.

Conclusion

The use of RWD for generating RWE in the development and application of AI in healthcare offers advantages of scale that

cannot easily be matched by traditional data creation methods, such as clinical trials or other ongoing efforts to create the learning healthcare system.[28] This is particularly true in cases involving rare events and diagnoses, atypical presentations of more common conditions, patterns of adverse events, or unexpected changes in incidence rates. Examples include post-market surveillance of clinical outcomes such as adverse events, identifying unique risk factors or predictors of disease progression, detecting emerging infectious diseases, and understanding environmental and social determinants of health. However, we must recognize that such data has inherent challenges to its use—including missing contextual data, inadequate documentation, and context-driven selection biases that may affect its representativeness. While bias can be introduced when using RWD through convenience sampling, barriers to access, and effective exclusion of underrepresented populations, a growing number of tools and techniques have become available to mitigate bias[15,23] that can be applied as long as there is adequate documentation to allow a developer or user to understand the level and the nature of the need in a given context.

The guidelines put forth in this paper synthesize expert, multi-disciplinary consensus group recommendations across key areas to promote responsible and ethical usage of RWD for AI in healthcare. Realizing the vision outlined herein will require sustained and multifactorial efforts by interdisciplinary public and private partnerships among data producers and consumers at the institutional level, health systems, academia, technology companies, government agencies, policymakers, and other organizational stakeholders. Individual-level stakeholders such as clinical experts, data scientists, developers, social scientists, ethicists, patient advocates, and end users are also necessary.

We believe it is essential to establish guidelines whereby RWD can be evaluated and used appropriately in AI to support various applications in medical research and healthcare. We propose to improve documentation of data provenance, context, and metadata; to develop and implement training and education for both data providers and data users; to establish criteria for evaluating a data set against the intended context of use; to define appropriate multi-stakeholder evaluation and validation of the AI output; to promote transparency about all aspects of tool development and evaluation; and finally to evaluate performance over time, with particular emphasis on the process required to apply tools created in one setting and context to a new setting or context.

Near-term priorities for ensuring responsible RWD-based AI are focused on approaches and methods for documentation, training, transparency, explainability, equity, validation, accountability, monitoring, and updates. Over time, as more applications of AI are developed, deployed, and evaluated, these guidelines will be advanced and expanded accordingly. In the longer term, avenues for voluntary compliance must be considered, potentially including a process for certification of users and models.

It must be stressed that this paper primarily focuses on the panel's deliberation of one challenge to AI development and deployment in healthcare. The panel also recognizes numerous critical issues related to the intended use and applications of AI technologies in varying contexts (e.g, clinical decision support, drug discovery, diagnostics); usability and workflow integration; patient consent, privacy, and ethics; and many more that will be critical to the responsible and effective use of AI in healthcare whether it involves the use of RWD or not. As such, we believe that a broad framework that addresses all aspects of AI in healthcare is also needed and that guidelines for using RWD represent one of the building blocks of such a framework.

Acknowledgments

We would like to thank Dr David Bray, Dr Tiffani J. Bright, and Dr Isaac Kohane, for their keynote talks at the conference. We would like to thank Rabbi Daniel Cohen and Reverend Greg Doll for their webinar on Bioethics and AI, Dr Luke Sato, Dr Paul R. DeMuro, and Kenneth E. White, JD for their webinar on AI in Healthcare: Risk Management, Trust, and Liability, Dr Amy Price and Dave deBronkart (ePatient Dave) for their webinar on AI in Healthcare: The Patient Perspective, and Dr Anne-Marie Meyer, Mark Shapiro, and Rob Stolper for their webinar: AI in Healthcare: Real World Data Generation And The Regulatory Perspective.

Author contributions

Eileen Koski (Conceptualization, Formal analysis, Methodology, Writing—original draft, Writing—review & editing), Amar Das (Writing—original draft, Writing—review & editing), Pei-Yun Sabrina Hsueh (Conceptualization, Formal analysis, Methodology, Writing—original draft, Writing—review & editing), Anthony Solomonides (Conceptualization, Formal analysis, Writing—original draft, Writing—review & editing), Amanda L. Joseph (Conceptualization, Formal analysis, Methodology, Writing—original draft, Writing—review & editing), Gyana Srivastava (Conceptualization, Data curation, Methodology, Project administration, Writing—original draft, Writing—review & editing), C. Erwin Johnson (Conceptualization, Formal analysis, Writing—review & editing), Joseph Kannry (Conceptualization, Formal analysis, Writing—review & editing), Amy Price (Conceptualization, Formal analysis, Writing—original draft, Writing—review & editing), Steven Labkoff (Conceptualization, Formal analysis, Methodology, Writing—original draft, Writing—review & editing), Gnana Bharathy (Writing—original draft, Writing—review & editing), Baihan Lin(Formal analysis, Writing—review & editing), Douglas B. Fridsma (Formal analysis, Writing—review & editing), Lee A. Fleisher (Formal analysis, Writing—review & editing), Monica Lopez-Gonzalez (Formal analysis, Writing—review & editing), Reva Singh (Writing—original draft, Writing—review & editing), Mark Weiner (Formal analysis, Writing—review & editing), Robert Stolper (Formal analysis, Writing—review & editing), Suzanne Sincavage (Formal analysis, Writing—review & editing), Tristan Naumann (Formal analysis, Writing—review & editing), Tayler Williams (Formal analysis, Writing—review & editing), Tien Thi Thuy Bui (Formal analysis, Writing—review & editing), and Yuri Quintana (Conceptualization, Formal analysis, Methodology, Project administration, Writing—original draft, Writing—review & editing)

Funding

None declared.

Conflicts of interest

None declared.

Data availability

All data used in this article has been reported in this paper. There is no other data associated with this paper. No datasets were generated or analyzed during the current study.

Code availability

No computer code was produced or analyzed for this article.

Disclosure statement

- Eileen Koski: Employed by & owns stock in IBM
- Amar Das: Employee of Guardant Health and owner of Guardant Health stock.
- Pei-Yun Sabrina Hsueh: Employee of Pfizer Inc.; owns stock and/or stock options for IBM, Bayesian Health, and Pfizer; on the Practitioners Board of Association for Computing Machinery (ACM)
- Anthony Solomonides: no conflicts or competing interests to disclose
- Amanda L. Joseph: no conflicts or competing interests to disclose
- Gyana Srivastava: no conflicts or competing interests to disclose
- Carl Johnson: Employed by and receive shares of stock from Merck & Co., Inc.
- Joseph Kannry: no conflicts or competing interests to disclose
- Bilikis Oladimeji: no conflicts or competing interests to disclose
- Amy Price: no conflicts or competing interests to disclose
- Steven Labkoff: 401(K) at BMS; Stockholder (<10% ownership), Quantori
- Gnana Bharathy: no conflicts or competing interests to disclose
- Baihan Lin: no conflicts or competing interests to disclose
- Douglas Fridsma: employee of Health Universe
- Lee A. Fleisher: Principal at Rubrum Advising which advises companies on coverage of new technologies including AI
- Monica Lopez-Gonzalez: no conflicts or competing interests to disclose
- Reva Singh: no conflicts or competing interests to disclose
- Mark G. Weiner: no conflicts or competing interests to disclose
- Robert Stolper: no conflicts or competing interests to disclose
- Russell Baris: no conflicts or competing interests to disclose
- Suzanne Sincavage: no conflicts or competing interests to disclose
- Tristan Naumann: Employee of Microsoft Corporation
- Tayler Williams: no conflicts or competing interests to disclose
- Tien Thi Thuy Bui: no conflicts or competing interests to disclose
- Yuri Quintana: no conflicts or competing interests to disclose

References

1. OECD AI Principles overview [Internet]. Accessed April 17, 2024. https://oecd.ai/en/ai-principles
2. FDA. Submitting Documents Using Real-World Data and Real-World Evidence to FDA for Drugs and Biologics Guidance for Industry | FDA. (n.d.). Accessed April 25, 2022. 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
3. FDA. Framework for FDA's Real-World Evidence Program. Accessed February 27, 2024. https://www.fda.gov/media/120060/download
4. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366:447-453.
5. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542:115-118. https://doi.org/10.1038/nature21056
6. Mahmood SS, Levy D, Vasan RS, Wang TJ. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *Lancet*. 2014;383:999-1008.
7. Gijsberts CM, Groenewegen KA, Hoefer IE, et al. Race/ethnic differences in the associations of the Framingham risk factors with carotid IMT and cardiovascular events. *PLoS One*. 2015;10:e0132321.
8. Solomonides AE, Koski E, Atabaki SM, et al. Defining AMIA's artificial intelligence principles. *J Am Med Inform Assoc*. 2022;29:585-591. https://doi.org/10.1093/jamia/ocac006
9. American Medical Informatics Association (AMIA). AI Evaluation Showcase. Accessed October 3, 2024. https://amia.org/education-events/amia-2024-artificial-intelligence-evaluation-showcase
10. DCI Network. (2023, September 21). Blueprints for Trust: Best practices and Regulatory Pathways for ethical AI in healthcare. Accessed April 17, 2024. https://www.dcinetwork.org/events/192
11. Labkoff S, Oladimeji B, Kannry J, et al. Toward a responsible future: recommendations for AI-enabled clinical decision support. *J Am Med Inform Assoc*. 2024;31:2730-2739. https://doi.org/10.1093/jamia/ocae209
12. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016;3:160018.
13. Schwartz Schwartz R, Vassilev A, Greene K, et al. *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*. Vol. 3. US Department of Commerce, National Institute of Standards and Technology; 2022.
14. Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. *JAMA*. 2019;322:2377-2378.
15. Zink A, Obermeyer Z, Pierson E. Race adjustments in clinical algorithms can help correct for racial disparities in data quality. *Proc Natl Acad Sci USA*. 2024;121:e2402267121. https://doi.org/10.1073/pnas.2402267121
16. Huang Y, Yuan W, Kohane IS, Beaulieu-Jones BK. Illustrating potential effects of alternate control populations on real-world evidence-based statistical analyses. *JAMIA Open*. 2021;4:ooab045. https://doi.org/10.1093/jamiaopen/ooab045
17. Conover MM, Ryan PB, Chen Y, Suchard MA, Hripcsak G, Schuemie MJ. Objective study validity diagnostics: a framework requiring pre-specified, empirical verification to increase trust in the reliability of real-world evidence. *J Am Med Inform Assoc*. 2025;32:ocae317-525. https://doi.org/10.1093/jamia/ocae317 Epub ahead of print. PMID: 39789670.
18. Wartella EA, Lichtenstein AH, Boon CS, eds. Institute of Medicine (US). Committee on Examination of Front-of-Package Nutrition Rating Systems and Symbols.
19. History of Nutrition Labeling. *Front-of-Package Nutrition Rating Systems and Symbols: Phase I Report*. National Academies Press

(US); 2010. Accessed February 24, 2024. https://www.ncbi.nlm.nih.gov/books/NBK209859
20. Sendak MP, Gao M, Brajer N, Balu S. Presenting machine learning model information to clinical end users with model facts labels. *NPJ Digital Med*. 2020;3:1-4.
21. Rozenblit L, Price A, Solomonides A, et al. Towards a Multi-Stakeholder process for developing responsible AI governance in consumer health. *Int J Med Inform*. 2025;195:105713. https://doi.org/10.1016/j.ijmedinf.2024.105713
22. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. 2015. Intelligible models for HealthCare: predicting pneumonia risk and hospital 30-day readmission. In: *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. Association for Computing Machinery, 1721-1730. https://doi.org/10.1145/2783258.2788613
23. Park Y, Singh M, Koski E, Sow DM, Scheufele EL, Bright TJ. Algorithmic Fairness and AI Justice in Addressing Health Equity. In: Kiel JM, Kim JR, Ball MJ. (eds.), *Healthcare Information Management Systems: Cases, Strategies & Solutions*, 5th edn. Springer Nature Switzerland AG; 2022, 223-234.
24. Standardized Data: The OMOP Common Data Model. Accessed January 22, 2025. https://www.ohdsi.org/data-standardization/
25. FDA. Data Standards for Drug and Biological Product Submissions Containing Real-World Data. 2023. Accessed March 10, 2024. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/data-standards-drug-and-biological-product-submissions-containing-real-world-data
26. Dignum V. *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Vol. 1. Springer; 2019.
27. Concato J, Corrigan-Curay J. Real-world evidence—where are we now? *N Engl J Med*. 2022;386:1680-1682. https://doi.org/10.1056/NEJMp2200089
28. Institute of medicine (IOM). *The Learning Healthcare System: Workshop Summary*. The National Academies Press; 2007.

Journal of the American Medical Informatics Association, 2025, 32, 1746–1755
https://doi.org/10.1093/jamia/ocaf133
Case Report