

AIRWEE: A Conceptual Framework for Evaluating AI-Enabled Real-World Evidence Workflows

Ashis Kumar Das¹, Zoe Wu², Yuqin Wei³

¹ Evidence Intelligence, Komodo Health, USA

² Biostatistics, Komodo Health, USA

³ Product Development, Komodo Health, USA

Background

AI-enabled real-world evidence (RWE) generation encompasses protocol development, cohort construction, variable derivation, statistical analysis, and interpretation. Established RWE guidance addresses study specification, reporting, reproducibility, and data quality, but does not explicitly operationalize evaluation of AI contributions across these activities.[1–5]

Complementary AI guidelines address biomedical large language model (LLM) research, clinical decision support, AI interventions, chatbot health advice, and health economics and outcomes research.[6–10] VALID evaluates the reliability of AI-extracted electronic health records (EHR) data through performance metrics, verification checks, and replication analyses.[11] Emerging approaches, including RWE-bench, FastOMOP, and applications of ELEVATE-GenAI, address aspects of AI-enabled analytical workflows, although their scope and evaluation objectives differ.[12–14] Regulatory guidance further establishes principles for AI credibility, oversight, and non-interventional research.[15–18]

We propose AIRWEE (AI-enabled Real-World Evidence Evaluation), a framework for evaluating AI-enabled RWE generation workflows in health economics and outcomes research (HEOR). AIRWEE integrates observational research requirements with assessment of methodological validity, implementation fidelity, expert oversight, and reproducibility. It is designed for workflows with varying degrees of AI involvement across OMOP-mapped databases and other real-world data environments.

Methods

A scoping review identified reporting guidelines, methodological and regulatory guidance, evaluation frameworks, and benchmarks relevant to AI-enabled RWE generation. Searches covered PubMed, Embase, EQUATOR, preprint repositories, ISPOR publications, and regulatory websites, supplemented by citation searching through May 30, 2026. Search strategies, dates, eligibility criteria, and publication status were documented. Each record was mapped against a structured gap matrix assessing coverage across the full RWE pipeline — from research question formulation and protocol development through cohort construction,

variable derivation, analytic execution, interpretation, and reporting — and evaluated for applicability to AI-enabled workflows operating on real-world data. The pattern of gaps across frameworks informed the domain structure of AIRWEE.

Results

Existing sources addressed complementary aspects of evaluation, including study specification, AI performance reporting, extracted-data reliability, workflow governance, and observational analysis benchmarking.[1–13,15–18] Their differing scopes informed an integrated approach linking protocol assessment, patient-level accuracy verification, analytical checks, expert oversight, and reproducibility across the RWE workflow.

This synthesis informed AIRWEE’s eight-domain structure and its emphasis on expert oversight throughout AI-enabled RWE generation. Qualified researchers retain responsibility for scientific decisions and interpretation. The domains are organized into Governance (D1), Workflow evaluation (D2–D5), and Scientific assessment (D6–D8). The domains comprise D1 Study governance, D2 Study specification, D3 Cohort ascertainment, D4 Variable derivation, D5 Statistical implementation, D6 Results validity, D7 External validation, and D8 Computational reproducibility. Evaluation addresses protocol validity and data feasibility, implementation accuracy, support for reported findings, applicability across settings, and reproducibility. The framework specifies independent accuracy benchmarking through patient-level comparison with expert-adjudicated reference analyses using the same data and approved protocol. Comparisons address cohort membership, derived variables, statistical outputs, and reported conclusions. Review checkpoints assess intermediate outputs, with cross-cutting criteria addressing provenance, traceability, error detection and propagation, expert intervention, and protocol deviations. Detailed assessment procedures and context-specific acceptance criteria remain under development.

Conclusions

AIRWEE provides an evaluation framework purpose-built for AI-enabled RWE generation under expert oversight. Consistent with OHDSI’s emphasis on transparent phenotyping and reproducible research, AIRWEE is designed for application across OMOP-mapped databases and non-OMOP data sources. It complements existing AI reporting guidelines by specifying step-level evaluation of methodological appropriateness, implementation accuracy, and evidentiary support throughout AI-enabled RWE generation.

Figure 1. AI-enabled Real-World Evidence Evaluation (AIRWEE) Conceptual Framework



References

1. Wang SV, Pinheiro S, Hua W, et al. STaRT-RWE: structured template for planning and reporting on the implementation of real world evidence studies. *BMJ*. 2021;372:m4856.
2. Langan SM, Schmidt SAJ, Wing K, et al. The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE). *BMJ*. 2018;363:k3532.
3. Wang SV, Pottegård A, Crown W, et al. HARmonized Protocol Template to Enhance Reproducibility of Hypothesis Evaluating Real-World Evidence Studies on Treatment Effects: A Good Practices Report of a Joint ISPE/ISPOR Task Force. *Value Health*. 2022;25(10):1663–1672.
4. Berger ML, Sox H, Willke RJ, et al. Good Practices for Real-World Data Studies of Treatment and/or Comparative Effectiveness: Recommendations from the Joint ISPOR-ISPE Special Task Force on Real-World Evidence in Health Care Decision Making. *Value Health*. 2017;20(8):1003–1008.
5. Heads of Medicines Agencies; European Medicines Agency. *Data Quality Framework for EU medicines regulation: application to Real-World Data*. EMA/503781/2024. March 27, 2026.
6. Gallifant J, et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nat Med*. 2025;31:60–69.
7. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat Med*. 2022;28:924–933.
8. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. 2020;26:1364–1374.
9. The CHART Collaborative. Reporting guideline for chatbot health advice studies: the Chatbot Assessment Reporting Tool (CHART) statement. *BMJ Med*. 2025;4:e001632.
10. Fleurence RL, Dawoud D, Bian J, et al. ELEVATE-GenAI: Reporting Guidelines for the Use of Large Language Models in Health Economics and Outcomes Research: An ISPOR Working Group Report. *Value Health*. 2025;28(11):1611–1625.
11. Estevez M, Singh N, Dyson L, et al. Ensuring Reliability of Curated Electronic Health Record–Derived Data: The Validation of Accuracy for Large Language Model–/Machine Learning–Extracted Information and Data (VALID) Framework. *JCO Clin Cancer Inform*. 2026;10:e2500215.
12. Li D, He Y, Hu Y, Tian Y, Li J. Can LLM Agents Generate Real-World Evidence? Evaluating Observational Studies in Medical Databases. *arXiv [preprint]*. 2026. arXiv:2603.22767.
13. Moeller-Grell N, Shenzhang S, Jiang ZJ, Dobson RJB, Chandrabalan VV. FastOMOP: A Foundational Architecture for Reliable Agentic Real-World Evidence Generation on OMOP CDM data. *arXiv [preprint]*. 2026. arXiv:2604.24572.

14. Berliner E, Santos J, Fronstin M. Application of the ELEVATE-GenAI framework to an AI agentic system for data analytics [conference abstract]. ISPOR 2026; Philadelphia, PA. Abstract RWD169. *Value Health*. 2026;29(S6).
15. US Food and Drug Administration. *Considerations for the Use of Artificial Intelligence To Support Regulatory Decision-Making for Drug and Biological Products*. Draft guidance for industry and other interested parties. January 2025.
16. European Medicines Agency. *Reflection paper on the use of Artificial Intelligence (AI) in the medicinal product lifecycle*. EMA/CHMP/CVMP/83833/2023. 2024.
17. European Medicines Agency; US Food and Drug Administration. *Guiding principles of good AI practice in drug development*. January 2026.
18. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. *General Principles on Planning, Designing, Analysing, and Reporting of Non-interventional Studies That Utilise Real-World Data for Safety Assessment of Medicines: M14*. Final guideline. September 4, 2025.