From algorithms to insights: The role of medical writers in AI-enhanced real-world evidence

Gomathi Priya Jeyapal, Prarthana Vigneshwari Reddy, Vrushabh Baburao Satav, Pattabhi Machiraju
Indegene, Bengaluru, Karnataka, India

doi: 10.56012/vysp1464

Correspondence to:
Gomathi Priya Jeyapal
priya.jeyapal@indegene.com

Abstract

Real-world evidence (RWE) complements randomised controlled trials (RCTs) by assessing treatment effectiveness in diverse populations. Integrating artificial intelligence (AI) and machine learning (ML) enhances RWE by enabling predictive modelling, risk stratification, and clinical decision support. ML techniques like supervised or unsupervised learning, logistic regression, decision trees, random forests, and XGBoost can help optimise regulatory decision-making and patient care. This paper explores how the AI/ML models help identify high-risk patients, predict disease progression, and assess healthcare burden. The medical writer's role in structuring findings into clinically meaningful insights is essential for bridging the gap between data science and clinical application. As AI advances, skilled medical writers will ensure transparency, ethical compliance, and effective communication of AI-driven RWE findings.

Introduction

Randomised controlled trials (RCTs) are the gold standard for establishing causality in controlled settings, yet their findings often lack applicability to diverse real-world populations with varying genetic backgrounds, comorbidities, and treatment regimens. This "efficacy-effectiveness gap" limits the generalisability of RCT outcomes, posing challenges for regulatory decision-making.[1] To bridge this gap, healthcare stakeholders, including pharmaceutical companies, regulatory agencies, and health technology assessment (HTA) organizations, increasingly integrate real-world data (RWD) with RCTs.[2] Enabled by technological advancements, RWD comprising electronic health records (EHRs), registries, claims data, and mobile health applications offers valuable insights into routine healthcare delivery and patient outcomes.

When analysed, RWD generates real-world evidence (RWE), informing treatment effectiveness, safety, and economic impact, thereby supporting data-driven healthcare decisions and enhancing clinical and regulatory strategies.[3,4] The healthcare industry is witnessing an unprecedented transformation driven by artificial intelligence (AI) and machine learning (ML), which are redefining RWE generation, interpretation, and utilisation.[5] RWD, encompassing EHRs, claims data, patient registries, and wearable device outputs, offers vast potential to complement traditional clinical trials.[6]

AI identifies the patterns, medical writers connect them to patients, policies, and practice.

AI/ML algorithms enable efficient processing, analysis, and predictive modelling of these complex datasets, leading to actionable insights that improve patient outcomes, optimise treatment pathways, and guide regulatory decisions.[7] AI is significantly transforming the use of RWE in healthcare by facilitating more precise data analysis and decision-making. AI technologies are instrumental in analysing vast and complex RWD sources. These AI-driven approaches help identify patterns in treatment responses, predict patient outcomes, and improve clinical decision-making by integrating RWD into the healthcare system.[8,9]

However, as AI continues to revolutionize healthcare research, a critical challenge has emerged: the communication of complex, algorithm-driven insights to various healthcare stakeholders.[10] Regulators, clinicians, pharmaceutical companies, and policymakers require clear, precise, and scientifically accurate interpretations of AI-generated RWE.[11] Medical writers serve as essential intermediaries, ensuring that interpretations of AI-generated RWE are not only methodologically sound but also comprehensible, regulatory-compliant, and aligned with healthcare decision-making frameworks.

This paper explores the evolving landscape of AI-driven RWE, highlighting key ML techniques, such as clustering techniques, dimensional reduction algorithms, logistic regression (LR), decision trees (DT), random forests (RF), and extreme gradient boosting (XGBoost). It also delves into the role of medical writers in bridging the gap between complex AI/ML outputs and stakeholder needs, ensuring that scientific narratives derived from AI/ML-driven analytics are both accessible and impactful.

AI/ML techniques in RWE generation

The integration of AI/ML into RWE research enables the extraction of valuable patterns and insights from vast datasets. Supervised, unsupervised, and reinforcement learning are core ML approaches applied in healthcare and RWE generation. Supervised learning uses labelled data with known outcomes to train predictive models – such as LR, DT, RF, and XGBoost – that are commonly applied in diagnosing conditions like diabetes or hypertension. This approach follows a defined, iterative process involving data selection, processing, model training, and evaluation using metrics like receiver operating characteristic (ROC) curves and confusion matrices.

Unsupervised learning, on the other hand, works with unlabelled data to identify hidden structures or patterns using techniques such as clustering algorithms (e.g., k-means, hierarchical clustering) and dimensionality reduction methods like principal component analysis (PCA), which help group patients or detect

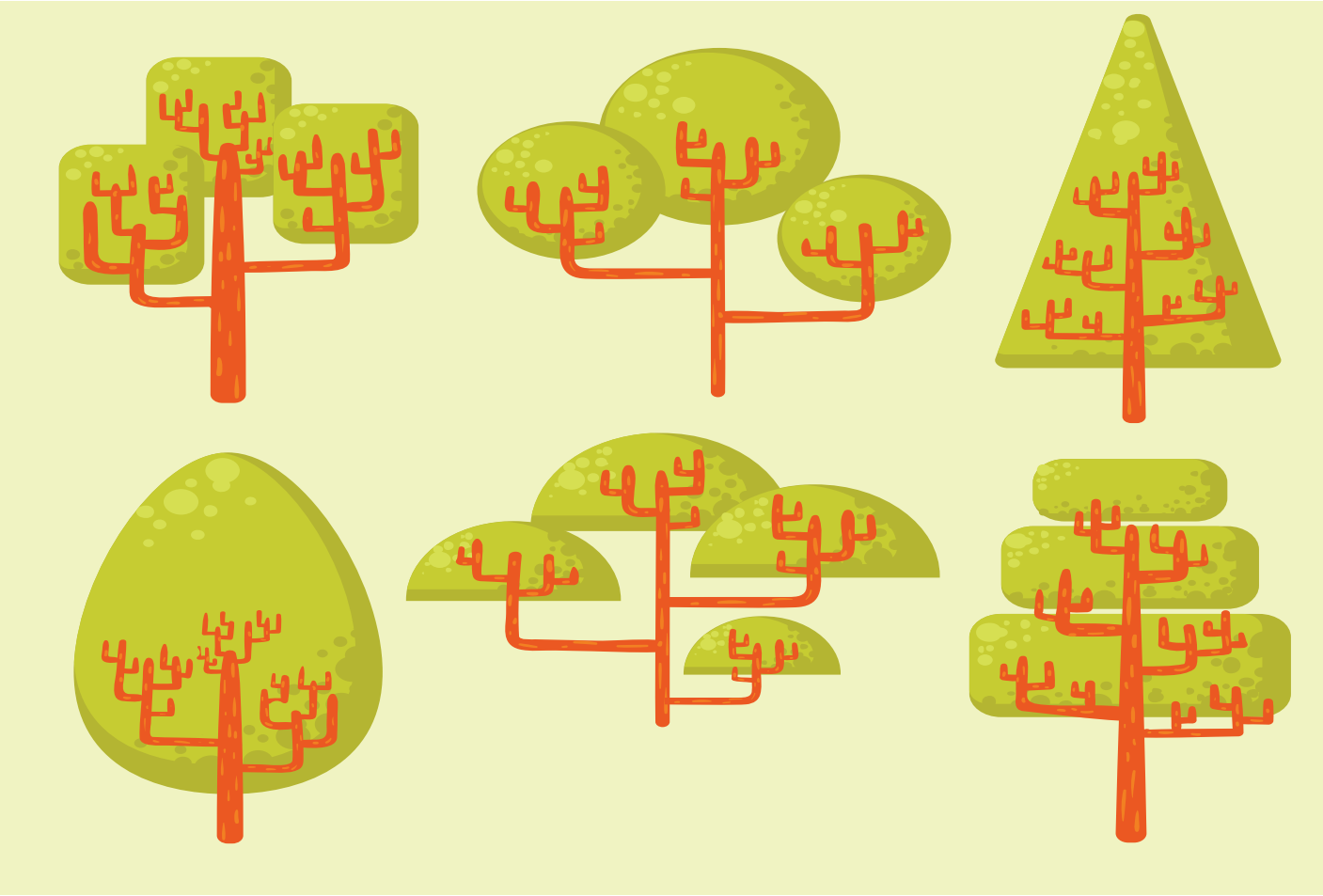

Image: Freepik

anomalies without predefined outcomes, though careful interpretation is necessary. Reinforcement learning allows systems to learn optimal decisions through trial and error, guided by feedback or rewards, making it promising for dynamic treatment decision support, despite challenges in defining rewards and causal pathways. The following section highlights key ML techniques and explores their applications in predicting clinical outcomes.

Supervised learning approaches

Logistic regression (LR)

LR is a foundational ML technique widely used in healthcare for binary classification tasks. It models the probability of an event occurring as a function of predictor variables, making it valuable for predicting clinical outcomes, adverse events, and patient risk stratification.[12] LR is a valuable tool for identifying high-risk patient groups by calculating a probability score that classifies patients into high-risk or low-risk categories, helping prioritise those needing immediate care. LR models are particularly useful in healthcare settings where predicting patient outcomes based on historical data can significantly enhance patient safety and clinical decision-making.[13] Performance metrics such as accuracy, precision, recall, F1-score (harmonic mean of precision and recall), and area under the curve (AUC)-ROC assess how well the model classifies patients.[14]

The model's coefficients show the contribution of each factor to a patient's risk level, with positive coefficients indicating a higher likelihood of being high-risk. The confusion matrix and ROC curve offer insights into how well the model distinguishes between high- and low-risk patients or slow and fast progressors. By analysing these results, healthcare providers can identify key risk factors and apply targeted interventions for better patient outcomes.[15]

> As real-world evidence evolves, medical writing expertise ensures that data-driven insights are ethically sound and clinically actionable.

Decision trees (DTs)

DTs use hierarchical structures to segment patient populations based on predictor variables, making them effective for classification and regression problems. A DT starts with a root node representing the entire dataset, which is then split into branches based on feature values.[16] These branches lead to decision nodes and eventually to leaf nodes, which represent the outcome. For example, a DT model was used to predict the risk of cardiovascular events in a large patient data set by analysing factors such as blood pressure and cholesterol levels, providing a clear pathway for targeted interventions.[17] Each decision helps categorise a patient as high-risk or low-risk. The model is easy to interpret, as it shows exactly how the decision is made at each branch.[18]

A DT approach may be better than LR due to its ability to capture non-linear relationships and handle mixed data types. While LR assumes a linear relationship between the predictors and the outcome, DTs can model complex interactions between variables without requiring

Image: Freepik

any assumptions about the data distribution. This flexibility allows DTs to provide more accurate predictions in cases where the relationships between variables are not straightforward.[18] One of the key strengths of DTs is their interpretability. The model visually represents how decisions are made at each branch, showing the exact conditions that lead to a particular classification. This transparency is particularly valuable in healthcare, where understanding the rationale behind risk predictions is crucial for clinical decision-making.

Healthcare providers can easily follow the tree structure to see how different patient-related factors contribute to the risk classification. DTs can also predict the importance of features, highlighting which variables have the most significant impact on the model's predictions. Through understanding the importance of different features, healthcare providers can focus on the most influential factors when designing interventions to reduce readmission rates. Healthcare providers can also gain deeper insights into patient-related factors and their impact on risk classification by leveraging DTs. This clarity helps in identifying patients who are likely to transition to higher risk categories, enabling targeted interventions that improve patient outcomes.[18]

Random forests (RFs)

RF is a powerful ensemble learning method that leverages multiple DTs to improve the accuracy and robustness of predictions. Unlike a single DT, which can easily overfit and perform poorly on unseen data, an RF builds many DTs using random sampling (bootstrap sampling) of both the data points and the features at each split.[19] This aggregation of multiple trees ensures that individual tree biases and variance are minimised, resulting in a more stable and reliable model.[20] An RF can handle complex interactions between features and make better predictions, especially when simpler models like LR or single DT do not yield satisfactory results.[21]

Each tree in an RF is built from a random subset of the data and features, which helps to reduce overfitting and improve generalization to new data. The final prediction is made by averaging the predictions of all the trees in the forest, which enhances the model's accuracy and robustness. For example, an RF model was used to predict phenotype transformations by analysing complex genetic and environmental interactions. This approach improved the accuracy of predictions and helped identify key factors driving phenotype changes, providing valuable insights for personalised medicine.[22] RF models can also provide insights into feature importance, indicating which variables have the most significant impact on the predictions.

XGBoost

XGBoost is a highly efficient and powerful boosting algorithm that builds trees sequentially, with each tree correcting the errors of the previous one by focusing on misclassified data points, thereby improving the model's predictive power and accuracy.[19] It optimises both speed and performance by using regularisation techniques to prevent overfitting and by implementing efficient tree-building algorithms.[23] XGBoost has several advantages compared to other algorithms, such as its ability to handle missing data and highly parallelizable code in large and complex datasets. It employs a novel sparsity-aware algorithm for sparse data and a weighted quantile sketch for approximate tree learning. This makes it particularly suited for applications in healthcare, where datasets can be vast and intricate.

In predicting patient risk for transitioning to an advanced disease stage, XGBoost can analyse a multitude of variables, including genetic factors, medical history, and lifestyle habits. By identifying small patterns and interactions that may be crucial for accurate predictions, XGBoost provides a robust tool for risk stratification. For instance, in a study predicting the progression of chronic kidney disease, XGBoost outperformed other models by accurately identifying patients at high risk of rapid disease progression.[24] XGBoost also provides insights into feature importance, highlighting which variables have

the most significant impact on the model's predictions. This information is valuable for healthcare providers as it helps identify key risk factors and prioritise interventions.

Unsupervised learning algorithms

Unsupervised learning algorithms are used in RWD analysis to detect clusters, reduce dimensionality, and uncover latent structures. They are increasingly featured in health economics and outcomes research (HEOR), pharmacovigilance, and post-market surveillance. Common techniques used are clustering (e.g., k-means, hierarchical clustering), which groups similar patients based on multiple clinical and demographic variables, and dimensionality reduction (e.g., PCA, t-SNE), which simplifies high-dimensional data to reveal visual patterns or key contributors.

The key outputs are cluster assignments, group labels, visualisations (heatmaps, dendrograms, scatter plots), and metrics (silhouette score, variance explained, cluster centroids). In one example, unsupervised clustering helped identify three distinct patient subtypes within the chronic kidney disease population.[25] One subgroup, comprising 30% of the population, showed frequent treatment switching and higher hospitalisation rates, indicating a high-risk phenotype with possible unmet needs.

The contribution of medical writers

Medical writing encompasses a broad and complex field, from clinical trials to regulatory submissions and from medical education to patient communication. Medical writers play a crucial role in translating complex medical information from research studies, clinical trials, and scientific articles into clear content. Balancing scientific accuracy using reliable evidence with clarity for the intended audience is a key challenge in healthcare. By adhering to ethical standards, medical writers maintain the trust of the scientific community and public while enhancing medical practices and knowledge.[26] Table 1 shows the key responsibilities of medical writers in translating RWE model results.[27]

Medical writing best practices in RWE studies

In the context of RWE generation, medical writers serve as critical knowledge translators responsible for accurately interpreting, contextualising, and communicating complex data to a wide array of stakeholders. Several case studies and whitepapers highlight the role of medical writers in successful AI-driven RWE projects.[28,29] For example, a study using AI to analyse EHRs for predicting cardiovascular outcomes required medical writers to translate complex ML models into actionable insights for clinicians.[30] Another case involved the use of natural language processing to extract RWE from unstructured clinical notes,[31] with medical writers ensuring the findings were accurately represented in regulatory submissions.

As well as RWE generation, AI/ML have been used for generating medical text. Despite advances in text generation, AI/ML cannot replace human medical writers and their use in medical writing raises ethical concerns.[32] AI-generated text has the potential to perpetuate bias, misinformation, and plagiarism. Furthermore, computer models need to be retrained regularly to ensure they are up to date, as the field of medicine is constantly evolving. Given these concerns, medical writers are indispensable for safeguarding the integrity of medical information and its compliance with ethical and regulatory standards.

By breaking down complex data, working with different teams, ensuring transparency, and getting results ready for publication, medical writers help make AI insights easier to under-

Table 1. Key responsibilities of medical writers in translating results of real-world evidence models

Key responsibility	Description
Bridging the language gap	Medical writers act as translators between data scientists and healthcare audiences. They interpret model assumptions and methodology, data inputs and limitations, and outputs, such as risk scores, clusters, or probabilities.
Distilling key insights	Writers identify and emphasise results that reveal clinically significant subgroups, suggest treatment response differences, and indicate burden of disease or health outcomes.
Crafting clinically meaningful narratives	An example: "The model identified a patient segment at threefold higher risk of hospitalization within 12 months, characterised by polypharmacy, diabetes, and advanced age. This suggests a target for early intervention." This kind of narrative brings model results into the clinical and strategic realm.
Visual interpretation support	Writers also help develop or adapt visualisations (e.g., Kaplan-Meier curves, cluster heatmaps), annotate plots to highlight clinically relevant findings, and ensure visual outputs are publication – or submission-ready.
Ensuring scientific and regulatory rigour	In RWE deliverables, especially for HTA or regulatory use, writers must clearly describe the model type, inputs, and limitations, avoid over-interpretation of exploratory analyses, and frame findings within accepted scientific standards.
Communicating data generation techniques using AI/ML	Developing peer-reviewed manuscripts and conference presentations that highlight the value of AI/ML techniques in RWE generation.[27]

Abbreviations: AI, artificial intelligence; HTA, health technology assessment; ML, machine learning; RWE, real-world evidence.

stand and use in healthcare. As such, medical writers are essential for clearly and accurately communicating complex scientific information and making sure that AI-derived findings meet ethical and regulatory standards. Since AI continues to shape healthcare, the need for skilled medical writers will grow.

Conclusion

Medical writers are indispensable in the AI-enhanced RWE ecosystem, ensuring that complex data is transformed into meaningful insights. By collaborating with data scientists, clinicians, and regulators, medical writers can help unlock the full potential of AI in healthcare. Medical writers do not need to be data scientists, but they must understand the fundamentals of analytic methodologies. As the field evolves, medical writers must embrace AI as a tool for innovation, while maintaining the highest standards of scientific integrity and ethical communication.

Acknowledgements

Not applicable.

Disclaimers

The opinions expressed in this article are the authors' own and do not necessarily reflect the position of the authors' affiliated organisations.

Disclosures and conflicts of interest

The authors declare no conflicts of interest.

References

1. Makady A, de Boer A, Hillege H,et al.; GetReal Consortium. What is real-world data? A review of definitions based on literature and stakeholder interviews. Value Health. 2017;20(7):858-65. doi:10.1016/j.jval.2017.03.008
2. Pongiglione B, Torbica A, Blommestein H, et al. Do existing real-world data sources generate suitable evidence for the HTA of medical devices in Europe? Mapping and critical appraisal. Int J Technol Assess Health Care. 2021;37(1):e62. doi:10.1017/S0266462321000301
3. McNair D, Lumpkin M, Kern S, et al. Use of RWE to Inform regulatory, public health policy, and intervention priorities for the developing world. Clin Pharmacol Ther. 2022;111(1):44-51. doi:10.1002/cpt.2449
4. Khosla S, White R, Medina J, et al. Real world evidence (RWE) – a disruptive innovation or the quiet evolution of medical evidence generation? F1000Res. 2018;7:111. doi:10.12688/f1000research.13585.2
5. Olczak J, Pavlopoulos J, Prijs J, et al. Presenting artificial intelligence, deep learning, and machine learning studies to clinicians and healthcare stakeholders: an introductory reference with a guideline and a Clinical AI Research (CAIR) checklist proposal. Acta Orthop. 2021;92(5):513-25. doi:10.1080/17453674.2021.1918389
6. Kandhare P, Kurlekar M, Deshpande T, et al. A review on revolutionizing healthcare technologies with AI and ML Applications in pharmaceutical sciences. Drugs Drug Candidates. doi:10.3390/ddc4010009
7. Kolluri S, Lin J, Liu R,et al. . Machine learning and artificial intelligence in pharmaceutical research and development: a review. AAPS J. 2022;24(1):19. doi:10.1208/s12248-021-00644-3
8. Wynants L, van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ. 2020;369:m1328. doi:10.1136/bmj.m1328
9. Knevel R, Liao KP. From real-world electronic health record data to real-world results using artificial intelligence. Ann Rheum Dis. 2023;82(3):306-11. doi:10.1136/ard-2022-222626
10. Alanazi A. Using machine learning for healthcare challenges and opportunities. Inform Med Unlocked. 2022;30:100924. doi:10.1016/j.imu.2022.100924
11. Dang A. Real-world evidence: A primer. Pharmaceut Med. 2023;37(1):25-36. doi:10.1007/s40290-022-00456-6
12. Shipe ME, Deppen SA, Farjah F, et al. Developing prediction models for clinical use using logistic regression: an overview. J Thorac Dis. 2019;11(Suppl 4):S574-84. doi:10.21037/jtd.2019.01.25
13. Coz E, Fauvernier M, Maucort-Boulch D. An overview of regression models for adverse events =analysis. Drug Saf. 2024;47(3):205-16. doi:10.1007/s40264-023-01380-7
14. Manikandan G, Pragadeesh B, Manojkumar V, et al. Classification models combined with Boruta feature selection for heart disease prediction.Inform Med Unlocked. . 2024;44:101442. doi:10.1016/j.imu.2023.101442.=
15. Maksabedian Hernandez EJ, Tingzon I, Ampil L, et al. Identifying chronic disease patients using predictive algorithms in pharmacy administrative claims: an application in rheumatoid arthritis. J Med Econ. 2021;24(1):1272-9. doi:10.1080/13696998.2021.1999132
16. Saha D, Manickavasagan A. Machine learning techniques for analysis of hyperspectral images to determine quality of food products: A review. Curr Res Food Sci. 2021;4:28-44. doi:10.1016/j.crfs.2021.01.002
17. Díaz JJS. Chapter 13 – Artificial intelligence in cardiovascular medicine: Applications in the diagnosis of infarction and prognosis of heart failure. In: Barh D, editor. Artificial Intelligence in Precision Health. Cambridge, MA: Academic Press; 2020:313-28. doi:10.1016/B978-0-12-817133-2.00013-6
18. Ozcan M, Peker S. A classification and regression tree algorithm for heart disease modeling and prediction. Healthc Anal (NY). 2023;3:100130. doi:10.1016/j.health.2022.100130
19. Mahajan P, Uddin S, Hajati F, et al. A comparative evaluation of machine learning ensemble approaches for disease prediction using multiple datasets. Health Technol. 2024;14(3):597-613. doi:10.1007/s12553-024-00835-w
20. Mienye ID, Sun Y, Wang Z. An improved ensemble learning approach for the prediction of heart disease risk. Inform Med Unlocked. 2020;20:100402. doi:10.1016/j.imu.2020.100402
21. Breiman L. Random forests. Machine learning. 2001;45(1):5-32. doi:10.1023/A:1010933404324
22. Matsuo R, Yamazaki T, Suzuki M, et al. A random forest algorithm-based approach to capture latent decision variables and their cutoff values. J Biomed Inform. 2020;110:103548 doi:10.1016/j.jbi.2020.103548

23. Hassoon IM. Boosting Learning Algorithms for Chronic Diseases Prediction: A Review. Iraqi J Computers Informatics. 2024;50(2):22-30. doi:10.25195/ijci.v50i2.506
24. Wu J, Gao Q, Tian M, et al. Explainable machine learning prediction of 1-year kidney function progression among patients with type 2 diabetes mellitus and chronic kidney disease: a retrospective study. QJM. 2025 Apr 24:hcaf101. Epub ahead of print. doi:10.1093/qjmed/hcaf101
25. Dashtban A, Mizani MA, Pasea L, et al. Identifying subtypes of chronic kidney disease with machine learning: development, internal validation and prognostic validation using linked electronic health records in 350,067 individuals. EBioMedicine, 2023;8. doi:10.1016/j.ebiom.2023.104489.
26. Khan DI. Unveiling the Vital Contributions of Professional Medical Writers in the Evolving Healthcare Landscape. Indian Pract. 2023;76(12):24-26.
27. Bergstrand S, Heddle C, Sabaté M, et al. Embracing artificial intelligence in medical writing: A new era of efficiency and collaboration. Med Writ. 2023;32(3):82-7. doi:10.56012/iamc1709
28. Franklin JM, Liaw KL, Iyasu S, et al. Real-world evidence to support regulatory decision making: New or expanded medical product indications. Pharmacoepidemiol Drug Saf. 2021;30(6):685-93. doi:10.1002/pds.5222
29. Vanderpuye J, Gordon M. AI in RWE: Key drivers for accelerating clinical development and patient access. 2024 [cited 2025 Jun 30]. Available from: https://www.ispor.org/conferences-education/conferences/past-conferences/ispor-2024/program/program/session/intl2024-3892/18096
30. Tsai ML, Chen KF, Chen PC. Harnessing electronic health records and artificial intelligence for enhanced cardiovascular risk prediction: A comprehensive review. J Am Heart Assoc. 2025;14(6):e036946. doi:10.1161/JAHA.124.036946
31. Zhao Y, Weroha SJ, Goode EL. Generating real-world evidence from unstructured clinical notes to examine clinical utility of genetic tests: use case in BRCAness. BMC Med Inform Decis Mak. 2021;21(1):3. doi:10.1186/s12911-020-01364-y
32. Kitamura FC. ChatGPT Is shaping the future of medical writing but still requires human judgment. Radiology. 2023;307(2):e230171. doi:10.1148/radiol.230171

Author information

Pattabhi Machiraju, PhD, has been a Senior Director at RWD & Biostats Department of Indegene Ltd since 2011. He leads all AI/ML-driven RWE projects globally and brings over two decades of leadership in biostatistics, real-world data analytics, innovative study designs, regulatory submissions, and cross-functional collaboration with global stakeholders and research partners.

Prarthana Reddy, BTech, MBA, has been a Data Scientist at the RWD & Biostats Department of Indegene Ltd since 2020. With an engineering background, she supports AI/ML-driven RWE projects globally, contributing to algorithm development, data-driven insights, and innovative analytical solutions across various therapeutic areas and RWE initiatives.

Gomathi Priya Jeyapal, PhD, has been a subject matter expert at the RWD & Biostats Department of Indegene Ltd since 2019. She supports AI/ML-driven RWE projects for global stakeholders and brings deep expertise in multiple therapeutic areas, real-world data analysis, evidence generation, protocol development, and cross-functional collaboration.

0000-0003-0529-2121

Vrushabh Satav, BHMS, MBA, has been a domain expert at the RWD & Biostats Department of Indegene Ltd since 2024. He supports AI/ML-driven RWE projects for global stakeholders, bringing strong domain knowledge and expertise in clinical data interpretation across multiple therapeutic areas