EVIDINNO
EVIDENCE NAVIGATION & SYNTHESIS

## BALANCING EFFICIENCY WITH ACCURACY:

# ARTIFICIAL INTELLIGENCE IN HEALTH ECONOMICS OUTCOMES RESEARCH

*Nishu, MBA (Pharm.) – Assistant Director, Evidence Synthesis*

# Introduction

The integration of AI into HEOR is transforming the process of how healthcare stakeholders assess value with the potential to optimize resource allocation to improve patient outcomes in a more timely manner. Traditionally, HEOR has relied on complex statistical models that incorporate clinical and real world data to conduct economic evaluations to inform healthcare decision-making. However, the rapid advancement of AI that incorporates machine learning (ML), natural language processing (NLP), and predictive analytics offers opportunities to enhance data-driven insights, improve efficiency, and refine decision-making processes.

AI enables HEOR professionals to process vast amounts of structured and unstructured healthcare data more efficiently, uncover hidden patterns, and generate predictive models that support cost-effectiveness analyses and comparative effectiveness research. From accelerating systematic literature reviews (SLRs) to improving disease modeling and forecasting, AI-driven methodologies are redefining how evidence is generated, analyzed, and applied in healthcare decision-making.[1]

## Enhancing Efficiency with AI in HEOR

**1) Literature Reviews:** Large language models (LLMs) are transforming the landscape of SLRs and meta-analyses by augmenting critical research tasks rather than replacing human expertise.[2] These AI-driven models assist in search strategy development by suggesting MeSH terms and keywords for databases like PubMed.[3] Recent studies have explored their effectiveness in abstract and full-text screening, with GPT-4 achieving performance levels comparable to human reviewers when provided with well-structured prompts in a fraction of the time.[4-6] Additionally, LLMs have been evaluated for their ability to explain exclusion reasoning[7] and assess the risk of bias using tools like the Cochrane Collaboration's risk assessment framework.[8] Moreover, LLMs have been tested for data extraction, with studies reporting high accuracy in replicating extracted data.[9-12]

**2) RWE Evidence Generation:** Real world evidence (RWE) generation can be significantly enhanced with generative AI by streamlining data extraction, reducing variability, and improving consistency in HTA processes. LLMs facilitate efficient extraction of critical information from unstructured electronic health record (EHR) data, such as radiology reports and physician notes[13,14] Domain specific LLMs, such as GatorTron, GatorTronGPT, and Me LLaMA, which are trained using large clinical texts, have been shown to improve the accuracy of outputs over human appraisers.[15-17] Techniques like "few-shot learning" further enhance AI's ability to accurately extract relevant variables from complex clinical texts.[18] Additionally, generative AI is being adapted to analyze various real-world data formats, including medical imaging, broadening its potential applications in evidence generation.[19]

## Enhancing Efficiency with AI in HEOR

**3. Predictive Modeling:** Generative AI has the potential to transform predictive modeling by using advanced algorithms to process large datasets and forecast outcomes with high precision. By analyzing data from sources like EHRs, genetic information, and imaging studies, AI-powered models can accurately predict patient risks, optimize treatment plans, and improve resource allocation. Hospitals use these tools to identify patients at risk of chronic conditions or readmissions, enabling early interventions and tailored care strategies.[20] Deep learning techniques have shown particularly high diagnostic accuracy in areas such as medical imaging, outperforming traditional methods in detecting diseases like diabetic retinopathy and lung cancer.[21]

**4. Comparative Effectiveness:** Comparative effectiveness research can be streamlined with the support of generative AI by enhancing data analysis, reducing bias, and improving evidence synthesis quality. Advanced ML algorithms efficiently process vast amounts of real-world data from claims databases and clinical trials to identify patterns and compare treatment outcomes across diverse patient populations.[22] Recent studies have shown that AI-powered clinical decision support systems demonstrate high concordance with physician recommendations, potentially improving the accuracy of comparative analyses.[22] NLP techniques enable AI to extract relevant information from unstructured medical texts, enhancing the comprehensiveness of comparative studies.[23,24] Furthermore, AI's capacity to simulate and evaluate numerous potential treatments optimizes effectiveness prediction against various diseases, leading to more robust comparative assessments.

**5. Health Economic Modelling:** Various stages of health economic model development can be supported with generative AI, including conceptualization, parameterization, implementation, and validation.[25-27] Studies have demonstrated that foundation models such as GPT-4 and Bing Chat can assist in building Markov models and partition survival models, with researchers employing advanced prompt engineering techniques like chain-of-thought prompting to improve accuracy.[11,28] Nevertheless while AI tools can generate model structures and code, human intervention remains crucial to ensure validity, particularly in complex models. Some studies, such as Ayer et al. 2023,[29] have demonstrated the feasibility of fully automating simpler health economic models, though further research is required to extend this capability to more sophisticated frameworks. Additionally, AI-driven automation could enhance structural uncertainty analysis, streamlining an otherwise resource-intensive process.[25]

## Ensuring Accuracy in AI Applications

Transparency is a critical component of generative AI's integration into scientific research workflows. It is a key focus of FDA regulatory guidance, as clear information about AI systems is essential for the agency to effectively evaluate AI-enabled medical devices and drug development tools. Ensuring transparency across all aspects of generative AI—from algorithmic design to output generation—is essential for enabling reproducibility, and mitigating biases which ultimately lead to fostering trust and increasing acceptance and more widespread use. This principle underpins the discussions on scientific validity and reliability, algorithmic fairness, and regulatory and ethical considerations.

**Scientific Validity and Reliability:** Ensuring the scientific validity and reliability of generative AI tools is crucial, particularly as they are integrated into research workflows. While these models can augment human expertise, researchers remain responsible for accuracy and reporting. LLMs trained on vast publicly available datasets, may introduce errors, especially in specialized fields like healthcare,[30] necessitating rigorous validation. A well-documented issue is the generation of "hallucinations," where models produce incorrect or fabricated information due to their statistical nature of learning.[31,32] Several strategies can mitigate these risks, including prompt engineering (e.g., chain-of-thought prompting,[33] few-shot learning[34]), retrieval-augmented generation,[35] and fine-tuning with domain-specific data.[17] Reproducibility also presents a challenge, as model outputs can vary due to user expertise, prompt quality, and inherent AI variability. Researchers have proposed frameworks to enhance reproducibility, such as repeated trial runs[11] and standardized reporting methods.[8]

**Algorithmic Bias and Fairness:** Bias can arise from multiple sources, including systemic bias due to historical underrepresentation of marginalized groups in training data,[36] as well as computational and statistical biases stemming from unrepresentative samples.[37,38] Exclusion of specific populations during data collection, model training, or evaluation further compounds these risks.[39] To address these challenges, researchers have proposed various mitigation strategies. Distributional approaches focus on improving data representativeness through techniques such as data augmentation, perturbation, reweighting, and synthetic data generation.[40] Federated learning, which enables models to be trained across multiple institutions without direct data sharing, offers potential for reducing localized biases.[39] Algorithmic approaches, including adversarial learning and loss-based methods, adjust model parameters to penalize biased predictions.[41] Ongoing research continues to explore and refine these techniques to enhance fairness in AI-driven applications.

**Regulatory and Ethical Considerations:** Regulatory and ethical considerations for generative AI in biomedical research are evolving, but existing data privacy laws, such as Health Insurance Portability and Accountability Act in the United States[42] and General Data Protection Regulation in the European Union,[43] remain relevant. Generative AI models need vast training data, but using data with protected health information poses reidentification risks, as absolute deidentification is not attainable.[44,45] The use of patient-level data in commercial LLMs presents additional privacy risks, while open-source models require stringent data security measures. Strategies such as synthetic data generation[46] and encrypted computations[47] are being explored to enhance privacy protections. Ethical concerns surrounding AI extend beyond privacy. They include issues such as ensuring informed consent for AI-driven research, addressing the risk of AI-generated misinformation influencing health decisions, and improving transparency in how AI models make decisions. .[48] Studies, such as Gichoya et al. 2023, have demonstrated that AI models can detect sensitive attributes like race from medical images, raising concerns about unintended biases and their impact on health equity.[49] As generative AI continues to advance, ongoing regulatory oversight and ethical considerations will be critical to ensuring responsible implementation.

## Balancing Efficiency and Accuracy

Overall, the need for balancing efficiency and accuracy in scientific research using generative AI is paramount, and requires human interaction to appropriately guide development and use of these tools. While AI can enhance efficiency and standardization, ensuring validity, reproducibility, and transparency requires careful **human oversight, methodological rigor, and collaboration across stakeholders.** Ensuring fairness requires **proactive strategies to address biases in data and model design**, integrating inclusive datasets and algorithmic safeguards. **Ethical considerations, regulatory oversight, and privacy protections** are crucial to fostering trust and ensuring equitable benefits across diverse populations.

Ongoing improvements in model reliability, explainability, and integration into research best practices will be key to leveraging AI responsibly while maintaining scientific integrity.

## The Path Forward with AI-enabled RWE in HEOR

The integration of generative AI into HEOR presents an opportunity to enhance efficiency, accuracy, and consistency across literature reviews, RWE generation, predictive modelling, comparative effectiveness, and health economic modeling. However, realizing its full potential requires a **balanced approach that prioritizes scientific validity, fairness, and ethical considerations.**

To ensure accuracy and reliability, researchers must employ best practices such as **responsible engineering, retrieval-augmented generation, and domain-specific fine-tuning.** Addressing bias and fairness necessitates proactive strategies, including **data augmentation, federated learning, and algorithmic adjustments** to mitigate disparities in AI-driven analyses. Regulatory and ethical considerations remain paramount, particularly regarding **data privacy, informed consent, and transparency** in decision-making processes.

**AI will not replace human expertise but will continue to serve as a powerful tool to support researchers, policymakers, and industry stakeholders. Collaboration among multidisciplinary experts including AI developers, clinicians, and health economists, will be crucial in developing guidelines that ensure responsible AI use.** Continuous refinement of AI models, integration with research best practices, and adherence to evolving regulatory frameworks will help maximize benefits while safeguarding scientific integrity and equity in healthcare decision-making.

# References


1. Fleurence RL, Bian J, Wang X, et al. Generative Artificial Intelligence for Health Technology Assessment: Opportunities, Challenges, and Policy Considerations: An ISPOR Working Group Report. Value in Health. 2025;28(2):175-183.
2. Use of AI in evidence generation: NICE position statement. 2024; https://www.nice.org.uk/about/what-we-do/our-research-work/use-of-ai-in-evidence-generation--nice-position-statement. Accessed 2025-02-10, 2025.
3. Qureshi R, Shaughnessy D, Gill KAR, Robinson KA, Li T, Agai E. Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? Systematic Reviews. 2023;12(1):72.
4. Guo E, Gupta M, Deng J, Park Y-J, Paget M, Naugler C. Automated Paper Screening for Clinical Reviews Using Large Language Models: Data Analysis Study. J Med Internet Res. 2024;26:e48996.
5. Huang W, Poojary V, Hofer K, Fazeli M. MSR217 Evaluating a Large Language Model Approach for Full-Text Screening Task in Systematic Literature Reviews With Domain Expert Input. Paper presented at: ISPOR EU2024; Barcelona, Spain
6. Khraisha Q, Put S, Kappenberg J, Warraitch A, Hadfield K. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. Research Synthesis Methods. 2024;15(4):616-626.
7. Robinson A, Thorne W, Wu BP, et al. Bio-sieve: exploring instruction tuning large language models for systematic review automation. arXiv preprint arXiv:230806610. 2023.
8. Hasan B, Saadi S, Rajjoub NS, et al. Integrating large language models in systematic reviews: a framework and case study using ROBINS-I for risk of bias assessment. BMJ evidence-based medicine. 2024;29(6):394-398.
9. Gartlehner G, Kahwati L, Hilscher R, et al. Data extraction for evidence synthesis using a large language model: A proof-of-concept study. Research synthesis methods. 2024;15(4):576-589.
10. Huang W, Poojary V, Kasireddy E, Fazeli M. MSR28 Evaluating the Performance of GPT-4o and Retrieval-Augmented Generation (RAG) in Extracting Data From Journal Articles: A Comparative Study. Paper presented at: ISPOR EU2024; Barcelona, Spain.
11. Reason T, Rawlinson W, Langham J, Gimblett A, Malcolm B, Klijn S. Artificial Intelligence to Automate Health Economic Modelling: A Case Study to Evaluate the Potential Application of Large Language Models. PharmacoEconomics - Open. 2024;8(2):191-203.
12. Schopow N, Osterhoff G, Baur D. Applications of the Natural Language Processing Tool ChatGPT in Clinical Practice: Comparative Study and Augmented Systematic Review. JMIR Med Inform. 2023;11:e48933.
13. Hu Y, Chen Q, Du J, et al. Improving large language models for clinical named entity recognition via prompt engineering. Journal of the American Medical Informatics Association. 2024;31(9):1812-1820.
14. Keloth VK, Banda JM, Gurley M, et al. Representing and utilizing clinical textual data for real world studies: An OHDSI approach. Journal of biomedical informatics. 2023;142:104343.
15. Peng C, Yang X, Chen A, et al. A study of generative large language model for medical research and healthcare. npj Digital Medicine. 2023;6(1):210.
16. Xie Q, Chen Q, Chen A, et al. Me-LLaMA: Foundation Large Language Models for Medical Applications. Research square. 2024.
17. Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records. NPJ digital medicine. 2022;5(1):194.
18. Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D. Large language models are few-shot clinical information extractors. December, 2022; Abu Dhabi, United Arab Emirates.
19. Xu H, Usuyama N, Bagga J, et al. A whole-slide foundation model for digital pathology from real-world data. Nature. 2024;630(8015):181-188.
20. Dixon D, Sattar H, Moros N, et al. Unveiling the Influence of AI Predictive Analytics on Patient Outcomes: A Comprehensive Narrative Review. Cureus. 2024;16(5):e59954.
21. Aggarwal R, Sounderajah V, Martin G, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ Digit Med. 2021;4(1):65.
22. Han R, Acosta JN, Shakeri Z, Ioannidis JPA, Topol EJ, Rajpurkar P. Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review. Lancet Digit Health. 2024;6(5):e367-e373.
23. Sezgin E, Hussain SA, Rust S, Huang Y. Extracting Medical Information From Free-Text and Unstructured Patient-Generated Health Data Using Natural Language Processing Methods: Feasibility Study With Real-world Data. JMIR Form Res. 2023;7:e43014.
24. Wieland-Jorna Y, van Kooten D, Verheij RA, de Man Y, Francke AL, Oosterveld-Vlug MG. Natural language processing systems for extracting information from electronic health records about activities of daily living. A systematic review. JAMIA Open. 2024;7(2):ooae044.
25. Briggs AH, Weinstein MC, Fenwick EAL, Karnon J, Sculpher MJ, Paltiel AD. Model Parameter Estimation and Uncertainty: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-6. Value in Health. 2012;15(6):835-842.
26. Caro JJ, Briggs AH, Siebert U, Kuntz KM. Modeling Good Research Practices—Overview: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-1. Value in Health. 2012;15(6):796-803.
27. Eddy DM, Hollingworth W, Caro JJ, Tsevat J, McDonald KM, Wong JB. Model Transparency and Validation: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-7. Value in Health. 2012;15(6):843-850.

# References


28. Chhatwal J, Yildirim I, Balta D, et al. EE355 Can Large Language Models Generate Conceptual Health Economic Models? Value in Health. 2024;27(6):S123.

29. Ayer T, Samur S, Yildirim I, Bayraktar E, Ermis T. JC Fully Replicating Published Health Economic Models Using Generative AI. Paper presented at: Annual Meeting of the Society for Medical Decision Making2024.

30. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nature medicine. 2023;29(8):1930-1940.

31. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus. 2023;15(2).

32. Jin Q, Leaman R, Lu Z. Retrieve, Summarize, and Verify: How Will ChatGPT Affect Information Seeking from the Medical Literature? Journal of the American Society of Nephrology. 2023;34(8):1302-1304.

33. Lin Z. How to write effective prompts for large language models. Nature Human Behaviour. 2024;8(4):611-615.

34. Wang S, Scells H, Zhuang S, Potthast M, Koopman B, Zuccon G. Zero-shot generative large language models for systematic review screening automation. Paper presented at: European Conference on Information Retrieval2024.

35. Wei C-H, Allot A, Lai P-T, et al. PubTator 3.0: an AI-powered literature resource for unlocking biomedical knowledge. Nucleic Acids Research. 2024;52(W1):W540-W546.

36. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-453.

37. Schwartz R, Schwartz R, Vassilev A, 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.

38. Parikh RB, Teeple S, Navathe AS. Addressing Bias in Artificial Intelligence in Health Care. JAMA. 2019;322(24):2377-2378.

39. Drukker K, Chen W, Gichoya J, et al. Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment. Journal of Medical Imaging. 2023;10(6):061104-061104.

40. Juwara L, El-Hussuna A, El Emam K. An evaluation of synthetic data augmentation for mitigating covariate bias in health data. Patterns. 2024;5(4):100946.

41. Yang Y, Lin M, Zhao H, Peng Y, Huang F, Lu Z. A survey of recent methods for addressing AI fairness and bias in biomedicine. Journal of biomedical informatics. 2024;154:104646.

42. McNair D, Price N. Health care artificial intelligence: law, regulation, and policy. In: The National Academies Press; 2022.

43. Sartor G, Lagioia F. The impact of the General Data Protection Regulation (GDPR) on artificial intelligence. 2020.

44. Benitez K, Malin B. Evaluating re-identification risks with respect to the HIPAA privacy rule. Journal of the American Medical Informatics Association. 2010;17(2):169-177.

45. Simon GE, Shortreed SM, Coley RY, et al. Assessing and minimizing re-identification risk in research data derived from health care records. Egems. 2019;7(1):6.

46. Mosquera L, El Emam K, Ding L, et al. A method for generating synthetic longitudinal health data. BMC Medical Research Methodology. 2023;23(1):67.

47. Brännvall R, Forsgren H, Linge H. HEIDA: Software Examples for Rapid Introduction of Homomorphic Encryption for Privacy Preservation of Health Data. Studies in health technology and informatics. 2023;302:267-271.

48. Duffourc MN, Gerke S. Health Care AI and Patient Privacy—Dinerstein v Google. JAMA. 2024;331(11):909-910.

49. Gichoya JW, Banerjee I, Bhimireddy AR, et al. AI recognition of patient race in medical imaging: a modelling study. The Lancet Digital Health. 2022;4(6):e406-e414.