The Critical Need to Build the Evidence Base for AI Implementations in Biomedicine and Health

William Hersh, MD
Professor
Department of Medical Informatics & Clinical Epidemiology
School of Medicine
Oregon Health & Science University
Portland, OR, USA
https://www.ohsu.edu/informatics
Email: hersh@ohsu.edu
Web: http://www.billhersh.info/
Blog: https://informaticsprofessor.blogspot.com/
Twitter: @williamhersh

References

Chen, R.J., Wang, J.J., Williamson, D.F.K., Chen, T.Y., Lipkova, J., Lu, M.Y., Sahai, S., Mahmood, F., 2023. Algorithmic fairness in artificial intelligence for medicine and healthcare. Nat Biomed Eng 7, 719–742. https://doi.org/10.1038/s41551-023-01056-8

Claburn, T., 2024. AI is changing search, for better or worse [WWW Document]. The Register. URL https://www.theregister.com/2024/01/30/ai_is_changing_search/ (accessed 2.10.24).

Dhar, S., Shamir, L., 2021. Evaluation of the benchmark datasets for testing the efficacy of deep convolutional neural networks. Visual Informatics 5, 92–101. https://doi.org/10.1016/j.visinf.2021.10.001

Donzé, J., John, G., Genné, D., Mancinetti, M., Gouveia, A., Méan, M., Bütikofer, L., Aujesky, D., Schnipper, J., 2023. Effects of a Multimodal Transitional Care Intervention in Patients at High Risk of Readmission: The TARGET-READ Randomized Clinical Trial. JAMA Intern Med 183, 658–668. https://doi.org/10.1001/jamainternmed.2023.0791

Dorr, D.A., Adams, L., Embí, P., 2023. Harnessing the Promise of Artificial Intelligence Responsibly. JAMA 329, 1347–1348. https://doi.org/10.1001/jama.2023.2771

Embi, P.J., 2021. Algorithmovigilance-Advancing Methods to Analyze and Monitor Artificial Intelligence-Driven Health Care for Effectiveness and Equity. JAMA Netw Open 4, e214622. https://doi.org/10.1001/jamanetworkopen.2021.4622

Finlayson, S.G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I.S., Saria, S., 2021. The Clinician and Dataset Shift in Artificial Intelligence. N Engl J Med 385, 283–286. https://doi.org/10.1056/NEJMc2104626

Greenhalgh, T., Fisman, D., Cane, D.J., Oliver, M., Macintyre, C.R., 2022. Adapt or die: how the pandemic made the shift from EBM to EBM+ more urgent. BMJ Evid Based Med 27, 253–260. https://doi.org/10.1136/bmjebm-2022-111952

Han, R., Acosta, J.N., Shakeri, Z., Ioannidis, J., Topol, E., Rajpurkar, P., 2023. Randomized Controlled Trials Evaluating AI in Clinical Practice: A Scoping Evaluation. https://doi.org/10.1101/2023.09.12.23295381

Hassan, C., Spadaccini, M., Mori, Y., Foroutan, F., Facciorusso, A., Gkolfakis, P., Tziatzios, G., Triantafyllou, K., Antonelli, G., Khalaf, K., Rizkala, T., Vandvik, P.O., Fugazza, A.,

Rondonotti, E., Glissen-Brown, J.R., Kamba, S., Maida, M., Correale, L., Bhandari, P., Jover, R., Sharma, P., Rex, D.K., Repici, A., 2023. Real-Time Computer-Aided Detection of Colorectal Neoplasia During Colonoscopy : A Systematic Review and Meta-analysis. Ann Intern Med. https://doi.org/10.7326/M22-3678

Heneghan, J.A., Walker, S.B., Fawcett, A., Bennett, T.D., Dziorny, A.C., Sanchez-Pinto, L.N., Farris, R.W.D., Winter, M.C., Badke, C., Martin, B., Brown, S.R., McCrory, M.C., Ness-Cochinwala, M., Rogerson, C., Baloglu, O., Harwayne-Gidansky, I., Hudkins, M.R., Kamaleswaran, R., Gangadharan, S., Tripathi, S., Mendonca, E.A., Markovitz, B.P., Mayampurath, A., Spaeder, M.C., Pediatric Data Science and Analytics (PEDAL) subgroup of the Pediatric Acute Lung Injury and Sepsis Investigators (PALISI) Network, 2023. The Pediatric Data Science and Analytics Subgroup of the Pediatric Acute Lung Injury and Sepsis Investigators Network: Use of Supervised Machine Learning Applications in Pediatric Critical Care Medicine Research. Pediatr Crit Care Med. https://doi.org/10.1097/PCC.0000000000003425

Hersh, W., 2024a. Search still matters: information retrieval in the era of generative AI. J Am Med Inform Assoc ocae014. https://doi.org/10.1093/jamia/ocae014

Hersh, W., 2024b. Translational AI: A Necessity and Opportunity for Biomedical Informatics and Data Science [WWW Document]. NLM Musings from the Mezzanine. URL https://nlmdirector.nlm.nih.gov/2024/02/07/translational-ai-a-necessity-and-opportunity-for-biomedical-informatics-and-data-science/ (accessed 2.10.24).

Hersh, W., 2023. Physician and Medical Student Competence in AI Must Include Broader Competence in Clinical Informatics. Informatics Professor. URL https://informaticsprofessor.blogspot.com/2023/09/physician-and-medical-student.html (accessed 9.15.23).

Lancaster, F.W., 1979. Information retrieval systems: Characteristics, testing, and evaluation, 2nd ed edition. ed. John Wiley & Sons, New York.

Liu, X., Rivera, S.C., Moher, D., Calvert, M.J., Denniston, A.K., SPIRIT-AI and CONSORT-AI Working Group, 2020. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. BMJ 370, m3164. https://doi.org/10.1136/bmj.m3164

Mangas-Sanjuan, C., de-Castro, L., Cubiella, J., Díez-Redondo, P., Suárez, A., Pellisé, M., Fernández, N., Zarraquiños, S., Núñez-Rodríguez, H., Álvarez-García, V., Ortiz, O., Sala-Miquel, N., Zapater, P., Jover, R., CADILLAC study investigators*, 2023. Role of Artificial Intelligence in Colonoscopy Detection of Advanced Neoplasias : A Randomized Trial. Ann Intern Med. https://doi.org/10.7326/M22-2619

Meyer, A., Benn, R., 2023. Hype Cycle for Healthcare Providers, 2023 [WWW Document]. Gartner. URL https://www.gartner.com/en/documents/4534899 (accessed 1.6.23).

Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S., 2019. Dissecting racial bias in an algorithm used to manage the health of populations. Science 366, 447–453. https://doi.org/10.1126/science.aax2342

Plana, D., Shung, D.L., Grimshaw, A.A., Saraf, A., Sung, J.J.Y., Kann, B.H., 2022. Randomized Clinical Trials of Machine Learning Interventions in Health Care: A Systematic Review. JAMA Netw Open 5, e2233946. https://doi.org/10.1001/jamanetworkopen.2022.33946

Russell, R.G., Lovett Novak, L., Patel, M., Garvey, K.V., Craig, K.J.T., Jackson, G.P., Moore, D., Miller, B.M., 2023. Competencies for the Use of Artificial Intelligence-Based Tools by

Health Care Professionals. Acad Med 98, 348–356. https://doi.org/10.1097/ACM.0000000000004963

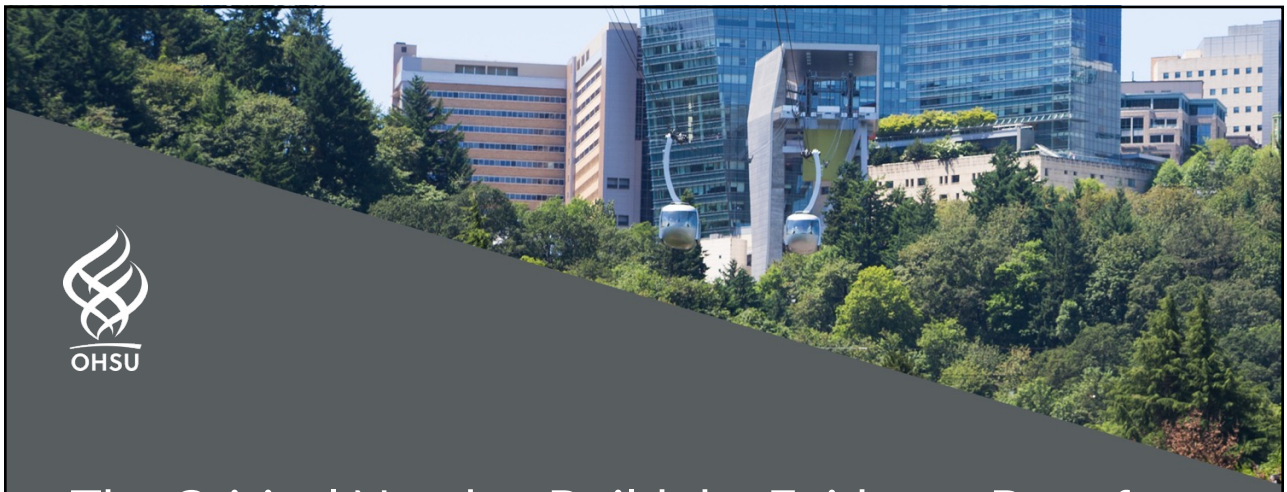
Vaid, A., Sawant, A., Suarez-Farinas, M., Lee, J., Kaul, S., Kovatch, P., Freeman, R., Jiang, J., Jayaraman, P., Fayad, Z., Argulian, E., Lerakis, S., Charney, A.W., Wang, F., Levin, M., Glicksberg, B., Narula, J., Hofer, I., Singh, K., Nadkarni, G.N., 2023. Implications of the Use of Artificial Intelligence Predictive Models in Health Care Settings : A Simulation Study. Ann Intern Med. https://doi.org/10.7326/M23-0949

Walker, S.C., French, B., Moore, R.P., Domenico, H.J., Wanderer, J.P., Mixon, A.S., Creech, C.B., Byrne, D.W., Wheeler, A.P., 2023. Model-Guided Decision-Making for Thromboprophylaxis and Hospital-Acquired Thromboembolic Events Among Hospitalized Children and Adolescents: The CLOT Randomized Clinical Trial. JAMA Netw Open 6, e2337789. https://doi.org/10.1001/jamanetworkopen.2023.37789

Wang, S., Scells, H., Koopman, B., Zuccon, G., 2023. Can ChatGPT Write a Good Boolean Query for Systematic Review Literature Search? https://doi.org/10.48550/arXiv.2302.03495

Youssef, A., Pencina, M., Thakur, A., Zhu, T., Clifton, D., Shah, N.H., 2023. External validation of AI models in health should be replaced with recurring local validation. Nat Med. https://doi.org/10.1038/s41591-023-02540-z

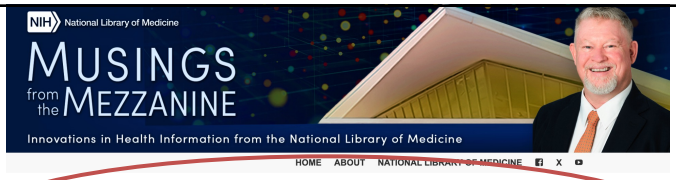
Zhou, Q., Chen, Z.-H., Cao, Y.-H., Peng, S., 2021. Clinical impact and quality of randomized controlled trials involving interventions evaluating artificial intelligence prediction tools: a systematic review. NPJ Digit Med 4, 154. https://doi.org/10.1038/s41746-021-00524-2



The Critical Need to Build the Evidence Base for AI Implementations in Biomedicine and Health

William Hersh, MD, FACMI
Professor
Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University
Portland, OR, USA

1



Translational AI: A Necessity and Opportunity for Biomedical Informatics and Data Science

Posted on February 7, 2024 by Guest Author

Journal of the American Medical Informatics Association, 2024, 1–3
https://doi.org/10.1093/jamia/ocae014
Perspective

AMIA OXFORD

Perspective

Search still matters: information retrieval in the era of generative AI

William Hersh, MD*

Department of Medical Informatics & Clinical Epidemiology, School of Medicine, Oregon Health & Science University, Portland, OR 97239, United States

*Corresponding author: William Hersh, MD, Department of Medical Informatics & Clinical Epidemiology, School of Medicine, Oregon Health & Science University, BICC, 3181 SW Sam Jackson Park Rd., Portland, OR 97239, United States (hersh@ohsu.edu)

Abstract

Objective: Information retrieval (IR, also known as search) systems are ubiquitous in modern times. How does the emergence of generative artificial intelligence (AI), based on large language models (LLMs), fit into the IR process?

Process: This perspective explores the use of generative AI in the context of the motivations, considerations, and outcomes of the IR process with a focus on the academic use of such systems.

Conclusions: There are many information needs, from simple to complex, that motivate use of IR. Users of such systems, particularly academics, have concerns for authoritativeness, timeliness, and contextualization of search. While LLMs may provide functionality that aids the IR process, the continued need for search systems, and research into their improvement, remains essential.

Key words: information storage and retrieval; generative artificial intelligence; large language models; ChatGPT.

ACMI 2024

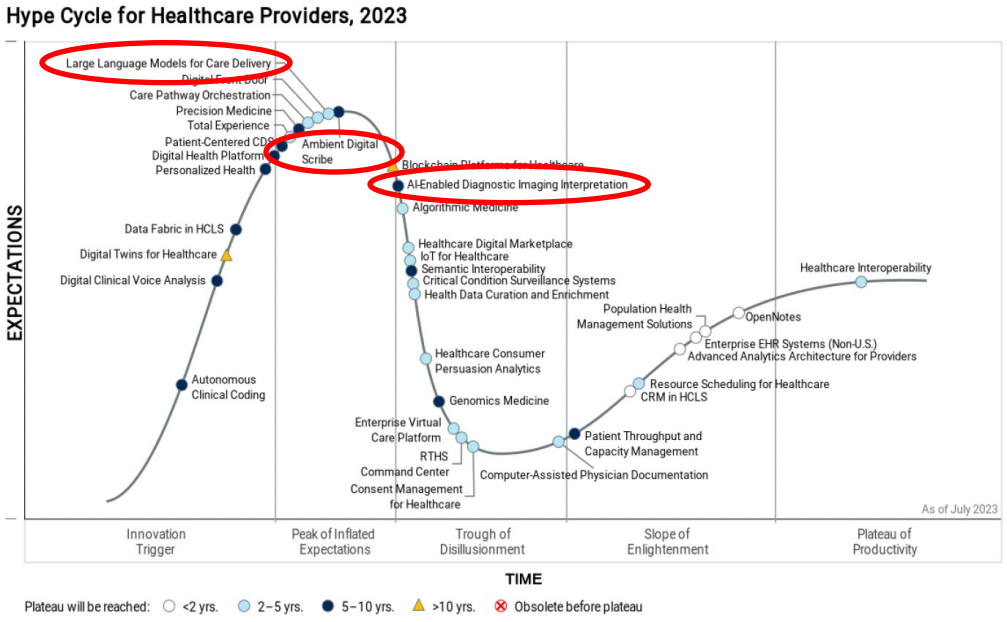
2

Overview of talk

- Translational AI
  - https://nlmdirector.nlm.nih.gov/2024/02/07/translational-ai-a-necessity-and-opportunity-for-biomedical-informatics-and-data-science/
  - https://informaticsprofessor.blogspot.com/2024/02/translational-ai-necessity-and.html
- Search still matters
  - https://doi.org/10.1093/jamia/ocae014
  - https://www.theregister.com/2024/01/30/ai_is_changing_search/
  - https://informaticsprofessor.blogspot.com/2024/01/whither-search-new-perspective-on.html

Overheard in a workshop on AI for innovative medical educators, September 2022

"I hope that AI won't be as awful as the EHR"



Gartner

OHSU

What do we need for AI applications to make it to the plateau of productivity?

- Translational AI
  - Show us the evidence
- Search still matters
  - In many circumstances, who said what is more important than providing a generated answer

OHSU

How do we "show the evidence?"

- From evidence-based medicine (EBM), best evidence for any clinical intervention is from randomized controlled trials (RCTs) or systematic reviews of RCTs
- Although not as easy to carry out as RCTs of drugs or devices (and placebos), AI must demonstrate benefit for patient outcomes and/or healthcare delivery improvement
  - Additional issues for RCTs of AI (Liu, 2020)
- As with drugs and devices, we need to move from "basic science" to "clinical science"
- Not everything can be studied in an RCT and RCTs cannot be done for every last clinical question (Greenhalgh, 2022)

7

What is the evidence so far?

- Many, many papers published about models and simulated use (basic science), including systematic reviews of those papers
- Very few RCTs demonstrating value from real-world use (clinical science) – systematic reviews of RCTs show (Zhou, 2021; Plana, 2022; Han, 2023)
  - Much smaller numbers of RCTs – about 100, depending on how we count
  - 65-82% of RCT showed positive outcomes
  - Many RCTs showed aspects of "risk of bias"
- Concerns about generalization beyond initial clinical settings
  - Biased data and algorithms (Obermeyer, 2019: Dhar, 2021; Chen, 2023)
  - Data and algorithm drift (Finlayson, 2021; Vaid, 2023)

8

Learning from some specific examples

- Computer-aided detection (CADe) of polyps in colonoscopy
  - One of earliest and most widely-studied applications of AI
  - Recent systematic review shows polyps missed by colonoscopists are discovered, but mostly small and clinically inconsequential (Hassan, 2023)
  - RCT of CADe found no increased detection of advanced neoplasias (Mangas-Sanjuan, 2023)
- 30-day hospital readmissions
  - After implementation of CMS penalty, proliferation of highly accurate predictive models published in mid-2010s
  - Recent RCT showed use of high-quality model and implementation of program around it did not reduce readmissions (Donzé, 2023)

Examples (cont.)

- RCT to assess whether use of previously validated hospital-acquired venous thromboembolism (HA-VTE) prognostic model, together with pediatric hematologist review, could reduce pediatric inpatient rates of HA-VTE (Walker, 2023)
  - No difference for intervention group randomized to use model
  - Reluctance to use model by primary care physicians – used only 26% of time
  - Even for children in intervention arm, model mostly not used, i.e., the "Cassandra Problem" (Wilson, 2023)
- Scoping review of ML in pediatric critical care (Heneghan, 2023)
  - *Publication of supervised machine learning models to address clinical challenges in pediatric critical care medicine has increased dramatically… While these approaches have the potential to benefit children with critical illness, the literature demonstrates incomplete reporting, absence of external validation, and infrequent clinical implementation.*

How do we get to "translational AI?"

- Singh, X, Feb 8 2024: *Researched models aren't implemented. Implemented models aren't researched*
- Clinician competence and education (Russell, 2023; Hersh, 2023)
- Postmarket surveillance, e.g., *algorithmovigilance* (Embi, 2021)
- Responsible use of AI (Dorr, 2023)

Search still matters

- Generative AI systems such as ChatGPT are cool and fun, but
  - For some tasks that many of us do, need more than answers, e.g.,
    - Clinical – patient-care questions
    - Research – methods and insights
    - Teaching – synthesizing knowledge for our students
  - Where the information comes from is as important what it says
- Information retrieval (IR) systems "do not inform user about a subject; indicate the existence (or nonexistence) and whereabouts of documents related to an information request" (Lancaster, 1978)

Search in the era of generative AI

- Another adage of EBM
  - Gen AI for background questions
  - Search and critical appraisal for foreground questions
- Retrieval-augmented generation (RAG) for improving Gen AI but do we need "generation-augmented retrieval" for LLMs to aid search?
  - Evidence modest so far, e.g., using ChatGPT for generating Boolean queries did not improve search results (Wang, 2023)
- Help us learn more
  - Text Retrieval Conference (TREC) 2024 – Biomedical Generative Retrieval, https://trec-cds.org/

Questions?

William Hersh, MD
Professor
Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University
Portland, OR, USA
Email
hersh@ohsu.edu
Web
http://www.billhersh.info
Blog
https://informaticsprofessor.blogspot.com/
Textbook
http://www.informaticsbook.info
What is Informatics?
http://informatics.health