

8.6b Evaluation and Future Directions

10x10 – Introduction to Biomedical Informatics & Artificial Intelligence
William Hersh, MD, FACMI, FAMIA, FIAHSI
Copyright 2026, All Rights Reserved



Applying IR to medical records for cohort discovery

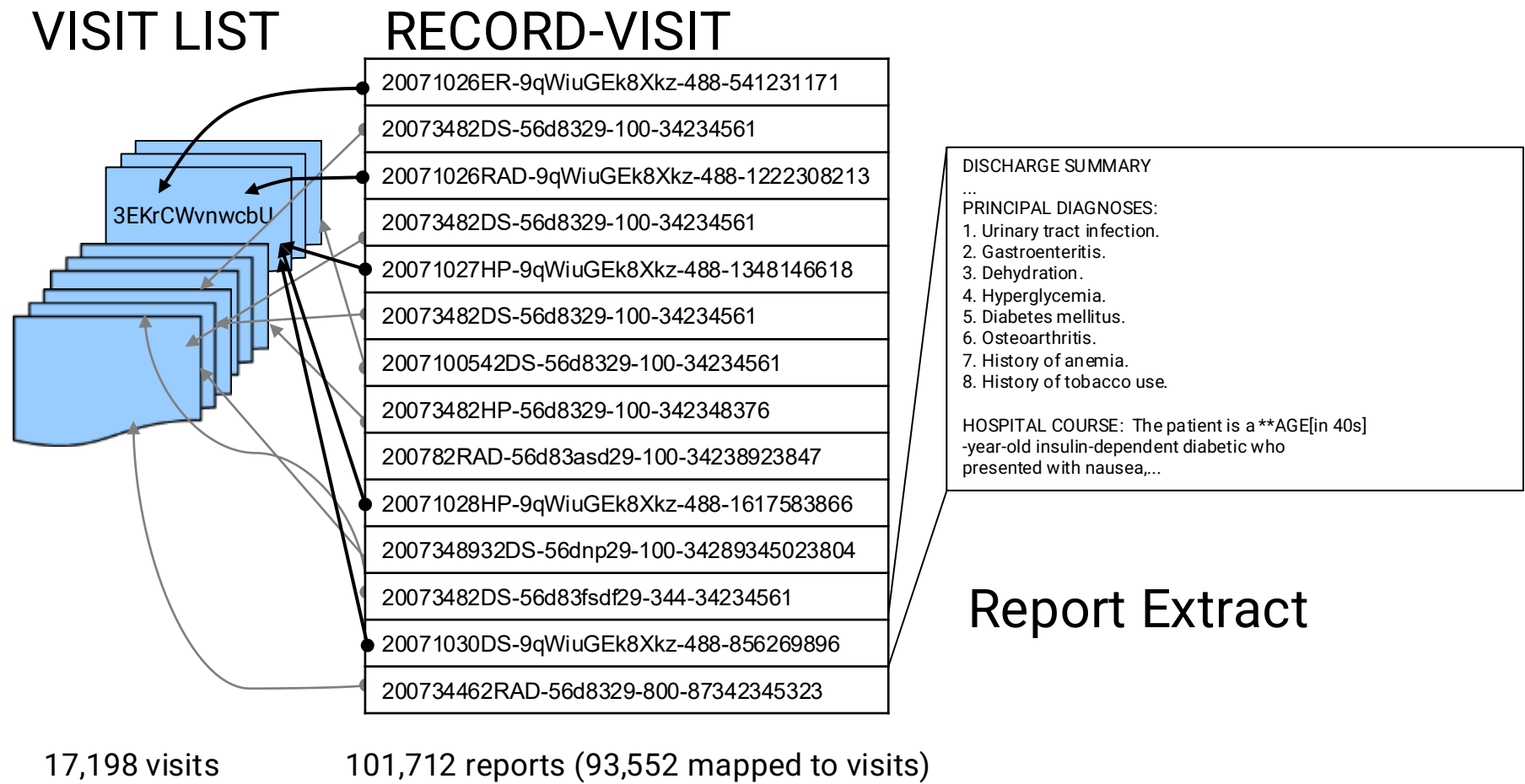
Most medical records still have large amounts of narrative documents

- Growth in EHR implementation has led to development and use of search tools and capabilities, e.g., EMERSE (Hanauer, 2020)

For some tasks, can we take an IR approach?

- TREC Medical Records Track used de-identified corpus of medical records in initial task of identifying patients as candidates for clinical research studies (Voorhees, 2012)

TREC Medical Records Track test collection



(Courtesy, Ellen Voorhees, NIST)

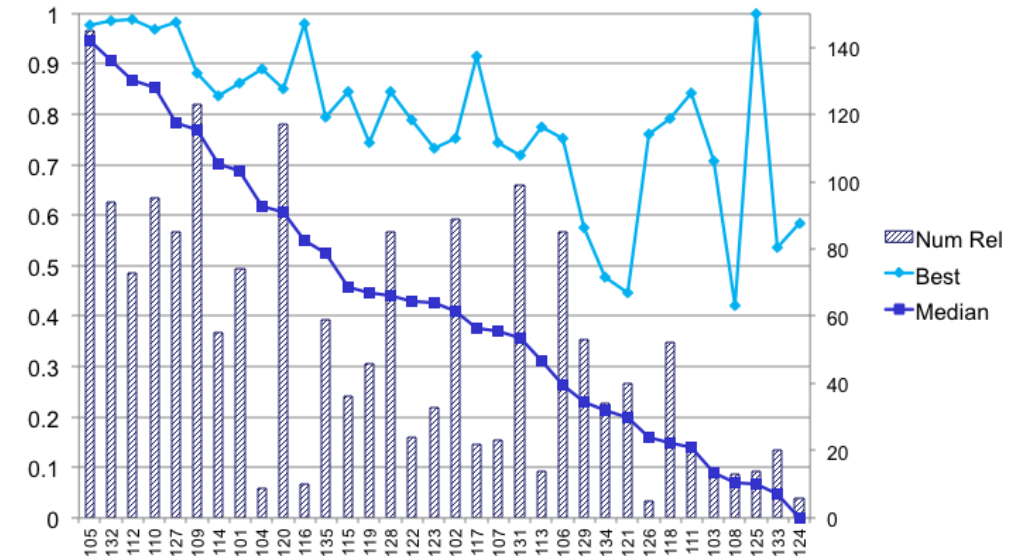
TREC Medical Records Track results (Voorhees, 2012)

Highly variable across different topics

- Easiest – consistently best results
 - 105: Patients with dementia
- Hardest – consistently worst results
 - 108: Patients treated for vascular claudication surgically
- Large differences between best and worst results
 - 125: Patients co-infected with Hepatitis C and HIV

Overall results show substantial room for improvement

- Best results involve manual modification of queries



Further work in EHR retrieval for cohort identification

Used larger and non-de-identified records for more realistic study

Developed set of 56 topics and evaluated over full records of 100,000 patients each from OHSU and Mayo Clinic (Wu, 2017; Chamberlin, 2020)

- Focused Boolean queries worked better than natural language queries

Challenges – data sets not sharable due to privacy concerns

- Developing standardized data format using Observational Medical Outcomes Partnership (OMOP) format so others can replicate with own data

IR evaluation in era of generative AI

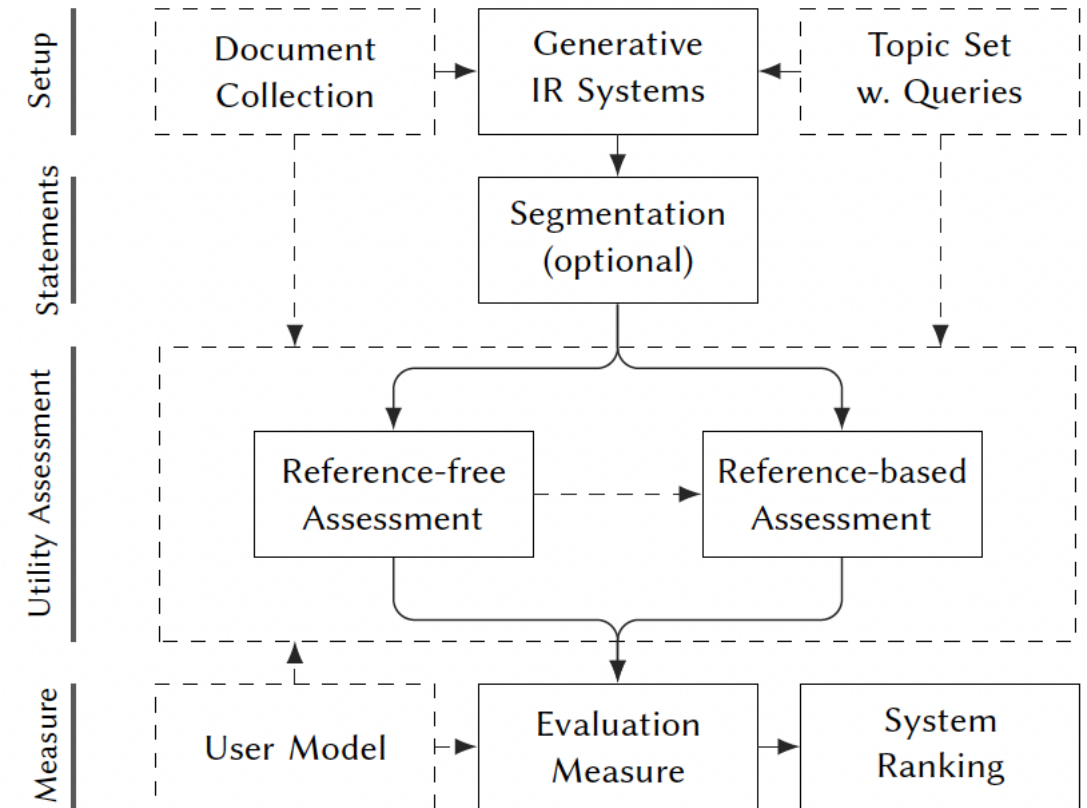
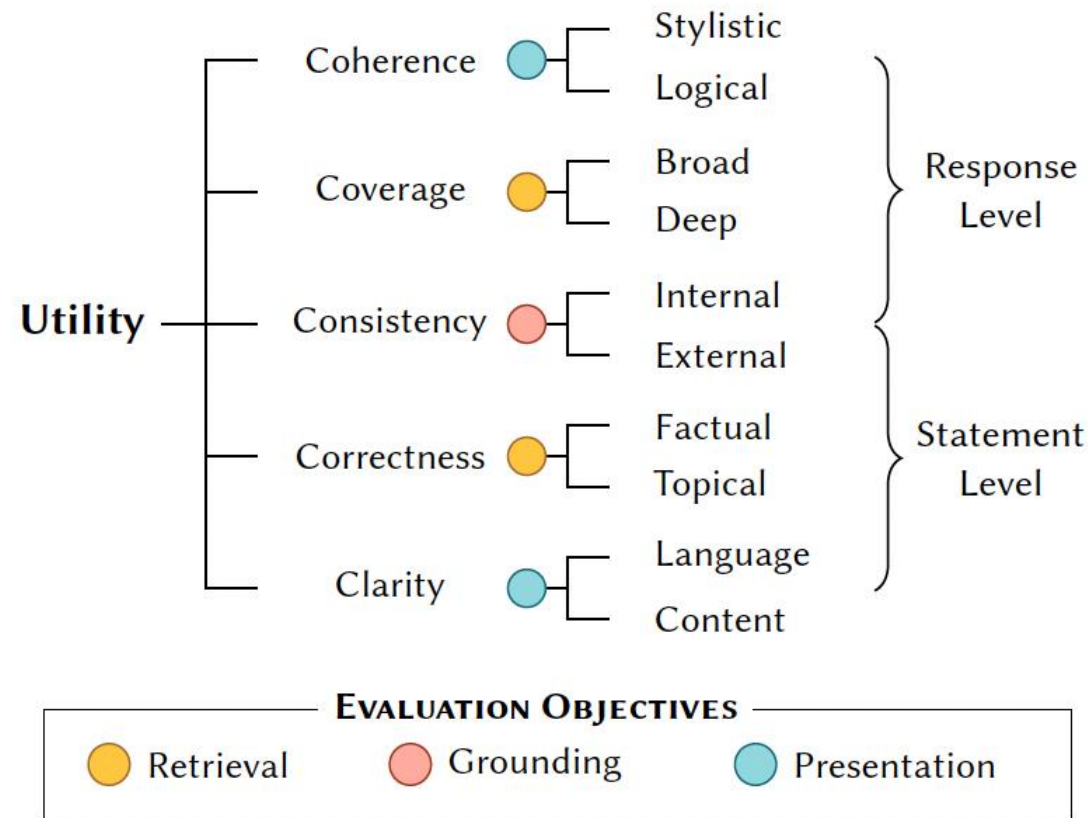
What to evaluate?

- GAR – use of generation to improve and/or complement retrieval
- RAG – use of retrieval for generation

Evaluation objectives (Gienapp, 2023)

- Retrieval – classic IR paradigm
- Grounding – veracity of generative output
- Presentation – output concise, coherent, and accessible

Evaluation of generative IR (Gienapp, 2024)



Other roles for generative AI in IR evaluation

Generating relevance labels comparable to humans (Faggioli, 2023; Thomas, 2023)

Can replace human raters but concerns over quality and bias (Anderson, 2023)

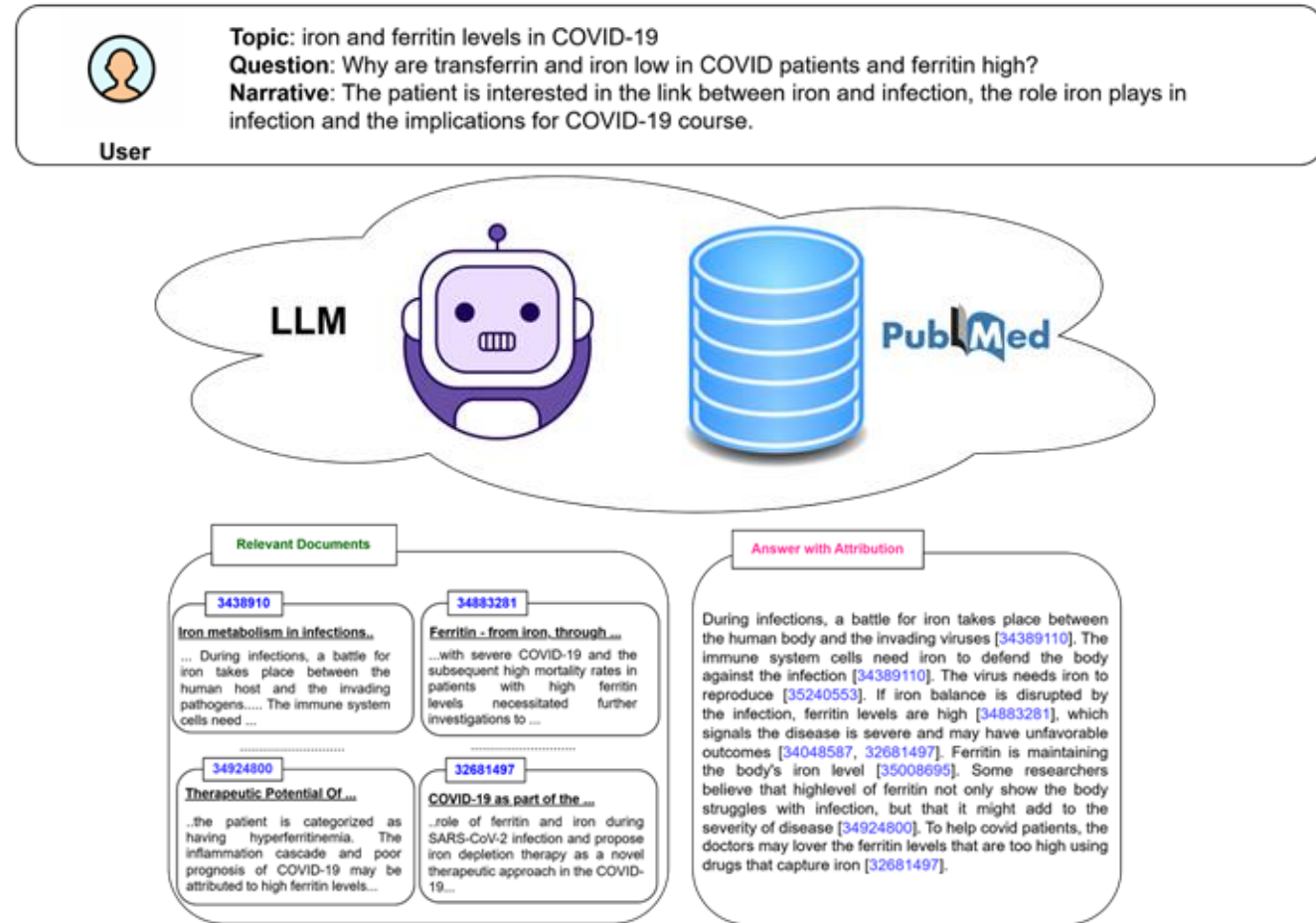
LLMs used in evaluation may favor their own LLM used in task (Liu, 2023)

Improve efficiency but at cost of bias? (Faggioli, 2024)

Use in TREC RAG Track – deemed to provide comparable results to human assessments (Rahmani, 2025; Upadhyay, 2024), but practical and theoretical concerns (Clarke, 2024; Soboroff, 2025)

TREC Biomedical Generative Retrieval (BioGen) Track (Gupta, 2024)

Generate answers using large language models (LLMs), where each answer sentence must be supported by appropriate references



TREC Biogen 2024 test collection

Documents

- Annual baseline snapshot of PubMed through end of 2023
- Provided pre-processed set of 20,727,695 PMIDs representing the abstracts in 2023 snapshot

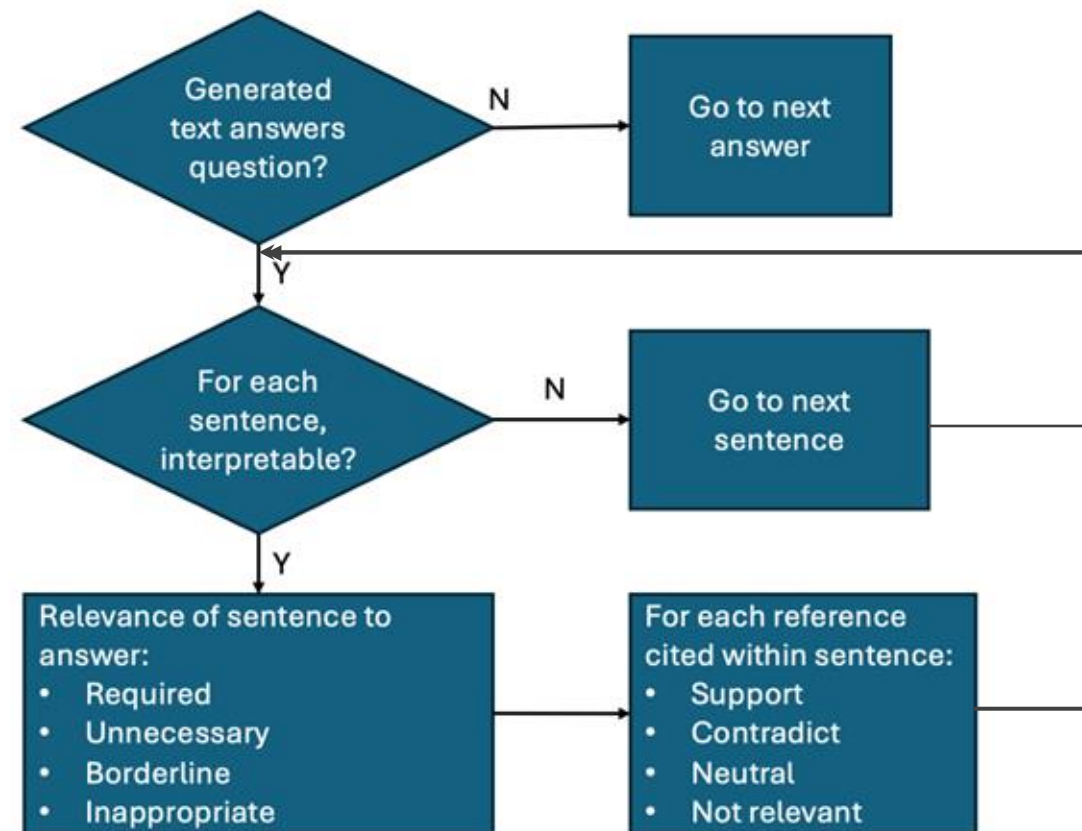
Topics

- 40 topics developed using information requests submitted by self-identified non-clinicians to NLM
- 25 topics developed based on the collection that contains questions generated using documents from Mayo Clinic, UpToDate , and Reddit r/AskDocs (Wu, 2025)

Relevance assessments

- By clinicians in informatics graduate programs

Relevance assessment process



Methods

Methods

Quality and factuality of text generated by LLMs

- Answer completeness (recall), precision, redundancy

Reference attribution

- Citation quality, support, contradict

Document relevance

- Recall, precision

Metrics

Most participants used variants of

- Retrieved documents with lexical search method (BM25)
- Reranking to obtain top-k relevant documents/snippets
- LLMs then used to generate answer citing appropriate documents

Overall results and future directions (Gupta, 2024)

Accuracy generally high ~90%

Citation coverage 70-90%

Citation support rate 60-80%

Document recall low and precision high – reflective of database as large as MEDLINE

Conclusion: LLMs not ready to replace human searching and information exploration

Future of IR and generative AI

Generative AI shows potential for tasks historically associated with IR

- Better for answering queries than providing resource attribution

Must also be cognizant of other knowledge-based information issues, e.g., access, production and quality, integration, etc.

Maybe an adage from evidence-based medicine (EBM)?

- Gen AI for background questions
- Search and critical appraisal for foreground questions