Navidence.

Poster Code: A-026

Searching for Signal in the Semantic Noise: Evaluating Augmented Search Strategies for Retrieval of Computable Operational Definitions in Pharmacoepidemiologic Study Design

Scan for poster + more

Aimee Harrison MFA,[1] Megan Somerday BA,[1] Isaac Parker,[1] Michael Buck PhD,[1] Aaron Kamauu MD MS MPH,[1] Craig Parker MD MS[1]

[1]Navidence, Inc., USA

DISCLOSURES: M Somerday, I Parker, M Buck, A Kamauu, and C Parker are employees of Navidence. Aimee Harrison was an employee of Navidence during this work.

So What?

Search strategies employing ontology-based expansion and dense retrieval produce reliable, reproducible, and more efficient results than LLMs. Dense retrieval and UMLS synonym expansion both show promise for generating improved precision and recall, with UMLS expansion showing more promise for precise concept matching.

BACKGROUND

Context

Efficient search strategies are an essential component of database design for clinical research tools. At the time of this study, Navidence's knowledge base contained >3,650 Computable Operational Definitions (CODefs). The results of four search strategies were compared for 15 keyword queries against database CODefs.

The Gap / Problem

Different search strategies produce highly variable results, and search recall is notably impacted by variance in terminology such as abbreviations and common lay synonyms for technical terms. While LLMs can be used to enhance searches, results are not strictly reproducible. Strategies employing ontology-based expansion and dense retrieval produce reliable, reproducible, and more efficient results than LLMs. While dense retrieval has been shown to improve search results given variance in terminology,[1,2] knowledge-based query expansion has not been uniformly shown to improve search results.[3,4] Few studies explore the various combinations of these search strategies.

Pharmacoepidemiologic Relevance

Reliable, reproducible search functionality is foundational to pharmacoepidemiologic study design. As CODef libraries grow, the ability to retrieve relevant operational definitions for exposures, outcomes, and covariates directly affects study efficiency and reproducibility. Findings can inform search and recommendation tools that support transparent, well-documented study design consistent with initiatives such as ENCePP and Good Pharmacoepidemiology Practice (GPP).

OBJECTIVES

Primary Objective

To compare the retrieval accuracy and ranking quality of (1) approximate string matching using RapidFuzz[5], (2) UMLS-expanded string matching, where queries were expanded with Metathesaurus[6] synonyms, (3) dense retrieval using S-BioBERT embeddings via Pinecone, and (4) UMLS-expanded dense retrieval (with S-BioBERT[1]) for retrieving entries in a structured database of computable operational definitions (CODefs) to inform the development of search and recommendation functionality in study design tools.

Secondary Objectives

Identify promising avenues for expansion and improvement of search functionality.

LIMITATIONS

Study size was small, at only fifteen search terms. Database users may reasonably expect a search to return related clinical concepts; while dense retrieval performs well on this task, there are challenges to defining relatedness for the purposes of creation of a list of expected results. Dense retrieval may therefore appear to underperform relative to expected results. Additionally, it is necessary to use distinct similarity metrics for dense retrieval and fuzzy search, but there are limitations to directly comparing search results at the same similarity threshold for these distinct metrics.

Computable Operational Definitions (CODefs™)

Study elements — exposures, outcomes, and covariates — were defined using Navidence CODefs™: rigorous, data source-agnostic algorithms that ensure precision and reproducibility across real-world data sources.

navidence.com

METHODS

Study Design & Data Source

For each of 15 keyword queries, four search strategies were compared against all labels in the structured Navidence CODef database, with abbreviations and synonyms stripped out. **For the two strategies incorporating UMLS synonyms, Concept Unique Identifiers (CUIs) were used to create a list of synonyms for the search term from the UMLS Metathesaurus (e.g. "GERD" for "gastroesophageal reflux disease"). The original search term was also included in the list of synonyms. Each term in the list was searched against the database, with highest similarity score for any term in the list retained as that query's overall score.** Similarity scores were calculated for results: cosine similarity for dense retrieval, and modified Indel similarity (Indel similarity calculated on strings split by word and sorted alphabetically, then normalized from percentage match to a decimal from 0 to 1)[5] for fuzzy search. Results with a similarity score of 0.7 or greater were included. Precision and recall were calculated using expert-curated expected results, defined as either exact conceptual match (e.g. "diagnosis of myocardial infarction" for "MI"), or clinical metrics associated with the concept (e.g. "LDL (any value)" for "high cholesterol").

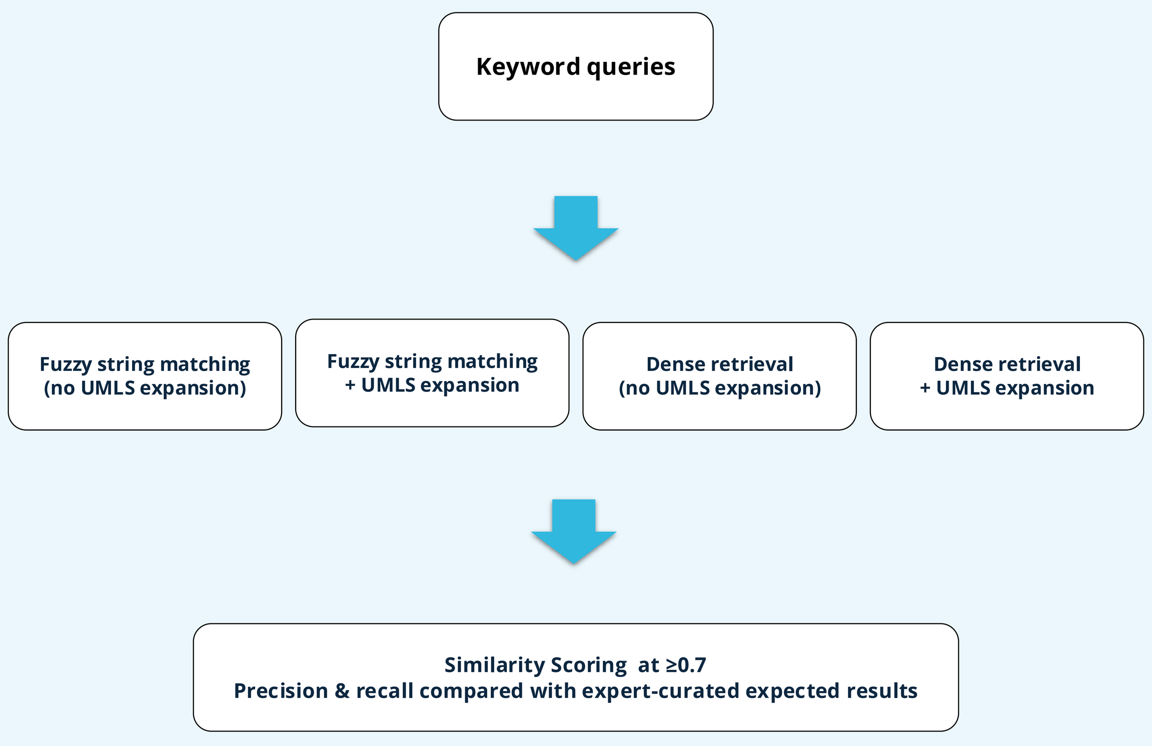

Figure. Search strategy comparison workflow: keyword queries were run through four retrieval strategies, scored by similarity, and evaluated against expert-curated expected results.

Example: Query Comparison

Search string: "respiratory syncytial virus immunization"

Expert-Curated Expected Results	Fuzzy + UMLS (top 5)	Dense + UMLS (top 5)
• Respiratory Syncytial Virus (RSV) Vaccination (≥1 px) • Respiratory Syncytial Virus (RSV) Vaccination (≥1 rx) • Monoclonal Antibody Therapy for RSV (≥1 rx) • Monoclonal Antibody Therapy for RSV (≥1 rx or px) • Monoclonal Antibody Therapy for Respiratory Syncytial Virus (RSV) (≥1 px) • Monoclonal Antibody Therapy for Respiratory Syncytial Virus (RSV) (≥1 rx or px) • Monoclonal Antibody Therapy for Respiratory Syncytial Virus (RSV) (≥1 rx)	• Respiratory Syncytial Virus (RSV) Infection Test (Order Procedure) (0.829) • Respiratory Syncytial Virus (RSV) Infection Test (any value) (0.829) • Respiratory Syncytial Virus (RSV) Infection Test (positive) (0.829) • Respiratory Syncytial Virus (RSV) Infection Test (negative) (0.829) • Respiratory Syncytial Virus (RSV) (≥2 any dx) (0.806)	• **Respiratory Syncytial Virus (RSV) Vaccination (≥1 px) (0.970)** • **Respiratory Syncytial Virus (RSV) Vaccination (≥1 rx) (0.970)** • Respiratory Syncytial Virus (RSV) (≥2 any dx) (0.806) • Respiratory Syncytial Virus (RSV) (Broad) (≥1 inpt or ≥2 any dx) (0.806) • Respiratory Syncytial Virus (RSV) (Broad) (≥1 dx) (0.806)

Table 3. Expert-curated expected results vs. top-5 outputs from Fuzzy + UMLS and Dense + UMLS search strategies (similarity scores shown; matches to expected results in green).

Because dense retrieval operates in a learned semantic space, it tends to outperform fuzzy matching when a query uses a lay synonym or alternate term without requiring the user to know or search every possible variant in advance (e.g., a search for "heart attack" can surface concepts indexed under "myocardial infarction"). Above, Dense + UMLS recognized "immunization" as a synonym for "vaccination," while Fuzzy + UMLS surfaced lexically similar but conceptually distinct results (RSV diagnostic tests). This same property also allows dense retrieval to surface closely related, but not exactly synonymous, concepts. For example, in a study of asthma, dense retrieval might also surface "treatment for asthma," concepts that were not included due to how expected results were defined, but that a researcher may want to include. This behavior may be more desirable for database search users but led to decreased precision in study results.

TRANSPARENCY & REPRODUCIBILITY

Protocol: study design, computable operational definitions (CODefs™), and code lists are documented internally and available from the corresponding author on request; no separate protocol was registered. **Pre-registration:** not registered. **Analytic code:** available from the authors on request. **Reporting:** consistent with STROBE and RECORD-PE principles.

RESULTS

>3,650
CODefs in Navidence Knowledge Base

0.47
Best Precision (Dense + UMLS)

0.69
Best Recall (Dense + UMLS)

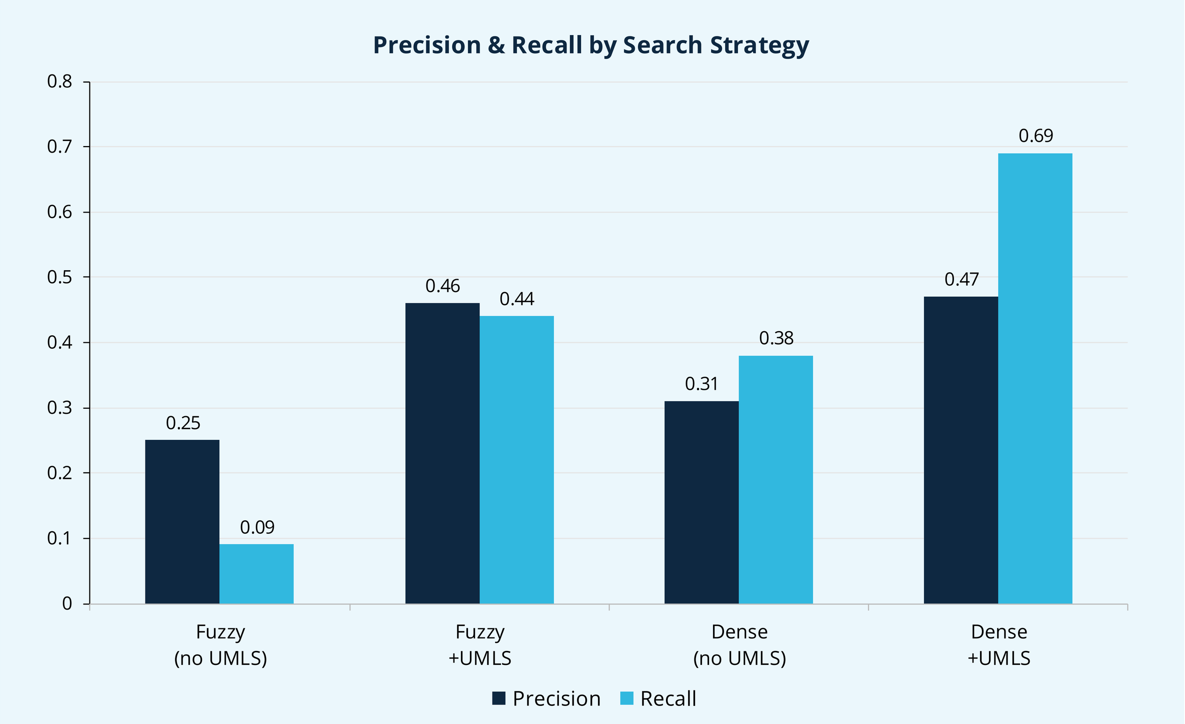

Figure 1. Precision and recall by search strategy, with and without UMLS synonym expansion (similarity threshold ≥0.7).

Search Strategy	Precision	Recall
Fuzzy string matching (no UMLS)	0.25	0.09
Dense retrieval – S-BioBERT (no UMLS)	0.31	0.38
UMLS-expanded fuzzy matching	0.46	0.44
UMLS-expanded dense retrieval (S-BioBERT)	0.47	0.69

Table 1. Precision and recall by search strategy (results with similarity score ≥0.7).

Results

Without UMLS expansion, dense retrieval outperformed fuzzy string matching on both precision (0.31 vs. 0.25) and recall (0.38 vs. 0.09). With UMLS expansion, dense retrieval and fuzzy matching performed similarly on precision (0.47 vs. 0.46) but dense retrieval outperformed fuzzy matching on recall (0.69 vs. 0.44). UMLS expansion resulted in better precision and recall across both search strategies. These results suggest that UMLS-based synonym expansion improves retrieval quality regardless of the underlying matching strategy, and that pairing it with dense retrieval yields the strongest overall recall of the four strategies tested.

CONCLUSIONS

Key Findings

Dense retrieval and UMLS synonym expansion both show promise for generating improved precision and recall, with UMLS expansion showing more promise for precise concept matching. Expected results were defined narrowly for the purposes of this assessment; for dense retrieval, this categorized closely related concepts as false positives, though retrieval of these items may be desirable in many use cases. In all search strategies, similarity score threshold may require adjustment to balance precision and recall for specific use cases.

Complex queries may benefit from further enhanced search strategies, such as multi-step, hybrid processes or expansion of semantic relationships using UMLS Semantic Network. Because the underlying database includes UMLS ontology type (e.g., diagnosis, procedure, prescription), potential extensions of functionality include allowing users to optionally filter results by concept type, for example restricting an asthma-related search to prescription concepts rather than returning all asthma-related concepts. This could offer a complementary way to increase search precision.

Clinical & Policy Implications

Findings can inform development of search and recommendation tools for CODef databases, helping researchers retrieve accurate, reproducible operational definitions more efficiently during pharmacoepidemiologic study design.

REFERENCES

1. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020 Feb 15;36(4):1234-1240. doi: 10.1093/bioinformatics/btz682. PMID: 31501885; PMCID: PMC7703786.
2. Ngo DH, Kemp M, Truran D, Koopman B, Metke-Jimenez A. Semantic Search for Large Scale Clinical Ontologies. AMIA Annu Symp Proc. 2022 Feb 21;2021:910-919. PMID: 35308904; PMCID: PMC8861757.
3. Hersh W, Price S, Donohoe L. Assessing thesaurus-based query expansion using the UMLS Metathesaurus. Proc AMIA Symp. 2000:344-8. PMID: 11079902; PMCID: PMC2244120.
4. Massonnaud, Clément & Kerdelhué, Gaétan & Grosjean, Julien & Lelong, Romain & Griffon, Nicolas & Darmoni, Stefan. (2020). Identification of the Best Semantic Expansion to Query PubMed Through Automatic Performance Assessment of Four Search Strategies on All Medical Subject Heading Descriptors: Comparative Study. JMIR Medical Informatics. 8. e12799. 10.2196/12799.
5. RapidFuzz 3.14.5 documentation. https://rapidfuzz.github.io/RapidFuzz/Usage/fuzz.html#token-sort-ratio. Accessed 08/2019.
6. National Institute of Health Unified Medical Language System Metathesaurus. https://uts.nlm.nih.gov/uts/umls/home

Contact: aaron@navidence.com