applied system innovation

MDPI

Article

Using Natural Language and Health Ontologies in Hope Recommender System: Evaluation of Use in Medicine

Hans Eguia [1,2], Carlos Sánchez-Bocanegra [1,*], Carlos Fernandez Llatas [3], Fernando Alvarez López [4] and Francesc Saigí-Rubió [1]

[1] Faculty of Health Sciences, Open University of Catalonia (UOC), 08018 Barcelona, Spain; heguia@uoc.edu (H.E.); fsaigi@uoc.edu (F.S.-R.)
[2] SEMERGEN New Technologies Working Group, 28009 Madrid, Spain
[3] ITACA-SABIEN, Institute of Applied Information Technologies and Advanced, Valencia Polytechnic University, 46022 Valencia, Spain; cfllatas@itaca.upv.es
[4] Faculty of Health Sciences, Manizales University, Manizales 170001, Colombia; falvarezlo@uoc.edu
* Correspondence: csanchezbo@uoc.edu; Tel.: +34-671-569-510

Abstract

Objectives: Despite the widespread availability of digital clinical information, timely access to relevant biomedical evidence during routine consultations remains limited in practice. Primary care clinicians, in particular, face significant time constraints that make it difficult to integrate comprehensive literature searches into everyday workflows. This study evaluates whether an ontology-based recommender system can support routine clinical workflows by reducing information retrieval time while preserving the clinically acceptable usefulness of retrieved evidence. We assessed the performance of the HOPE (Health Operation for Personalised Evidence) system compared with realistic manual PubMed searches conducted by physicians. **Materials and Methods:** We conducted an observational evaluation involving 50 primary care physicians, who independently assessed 30 anonymised, rewritten clinical cases representative of common primary care scenarios. HOPE automatically extracted biomedical concepts from case descriptions using natural language processing and mapped them to Unified Medical Language System (UMLS) ontologies to generate ranked PubMed recommendations. A subset of 10 physicians also conducted manual PubMed searches in line with their usual clinical practice. Article relevance was assessed using a predefined binary criterion, and a reference relevance set was established by consensus among three senior physicians using a pooled document set. Retrieval performance was evaluated using Precision@k, relative Recall@k, and Normalised Discounted Cumulative Gain (NDCG@k). Manual search time was measured using a standardised stopwatch protocol, whereas HOPE response time was logged automatically by the system. **Results:** Inter-physician agreement in relevance assessment was substantial (Fleiss' $\kappa = 0.66$; 95% CI: 0.61–0.70). HOPE achieved moderate-to-high precision within the top-ranked results (Precision@3 = 0.72), with relative recall increasing as additional documents were considered. Ranking metrics indicated that relevant articles were generally positioned early in the result lists. The mean total retrieval time for manual PubMed searches was 13.3 ± 1.7 min per case, compared with 17.4 ± 2.1 s for HOPE-assisted retrieval ($p < 0.001$). **Conclusions:** In a controlled, workflow-oriented evaluation using synthetic clinical cases, HOPE substantially reduced information retrieval time while maintaining clinically acceptable relevance in the retrieved literature. These findings support the use of ontology-based, AI-assisted systems as workflow-support tools to facilitate timely access to biomedical evidence, without replacing clinical judgment.

Academic Editor: Lu Meng

Received: 21 March 2026
Revised: 11 April 2026
Accepted: 20 April 2026
Published: 27 April 2026

Keywords: ontologies; clinical decision support; natural language processing; artificial intelligence; evidence-based medicine

1. Introduction

The increasing availability of digital clinical information has changed everyday medical practice. Electronic health records (EHRs) now contain large volumes of structured and unstructured data, including clinical notes, laboratory results, and diagnostic reports. The growing volume of unstructured data in EHR systems often requires manual review, which is time-intensive and increases the risk of diagnostic errors. While these data have the potential to support evidence-based medicine, they also contribute to a growing cognitive burden for clinicians, particularly in primary care settings where consultation time is limited (professional burnout can increase by up to 30%) [1].

In routine clinical settings, evidence-based medicine depends not only on the availability of high-quality research but also on clinicians' ability to access and interpret that evidence within limited consultation times. This operational gap remains a major barrier to the practical implementation of Evidence-based medicine (EBM), particularly in primary care. The information used to support clinical decisions must be reliable and, preferably, derived from sources such as certified medical databases [2]. EBM involves the integration of the best available evidence with clinical expertise in real-world decision-making [3], prioritising controlled clinical observations over anecdotal experiences, biological experiments, or personal intuition.

However, accessing relevant biomedical literature during routine consultations remains difficult. Manual searches typically involve translating a clinical problem into searchable terms, constructing and refining database queries, and screening multiple results to identify potentially useful articles. These steps are time-consuming [4] and cognitively demanding, and they often compete with other clinical tasks. Automating this process with artificial intelligence (AI) could substantially reduce time and improve consistency in clinical decision-making [5]. Studies have already demonstrated the feasibility of extracting information from clinical narratives [6]. These automated methods are emerging due to the increasing volume and accessibility of electronic information available [7], with AI presenting as a promising solution.

Natural language processing (NLP) and biomedical ontologies have been proposed as tools to support information retrieval by automating parts of this process. By extracting clinically meaningful concepts from free-text documentation and mapping them to standardised terminologies, such systems can assist in generating structured queries and retrieving relevant literature from biomedical databases. Several existing platforms have demonstrated the technical feasibility of combining NLP and ontology-based approaches for information retrieval.

Despite these advancements, challenges remain, particularly in semantic and structural data integration [8]. This is where ontologies come into play. Ontologies provide a structured framework for integrating biomedical data, as demonstrated by standards such as ICD-11 (International Classification of Diseases) [9,10], MeSH (Medical Subject Headings) [11], and SNOMED-CT, considered the most comprehensive clinical terminology currently available, with more than 300,000 medical concepts [9]. Many studies focus on algorithmic accuracy or domain-specific applications, while fewer evaluate the practical trade-off between retrieval speed and clinical usefulness. Comparisons with realistic manual searches performed by clinicians are particularly scarce.

The HOPE system (Health Operation for Personalised Evidence) was developed to address this gap. It is an ontology-driven recommender system designed to support clinicians in identifying potentially relevant sources of biomedical evidence during routine practice. HOPE uses NLP to extract biomedical concepts from EHR-derived clinical text, maps these concepts to the Unified Medical Language System (UMLS) [10], and generates ranked queries to retrieve literature from established medical databases. This signifies that NLP is the task of converting unstructured human language data into structured data that a machine can understand [11–13]. Modern NLP methods can also be based on statistical Machine Learning [14], which enables the examination and analysis of data patterns, thereby improving the system/program [12]. In summary, HOPE not only integrates UMLS terms with NLP but also leverages real-time clinical data from EHRs to produce tailored recommendations.

Existing systems—e.g., SemEHR, CLAMP, or IBM Watson for Oncology—already demonstrate the potential of NLP-driven retrieval. However, many are domain-specific or computationally intensive. The HOPE system (Health Operation for Personalised Evidence) was developed as a lightweight, ontology-driven tool that provides real-time, personalised, evidence-based recommendations during medical consultations.

The aim of HOPE is not to improve diagnostic accuracy or to automate clinical decisions. Instead, it seeks to reduce the mechanical and cognitive effort associated with literature searching, enabling clinicians to access evidence more efficiently within the constraints of everyday workflows. From this perspective, the central question is not whether automation is faster than manual searching, but whether time savings can be achieved without substantially compromising the usefulness of the retrieved information. Also, HOPE focuses on identifying where to find information related to the case presented in the EHR, allowing physicians not only to obtain the information required for diagnosis and treatment but also to avoid the automation bias present in other clinical practice management systems [15].

The objective of this study was to evaluate whether an ontology-based recommender system can reduce information retrieval time during routine clinical workflows while maintaining a clinically acceptable level of relevance, compared with realistic manual PubMed searches performed by physicians.

2. Materials and Methods

2.1. Study Design

This observational, workflow-oriented evaluation assessed whether an ontology-based recommender system could reduce information retrieval time while maintaining the clinical acceptability of retrieved evidence. The evaluation included a benchmark comparison with manual PubMed searches performed by a predefined subgroup of physicians, reflecting realistic information-seeking behaviour in routine clinical practice rather than an exhaustive or expert-level search strategy. The study focused on inter-physician agreement, retrieval performance within top-ranked results, and retrieval time, comparing HOPE-assisted searches with realistic manual PubMed searches. The evaluation was preclinical in nature and did not aim to assess diagnostic accuracy, treatment decisions, or patient outcomes.

The study was carried out between February and July 2023. The study adhered to ethical standards and received institutional approval.

2.2. Participants

Fifty primary care physicians from different regions of Spain participated. Mean professional experience was 12.4 $\pm$ 6.3 years. All participants reported regular use of

biomedical literature databases, including PubMed. Standardised written instructions were provided before the evaluation.

A randomly selected subgroup of participants additionally performed manual PubMed searches to provide a reference benchmark for retrieval time and perceived clinical usefulness.

2.3. Clinical Cases

Thirty synthetic clinical cases representative of common and moderately complex primary care scenarios were used. The cases were derived from real clinical encounters, rewritten to remove identifiers and standardise their structure, and designed to include sufficient clinical uncertainty to motivate an information search.

Each case included:

- Patient age and sex;
- Relevant symptoms and clinical findings;
- Key elements of medical history;
- Diagnostic uncertainty motivating an information search.

No identifiable personal data was processed at any stage.

The number of cases (n = 30) was selected to balance variability in clinical scenarios with feasibility for physician evaluation, in line with prior workflow-oriented retrieval studies. Case complexity was calibrated to reflect real-world primary care conditions, including a mix of common, moderately complex, and diagnostically ambiguous presentations. Cases were purposively selected to ensure variability across scenarios, approximating the range of clinical uncertainty typically encountered in routine primary care practice.

2.4. System Description

HOPE operates by extracting free-text clinical information, identifying biomedical concepts using MetaMap Lite v3.6.2, mapping these concepts to UMLS vocabularies (including MeSH, SNOMED CT, and ICD-11), and generating structured queries.

Candidate documents are retrieved from PubMed and ClinicalTrials.gov and ranked using a weighted combination of cosine similarity between concept vectors and MeSH term overlap. The final ranking score was computed as a weighted sum of these two components, allowing simultaneous prioritisation of semantic similarity and controlled vocabulary matching. The system was executed in a secure, on-premise environment compliant with GDPR requirements.

The HOPE system was developed by a multidisciplinary research team at Universitat Oberta de Catalunya (Barcelona, Spain) and collaborating institutions. At the time of this study, the system was implemented as a research prototype and was not publicly available. The system source code and full implementation details are not publicly available at this stage. This may limit external reproducibility and independent validation of the system, although sufficient methodological detail is provided to support conceptual replication. To our knowledge, this is the first formal evaluation of the system in a clinical workflow context.

The technical details of the ranking strategy are provided for transparency but were not individually evaluated as outcome variables in this study, which focused on workflow-level performance and retrieval usefulness.

2.5. Reference Standard

The use of three senior physicians was considered appropriate for this preclinical, workflow-oriented evaluation, as the objective was not to establish an exhaustive gold standard but to obtain a consistent and clinically grounded reference for comparative

purposes. Similar expert panel sizes have been used in prior workflow-level evaluations of clinical information retrieval systems, where the focus is on relative usefulness rather than absolute completeness. Physicians evaluated only the top-ranked results ($k \leq 5$), reflecting realistic clinical behaviour where only a limited number of articles are reviewed.

Documents were labelled as "relevant" or "not relevant" based on perceived clinical usefulness. Relevance was assessed binary (yes/no) to reduce ambiguity and improve inter-rater consistency. A document was included in the reference relevance set if at least two reviewers agreed. The reference relevance set was intentionally limited to the pooled evaluation documents and was not intended to represent the full set of potentially relevant publications available in PubMed for a given case.

2.6. Comparative Retrieval Procedure

To enable a realistic comparison between automated and manual information retrieval under routine clinical conditions, two complementary retrieval approaches were evaluated. All participants assessed clinical cases using the HOPE system, while a randomly selected subgroup of physicians also conducted manual PubMed searches in line with their usual clinical practice.

For each clinical case, HOPE retrieved and ranked a fixed number of candidate articles (top-k results), of which only the first k results (k = 3, 4, 5) were evaluated for performance metrics.

For the manual PubMed search, a random sample of 10 physicians independently conducted searches using their usual clinical strategy, including keyword selection, Boolean operators, and iterative query refinement. The use of a smaller subgroup for manual searches was intended to balance feasibility with representativeness. Although randomly selected, this subgroup may not fully capture variability in information-seeking behaviour across the entire cohort, which should be considered when interpreting comparative results. Physicians were instructed to identify up to five articles they considered clinically relevant for each case.

Manual search time was decomposed into three consecutive tasks:

1. **Clinical case review and interpretation**, including identification of key clinical concepts (mean $\approx$ 5 min).
2. **Query formulation and refinement in PubMed**, including modifications after initial results (mean $\approx$ 2 min).
3. **Screening of retrieved results**, limited to the first 10 articles (titles and abstracts) (mean $\approx$ 6 min). Screening was limited to the first 10 results to reflect typical clinician behaviour, as physicians rarely review beyond the first page of search results during routine consultations.

Total manual search time was measured from case presentation to final article selection using a standardised stopwatch protocol.

For the **HOPE-assisted workflow**, the system automatically logged search time from case submission to the display of ranked results. This included NLP processing, UMLS concept mapping, and querying external databases.

Manual searches were intentionally designed to reflect routine clinical information-seeking behaviour rather than optimal or expert-level literature retrieval. The aim was to compare HOPE-assisted retrieval with realistic conditions under which clinicians typically search for evidence during practice, rather than with idealised search strategies. The detailed breakdown of retrieval time components is presented in Table 1.

Table 1. Comparison of information retrieval time between manual PubMed search and HOPE-assisted retrieval.

Retrieval Approach	Task Description	Mean Time	SD
Manual PubMed	Clinical case review	5.1 min	0.9
Manual PubMed	Query formulation and refinement	2.2 min	0.6
Manual PubMed	Screening of first 10 results	6.0 min	1.1
Manual PubMed (total)	—	13.3 min	1.7
HOPE-assisted	Automated NLP + ontology-based retrieval	17.4 s	2.1 s

Paired comparison: manual vs. HOPE, $p < 0.001$.

2.7. Metrics

Retrieval performance was assessed using Precision@k, relative Recall@k, and NDCG@k (k = 3, 4, 5). Recall was reported descriptively relative to the reference set and not interpreted as a measure of completeness. NDCG was used as a secondary ranking metric. Inter-physician agreement was assessed using Fleiss' Kappa. Descriptive statistics and paired *t*-tests were used for analysis.

NDCG@k was used to assess ranking quality by accounting for both relevance and position of documents. It is defined as the ratio between the discounted cumulative gain (DCG@k) and the ideal discounted cumulative gain (IDCG@k), where higher values indicate that relevant documents are ranked closer to the top of the list.

$$DCG@K = \sum_{K=1}^{K} \frac{rel_i}{log_2(i+1)}$$

All metrics were averaged across cases and raters. Inter-rater agreement was evaluated using Fleiss' Kappa (κ):

$$k = \frac{P0 - Pe}{1 - Pe}$$

where P_0 is the observed agreement and P_e the expected chance agreement. The interpretation of Fleiss' Kappa values follows standard categories, as shown in Table 2.

Table 2. Interpretation followed Landis & Koch's categories.

Kappa Results	Concordance
<0	Poor
0.01–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–0.99	Almost perfect

Statistical analyses were performed in R (version 4.3). Confidence intervals were computed via bootstrapping (10,000 resamples), and paired *t*-tests were used to compare retrieval times between methods.

3. Results

The results are presented in three main domains: inter-physician agreement in relevance assessment, retrieval performance of the HOPE system, and comparative time efficiency between manual PubMed searches and HOPE-assisted retrieval.

All 50 physicians completed the evaluation of the 30 clinical cases using the HOPE system. A randomly selected subgroup of 10 physicians additionally completed the manual

PubMed search benchmark. No relevant differences in years of professional experience were observed between the full cohort and the manual-search subgroup.

Inter-rater reliability for relevance assessment was substantial. The overall Fleiss' Kappa was $\kappa = 0.66$ (95% CI: 0.61–0.70), indicating consistent physician agreement regarding the clinical relevance of retrieved articles.

This level of agreement supports the consistency of the reference relevance set used in subsequent retrieval performance analyses.

3.1. Retrieval Performance

Thresholds of 0.5 and 0.75 were used to categorise performance levels, reflecting commonly used cut-offs for moderate and high precision in information retrieval (see Figure 1 legend).

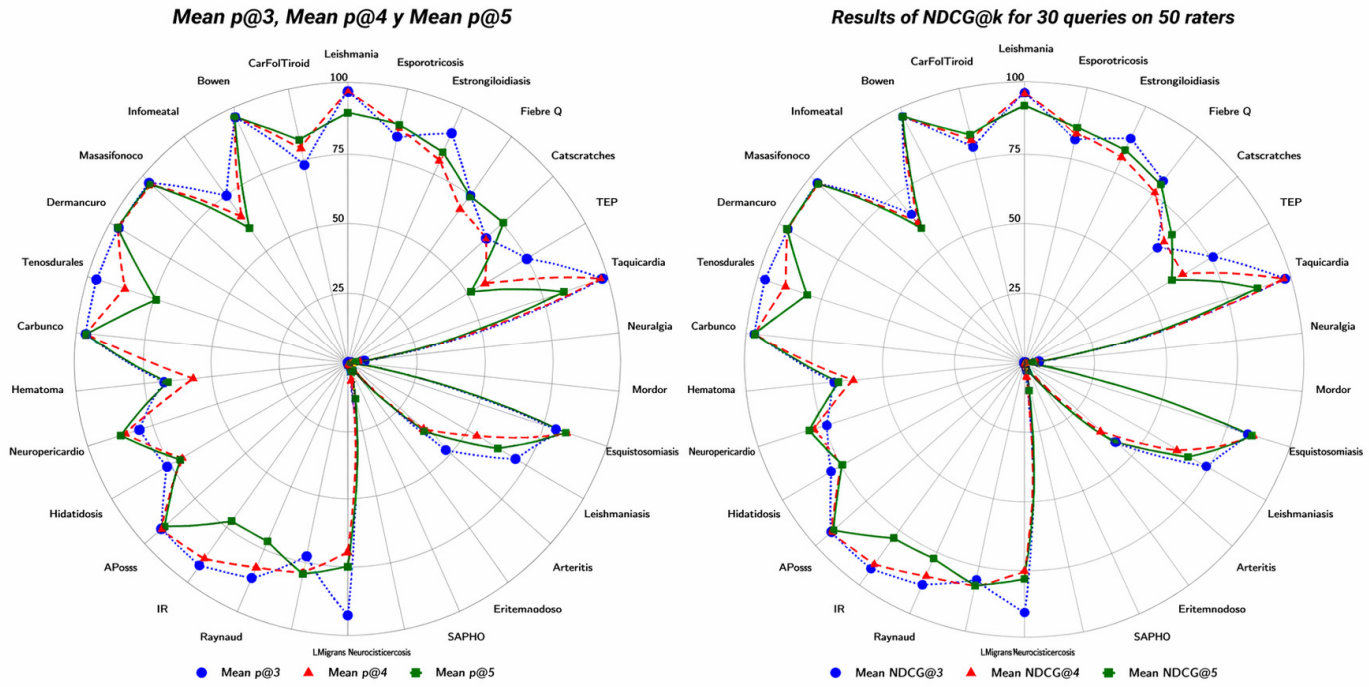

Figure 1. Representation of the mean precision@k and mean NDCG@k for each clinical case and the evaluations by the raters of the recommendations given by HOPE. Red, blue, and green categories represent precision thresholds of <0.5, 0.5–0.75, and >0.75, respectively, corresponding to low, moderate, and high retrieval performance.

Retrieval performance of the HOPE system was primarily assessed using Precision@k as the main indicator of clinical usefulness within the top-ranked results; it was calculated relative to the relevant reference set. The mean Precision@k values are shown in Table 3. Relative Recall@k and NDCG@k were used as secondary, descriptive metrics to characterise coverage and ranking tendencies, respectively.

Table 3. Summary of retrieval performance metrics for HOPE-assisted retrieval.

Metric	k = 3	k = 4	k = 5
Precision@k	0.72	0.72	0.68
Relative Recall@k	0.68	0.71	0.74
NDCG@k	0.63	0.60	0.60

These results indicate that a substantial proportion of the top-ranked articles retrieved by HOPE were considered clinically useful by physicians.

Relative Recall@k was calculated descriptively to characterise coverage of reference-relevant documents within the evaluated document pool. Relative recall increased as the number of retrieved documents increased (Relative Recall@3 = 0.68; Relative Recall@5 = 0.74), but was not interpreted as a measure of completeness of literature retrieval.

These results indicate moderate-to-high retrieval usefulness within the top-ranked results, with incremental gains in recall as k increased.

Relative recall is reported with respect to the reference relevance set and does not represent completeness of PubMed retrieval. NDCG values are used descriptively to illustrate tendencies in ranking positions.

Ranking position analysis

Ranking position was evaluated using Normalized Discounted Cumulative Gain (NDCG@k) as a secondary descriptive metric.

Given the binary relevance scheme, NDCG values were used descriptively to assess the tendency of relevant documents to appear earlier in the ranked lists.

Retrieval performance varied across clinical cases. Higher precision and ranking scores were observed in cases involving common or moderately complex conditions. Lower performance was observed in cases involving rare or diagnostically ambiguous conditions, which accounted for approximately 15–20% of cases.

Performance varied across cases, with most cases showing moderate to high precision. Lower values were observed in a minority of cases. Detailed case-level distributions are illustrated in Figure 1.

NDCG@k was calculated to account for both relevance and ranking position. Results showed consistent ranking performance across cases.

3.2. Time Efficiency Comparison

The mean total retrieval time for manual PubMed searches was 13.3 ± 1.7 min per case, comprising clinical case review, query formulation, and result screening. In contrast, the HOPE-assisted retrieval time was 17.4 ± 2.1 s, including NLP processing, ontology mapping, and database querying.

This represents a time reduction exceeding 95%, which was statistically significant ($p < 0.001$). These data are included for completeness but were not used for the primary interpretation of system performance, which focused on aggregated workflow-level outcomes.

4. Discussion

This workflow-oriented evaluation shows that an ontology-based recommender system can substantially reduce information retrieval time while maintaining a level of clinical usefulness comparable to that achieved through manual PubMed searches performed under routine practice conditions.

Although the speed advantage of automation is expected, its relevance lies in the context of routine practice, where time constraints often limit access to evidence. The observed precision within top-ranked results suggests that automated retrieval can still provide information that clinicians consider useful under these constraints. Agreement among physicians further supports the consistency of relevance judgments, despite the inherent subjectivity of such assessments.

From a retrieval perspective, HOPE demonstrated moderate to high precision within the top-ranked results. On average, more than two-thirds of the first three articles presented were judged to be clinically useful, which is important in settings where physicians typically consult only a small number of references.

Performance was not uniform across all clinical cases. As expected, cases involving common conditions or well-established clinical entities yielded higher retrieval accuracy, whereas cases involving rare diseases or diagnostically ambiguous presentations yielded lower performance. This variability likely reflects limitations inherent to literature-based retrieval systems, including uneven representation of rare conditions in biomedical databases and variability in terminological usage, and highlights areas for future refinement, particularly in improving concept mapping and ontology coverage. These constraints include reduced ontology coverage for rare entities, limitations in NLP concept extraction when clinical descriptions are less standardised, and lower representation of high-quality indexed literature in PubMed for low-prevalence conditions.

A key finding of this study is the difference in time efficiency. Manual PubMed searches required more than 13 min per case on average. This duration reflects not only the mechanics of database querying, but also the cognitive effort needed to translate a clinical problem into an effective search strategy. In contrast, HOPE delivered ranked results in under 20 s. This reduction (>95%) was both statistically significant and clinically meaningful.

HOPE does not replace clinical reasoning; instead, it automates repetitive and technically demanding steps involved in information seeking, such as query construction and iterative refinement. In everyday clinical practice, these tasks often compete with direct patient care for limited time and attention. By streamlining access to evidence, systems such as HOPE may help bridge the gap between the principles of evidence-based medicine and their practical application in real-world settings.

It is important to emphasise that HOPE is not intended to replace clinical judgment or to generate autonomous diagnostic or therapeutic recommendations. Instead, the system is designed to support clinicians by directing them toward potentially relevant sources of evidence, while leaving interpretation and decision-making firmly in human hands. This design approach may reduce the risk of automation bias and aligns with current perspectives on responsible deployment of AI in healthcare.

Recent AI-driven scientific search tools, such as AI Discovery (Scopus AI) or Perplexity AI, also aim to streamline literature retrieval using large language models. However, these systems are generally not designed for structured integration with clinical ontologies or EHR-derived narratives. In contrast, HOPE focuses on ontology-driven concept mapping and workflow integration, which may offer advantages in clinical interpretability and reproducibility. Direct comparative evaluations with these emerging tools represent an important area for future research. However, the absence of direct benchmarking limits conclusions regarding comparative performance.

Several limitations should be acknowledged. The evaluation relied on synthetic clinical cases derived from real scenarios, which, while necessary for ethical and methodological reasons, may not capture the full complexity of live EHR data. Furthermore, the reference relevance set was not exhaustive, and recall was therefore interpreted only in relative terms. Additionally, the manual search benchmark was conducted by a subset of participating physicians, although this approach was sufficient to demonstrate clear differences in time requirements. Finally, relevance judgments were based on expert consensus rather than exhaustive identification of all relevant publications, a common but recognised limitation in information retrieval research. The evaluation was conducted with primary care physicians trained in the Spanish healthcare system. Differences in clinical workflows, information-seeking behaviour, and familiarity with biomedical databases across healthcare systems may influence the generalisability of the findings. Future studies should assess system performance in other clinical and cultural contexts.

Within the constraints of this preclinical evaluation, the findings suggest that ontology-based recommender systems can meaningfully support clinical information seeking by

reducing time demands without substantially compromising perceived relevance. From a clinical perspective, reducing evidence retrieval time from minutes to seconds may directly influence consultation efficiency, cognitive load, and diagnostic confidence in time-constrained primary care settings.

5. Conclusions

In this preclinical, workflow-oriented evaluation, the HOPE system demonstrated a substantial reduction in information retrieval time compared with realistic manual PubMed searches, while maintaining a clinically acceptable level of relevance in the retrieved literature.

By supporting access to evidence rather than generating automated clinical decisions, the system aligns with current principles for the responsible use of artificial intelligence in healthcare.

Further studies using real-world clinical data will be required to assess the integration of such systems into everyday practice and their impact on clinical workflows and decision-making.

Author Contributions: Conceptualization, F.S.-R., C.S.-B. and H.E.; methodology, C.S.-B.; software, C.S.-B.; validation, F.S.-R., F.A.L. and F.S.-R.; formal analysis, H.E. and F.S.-R.; investigation, H.E. and C.F.L.; resources, C.S.-B.; data curation, H.E. and C.S.-B.; writing—original draft preparation, H.E. and C.S.-B.; writing—review and editing, H.E. and F.S.-R.; supervision, F.S.-R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Research Ethics Committee of the Universitat Oberta de Catalunya (reference CEIC-2023-042).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on reasonable request from the corresponding author. The data are not publicly available due to privacy and ethical restrictions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Budd, J. Burnout Related to Electronic Health Record Use in Primary Care. *J. Prim. Care Community Health* **2023**, *14*, 21501319231166920. [CrossRef] [PubMed]
2. Atallah, Á.N. Evidence-based medicine. *Sao Paulo Med. J.* **2018**, *136*, 99–100. [CrossRef] [PubMed]
3. Djulbegovic, B.; Guyatt, G.H. Progress in evidence-based medicine: A quarter century on. *Lancet* **2017**, *390*, 415–423. [CrossRef] [PubMed]
4. Lin, C.; Hsu, C.J.; Lou, Y.S.; Yeh, S.J.; Lee, C.C.; Su, S.L.; Chen, H.C. Artificial Intelligence Learning Semantics via External Resources for Classifying Diagnosis Codes in Discharge Notes. *J. Med. Internet Res.* **2017**, *19*, e380. [CrossRef] [PubMed]
5. Olex, A.L.; McInnes, B.T. Review of Temporal Reasoning in the Clinical Domain for Timeline Extraction: Where we are and where we need to be. *J. Biomed. Inform.* **2021**, *118*, 103784. [CrossRef] [PubMed]
6. Suh, H.S.; Tully, J.L.; Meineke, M.N.; Waterman, R.S.; Gabriel, R.A. Identification of Preanesthetic History Elements by a Natural Language Processing Engine. *Anesth. Analg.* **2022**, *135*, 1162–1171. [CrossRef] [PubMed]
7. Wong, A.; Plasek, J.M.; Montecalvo, S.P.; Zhou, L. Natural Language Processing and Its Implications for the Future of Medication Safety: A Narrative Review of Recent Advances and Challenges. *Pharmacother. J. Hum. Pharmacol. Drug Ther.* **2018**, *38*, 822–841. [CrossRef] [PubMed]
8. Duncan, J.; Eilbeck, K.; Narus, S.P.; Clyde, S.; Thornton, S.; Staes, C. Building an Ontology for Identity Resolution in Healthcare and Public Health. *Online J. Public Health Inform.* **2015**, *7*, e61704. [CrossRef] [PubMed]
9. Home | SNOMED International. Available online: https://www.snomed.org/ (accessed on 8 April 2023).

10. Kühl, N.; Schemmer, M.; Goutier, M.; Satzger, G. Artificial intelligence and machine learning. *Electron. Mark.* **2022**, *32*, 2235–2244. [CrossRef]
11. Voytovich, L.; Greenberg, C. Natural Language Processing: Practical Applications in Medicine and Investigation of Contextual Autocomplete. In *Machine Learning in Clinical Neuroscience*; Springer: Cham, Switzerland, 2022; pp. 207–214. [CrossRef]
12. Alsawas, M.; Alahdab, F.; Asi, N.; Li, D.C.; Wang, Z.; Murad, M.H. Natural language processing: Use in EBM and a guide for appraisal. *BMJ Evid.-Based Med.* **2016**, *21*, 136–138. [CrossRef] [PubMed]
13. Yim, W.W.; Yetisgen, M.; Harris, W.P.; Sharon, W.K. Natural Language Processing in Oncology: A Review. *JAMA Oncol.* **2016**, *2*, 797–804. [CrossRef] [PubMed]
14. Deo, R.C. Machine Learning in Medicine. *Circulation* **2015**, *132*, 1920–1930. [CrossRef] [PubMed]
15. Kücking, F.; Hübner, U.; Przysucha, M.; Hannemann, N.; Kutza, J.O.; Moelleken, M.; Erfurt-Berge, C.; Dissemond, J.; Babitsch, B.; Busch, D. Automation Bias in AI-Decision Support: Results from an Empirical Study. *Stud. Health Technol. Inform.* **2024**, *317*, 298–304. [CrossRef] [PubMed]