Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Comparative Study of Medical Information Retrieval Approaches: A PRISMA-Based Systematic Review of Query Expansion, Term Weighting, Knowledge Modeling, and Bayesian Methods for Enhanced Clinical Decision Support

Grace Katuri

grace.katuri@accuityhealthcare.com

Accuity Healthcare USA https://orcid.org/0009-0002-3178-5560

Systematic Review

Keywords: medical information retrieval, PRISMA, query expansion, term weighting, knowledge modeling, Bayesian inference, clinical decision support

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11802469/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Comparative Study of Medical Information Retrieval Approaches: A PRISMA-Based Systematic Review of Query Expansion, Term Weighting, Knowledge Modeling, and Bayesian Methods for Enhanced Clinical Decision Support

Grace Katuri[1]

Accuity Healthcare USA, Mount Laurel, New Jersey, USA
grace.katuri@accuityhealthcare.com

Abstract. Medical information retrieval (IR) enables clinicians, scientists and policymakers to find relevant, reliable, and contextual evidence from clinical practice guidelines, diagnostic tools, systematic reviews and patient records. Retrieval in this field is challenged by semantic variability, abbreviations, variability in data formats and patient privacy. This review adheres to the PRISMA standards and describes four exemplary directions for methods that advance medical IR: 1) external collection cluster-based query expansion, 2) part-of-speech pre-weighting of terms, 3) domain knowledge modeling via data-driven mining of sublanguage patterns, and 4) knowledge fusion of domain's heterogeneous knowledge bases with a Bayesian integration model for precision medicine. All four methods are shown to outperform common IR benchmarks with respect to mean average precision, normalized discounted cumulative gain, recall, and annotation quality. We organize the four directions into a taxonomy named MERIT, the Medical Evidence Retrieval Improvement Taxonomy. Based on the reviewed evidence, a reliable evaluation of retrieval techniques requires the use of standard medical IR benchmarks as well as standard baselines. Hybrid IR methods that fuse external evidence, linguistic structure, mined patterns, and probabilistic knowledge may promote clinical decision support in the future.

Keywords: medical information retrieval, PRISMA, query expansion, term weighting, knowledge modeling, Bayesian inference, clinical decision support

1 Introduction

Medical IR is a core task of medical informatics, because it is the link between evidence-seeking decision-makers and vast, ever-expanding collections of clinical and biomedical knowledge [1]. Immediate access to guideline statements, diagnostic support, systematic reviews, and patient records must be made available to busy clinicians, while evidence for policy makers and subject matter experts

must be available in such a way as to be well filtered for evidence quality and clinical decision support [2]. The usefulness of a medical IR system, beyond ranking quality, lies in its ability to understand domain language, maintain clinical significance, and operate within privacy and governance constraints. These issues set medical retrieval apart from open web search and suggest the need for domain adaptation.

The first obstacle is the language of medicine. The clinical notes, lab results, imaging descriptions, and abstracts we read are rife with acronyms, sets of synonyms, morphological variants, and author-specific shorthand. A search using one set of words may not find a document expressing the same idea in another way. Recent biomedical language models illustrate both the promise and limits of purely lexical matching [3, 4]. If we add to that the ambiguity of words with multiple meanings, we realize that under those conditions, models that depend solely on noun overlap will likely leave us short of the evidence we want.

A second challenge is the heterogeneity of evidence. Medical IR needs to perform across various sources of evidence, such as semistructured electronic health records, unstructured scientific articles, and well-structured databases. Sources differ in their meta data properties, levels of granularity, value, and in the legal and ethical considerations with regard to data sharing, model training, and evaluation. This is especially relevant when automating evidence synthesis and systematic review retrieval [5]. Any evidence retrieval system for medicine needs to be effective and at the same time controlled and traceable across heterogeneous sources of evidence.

In the face of these difficulties, much recent work has responded by adapting (rather than inventing new or disconnected) IR models: Query expansion methods may seek to mitigate vocabulary mismatches by adding related terms from external collections or from comprehensive knowledge bases [6]. Term weighting methods may seek to better model linguistic structure than is possible in a bag of words assumption. Knowledge modeling methods may seek to induce a domain semantic structure from corpora, while probabilistic methods may seek to combine genomic, pharmacological, and textual evidence in a single ranking model. Clinical trial search re-ranking demonstrates the value of biomedical language understanding for evidence ranking [7]. In all cases, the prevailing suggestion is that the assumptions that make retrieval effective for general purpose IR may simply not hold for medical decision support.

This paper conducts a systematic review of representative methodological advances in medical IR according to the PRISMA framework [8]. Our review addresses three questions. The first concerns the distinctions among the techniques with regard to handling of vocabulary mismatch, semantic complexity, and heterogeneous evidence. The second addresses the comparability of the results reported: are they of value in comparing well-established IR baselines and using standard IR experimental datasets? The third addresses what performance improvement are achieved when using modifications or ensembles of existing approaches.

The review provides three services. It classifies the selected methods by the main enhancement mechanism, which includes query expansion, term weighting, knowledge modeling, and Bayesian knowledge combination. It compares the evaluation approaches of the selected papers, including the benchmark datasets, baselines, and metrics reported. It further synthesizes the gains reported across those papers to infer the patterns of design choices that best lead to better performance of the methods in medical retrieval. The review is intentionally limited to the quality of methodology, not breadth of references.

The rest of the paper is organized as follows. The review methodology and PRISMA-based selection procedure are discussed in Section II. Query expansion and term weighting techniques are explored in Section III. Knowledge modeling and Bayesian analysis is discussed in Section IV. Comparative analysis, benchmarking discussion, and integration of research questions are discussed in Section V. We conclude the review and suggest future research directions in VI.

2 Review Methodology and Research Framework

This review was conducted as a systematic literature review according to the PRISMA guideline. The protocol was set to be transparent about records finding, screening, eligibility, and ultimately selection. The review examined research papers in computer science and engineering venues in the English language published after 2010. The concept of information retrieval was combined with the medical or health-related subjects and the enhancement-related subjects to effectively identify records that describe adopted retrieval approaches in various digital libraries.

The identification process searched the following four bibliographic databases: Science Direct, Scopus, arXiv, and Springer. The search string combined the following two concepts: "Information Retrieval," and the concept words "Medic" and "Enhance," where the concept of knowledge was incorporated within 'Medic' and the concept of improvement was incorporated within the word 'Enhance' with the same effect. The output from the first searches was 811 (Science Direct), 95 (Scopus), 123 (arXiv), and 79 (Springer). As the total pool of records was large and unevenly distributed, the first 40 records from each source were kept, so the working set was 160 records to screen.

Duplicates were removed with Zotero. Title and abstract screening was conducted based on the review scope; after duplicate removal 59 papers remained. Title and abstract screening yielded a remaining set of 19 papers. A manual relevance check on the remaining set produced a set of 7 papers, with a further full text eligibility check producing the final set of 4 papers for inclusion in the review. The flow of papers is summarised in Fig. 1. The final set contains four different enhancement strategies, with enough variety in the methodologies to allow for comparison.

The inclusion criteria screening focused on studies that suggested a specific modification to a medical IR component and provided empirical evaluation on a standard benchmark or a standard baseline. Studies that used general biomedical

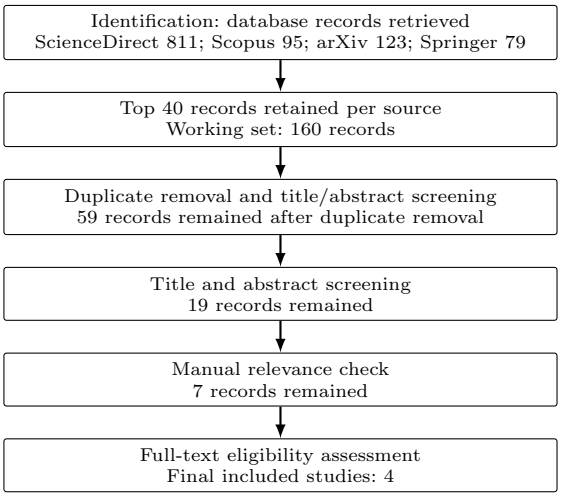

Fig. 1: PRISMA-based record selection flow used in this review.

NLP tools without retrieval evaluation, or that only performed non-retrieval clinical prediction tasks, were excluded. The four included studies are related to cluster-based query expansion [13], part-of-speech term weighting [14], data-driven sublanguage pattern mining [15], and Bayesian knowledge integration [16], respectively. Such distribution lends itself to the comparative evaluation of distinctive retrieval enhancement mechanisms.

The review questions were designed to facilitate structured synthesis. The first question was about comparison of IR techniques in the setting of medical retrieval. The second was about comparison to standard datasets, or to known IR methods in general. The third was about improvement gained by combining or modifying existing IR methods. The extraction process focused on extracting details about method, evaluation setting and performance values reported by the authors. It was structured around the information required to compare each included study in terms of their target of improvement, modeling principle, benchmark dataset used, baseline technique and gain reported.

The synthesis was qualitative and comparative rather than a statistical meta-analytic synthesis due to the heterogeneity of the approaches, datasets, and evaluation metrics included. For instance, query expansion approaches report mean average precision and normalized discounted cumulative gain; term weighting approaches report mean average precision improvements compared to bag-of-words and Markov random field baselines; knowledge modeling approaches report entity and relationship coverage together with annotation precision, recall, and F-score; the Bayesian approach to precision medicine approaches report total recall. A pooled effect size would not reflect differences in evaluation aims.

For transparent comparison, the review notes the benchmarked role of each evaluation set. Most general medical IR sets, including TREC CDS, OHSUMED,

the TREC Medical Records tracks, the TREC Genomics track, and the TREC Precision Medicine set [9] all contain externally-derived relevance assessments and thus are suitable for comparison across studies. Similarly, query likelihood, external expansion, bag-of-words, Markov random field, tf-idf, and BM25 are all well-developed baseline methods for comparison within each study. The use of both benchmark sets and well-established baselines was considered when judging the overall strength of each paper's methods.

There are three limitations with this approach. First, the initial screening retained the first 40 records returned by each database rather than the full result sets, so the pool screened was not exhaustive, and the review is best read as a focused comparative review of representative methods rather than a complete census of the field. Second, because our review considers only four illustrative papers, we do not claim to cover all of the work in this domain. Third, the different values of metrics and task formulations in the studies preclude numerically combining the results across them. We address these limitations by focusing on comparative interpretation, separating quantitative gains by metric and dataset, and capturing recurring common patterns in experimental design principles (rather than claiming one effect size).

3 Query Expansion and Term Weighting Enhancements

Expansion is an almost inevitable consequence of vocabulary mismatch in medical retrieval. When the normal way a user expresses a clinical concept differs from the language used in a relevant document, expansion can add similar terms to increase recall while preserving the initial query's spirit. Pseudo-relevance feedback is a typical example of expansion, where an initial retrieval run helps identify expansion terms, which can then be supplemented with evidence of co-occurrence of related terms drawn from external collections. The design question is whether such external evidence can be used without moving too far from the original topic.

The cluster-based external expansion model to this situation is similar to the previous model as it combines an external expansion model with a cluster-based document model. Instead of not considering an external collection as a simple document pool, the model considers it as a collection which is further exploited as a cluster for estimation of document relevance. This removes one clustering phase (e.g., K-means) at the same time using the grouping knowledge that exists in the external sources.

Formally, let q denote the original query and let D denote a target document. The retrieval score can be written as a weighted combination of a target-collection relevance score and an external-collection expansion score,

$$S(D, q) = \lambda S_{\text{target}}(D, q) + (1 - \lambda) S_{\text{ext}}(D, q), \tag{1}$$

where λ controls the influence of the target collection and S_{ext} represents evidence estimated from external clusters. The cluster-based component allows S_{ext} to reflect collection-specific term distributions rather than a single global

expansion distribution. This design is particularly relevant in medical IR, where external collections may differ in terminology, genre, and granularity.

Empirical evaluation of the model was conducted on TREC CDS, CLEF eHealth, and OHSUMED, and the gains reported here are those for TREC CDS and OHSUMED. When compared to the original external expansion model and query likelihood, the cluster-based model yielded consistent gains on both collections, with improvements of 15.19% and 10.32% for mean average precision and normalized discounted cumulative gain respectively on TREC CDS and 15.51% and 12.33% respectively on OHSUMED. These results demonstrate that the structured application of external collections can significantly address vocabulary mismatch in medical retrieval.

Term weighting. Bag-of-words representations do not take into account syntactic information and do not differentiate between multiple occurrences of the same term. tf-idf remains a common weighted baseline [10]. Markov random field models capture word proximity and co-occurrence constraints, yet may not differentiate between multiple occurrences of a term within a clique. For instance, this aspect can be seen in a particular part-of-speech scheme, which describes how part-of-speech categories should impact upon weights. The hypothesis underlying it is that each part-of-speech category: belongs to a different weight set, has a different contribution to biomedical relevance, and that these differences are learned, and not fixed.

The POS-based models, denoted POS-BoW and POS-MRF, extend bag-of-words and Markov random field retrieval respectively. The term weights are optimized through supervised learning using mean average precision as the objective. Golden section search and cyclic coordinate descent are used to search the weight space. A generic objective can be expressed as

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \ \mathrm{MAP}(\mathbf{w}), \tag{2}$$

where $\mathbf{w}$ denotes the vector of part-of-speech weights and $\mathrm{MAP}(\mathbf{w})$ is the retrieval effectiveness obtained with those weights. This formulation converts term weighting from a manually specified heuristic into a data-driven optimization problem.

The evaluation utilised the TREC 2011 and 2012 Medical Records tracks and the TREC 2004 Genomics track. On the Medical Records tracks, POS-BoW achieved a relative improvement of 10.88% in mean average precision, and POS-MRF 13.59%, over regular bag-of-words and Markov random field baselines respectively. On the Genomics track, POS-MRF additionally produced a 10.04% gain. These results indicate that syntactic category information is a useful signal for medical retrieval, provided the weighting is optimised against a task-relevant relevance metric rather than pre-determined.

In summary, query expansion and term weighting constitute two separate methods for achieving the adaptation of IR models to medical language. The expansion approach remedies mismatch by expanding the vocabulary coverage (at the level of expressions), whilst the weighting approach allows for finer grained rank ordering by contextual sensitivity of the conceptual units (at the level of

Table 1: Representative enhancement strategies included in the review.

Enhancement strategy	**Core method**	**Primary retrieval role**
External query expansion	Cluster-based external expansion	Reduces vocabulary mismatch using external collections
Term weighting	POS-BoW and POS-MRF	Introduces syntactic category information into term importance
Knowledge modeling	Data-driven sublanguage mining	Induces domain-specific entities and relations
Bayesian expansion	Precision medicine framework	Integrates genomic, pharmacological, and textual evidence

representation) and/or their importance. The expansion technique enhances recall using evidence external to the test collection while the weighting technique enhances rank ordering by making individual items' contribution more linguistically sensitive. Both maintain the existing retrieval framework and have been implemented by modifying individual components, thus making them practical for use as part of a combined approach with additional medical domain adaptations.

4 Knowledge Modeling and Bayesian Integration

Knowledge models offer structured representations of concepts and relations in a domain. They have been used in medical IR for understanding, matching semantics, and organizing evidence by providing a mapping of unstructured text to interpretable concepts and relations. Ontology-based and Machine-learning-based models are among the most popular knowledge models used, yet they still present a number of challenges: manually-created ontologies need significant effort from experts and may not generalize well to new subdomains; automatically learned models have low coverage of domain semantics and low accuracy. Knowledge graphs are increasingly used to organize precision-medicine evidence [11]. This is especially problematic in medical image reports and clinical notes, where domain-specific language and implicit relations are often used.

The data-driven sublanguage pattern mining method overcomes these shortcomings by merging the advantages of the syntactic and semantic network analysis methods in a bottom-up, context-dependent process. Instead of deriving an a priori ontology, this method inducts patterns from a domain corpus to form a knowledge model. It is based on the assumptions that patterns are recurrent in specialized sublanguage and that these regularities can be used to discover entities and relations with a high degree of coverage. The resulting model is aimed

at enabling a computational representation of the domain knowledge based on empirical foundations.

The sublanguage mining technique yielded a coverage of 98% of the entities and 97% of the relationships in the target domain [15]. Precision, recall and F-score on machine annotated input were 87%, 79% and 82%, respectively. These numbers suggest that the induced model is able to represent a high proportion of the relevant semantics, while maintaining an acceptable level of annotation quality; the ratios of precision to recall are especially crucial here, as a knowledge model that covers many entities, but incorrectly proposes too many relations is of little help in retrieval and decision support tasks.

The methodological importance of the method is in the treatment of the semantic complexity: by not presupposing any a priori ontologies, the method is considering the knowledge model as something to be induced by corpus evidence, thus mitigating the manual intervention, making adaptation to new data sources more flexible. Yet, the use of syntactic and semantic patterns involves loss of model quality, dictated by corpus representativeness and rigor of pattern mining procedure, the trade-offs being fundamental to the application of knowledge models in retrieval.

Bayesian integration provides an alternative approach to domain adaptation. In the context of precision medicine, the critical capability for clinical decision support is retrieval of evidence relating patient factors and genomic context to drug interactions and published literature. Structured queries may make use of genomic and textual constraints, but typical retrieval approaches cannot exploit these heterogeneous signals effectively. Variance in semantics and vocabulary mismatch between information sources can lead to topic drift during query expansion if the expansion terms are derived from information sources with disjoint terminologies. A probabilistic framework can make the integration of these sources explicit.

The Bayesian approach for precision medicine generates the query expansion from several knowledge bases, such as the unified medical language system, the drug-gene interaction knowledge base, or biomedical articles. Prior likelihoods are estimated for the conjunction of two words and the genomic context, then combined with a Bayesian inference. A simplified posterior expansion weight for a candidate term t can be written as

$$P(t \mid q) = \frac{P(q \mid t)P(t)}{\sum_{t'} P(q \mid t')P(t')}, \tag{3}$$

where $P(t)$ is the prior probability of the candidate term and $P(q \mid t)$ represents the likelihood of the query given that term. This formulation allows heterogeneous evidence sources to contribute to expansion in a principled manner.

On the TREC 2017 Precision Medicine dataset, the Bayesian framework [16] obtained a total recall of 0.4929, compared with 0.4593 for the current system. The enhancement in terms of absolute recall numbers was not large, but was considered clinically significant in the context of precision medicine, where recall of relevant evidence is the key goal. From the results, we also observed that MEDLINE concepts were very effective, as well as disease and gene attributes. This

suggests that the probabilistic integration of heterogeneous knowledge bases can yield better retrieval for situations in which the evidence sources are conceptually related to the question.

The two knowledge modeling and Bayesian methods extend query expansion and term weighting as discussed above. For knowledge modeling, the controlled representations offer a potentially useful aid to semantic matching, while the Bayesian integration offers a sound principle of integrating competing sources of evidence under uncertainty. Both introduce something new relative to lexical matching. However, they operate on different levels of abstraction: the sublanguage mining technique induces structure from corpus data, while the Bayesian method applies a probabilistic scheme to combine textual resources with previously formalized knowledge. They exemplify the two directions, which medical IR research could take to grapple with the semantic intricacies.

5 Comparative Analysis, Benchmarking, and Discussion

The comparison is structured around the three review questions. The methods in question differ in where they intervene in the retrieval pipeline: Cluster based external expansion intervenes at the query formulation step by augmenting the query with cues extracted from external collections; POS weighting intervenes at the representation step by scoring terms with different weightings; Sublanguage pattern mining intervenes at the knowledge representation step by inducing domain-specific lexemes and their relations; Bayesian precision medicine intervenes at the evidence integration step by combining textual and genomic signals via probabilistic inference. This variation is relevant because failures in medical IR can occur at any of these stages.

In the case of Question 2, there is considerable dependance on typical benchmarks and comparisons. The cluster based expansion model was tested on TREC CDS and OHSUMED, compared to query likelihood and the original external expansion model. The POS weighting models were tested on the TREC 2011 and 2012 Medical Records tracks and TREC 2004 Genomics track and tf-idf and BM25-style baselines are used as points of comparison. The Bayesian framework was tested on the TREC 2017 Precision Medicine dataset. The use of these standard comparisons allows for a measure of comparability, although the differences in task definitions and the tasks' various metrics necessitates careful interpretation. Recent contrastive pre-trained transformers such as MedCPT further illustrate the value of domain-specific representation for zero-shot biomedical retrieval [12].

The results are consistently in favor of modified or combined IR approaches, so the answer to the third question is affirmative. The cluster-based model improved mean average precision and normalized discounted cumulative gain on the TREC CDS and OHSUMED collections. The POS weighting models improved mean average precision over bag-of-words and Markov random field baselines, with further improvements on the Genomics track. The sublanguage mining approach achieved high entity and relationship coverage at an acceptable level of

Table 2: Reported performance gains of the reviewed methods.

Method	**Dataset**	**Metric**	**Reported gain**
Cluster-based expansion	TREC CDS	MAP	+15.19%
	TREC CDS	NDCG	+10.32%
	OHSUMED	MAP	+15.51%
	OHSUMED	NDCG	+12.33%
POS-BoW	TREC 2011/2012 Records	MAP	+10.88%
POS-MRF	TREC 2011/2012 Records	MAP	+13.59%
	TREC 2004 Genomics	MAP	+10.04%
Bayesian precision medicine	TREC 2017 PM	Recall	0.4929 vs. 0.4593

annotation quality. The Bayesian scoring model improved total recall on the data-set for the precision medicine benchmark. These results do not suggest that any one approach is always better than the others, but they do suggest that domain-aware adaptation makes a difference.

Another comparison is the measure(s) used. Mean average precision and normalized discounted cumulative gain prioritize ranking quality; recall reflects the quantity of relevant evidence retrieved. Precision, recall, and F-score are appropriate for annotation of knowledge models, because you are interested in the extent to which the entities and relations are correctly identified. Total recall is appropriate for precision medicine, because you are interested in the extent to which evidence coverage for clinical decision support is achieved. Thus the measure used will depend on your intended application.

Benchmarking alone also emphasizes the need for the base to be chosen carefully. A technique that beats an inferior baseline does not constitute very convincing evidence of an effect in the presence of domain adaptation. While all of the studies reviewed use some baseline of known retrieval models or of the proposed model without the modification, this should theoretically strengthen the internal validity of the gains, while the fact that no single evaluation standard spans all of the articles reviewed means that the gains should be taken to be within-study effects, not effect sizes in and of themselves. The synthesis prefers these studies' ascertained effects in the same direction over identical effects in size.

Fig. 2 concludes this review by summarizing the taxonomy of methods the clinical question examined. We name this taxonomy MERIT, the Medical Evidence Retrieval Improvement Taxonomy. Its four limbs correspond to Query expansion, Term weighting, Knowledge modeling, and Bayesian combination. This taxonomy makes clear that rather than a one-method family of medical

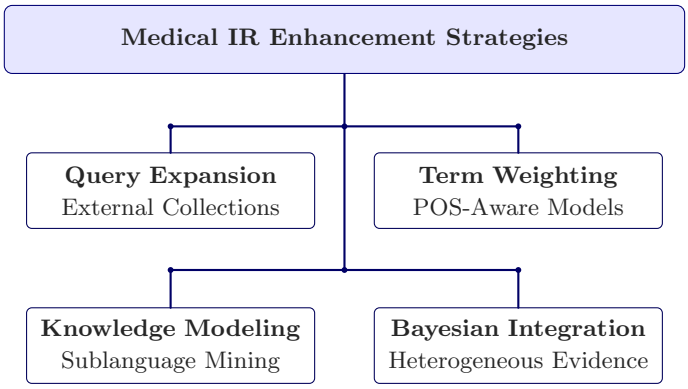

Fig. 2: Taxonomy of the medical IR enhancement strategies reviewed in this paper.

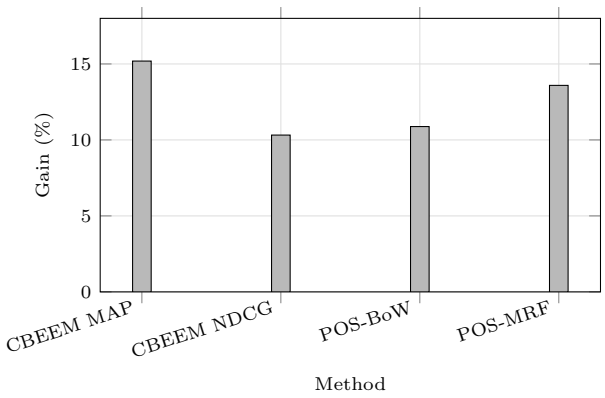

Fig. 3: Representative reported gains for cluster-based expansion and POS weighting methods.

IR improvements, there are several types, each one addressing a specific source of failure in the IR process. Query expansion addresses vocabulary mismatch, Term weighting addresses representational mismatch, Knowledge modeling addresses semantic structure, Bayesian combination addresses heterogeneous evidence. For each of these failure sources, the reviewed literature offers a method with reported evidence of effectiveness.

Fig. 3 shows a selection of reported gains for the two approaches. They are not directly pooled as their datasets and metrics are dissimilar, but they demonstrate the overall trend that adaptation has positive benefits. The gains from cluster-based expansion are significant in both TREC CDS and OHSUMED, and the gains from POS weighting are also significant on the Medical Records and Genomics tracks. These results reinforce that domain adaptation can lead to increases in effectiveness.

The discussion also shows the tension between specificity and generality. Domain-specific methods, like sublanguage pattern mining for image report annotation, achieve complete coverage and quality when constrained to a medical subdomain, but need further validation of portability to other subdomains. Collection-based methods, like cluster-based expansion, generalize easily to other collections if the external evidence sources are representative, but require a collection of similar size and quality. Bayesian integration, while very flexible in principle, depending on an external curated collection of knowledge bases, also can be tailored to specific knowledge sets. These are all important trade-offs in translating research methods to operational CDS.

The last comparison of models is in regard to hybrids. The literature reviewed here shows the benefits of making improvements to a model and the potential to make improvements through the combination of models. External expansion could add more terms to a POS weighted model, a knowledge mined model could provide semantic features for Bayesian expansion, and a Bayesian model could estimate confidence for expansion terms mined from external clusters. The evidence reviewed here does not test any such combinations directly, but does show a way to proceed toward hybrid designs to overcome multiple sources of retrieval error at once.

6 Conclusions and Future Directions

This review considered four representative approaches to medical information retrieval, selected and reported following the PRISMA framework. The papers they reviewed relate to: query expansion using external collections, part-of-speech term weighting, data-driven sublanguage pattern mining and Bayesian combinations of heterogeneous knowledge sources. The findings show consistently that adapting retrieval models to the language, structure, and evidence of the medical domain improves performance across ranking, annotation, and recall measures, when those adaptations are evaluated empirically on standard benchmarks.

In conclusion, the cluster-based external expansion model shows that external collections can be exploited better when their collection level structure is explicitly modeled. The POS weighting models show that linguistic categories can offer useful term importance signals when term weights are learned against a retrieval target. The sublanguage mining approach shows that knowledge models can be induced from evidence in a corpus with high coverage of entities and relations and reasonable annotation accuracy. The Bayesian precision medicine approach shows that heterogeneous knowledge bases can be integrated in a probabilistic way to enhance recall on a precision medicine benchmark.

The review further demonstrates that benchmarking is crucial for genuine improvement. The four studies examined use widely used medical IR datasets and established baselines, which supports validity within each study and allows comparison along accepted lines. However, the differing metrics and task descriptions make direct numerical synthesis impossible; work in the future should

include as many metrics as possible when sharing evaluation protocols so that individual benefits can be better evaluated.

Many implications follow. The first is that hybrid retrieval architectures should be compared on a systematic basis with combining external expansion, linguistic weighting, the mined knowledge structures and the Bayesian evidence integration. The second is that the portability of sublanguage mining across medical subdomains should be investigated with cross-domain corpora and transfer learning. The third is that the calibration and robustness of the Bayesian expansion should be studied on incomplete or noisy knowledge bases. The fourth is that privacy-preserving evaluation protocols should be developed so that the medical IR methods can be evaluated without risking patient confidentiality.

Explainability and governance must be considered when deploying clinically. When clinicians retrieve a document, they need to understand the rationale behind the retrieval, and how confident they should be in the retrieved evidence. Query expansion terms, term weights, mined relations and Bayesian priors are all potential explainability signals if revealed, but these signals also need to be audited for bias and validity. In future medical IR systems, transparency and auditability should be built into our system design principles.

Briefly, the literature reinforces the clear conclusion that medical IR can profit from the tailored adaptation of retrieval models to domain language, domain structure, and domain evidence. The reviewed methods outperform standard IR baselines and supply orthogonal mechanisms to deal with vocabulary mismatch, semantics, and heterogeneous evidence. Hybrid and XAI designs leveraging these mechanisms in combination, and with continued rigorous evaluation, will be the most promising path to improving clinical decision support from medical information retrieval.

Declarations

Author contributions. The author conceived, conducted, and wrote this review in full. **Funding.** This work received no external funding. **Competing interests.** The author declares no competing interests. **Data availability.** No new data were created or analysed in this study. All studies reviewed are published works cited in the references. **Ethics approval.** Not applicable, as this is a review of published literature.

References

1. W. Hersh, *Information Retrieval: A Biomedical and Health Perspective*, 4th ed. Cham, Switzerland: Springer, 2020, doi: 10.1007/978-3-030-37584-3.
2. R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, "An overview of clinical decision support systems: benefits, risks, and strategies for success," *NPJ Digital Medicine*, vol. 3, no. 1, p. 17, 2020, doi: 10.1038/s41746-020-0221-y.

3. J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020, doi: 10.1093/bioinformatics/btz682.
4. Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Transactions on Computing for Healthcare*, vol. 3, no. 1, pp. 1–23, 2021, doi: 10.1145/3458754.
5. W. Kusa, A. Lipani, P. Knoth, and A. Hanbury, "Automation of systematic reviews: A systematic review of information retrieval and machine learning approaches," *Information Processing & Management*, vol. 60, no. 4, p. 103393, 2023, doi: 10.1016/j.ipm.2023.103393.
6. H. K. Azad and A. Deepak, "Query expansion techniques for information retrieval: A survey," *Information Processing & Management*, vol. 56, no. 5, pp. 1698–1735, 2019, doi: 10.1016/j.ipm.2019.05.009.
7. M. Rybinski, J. Xu, and S. Karimi, "Clinical trial search: Using biomedical language understanding models for re-ranking," *Journal of Biomedical Informatics*, vol. 109, p. 103530, 2020, doi: 10.1016/j.jbi.2020.103530.
8. M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, and others, "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *BMJ*, vol. 372, p. n71, 2021, doi: 10.1136/bmj.n71.
9. K. Roberts, D. Demner-Fushman, E. M. Voorhees, S. Bedrick, and W. R. Hersh, "Overview of the TREC 2020 Precision Medicine Track," in *Proceedings of the Twenty-Ninth Text REtrieval Conference (TREC 2020)*, 2020, doi: 10.6028/NIST.SP.1266.pm.
10. M. Das, S. Kamalanathan, and P. J. A. Alphonse, "A comparative study on TF-IDF feature weighting method and its analysis using unstructured dataset," *Journal of Ambient Intelligence and Humanized Computing*, 2021, doi: 10.1007/s12652-021-03097-w.
11. P. Chandak, K. Huang, and M. Zitnik, "Building a knowledge graph to enable precision medicine," *Scientific Data*, vol. 10, no. 1, p. 67, 2023, doi: 10.1038/s41597-023-01960-3.
12. Q. Jin, W. Kim, Q. Chen, D. C. Comeau, L. Yeganova, W. J. Wilbur, and Z. Lu, "MedCPT: Contrastive Pre-trained Transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval," *Bioinformatics*, vol. 39, no. 11, p. btad651, 2023, doi: 10.1093/bioinformatics/btad651.
13. H.-S. Oh and Y. Jung, "Cluster-based query expansion using external collections in medical information retrieval," *Journal of Biomedical Informatics*, vol. 58, pp. 70–79, 2015, doi: 10.1016/j.jbi.2015.09.017.
14. Y. Wang, S. Wu, D. Li, S. Mehrabi, and H. Liu, "A part-of-speech term weighting scheme for biomedical information retrieval," *Journal of Biomedical Informatics*, vol. 63, pp. 379–389, 2016, doi: 10.1016/j.jbi.2016.08.026.
15. Y. Zhao, N. J. Fesharaki, H. Liu, and J. Luo, "Using data-driven sublanguage pattern mining to induce knowledge models: application in medical image reports knowledge representation," *BMC Medical Informatics and Decision Making*, vol. 18, no. 1, p. 61, 2018, doi: 10.1186/s12911-018-0645-3.
16. S. Balaneshinkordan and A. Kotov, "Bayesian approach to incorporating different types of biomedical knowledge bases into information retrieval systems for clinical decision support in precision medicine," *Journal of Biomedical Informatics*, vol. 98, p. 103238, 2019, doi: 10.1016/j.jbi.2019.103238.