Information Retrieval Research 1 (2025) 93–108 Submitted 10/24; Revised 02/25; Published 05/25

# Supporting Evidence-Based Medicine by Finding Both Relevant and Significant Works

**Sameh Frihat** SAMEH.FRIHAT@UNI-DUE.DE
*University of Duisburg-Essen*
*Duisburg, Germany*

**Norbert Fuhr** NORBERT.FUHR@UNI-DUE.DE
*University of Duisburg-Essen*
*Duisburg, Germany*

**Editor:** Haiming Liu

## Abstract

In this paper, we present a new approach to improving the relevance and reliability of medical information retrieval, which builds upon the concept of Level of Evidence (LoE). The LoE framework categorizes medical publications into seven distinct levels based on the underlying empirical evidence. Despite LoE framework's relevance in medical research and evidence-based practice, only few medical publications explicitly state their LoE. Therefore, we develop a classification model for automatically assigning LoE to medical publications, which successfully classifies over 26 million documents in MEDLINE database into LoE classes. The subsequent retrieval experiments on the TREC Precision Medicine datasets show substantial improvements in retrieval relevance, when LoE is used as a search filter.

**Keywords:** Medical Document Facade, Level of Evidence, Evidence-Based Medicine, Medical Search Engines

## 1 Introduction

In medical research and practice, where findings and decisions directly impact human lives, successful retrieval of relevant and reliable information from scientific literature is paramount. Relevant information includes findings that are directly applicable to a condition under study, whereas reliable means that the findings are consistent under similar conditions (Strage et al., 2023). These concepts contribute to identifying significant information, which implies that findings have a practical and meaningful impact that is not due to chance in terms of its effect on patient care or outcomes (Sathian et al., 2010).

Modern evidence-based medicine (EBM) relies on a systematic approach to guide medical decisions using scientific evidence (Burns et al., 2011; Patrick et al., 2004). A key component of EBM is the Level of Evidence (LoE) framework, which categorizes medical research papers into 7 main distinct levels based on the strength and reliability of evidence reported (Rosner, 2012; Desai et al., 2019; Van de Vliet et al., 2023). This stratification, exemplified by the OCEBM (Oxford Centre for Evidence-Based Medicine `https://www.cebm.net/`) framework (Howick, 2011), ranges from highly rigorous and reliable systematic reviews of randomized controlled trials (Level 1a) to case studies with limited evidential value (Level 4) (Borawski et al., 2007; Group et al., 2002).

WWW: https://irrj.org DOI: 10.54195/irrj.19784

Within this framework, each level holds unique significance, representing a specific study design and methodology (Borawski et al., 2007). The hierarchy includes the following Levels of Evidence (LoEs):

- **Level 1a: Systematic Reviews of Randomized Controlled Trials (RCTs)**. At the apex of the LoE pyramid are systematic reviews and meta-analyses of well-conducted RCTs. Renowned for their comprehensive analysis of rigorous research, these reviews yield the most authoritative evidence.
- **Level 1b: Individual Randomized Controlled Trials (RCTs)**. This level features individual RCTs that contribute crucial insights into causal relationships by evaluating interventions in controlled settings.
- **Level 2a: Systematic Reviews of Cohort Studies**. Systematic reviews of cohort studies provide valuable evidence regarding associations between interventions and outcomes in real-world settings.
- **Level 2b: Individual Cohort Studies**. Individual cohort studies at this level offer meaningful evidence about interventions' effects within specific populations.
- **Level 3a: Systematic Reviews of Case-Control Studies**. Systematic reviews of case-control studies extend insight into the associations between interventions and outcomes, offering a broader perspective.
- **Level 3b: Individual Case-Control Studies**. Individual case-control studies contribute evidence by exploring the relationships between interventions and outcomes within well-defined contexts.
- **Level 4: Case Series**. At this level, case series provide preliminary evidence about interventions' effects, although they are limited by their susceptibility to biases and confounding factors.

Although LoE is a crucial parameter for assessing a medical publication's significance, it is often not explicitly stated in publications, creating a problem for medical information retrieval (IR), where the aim is to retrieve significant medical publications or their content.

Our work addresses the 'Acquiring' stage of the 5A's model (Ask, Acquire, Appraise, Apply, and Assess) in EBM (Leung, 2001), which focuses on retrieving relevant literature to help users find the best available evidence. While LoE and the 5A's model are distinct frameworks, enabling users to filter retrieved information based on LoE supports the 'Acquiring' stage. Future work could explore integrating automated evidence appraisal to complement our retrieval approach.

In this article, we propose an automatic approach to identifying and prioritizing significant works in medical research. First, we develop a classification method for automatically assigning LoE to medical publications, then we use the identified LoE as a search filter in an IR setting. We demonstrate on the TREC PM (Precision Medicine) 2017–2019 collections (Roberts et al., 2017) that using LoE as a filter when retrieving medical papers leads to improved retrieval results, and that the gain is highest for highly evidential medical papers.

## 2 Related Work

Recent advancements in Evidence-Based Medicine (EBM) have emphasized the role of automation in enhancing the classification and credibility assessment of Clinical Trials and RCTs. A key development in this area is the RobotReviewer system introduced by (Marshall et al., 2014, 2016), which automates the risk of bias assessment in RCTs and provides quality supporting text for bias assessments. This is vital for individual RCTs and also applicable to systematic reviews and meta-analyses of RCTs. The evaluation results indicate that RobotReviewer could match the performance of human reviewers in assessing the risk of bias (Marshall et al., 2016; Marshall and Wallace, 2019), which has been confirmed by several subsequent studies (Soboczenski et al., 2019; Hirt et al., 2021; Arno et al., 2022). Further, contributions from Hartling and Gates (2022) highlight the potential of such automation technologies to refine the quality and efficiency of systematic reviews, particularly in evaluating RCTs.

These advancements mark a significant shift in EBM, offering effective solutions for processing and categorizing extensive medical literature. However, these studies do not cover the full range of evidence levels of medical publications. Instead, they focus only on RCTs and their systematic reviews (Levels 1b and 1a in the LoE framework) and are possibly also applicable to levels 2b and 2a (cohort studies and their systematic reviews).

While large-scale metadata sources such as PubMed's "publication type" field offer broader coverage (e.g., labeling studies as "Clinical Trial" or "Review"), they lack explicit evidence-hierarchy distinctions (e.g., differentiating high-quality RCTs from lower-quality observational studies) required for direct alignment with the LoE framework (Pasche et al., 2020). Several machine learning-based tools are developed and used for predicting the "publication type" field such as Anne O'Tate, RCT Tagger, Multi-Tagger, etc (Cohen et al., 2021).

No other automation effort to date has explicitly attempted to incorporate the LoE framework, despite its central place in EBM practice. This work's main contribution is in providing a fully automatic retrieval system for medical publications by automatizing the EBM practice of assigning LoE to medical publications and then using LoE to decide on the relevance of a publication in a given context.

## 3 LoE Classifier

We view the problem of assigning LoE to medical publications as a classification task and explain in this section the training and the evaluation of the LoE classifier.

### 3.1 Data

We use a dataset derived from the Oncology Guidelines of the German Association of Scientific Medical Societies (Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften[1]). This dataset is unique in that it explicitly mentions the LoE of various medical publications as per the OCEBM framework. It includes 2816 publication–LoE pairs, extracted from unstructured PDFs[2]. The distribution of LoE levels in the dataset is

[1] `https://www.awmf.org/`

[2] A structured format of the dataset is available on `https://github.com/samehfrihat/LevelOfEvidence`

as follows: 14% in 1a, 18% in 1b, 10% in 2a, 24% in 2b, 12% in 3a, 7% in 3b, and 15% in 4. In Section 4.1, we compare this dataset distribution with the rest of the medical literature.

The Oncology Guidelines mention publications as citations, which include the authors names, publication year, and publication title. This information is not sufficient for automatic LoE classification, which additionally requires some of the methodology, interventions, and clinical outcomes. This information can only be found in publication abstracts or full texts. Therefore, we leverage the PubMed API[3] to enrich the initial dataset with abstracts and PubMed IDs.

The average word count in the abstracts is 263 (SD=97), slightly above the typical range for medical articles (Andrade, 2011). The prevalence of longer abstracts can be attributed to the frequent use of structured abstract formats within the medical literature (Hartley, 2004). Notably, we observe a positive correlation between the abstract length and the LoE classification: publications with higher evidence levels tend to have longer abstracts (e.g. LoE 1a with a mean of 325 words (SD=163) than those with lower levels (LoE 3b and 4 with a mean of 233 words (SD=71)).

We split this data into a training dataset containing 1690 instances (60%) and a validation and testing dataset containing 563 instances (20%) each, ensuring a stratified representation across all classes.

### 3.2 Experimental Setup

For the task of LoE classification, we focus on fine-tuning PubMedBERT (Gu et al., 2021). PubMedBERT is a natural choice for this domain-specific classification task as it is a transformer-based model pre-trained using abstracts sourced directly from PubMed. Its efficacy has been well-established: It currently holds the top score on the Biomedical Language Understanding and Reasoning Benchmark (Gu et al., 2021), it excels in accurately interpreting the unique terminologies and context of biomedical texts, and it is proficient in handling the complexities of biomedical literature. The model is fine-tuned using the training set and hyperparameters are optimized using the validation set. We develop the following classifiers:

**Random Forest (RF)** RF serves as our baseline. It is trained on the training set for multi-class classification. We use TF-IDF vectorization and chi-squared feature selection, and K-Fold cross-validation using the validation dataset, evaluating its performance with the macro-F1 score.

**Multi-Class-PubMedBERT** This classifier is directly fine-tuned on the training set to classify texts into specific LoE classes, with the macro-F1 score as the evaluation matrix.

**Reg-PubMedBERT** This is a regression approach, which assigns numeric values to LoE classes. PubMedBERT is fine-tuned to predict these values, by mapping different LoEs (1a, 1b, 2a, 2b, 3a, 3b, 4) to the numeric values (0, 1, 2, 3, 4, 5, 6). We used root-mean-square error (RMSE) for evaluation. To align the model's predictions with the original LoE classes and to facilitate comparison with other classifiers using the F1 matrix, we mapped the predicted value to the nearest integer value and then used the same map to get predictions back to their corresponding LoE classes.

[3] `https://pubmed.ncbi.nlm.nih.gov/`

**Multi-Label-PubMedBERT** This classifier incorporates the multi-label classification approach, i.e. we transform the LoE categorization into a set of binary labels. Each label corresponds to a specific LoE class, effectively converting the problem into a multi-label classification task. This version enabled PubMedBERT to predict multiple labels simultaneously, accommodating the scenario where only one of the labels should be true while others are false. By modelling the LoE classification as a multi-label task, we aim to capture potential overlap between LoE classes and assess the model's capacity to handle such nuances by looking at the prediction list that might contain multiple levels of evidence. For proper evaluation, we assigned the highest confidence value when multiple positive predictions.

**Ensemble Majority Vote** Ensemble methods are a well-established technique in classification that capitalizes on the strengths of diverse classifiers to enhance prediction accuracy and generalization (Polikar, 2012). We employed an Ensemble Majority Vote strategy, combining the strengths of the three PubMedBERT models (Multi-Class, Reg, and Multi-Label). This approach used majority voting to aggregate predictions from each model, enhancing the overall classification accuracy and robustness (Zhou and Zhou, 2021; Dang et al., 2020).

### 3.3 Classifier Evaluation

We evaluate our LoE document classifiers using Macro F1 score, RMSE, and Confusion matrices. RMSE treats the LoE classification as a regression task by mapping each LoE category to a corresponding integer. The RMSE score reflects the average squared difference between the predicted and true LoE values, providing important insight into how closely the model captures the ordinal nature of the LoE hierarchy. Lower RMSE values indicate that the model's predictions are closer to the true LoE, particularly emphasizing the reduced impact of misclassifications to adjacent levels.

#### 3.3.1 Individual Classifiers Performance

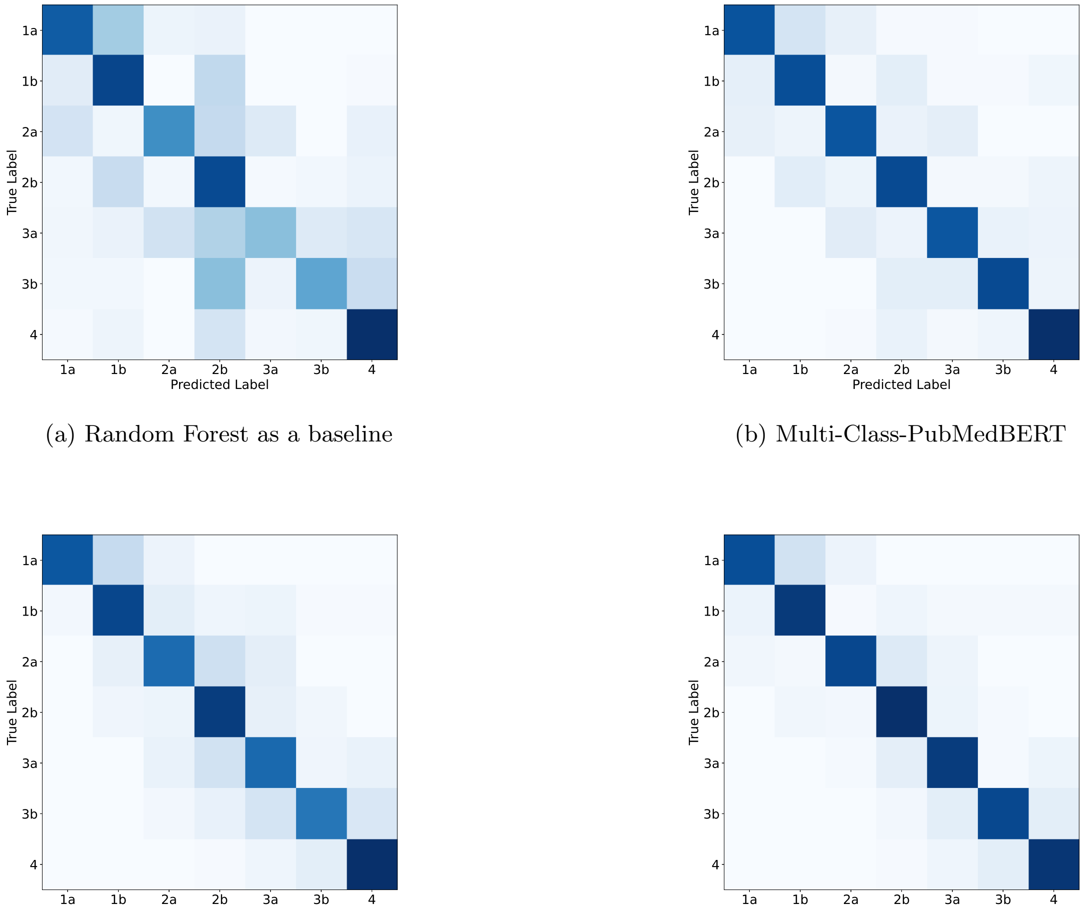
Table 1 summarizes the performance of each classifier on the test dataset.

| Model | F1 score | RMSE |
|---|---|---|
| Random Forest (RF) | 0.59 | 1.30 |
| Multi-Class-PubMedBERT | 0.78 | 0.90 |
| Reg-PubMedBERT | 0.74 | 0.69 |
| Multi-Label-PubMedBERT | 0.79 | 0.90* |
| Majority voting | **0.83** | **0.65** |

Table 1: Level of Evidence Classifiers Performance on our test set. Macro F1 Score.
* By considering the label of the highest confidence score as predicted class

**RF Baseline** The RF model's performance with a macro-F1 score of 0.59 and an RMSE of 1.30 did not surpass the deep learning models' results. Nevertheless, the RF model shows robustness in effectively handling the challenges of multi-class LoE classification. Analyzing the confusion matrix in Figure 1a, we see that the misclassifications are rather scattered, they are not clustered in any particular class or in neighboring classes.

**Multi-Class-PubMedBERT** Multi-Class-PubMedBERT scored 0.78 in F1 (+0.19 compared with baseline) and 0.90 in RMSE, showing effectiveness in multi-class categorization. However, after we analysed misclassification in Figure 1b, we found that the model has some difficulties distinguishing closely related LoE classes. This suggests considering the problem as a regression task, since misclassification with neighbor classes is not as bad as assigning distant classes such as replacing 1a with 4.



(a) Random Forest as a baseline

(b) Multi-Class-PubMedBERT

(c) Reg-PubMedBERT

(d) Majority voting

Figure 1: Confusion Matrices using the test set per model.

**Reg-PubMedBERT** exhibited strengths in capturing the ordered nature of LoE with an F1 score of 0.74 and the second-best RMSE of 0.69, indicating proficiency in differentiating between levels. This makes misclassified documents closer to the true labels, which is reflected in the smaller RMSE and highlighted in Figure 1c. This makes misclassification to neighboring classes less harmful than assigning far classes.

**Multi-Label-PubMedBERT** The model performed best among individual classifiers with an F1 score of 0.79, adeptly handling documents with multiple LoE categories. A closer qualitative examination of this model's performance revealed that some documents were assigned into multiple LoE classes. This is a well known phenomenon, which was explored in the work of Murad et al. (2016) bringing into question the clear demarcation between the evidence levels of the EBM pyramid. Instead, a nuanced perspective on LoEs has been proposed to align with the flexibility of multi-label classification as demonstrated by Multi-Label-PubMedBERT.

### 3.3.2 Ensemble Majority Vote Performance

The Ensemble Majority Vote method combines the predictions of all three PubMedBERT models and demonstrates the best performance. It scores highest in F1 (0.83) and achieves an RMSE of 0.65, indicating its effectiveness in accurately categorizing medical literature by LoE. This result emphasizes the significant role of collaborative intelligence in enhancing classification outcomes. It also benefited from the power of the regression model, where misclassification resulted in neighboring classes as shown in Figure 1d and the smallest RMSE.

### 3.3.3 Statistical Significance Analysis

We performed a statistical significance analysis on our machine learning models using a paired t-test. After applying Bonferroni correction ($\alpha = 0.05/10$), we found that all deep learning models significantly outperformed the Random Forest baseline, indicating their effectiveness in LoE classification. However, no significant performance differences were observed among the deep learning models themselves, highlighting their comparable efficacy in evidence-based classification.

### 3.3.4 Identifying Significant Terms

We utilized the LIME (Local Interpretable Model-Agnostic Explanations) explainer (Ribeiro et al., 2016) to identify key terms influencing our model's predictions for different Levels of Evidence (LoE) categories. This method provides insights by aggregating term scores, helping us to determine significant terms for each LoE level. Such an approach enhanced the interpretability and transparency of our model, highlighting LoE-specific terms in the analyzed documents.

Table 2 presents the top 10 contributing terms across the LoE levels in the test set. The results highlighted that our model was able to identify discriminating terms for each class. Moreover, we discovered common terms shared across multiple levels, such as "systematic review" in 1a (systematic reviews of RCTs), 2a (systematic reviews of cohort studies), and 3a (systematic reviews of case-control studies), and "RCT" in 1a and 1b (individual RCTs). Additionally, some less expected terms, like "risk" in 2a, 2b (individual cohort studies), 3a, and 3b (individual case-control studies), and "accuracy study" in 1a, 2a, and 3a (pertaining to Diagnostic Test Accuracy studies), emerged as significant classifiers. Interestingly, a specific therapy ("acupuncture") only occurs among the terms of level 4, possibly indicating the lack of stronger evidence for this method.

| 1a | | 1b | | 2a | | 2b | |
|---|---|---|---|---|---|---|---|
| term | score | term | score | term | score | term | score |
| accur predict | 2.11 | achiev complet | 1.92 | cohort studi | 1.30 | cohort studi | 1.62 |
| accur stage | 1.85 | achiev patient | 1.91 | accuraci detect | 1.14 | accrual | 1.42 |
| accuraci respect | 1.72 | activ control | 1.58 | systemat review | 1.09 | acquisit | 1.14 |
| rct | 1.42 | activ intervent | 1.56 | meta analysi | 1.02 | accept | 1.11 |
| meta analysi | 1.31 | activ surveil | 1.25 | exposur | 0.98 | access | 1.08 |
| systemat review | 1.30 | rct | 1.21 | longitudin | 0.95 | accru | 1.01 |
| accuraci studi | 1.17 | control set | 1.12 | access | 0.74 | longitudin | 0.89 |
| accuraci clinic | 1.16 | acut delay | 0.98 | accur stage | 0.73 | risk | 0.61 |
| achiev | 1.15 | acut | 0.79 | accuraci studi | 0.64 | administr | 0.21 |
| activ treatment | 1.02 | adjuv | 0.71 | risk | 0.59 | affect patient | 0.14 |

| 3a | | 3b | | 4 | |
|---|---|---|---|---|---|
| term | score | term | score | term | score |
| systemat review | 1.24 | case control | 1.60 | small sampl | 1.69 |
| epidemiolog | 1.21 | case definit | 1.41 | preliminari evid | 1.32 |
| case definit | 1.17 | exposur | 1.02 | exploratori research | 0.99 |
| abnorm | 1.12 | risk | 0.49 | uncontrol studi | 0.98 |
| exposur | 1.11 | advers reaction | 0.31 | acupunctur treatment | 0.68 |
| absent | 0.98 | affect patient | 0.30 | patient characterist | 0.60 |
| accuraci respect | 0.88 | age | 0.29 | acupunctur effect | 0.51 |
| accuraci studi | 0.71 | age diagnosi | 0.23 | analysi reveal | 0.22 |
| risk | 0.64 | advers effect | 0.19 | analysi identifi | 0.22 |
| accur stage | 0.51 | affect surviv | 0.10 | affect | 0.13 |

Table 2: Significant Terms in the Level of Evidence Classifier.

## 4 Levels of Evidence as a filter in medical IR

For the retrieval experiments, the 7-class LoE model was simplified into a 4-class setup by grouping related evidence levels. This decision reflects real-world usage patterns where users often prioritize broader evidence categories, such as high-quality studies (e.g., systematic reviews and RCTs) or intermediate-level evidence (e.g., cohort and case-control studies). This reduction not only simplifies classification, but also improves retrieval effectiveness without compromising performance.

In this experiment, we investigate the benefit of LoE classification for the IR of medical publications using TREC Precision Medicine (PM) datasets from 2017 to 2019 (Roberts et al., 2017, 2018, 2019).

### 4.1 Data

The TREC PM datasets, sourced from the Medline collection[4], consist of over 26 million research article abstracts accessible via PubMed and designed to enhance biomedical IR. Topics/queries were constructed based on disease and gene fields from the dataset, omitting demographic data to focus specifically on abstract retrieval. Relevance judgements were

[4] `https://www.nlm.nih.gov/medline/medline_overview.html`

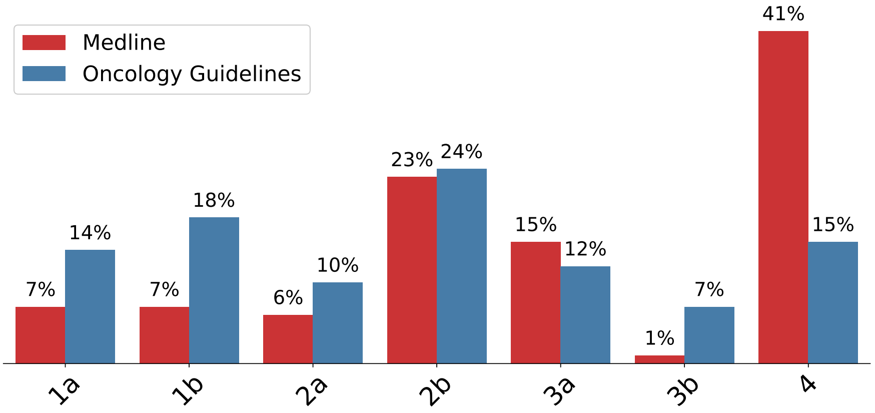


Figure 2: The distribution of LoE Classes in the Medline Dataset and Oncology Guidelines (Classifier Dataset).

performed by expert assessors on a scale of ‘not relevant (0)’, ‘partially relevant (1)’, and ‘definitely relevant (2)’, based on alignment with a given topics (Roberts et al., 2017). The criteria for relevance did not include the LoE of the documents.

We categorize each abstract in the Medline collection into its respective LoE category using our ensemble classifier. Figure 2 shows the distribution of LoE classes in Medline data. Most frequent are Level 4 documents (41% of the collection), which require the smallest empirical basis. The highest LoE 1a and 1b each represent only 7% of the documents. This unbalanced distribution reflects the inherent nature of the biomedical literature, where expert opinion and hypothesis-generating studies far outnumber high-evidence clinical research, which is usually conducted after several studies have confirmed the same observations.

In contrast, the oncology guidelines exhibit a reversed pattern, with only 15% of low evidence documents, suggesting a higher balance ratio. This difference is due to the selection process during the formulation of new clinical guidelines, where publications with higher evidence are prioritized.

### 4.2 Experimental Setup

In this subsection, we present the IR methods used for performing this task and the evaluation metric. As a core retrieval algorithm, we use BM25 (Robertson et al., 2009), which is widely used in IR for scoring and ranking documents based on their relevance to a user query. It is a probabilistic-based approach that builds on the classic TF-IDF (Term Frequency-Inverse Document Frequency) model, refining it by incorporating factors like term frequency saturation and document length normalization. The algorithm calculates a score for each document by considering how frequently the query terms appear in the document, adjusting for the overall document length and the rarity of the terms across the entire corpus. BM25 is particularly valued for its ability to effectively balance term frequency and inverse document frequency, making it one of the most robust and popular methods for ranking search results.

In the retrieval experiment, we first indexed the entire Medline collection (Sec. 4.1) alongside their assigned LoE classes. The BM25 algorithm, parameterized with $K_1 = 1.2$ and $b = 0.75$ as recommended in (Connelly, 2019), was then applied to retrieve and rank documents based exclusively on textual relevance (abstracts and titles) to the query, without integrating LoE into the ranking process. This methodology ensures a fair comparison with the baseline.

### 4.2.1 Retrieval Methods

Our experiment utilizes the BM25 retrieval method applied to documents of all LoE classes (‘All’) as a baseline for our IR process. The impact of LoE classification is tested by filtering the documents based on their LoE as follows:

- *LoE3+:* LoE categories 3b to 1a, i.e. case-control studies or higher LoE.
- *LoE2+:* LoE categories 2b to 1a, i.e. cohort studies or higher LoE.
- *LoE1:* LoE categories 1a and 1b, i.e. RCTs only.

### 4.2.2 Evaluation Metric

The performance of each model's effectiveness was assessed using infNDCG, R-Prec, and P@10 matrices, as these are the official matrices used to report on the datasets. Also, we report the “Normalized discounted cumulative gain @10” (NDCG@10) metric as our core metric (Järvelin and Kekäläinen, 2000). This measure allows for considering relevance grades $0 \ldots n$, where, in our case, irrelevant documents receive a score of 0, while partially relevant and definitely relevant documents receive scores of 1 and 2, respectively. For a ranked document list, let $r_j$ denote the relevance grade of the document at rank $j$. Then the (unnormalized) discounted cumulative gain for a ranked list of length $k$ is defined as:

$$DCG(k) = \sum_{j=1}^{k} \frac{r_j}{max(1, \log_b(j))}.$$

With the denominator in the summation elements, DCG simulates a stochastic user stopping behavior, where not every user checks all documents up to the final rank $k$, but some users might stop at earlier ranks; the fraction of users reaching a certain rank is controlled by the logarithm base $b$ (usually chosen as $b = 2$). As the DCG values for a query depend heavily on the number of relevant documents in the collection, they are normalized by comparing them with the value $DCG_{opt}(k)$ of the optimum retrieval result (i.e. ranking documents by decreasing relevance grades), thus arriving at the normalized discounted cumulative gain:

$$NDCG@k = DCG(k)/DCG_{opt}(k).$$

Besides incorporating a fairly realistic user stopping behavior and being one of the few retrieval metrics considering different relevance grades, NDCG also has a nice theoretic property: (Ferrante et al., 2021) showed that NDCG comes closest to an interval scale (which is a requirement for computing means and effect sizes), while other popular measures with stochastic stopping behavior (like average precision or rank-biased precision) clearly violate this property.

### 4.3 Results

As shown Table 3 using LoE to filter out document set to be searched improves the retrieval effectiveness as measured by NDCG@10 score. The retrieval of RCT documents with highest LoEs is the most successful. Moreover, there is a clear trend in improving NDCG when the minimum LoE is increased. For all three collections, the strictest filter (LoE1 with only 14% of the collection) outperformed all other methods, with substantial NDCG improvements (0.08 ... 0.11) over the baseline. As we are re-using a test collection, performing statistical tests here would contradict statistical testing theory (Fuhr, 2017). Instead, we give the effect sizes, which indicate substantial improvements over the baseline.

Moreover, as shown in Table 4, our LoE1 model improved the performance of the baseline on all matrices. It also outperformed each of the best-reported runs on infNDCG matrix and provided comparable results on R-Prec[5]. In addition, the retrieval quality of our method is accompanied with the guarantee of returning only documents of the highest evidence. On the other hand, as the results for P@10 show, LoE seems to be too strict when the user is looking at all 10 top-ranking documents.

| Exp./Year | size* | 2017 | 2018 | 2019 |
|---|---|---|---|---|
| *All* | 100% | 0.46 | 0.59 | 0.54 |
| *LoE3+* | 59% | 0.48 (0.02)** | 0.60 (0.01) | 0.57 (0.03) |
| *LoE2+* | 43% | 0.49 (0.03) | 0.64 (0.05) | 0.58 (0.04) |
| *LoE1* | 14% | **0.54** (0.08) | **0.69** (0.10) | **0.65** (0.11) |

Table 3: Models' NDCG@10 performance on TREC PM datasets

* Size denotes the percentage of the collection that was considered in retrieval.

** Numbers in parentheses show the effect size when comparing with the baseline "All".

| Exp./Year | 2017 | 2018 | 2019 |
|---|---|---|---|
| *All* | 0.43 / 0.27 / 0.52 | 0.50 / 0.32 / 0.58 | 0.47 / 0.30 / 0.57 |
| *LoE3+* | 0.45 / 0.28 / 0.54 | 0.52 / 0.34 / 0.60 | 0.50 / 0.31 / 0.58 |
| *LoE2+* | 0.47 / 0.28 / 0.54 | 0.55 / 0.36 / 0.61 | 0.52 / 0.31 / 0.61 |
| *LoE1* | **0.52** / **0.30** / 0.55 | **0.57** / **0.38** / 0.61 | **0.58** / 0.34 / 0.61 |
| *Top run* | 0.46 / **0.30** / **0.64** | 0.56 / 0.37 / **0.71** | **0.58** / **0.36** / **0.65** |

Table 4: Models' InfNDCG/R-Prec/P@10 performance on TREC PM datasets.*

* Best reported runs per matrix, meaning the model performing best on P@10 is not the same as the model performing best on infNDCG.

[5] Note that these are pessimistic estimates, as unjudged documents only retrieved by our method are treated as irrelevant

## 5 Discussion

In this paper, we have effectively demonstrated the automated application of the LoE framework for improving the retrieval of relevant medical publications. Our approach, leveraging fine-tuned PubMedBERT models, has proven adept at classifying medical publications based on their LoE with a high degree of accuracy (macro F1 = 0.83). This advancement addresses a significant gap in existing literature, where previous studies have largely focused on specific evidence levels, particularly RCTs and their systematic reviews. The higher transparency of our approach gives users full control over the LoE of the documents returned. Moreover, the method investigated here could be directly integrated into the existing PubMed search engine, by simply adding estimated LoE as an additional document attribute that can be referred to in the query.

A key finding of our work is the effect of LoE filtering in directing attention towards the most reliable 14% of documents, while enhancing retrieval quality at the same time. This aspect is particularly crucial in the medical domain, where accessing accurate and high-quality information rapidly can make a pivotal difference in patient care and medical research. On the other hand, LoE2 or LoE3 papers may also be searched for in case there are no relevant answers in the top level, e.g. when the user is interested in more recent methods for which higher level studies are not available yet. Therefore, we acknowledge that clinical decision-making often requires synthesizing multiple sources across different LoE levels.

In our study, the LoE1 model outperformed the best-reported runs on the three datasets (Roberts et al., 2017, 2018, 2019) in terms of infNDCG and provided comparable results in R-Prec matrix. This demonstrated the effectiveness of using LoE as a filter in medical IR, improving the relevance and reliability of retrieved documents. These improvements over integrating the LoE filter in the BM25 baseline suggest that these benefits could extend to the other stronger baselines.

Although our study shows the potential of using LoE in Medline, one limitation that needs to be considered is the potential bias from using the oncology guideline dataset for training the classifiers. Medline collection contains publications where LoE can not be applied, such as bioinformatics. To apply it in real-world applications, we could introduce a new class, ”others”, where the model confidence score is below the seine threshold or when multiple positive labels are in the multi-label classifier.

Moverover, the LoE framework prioritizes study design rigor but does not assess study quality (e.g., risk of bias). Future work should integrate tools like GRADE (Guyatt, 2009) or Cochrane's risk of bias assessment to enhance reliability. This requires expanding the research article and analyzing the full text rather than the title and abstract, which are enough for assigning LoE.

In our recently published user study (Frihat et al., 2024), we present findings from an evaluation with medical professionals testing a clinical search engine that integrates LoE classification with biomedical concepts as a semantic layer (Frihat and Fuhr, 2025).

The results demonstrated strong user engagement with LoE: 93% of participants reported prior familiarity with LoE frameworks, and 85% actively filtered search results based on high LoE levels, noting that this feature facilitated their ability to prioritize high-quality evidence. Their feedback also highlighted the added value of biomedical concept extraction (e.g., gene-disease relationships) in contextualizing evidence.

## 6 Conclusion

Our research addresses the challenge faced by current search engines in identifying significant, evidence-backed medical publications. Although relevant and widely used in evidence-based medical practice, the LoE framework has not yet been fully automatised and tested for medical IR. We introduce a classification model for tagging medical research abstracts with LoE levels and demonstrate that a vast number of medical publications without LoE tags can be successfully and fully automatically enriched with this crucial information. Our retrieval results confirm that LoE is an effective filter that improves results in a fully automatic retrieval scenario. These results suggest that our LoE based approach to medical IR is a viable and robust tool to evidence-based medical practice, which can facilitate and improve medical decision-making, leading to better patient care. However, effective decision-making often requires synthesizing multiple studies and integrating clinical practice guidelines, which remains an important area for future work.

## Acknowledgments and Disclosure of Funding

This work was funded by a PhD grant from the DFG Research Training Group 2535 Knowledge- and data-based personalization of medicine at the point of care (WisPerMed), University of Duisburg-Essen, Germany. We also acknowledge support by the Open Access Publication Fund of the University of Duisburg-Essen.

## References

Chittaranjan Andrade. How to write a good abstract for a scientific paper or conference presentation. *Indian Journal of Psychiatry*, 53(2):172, 2011.

Anneliese Arno, James Thomas, Byron Wallace, Iain J Marshall, Joanne E McKenzie, and Julian H Elliott. Accuracy and efficiency of machine learning–assisted risk-of-bias assessments in “real-world” systematic reviews: A noninferiority randomized controlled trial. *Annals of Internal Medicine*, 175(7):1001–1009, 2022.

Kristy M Borawski, Regina D Norris, Susan F Fesperman, Johannes Vieweg, Glenn M Preminger, and Philipp Dahm. Levels of evidence in the urological literature. *The Journal of Urology*, 178(4):1429–1433, 2007.

Patricia B Burns, Rod J Rohrich, and Kevin C.Chung. The levels of evidence and their role in evidence-based medicine. *Plastic and Reconstructive Surgery*, 128(1):305, 2011.

Aaron M Cohen, Jodi Schneider, Yuanxi Fu, Marian S McDonagh, Prerna Das, Arthur W Holt, and Neil R Smalheiser. Fifty ways to tag your pubtypes: Multi-tagger, a set of probabilistic publication type and study design taggers to support biomedical indexing and evidence-based medicine. *medRxiv*, pages 2021–07, 2021.

Shane Connelly. Practical BM25 – Part 3: Considerations for picking b and k1 in Elasticsearch, 2019. URL `https://www.elastic.co/blog/practical-bm25-part-3-considerations-for-picking-b-and-k1-in-elasticsearch`.

Huong Dang, Kahyun Lee, Sam Henry, and Özlem Uzuner. Ensemble BERT for classifying medication-mentioning tweets. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 37–41, Barcelona, Spain, 2020. Association for Computational Linguistics. URL `https://aclanthology.org/2020.smm4h-1.5`.

Vishal S Desai, Christopher L Camp, and Aaron J Krych. What is the hierarchy of clinical evidence? *Basic Methods Handbook for Clinical Orthopaedic Research: A Practical Guide and Case Based Research Approach*, Springer, pages 11–22, 2019.

Marco Ferrante, Nicola Ferro, and Norbert Fuhr. Towards meaningful statements in IR evaluation: Mapping evaluation measures to interval scales. *IEEE Access*, 9:136182–136216, 2021. doi: 10.1109/ACCESS.2021.3116857.

Sameh Frihat and Norbert Fuhr. Integration of biomedical concepts for enhanced medical literature retrieval. *International Journal of Data Science and Analytics*, pages 1–24, 2025.

Sameh Frihat, Papernmeier, and Norbert Fuhr. Enhancing biomedical literature retrieval with level of evidence and bio-concepts: A comparative user study. The ACM/IEEE Joint Conference on Digital Libraries (JCDL), 2024.

Norbert Fuhr. Some common mistakes in IR evaluation, and how they can be avoided. *SIGIR Forum*, 51(3):0 32–41, 2017. URL `http://sigir.org/wp-content/uploads/2018/01/p032.pdf`.

Evidence-Based Medicine Working Group, Gordon Guyatt, Drummond Rennie, et al. *Users' guides to the medical literature: a manual for evidence-based clinical practice.* AMA Press, 2002.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

GH Guyatt. Grade: an emerging consensus on rating quality of evidence and strength of recommendations. *Chinese Journal of Evidence-Based Medine*, 9:8, 2009.

James Hartley. Current findings from research on structured abstracts. *Journal of the Medical Library Association*, 92(3):368, 2004.

Lisa Hartling and Allison Gates. Friend or foe? the role of robots in systematic reviews. *Annals of Internal Medicine*, 175(7):1045–1046, 2022.

Julian Hirt, Jasmin Meichlinger, Petra Schumacher, and Gerhard Mueller. Agreement in risk of bias assessment between robotreviewer and human reviewers: An evaluation study on randomised controlled trials in nursing-related cochrane reviews. *Journal of Nursing Scholarship*, 53(2):246–254, 2021.

Jeremy Howick. The Oxford 2011 levels of evidence. *Centre for Evidence-Based Medicine*, 2011. URL `https://www.cebm.ox.ac.uk/resources/levels-of-evidence/ocebm-levels-of-evidence`

Kalervo Järvelin and Jaana Kekäläinen. IR evaluation methods for retrieving highly relevant documents. In Emmanuel J. Yannakoudakis, Nicholas J. Belkin, Peter Ingwersen, and Mun-Kew Leong, editors, *SIGIR 2000: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 41–48, ACM, 2000. doi: 10.1145/345508.345545.

Gabriel M Leung. Evidence-based practice revisited. *Asia Pacific Journal of Public Health*, 13(2):116–121, 2001.

Iain J Marshall and Byron C Wallace. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Systematic reviews*, 8:1–10, 2019.

Iain J Marshall, Joël Kuiper, and Byron C Wallace. Automating risk of bias assessment for clinical trials. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 88–95, 2014.

Iain J Marshall, Joël Kuiper, and Byron C Wallace. Robotreviewer: evaluation of a system for automatically assessing bias in clinical trials. *Journal of the American Medical Informatics Association*, 23(1):193–201, 2016.

M Hassan Murad, Noor Asi, Mouaz Alsawas, and Fares Alahdab. New evidence pyramid. *BMJ Evidence-Based Medicine*, 21(4):125–127, 2016.

Emilie Pasche, Déborah Caucheteur, Luc Mottin, Anaïs Mottaz, Julien Gobeill, and Patrick Ruch. SIB text mining at TREC precision medicine 2020. In *Proceedings of the 29th Text REtrieval Conference (TREC 2020)*. 16–20, 2020.

Timothy B Patrick, George Demiris, Lillian C Folk, David E Moxley, Joyce A Mitchell, and Donghua Tao. Evidence-based retrieval in evidence-based medicine. *Journal of the Medical Library Association*, 92(2):196, 2004.

Robi Polikar. Ensemble learning. *Ensemble machine learning: Methods and applications*, pages 1–34, 2012.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Kirk Roberts, Dina Demner-Fushman, Ellen M Voorhees, William R Hersh, Steven Bedrick, Alexander J Lazar, and Shubham Pant. Overview of the TREC 2017 precision medicine track. In *Proceedings of the 28th Text REtrieval Conference (TREC)*, 2017.

Kirk Roberts, Dina Demner-Fushman, Ellen M Voorhees, William R Hersh, Steven Bedrick, and Alexander J Lazar. Overview of the TREC 2018 precision medicine track. In *Proceedings of the 29th Text REtrieval Conference (TREC)*, 2018.

Kirk Roberts, Dina Demner-Fushman, Ellen M Voorhees, William R Hersh, Steven Bedrick, Alexander J Lazar, Shubham Pant, and Funda Meric-Bernstam. Overview of the TREC 2019 precision medicine track. In *Proceedings of the 30th Text REtrieval Conference (TREC)*, 2019.

Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

Anthony L Rosner. Evidence-based medicine: revisiting the pyramid of priorities. *Journal of Bodywork and Movement Therapies*, 16(1):42–49, 2012.

Brijesh Sathian, Jayadevan Sreedharan, Suresh N Baboo, Krishna Sharan, ES Abhilash, and E Rajesh. Relevance of sample size determination in medical research. *Nepal Journal of Epidemiology*, 1(1):4–10, 2010.

Frank Soboczenski, Thomas A Trikalinos, Joël Kuiper, Randolph G Bias, Byron C Wallace, and Iain J Marshall. Machine learning to help researchers evaluate biases in clinical trials: a prospective, randomized user study. *BMC Medical Informatics and Decision Making*, 19:1–12, 2019.

Katya Strage, Stephen Stacey, Cyril Mauffrey, and Joshua A Parry. The interobserver reliability of clinical relevance in medical research. *Injury*, 54:S66–S68, 2023.

Peter Van de Vliet, Tobias Sprenger, Linde FC Kampers, Jennifer Makalowski, Volker Schirrmacher, Wilfried Stücker, and Stefaan W Van Gool. The application of evidence-based medicine in individualized medicine. *Biomedicines*, 11(7):1793, 2023.

Zhi-Hua Zhou and Zhi-Hua Zhou. *Ensemble learning*. Springer, 2021.