TU WIEN Informatics

Automated Eligibility Screening and its Evaluation in the Medical Domain

DISSERTATION

submitted in partial fulfillment of the requirements for the degree of

Doktor der Technischen Wissenschaften

by

Wojciech Kusa
Registration Number 12044810

to the Faculty of Informatics

at the TU Wien

Advisor: Prof. Dr. Allan Hanbury
Second advisor: Prof. Petr Knoth

The dissertation has been reviewed by:

Evangelos Kanoulas	Byron Wallace

Vienna, 8th April, 2024

Wojciech Kusa

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Wojciech Kusa

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 8. April 2024

Wojciech Kusa

Abstract

Eligibility screening in the medical field is a critical process that involves assessing whether certain information meets predefined criteria for research or clinical use. The large amount of available data, the complexity and diversity of medical information, and the absence of standardised formats for organising and presenting data complicate the screening task. These barriers make it difficult for researchers, healthcare professionals, and patients to quickly and effectively access the required information for informed decision-making.

A relevant task in the medical domain that requires eligibility screening is clinical trial recruitment. This thesis begins by examining the challenges in matching patients to clinical trials. Clinical trials are crucial in advancing medical research and providing patients with potential treatment options. However, identifying suitable clinical trials for individual patients can be a complex and time-consuming task. We explore eligibility screening approaches that consider patient characteristics and trial eligibility criteria, showing the improvement in the retrieval precision over baseline models.

Next, the thesis focuses on systematic literature reviews, often regarded by researchers and practitioners as the cornerstone of evidence-based medicine. Systematic literature reviews aim to provide a comprehensive and unbiased summary of existing research on a specific topic. However, conducting a systematic literature review can be a labour-intensive undertaking, particularly when it comes to screening and selecting relevant citations from a large pool of potentially eligible studies.

We investigate automation techniques to tackle the challenge of citation screening in systematic literature reviews. Citation screening involves determining whether a study meets specific eligibility criteria for inclusion in the review. Automatic citation screening regards the development of machine learning and natural language processing methods to identify relevant studies and exclude irrelevant ones, thus streamlining the review process. These approaches have the potential to save researchers time and effort, enabling them to focus on analysing and synthesising the most pertinent information. Our contributions in this domain focus on three key factors: *datasets*, *evaluation measures* and *automation approaches*.

First, in terms of datasets, we extensively evaluate available citation screening resources. We identify limitations in the available datasets related to their small size, poor documentation, dataset overlap and lack of common evaluation. To tackle these issues,

we introduce two comprehensive citation screening datasets: CSMED and CSMED-FT. CSMED is a meta-dataset containing more than 300 systematic literature reviews—the largest publicly available citation screening dataset. We further extend CSMED with systematic review description, eligibility criteria and search strategy information. In contrast, the CSMED-FT is the first dataset specifically designed to model the screening of full text documents.

Second, the thesis also contributes to improving evaluation methods for automated citation screening approaches. Evaluating the performance of these algorithms is crucial to ensure their reliability and effectiveness. The thesis proposes new evaluation measures and experimental designs to facilitate a more rigorous and standardised assessment of automated citation screening systems. We examine Work Saved over Sampling, the most popular evaluation measure in this field, showing its problems and proposing improvements. Additionally, we present an evaluation approach that shifts the focus to systematic review outcomes instead of Recall. We find that the evaluation based on individual publications' impact changes the ranking of compared models.

Finally, in terms of automation approaches, this work focuses on techniques based on neural networks and large language models to enhance the efficiency and accuracy of eligibility screening. We reproduce three neural network-based architectures for screening as binary classification, showing significant variability in results. We demonstrate how eligibility criteria can be used to model screening as a question-answering or natural language inference task. We also present results on the full text screening task showing that popular models still struggle with inference based on long documents with more than 4,000 words. To showcase how our findings can be used in practice, we introduce CRUISE–*Screening*, a tool combining search and screening capabilities, helping researchers conduct literature reviews more systematically.

Contents

Abstract v

Contents vii

1 Introduction 1
- 1.1 Eligibility Screening in the Medical Domain . . . 3
- 1.2 Systematic Literature Reviews . . . 5
- 1.3 Research Questions . . . 7
- 1.4 Published Research . . . 12
- 1.5 Thesis structure . . . 15

2 Matching Patients to Clinical Trials with Eligibility Criteria; A Pilot Study 17
- 2.1 Related Work . . . 18
- 2.2 Methodology . . . 21
- 2.3 Experiment Setup . . . 24
- 2.4 Results . . . 25
- 2.5 Summary . . . 33

3 Background and Literature Review 35
- 3.1 Notation . . . 35
- 3.2 Systematic Literature Reviews . . . 38
- 3.3 Citation Screening . . . 43
- 3.4 Citation Screening Datasets . . . 52
- 3.5 Evaluation of Citation Screening Automation . . . 59
- 3.6 Digital Tools and Resources in Literature Reviewing . . . 62

4 Citation Screening Datasets 67
- 4.1 CSMED: Citation Screening Meta-Dataset . . . 67
- 4.2 CSMED-FT: Full Text Classification Dataset . . . 74
- 4.3 Discussion . . . 77
- 4.4 Summary . . . 79

5 Relevance-based Evaluation Measures for Citation Screening 81

5.1 Screening Evaluation Measures Based on the Confusion Matrix 82
5.2 Work Saved over Sampling Measure 85
5.3 Analysis of the Work Saved over Sampling Measure 86
5.4 The Normalised WSS 89
5.5 Additional Evaluation Measures at Specific Recall Levels 91
5.6 VoMBaT Visual Analytics Tool 95
5.7 Discussion 100
5.8 Summary 104

6 Impact-based Evaluation Measures for Citation Screening 105
6.1 Evaluation Framework 107
6.2 Experiment Setup 112
6.3 Outcome-based Evaluation 115
6.4 Review-based Evaluation 120
6.5 Discussion 121
6.6 Summary 126

7 Automated Citation Screening as Binary Classification 129
7.1 Experiment Setup 130
7.2 Results 133
7.3 Discussion 137
7.4 Summary 142

8 Automated Citation Screening with Eligibility Criteria 143
8.1 Citation Screening with External Information 144
8.2 Full Text Screening 146
8.3 CRUISE–*Screening* 150
8.4 Discussion 157
8.5 Summary 158

9 Conclusion 161
9.1 Research Questions and Contributions 161
9.2 Future Research 166

List of Figures 169

List of Tables 173

Bibliography 177

CHAPTER 1

Introduction

Evidence-Based Medicine (EBM) has emerged as an approach that emphasises the use of the best available evidence to guide clinical decision-making [221, 222]. EBM integrates clinical expertise, patient values and preferences, and the most current and relevant research evidence. Its primary objective is to improve patient outcomes by ensuring that medical practices and interventions are based on solid scientific evidence.

One of the key roles of EBM is to bridge the gap between scientific research and clinical practice and its ability to enhance the quality and effectiveness of medical interventions [40]. EBM provides a systematic and structured approach to evaluate the available evidence, empowering healthcare practitioners to make informed decisions about patient care. By relying on rigorous scientific evidence, EBM minimises the influence of bias, personal opinions, and anecdotal experiences in medical decision-making. It promotes the use of interventions that have been proven effective through robust research methods while discouraging the adoption of interventions lacking supporting evidence or proven to be ineffective or harmful.

Furthermore, EBM encourages shared decision-making between healthcare providers and patients [60, 13]. It recognises that individual patients have unique preferences, values, and circumstances that should be considered when making medical decisions. EBM supports personalised and patient-centred care by integrating patient's preferences, beliefs, cultural perspectives, and life priorities with the best available evidence.

Figure 1.1 presents the evidence pyramid for EBM, which serves as a framework for evaluating the quality and reliability of different types of scientific evidence [91, 250]. Systematic literature reviews (SLRs) and meta-analyses are at the peak of the pyramid, providing the highest evidence level by synthesising data from multiple studies. These comprehensive analyses offer more reliable conclusions than individual studies alone. Just below, critically appraised sources offer expert evaluations of individual studies, enhancing the applicability of research findings. Following are randomised controlled trials (RCTs),

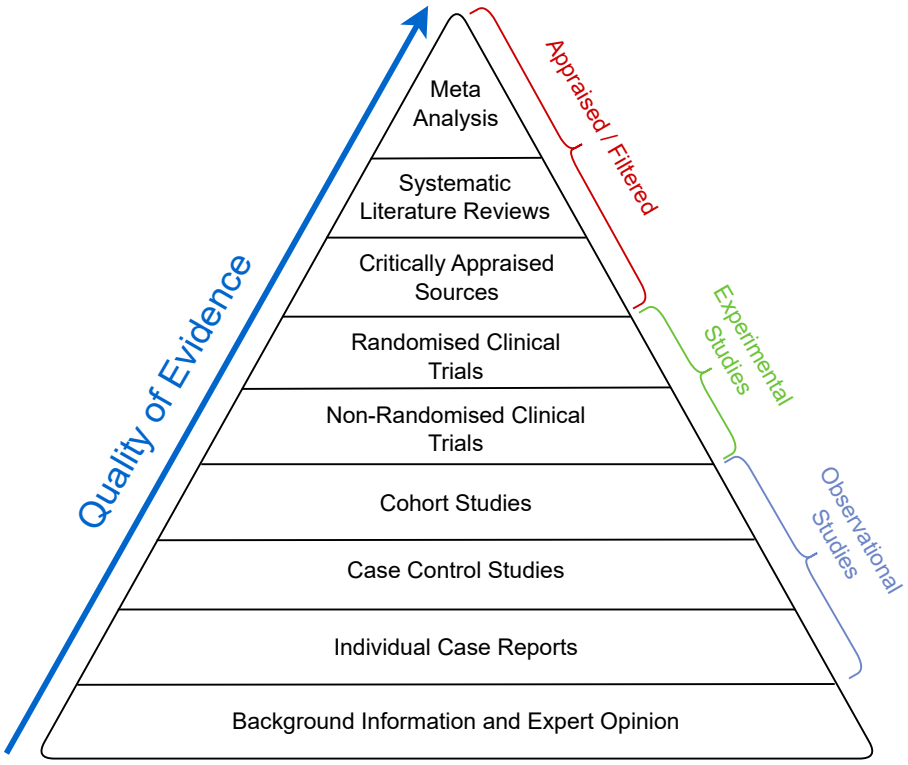


Figure 1.1: The Hierarchy of Evidence in Evidence-Based Medicine (EBM), detailing the quality of evidence from expert opinion to meta-analyses. This pyramid illustrates the increasing reliability and rigor of study designs as one moves up the levels, with meta-analyses representing the pinnacle of evidence quality due to their comprehensive review and analysis of literature. The differentiation between 'filtered' and 'unfiltered' information signifies the degree of critical appraisal and synthesis of evidence.

considered the gold standard for evaluating the effectiveness of interventions. Through the random assignment of participants to diverse treatment groups, RCTs aim to mitigate potential confounding variables. Non-randomised controlled trials also contribute valuable evidence, particularly where RCTs are not viable, despite their increased susceptibility to biases. Observational studies, such as cohort and case-control studies, are positioned further down the pyramid. While they may not provide evidence as robust as controlled trials, they can still offer valuable insights, especially when RCTs are not feasible or ethical. At the base of the pyramid are case reports, expert opinions and anecdotal evidence, which are considered the weakest forms of evidence. They often carry subjectivity and bias, making them less reliable for conclusive decisions. However, it is vital to recognise that even SLRs and RCTs are not immune to biases; these can skew results if the studies they incorporate are biased or if publication and selection biases are present.

The pyramid of evidence serves as a valuable tool for clinicians and researchers to

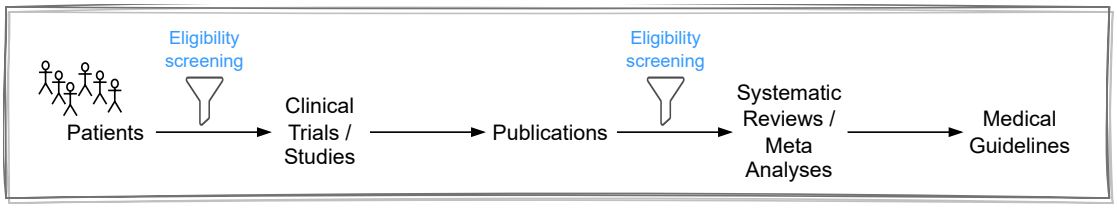


Figure 1.2: High-level data flow in Evidence-Based Medicine, depicting how individual patient data contributes to the creation of medical guidelines. Highlighted are sections where the task of eligibility screening is performed.

prioritise the most reliable evidence when making informed decisions about patient care and treatment strategies. Transitioning from understanding the broader context of EBM and its reliance on robust evidence hierarchies, we now delve into the nuances of eligibility screening in the medical domain.

1.1 Eligibility Screening in the Medical Domain

Eligibility screening is a systematic process of assessing whether a specific entity—a clinical study, patient, or any other subject—meets predetermined criteria for inclusion in a particular program, research, or study [244]. This ensures that only pertinent and suitable data or participants are included, which significantly impacts the quality and relevance of the final output, whether it is a research finding or a medical intervention. The term "eligibility screening" encapsulates various processes in the medical domain, and though they share a common objective, the specifics and implications differ based on the context. In this section, we examine how eligibility screening plays a pivotal role in EBM.

Figure 1.2 presents a simplified flow of data in EBM, illustrating how it all starts with individual patients data contributing to the creation of medical guidelines. Patients, by participating in clinical trials (CTs), enable their data to form a part of the evidence base. The results of these trials and research studies are then published as peer-reviewed publications in journals and conferences. Medical researchers search, screen and summarise all available evidence to create systematic literature reviews. These findings of SLRs become part of medical guidelines used by doctors [127].

This diagram highlights two steps where the task of eligibility screening is crucial: *matching patients to clinical trials* and *selecting publications for systematic literature reviews*. In both contexts, the essence of the process is to ensure the inclusion of relevant and appropriate data or participants, thereby enhancing the reliability and relevance of the outcome. However, while they share this foundational similarity, the specifics, methodologies, and implications can vary based on the context.

Selecting publications (often referred in this context as primary studies) for systematic literature reviews (Figure 1.3a) is an essential component of evidence synthesis [174, 94].

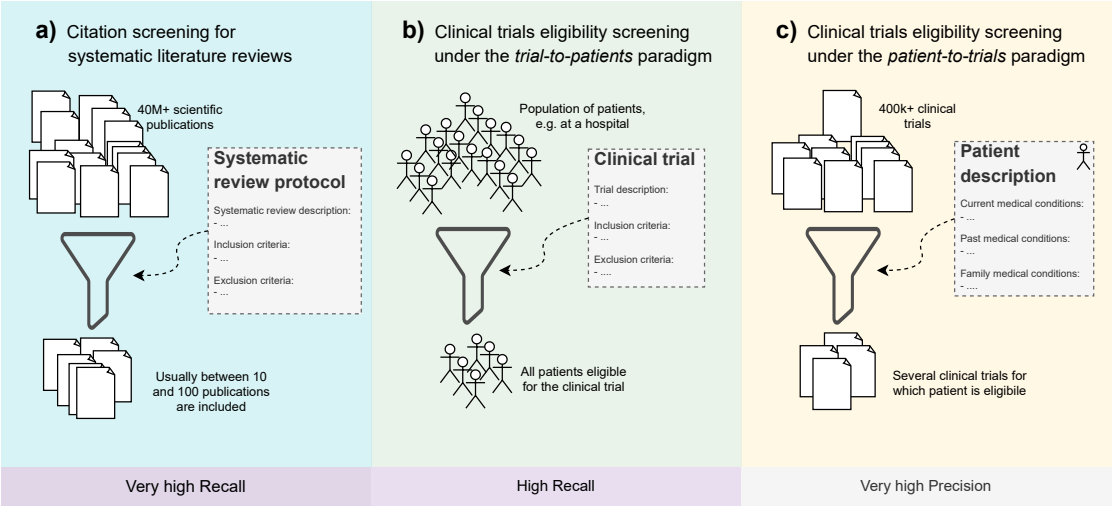


Figure 1.3: Three examples of eligibility screening in the medical domain. From the left: (a) citation screening for systematic literature reviews; (b) trial-to-patients clinical trials matching; (c) patient-to-trials clinical trials matching. The funnel in the middle of each diagram represents the eligibility screening process.

Systematic literature reviews aim to summarise and synthesise existing research on a specific topic to provide an overall understanding of the available evidence. During the selection process, researchers conduct a comprehensive search of various databases and sources to identify relevant studies that meet predetermined inclusion criteria. It is vital to conduct this search in a formal, broad manner to avoid bias by incorporating all relevant evidence, where relevance is well-defined based on the research question. The comprehensiveness of this search ensures that the review covers a wide range of study designs and methodologies, thus enhancing the reliability and generalisability of the findings. The eligibility screening involves assessing the study's methodology, quality, and relevance to the research question. The selected studies serve as the foundation for the systematic literature review, enabling researchers to draw conclusions and make evidence-based recommendations.

Selecting primary studies for systematic literature reviews aims to include all relevant research that addresses the research question. The eligibility criteria for systematic literature reviews are typically broader to encompass a wide range of study designs and methodologies. This inclusiveness allows for a comprehensive analysis of the available evidence and increases the generalisability of the findings. As a result, citation screening, and also the systematic literature review process overall, is a Recall-oriented task.

On the other hand, matching patients to CTs (Figures 1.3b and c) is a crucial step in medical research [171, 244]. The eligibility criteria for clinical trials are often designed to create a controlled environment that can isolate and evaluate the impact of the studied intervention. To conduct a clinical trial, researchers must recruit participants who meet

these criteria, such as age, gender, medical condition, and other relevant factors. The eligibility screening process involves identifying potential participants who meet the predetermined criteria and are willing to participate in the trial. This step ensures that the trial's results are applicable to the targeted patient population.

Two main paradigms for this type of screening can be presented: matching all eligible patients for a given trial or finding eligible trials for a given patient. The *trial-to-patients* paradigm (Figure 1.3b) is Recall-oriented as it prioritises broad inclusion. This approach aims to ensure diverse participation in studies for generalisable research results and balanced risk-benefit distribution. Researchers want to make sure that there is enough diversity in their studies [66, 120]. However, due to the HIPAA,[1] research access to the patient's medical records in the United States is limited, and therefore, this paradigm has rarely been studied [265, 216].

Conversely, the *patient-to-trials* paradigm (Figure 1.3c) focuses on helping single patients by providing them with the most relevant clinical trials. One patient realistically can only enrol and contribute to a few experiments (for instance, due to time constraints, excluding criteria or risks of confounding factors), but what matters most is the quality and relevance of those few trials. This means that screening clinical trials for a patient is a Precision-oriented task.

Eligibility screening is a fundamental process in the medical domain, ensuring that interventions, studies, and treatments are safe, relevant, and effective. Beyond clinical trials and systematic literature reviews, crucial and impactful areas where eligibility screening plays a pivotal role include organ transplantation, specialised treatments, genetic counselling, blood donation, and vaccination, among others. These processes collectively ensure that individuals receive appropriate and safe care tailored to their specific needs and conditions. As the main focus of the thesis is on systematic literature reviews, in the next section, we define this process in more detail.

1.2 Systematic Literature Reviews

The amount of scientific information, especially in the medical domain, is growing exponentially [159, 142]. Data from 2010 shows that 75 clinical trials and 11 systematic literature reviews were published per day [16]. Hoffmann et al. [97] estimated that, as early as 2019, a total of 80 systematic reviews were being published each day, demonstrating a significant rise compared to 2010. The Cochrane Library[2] listed 109,105 trials published in 2023 alone, which amounts to nearly 300 trials published daily. Scientific information, despite its growth in size, quickly becomes obsolete. New publications provide more recent experimental results that can change underlying views of a topic or even invalidate former findings. Typically, a single scientific paper focuses on a specific experiment, representing only a fraction of a broader research context. The tendency of dividing

[1] Health Insurance Portability and Accountability Act
[2] https://www.cochranelibrary.com/central

a larger study into several smaller publications not only fragments the literature but also, over time, encourages deeper exploration of specific facets of the main topic. Such fragmentation of literature is also caused by researchers trying to maximise the number of publications for a piece of work. This focused exploration results in specialisation and segregation into subfields of research [93].

Considering the characteristics mentioned above, it is almost impossible to stay up to date on the general research levels of all primary studies. Especially since human life is at stake in the medical domain, data available to specialists must provide recent and accurate results. For this purpose, systematic literature reviews (SLRs) are a standard process of analysing primary studies in the healthcare domain [42, 177]. They have a well-established and rigorous methodology for synthesising and evaluating the evidence on a specific research question [108].

Unfortunately, conducting SLRs is slow, labour-intensive and time-consuming as this relies primarily on human effort. A recent estimate shows that conducting a complete SLR takes, on average, 67 weeks [25], although another past study reports that the median time to publication was 2.4 years (125 weeks) [252]. Furthermore, according to Shojania et al. [236], 23% of published SLRs need updating within two years after completion.

This problem became very evident during the beginning of the COVID-19 pandemic. Only five SLRs on COVID-19 questions had been published worldwide in English a month after the World Health Organization (WHO) had declared the COVID-19 outbreak a public health emergency of international concern [284]. Even though the number of published SLRs increased quickly, many countries needed to decide on lockdowns, travel restrictions and quarantines without access to the systematically assessed scientific evidence about this topic. By now, the growth in produced documentation about COVID-19 generates another problem of a flood of evidence resulting in meta-SLRs [88]. Furthermore, there were examples of multiple systematic literature reviews conducted on the same topic by different research groups, raising concerns about research waste [238, 258].

Systematic literature reviews consist of multiple steps which do not necessarily follow a linear order. Depending on the granularity, previous studies enumerated between four and up to 15 tasks that might be included in the SLR process [254]. High-level tasks include steps of preparation, followed by retrieval and appraisal of primary studies and then synthesis and write-up of the evidence.

Citation screening (also known as a selection of primary studies) is a step in the SLR process that follows the retrieval of potentially relevant publications [254]. During the screening, reviewers read and assess hundreds (or thousands) of documents for eligibility with respect to the study criteria and decide whether or not these papers should be included in the SLR. These decisions are made based on considering each article's content with respect to predefined exclusion and inclusion criteria. Traditionally the screening consists of two stages. The first round of screening involves assessing titles and abstracts, which is supposed to narrow down the list of potentially relevant items. It is followed by

a task appraising the full texts, a more detailed (but also more time-consuming) revision of all included papers from the first stage based on the full text of articles.

Citation screening is among the most time-consuming steps of the SLR process, involving making thousands of eligibility decisions [186, 34]. Given the importance of citation screening in systematic literature reviews, there have been numerous attempts to automate the process [190]. Previous studies have investigated the use of automated citation screening methods for systematic literature reviews by using various natural language processing (NLP), machine learning (ML), and information retrieval (IR) methods to rank, retrieve, or classify papers [190, 263, 130, 99, 227, 114]. Already several commercial systems offer, to some degree, automation of the screening process.

Traditional ranking and classification models follow a similar approach and use text mining and machine learning algorithms to train a supervised model on an annotated dataset sample. This model is later used on the remaining part of the publication list to determine whether each article should be included or excluded from the review. More recently, the approaches include few- and zero-shot algorithms, often based on (large) language models, which decreased the need for a large annotated dataset sample. A successful automated citation screening algorithm should miss as few relevant papers as possible and also save time for the reviewers by removing irrelevant papers.

With the exponentially growing list of publications, manual screening is no longer feasible on a large scale. Having underscored the need to decrease the workload, as well as to improve the timeliness of published systematic literature reviews, this thesis focuses on models and evaluation approaches to improve this process. We assess available evaluation measures and datasets in citation screening for SLRs. We then establish a reliable and reusable benchmarking approach and propose novel evaluations of this task. Furthermore, we focus on citation screening in a zero-shot setting, i.e., without additional manual annotations. Moreover, we go beyond the systematic literature reviews and evaluate our algorithms in another medical application of clinical trial matching. In the next section, we discuss the research questions.

1.3 Research Questions

High-level research question: ***How can machine learning models help to automate the eligibility screening step in systematic reviews and clinical trial matching?***

To answer this question, we investigated the following four research questions:

RQ1: How should a comprehensive benchmark dataset for citation screening in systematic literature reviews be constructed to ensure robustness against large language models?

In research, a dataset refers to a structured collection of data. Typically, datasets for systematic literature reviews comprise individual literature items, their metadata, and often annotations or classifications based on research criteria. On the other hand, a benchmark is a standard or point of reference against which datasets or algorithms can be compared or assessed. In this context, a benchmark often involves a specific dataset, evaluation metrics, and set procedures that allow for consistent evaluation of different algorithms.

The emergence of deep learning and large language models (LLMs) has revolutionised NLP and ML research. However, the rapid progress in language models poses unique challenges in the creation of benchmark datasets that can endure through evolving research paradigms. This research question explores effective strategies for constructing robust benchmark datasets tailored for evaluating systematic literature review automation. By investigating methodologies and techniques that can enhance dataset diversity, complexity, and generalisability, this research seeks to establish guidelines for creating benchmark datasets. These guidelines aim to evaluate the true performance and capabilities of LLMs, while minimising the risks of issues like overfitting and data memorisation, terms which refer to models excessively tuning to specific data patterns and retaining specific data, respectively. To address RQ1, we defined three sub-questions, briefly described below.

***RQ1.1**: What is the current overview of benchmark datasets for automated citation screening?*

To develop effective and accurate automated screening systems, benchmark datasets serve as essential resources for training and evaluating machine learning models. This research question investigates the existing landscape of benchmark datasets designed explicitly for evaluating automated citation screening methods. By exploring the characteristics, size, evolution and quality of these datasets and their specific domain or research focus, this research seeks to provide a comprehensive understanding of the current state of benchmark datasets in this domain. The findings contribute to identifying potential gaps, limitations, and areas for improvement in the available datasets, thereby guiding future research efforts in developing more reliable and representative benchmark datasets for automated citation screening.

***RQ1.2**: Which properties should a benchmark collection have for a valid assessment of citation screening algorithms?*

As most of the previous research focused on using cross-validation to evaluate their algorithms, their effective subset of training and testing data differed between studies. This can generate problems, especially in datasets from citation screening exhibiting very high variability. Moreover, with a recent shift to the "pre-train, prompt and predict" paradigm, more and more datasets started to be used in a zero-shot setting. In this research question, we establish a set of properties for a new benchmark collection to

ensure a fair and detailed comparison between models. We then construct CSMED, a large meta-dataset consisting of 9 datasets comprising more than 300 historical systematic reviews. This dataset was rigorously cleaned from duplicates, updated with more detailed documentation and expanded available metadata with a systematic review description. In our approach, we emphasise the applicability of CSMED both in the traditional supervised classification, active learning, and zero-shot techniques.

***RQ1.3**: How should a dataset for full text publication screening in systematic literature reviews be constructed?*

Full text publication screening is a critical phase in systematic literature reviews where entire publications are evaluated in detail to determine their relevance to the research question. This process is more complex than title and abstract screening due to the depth and volume of information in full texts. Traditionally, automation efforts in this area were limited due to challenges in processing long documents with limited labelled data. However, the increasing use of LLMs is transforming this landscape. A comprehensive dataset for full text publication screening must address several unique challenges: it should capture the diversity in publication sources and formats, incorporate essential metadata, and include relevant annotations for specific criteria. Furthermore, the dataset requires scalability, continual updates to reflect the latest research, and user-friendly accessibility, complemented by thorough documentation.

RQ2: How should citation screening automation approaches for systematic literature reviews be evaluated?

If we assume an offline evaluation scenario, estimation of the quality might seem trivial at first sight, as the problem can be posed as a binary classification. However, the unique challenges of this domain, such as the very high class imbalance typically found in datasets, make evaluation more challenging. This imbalance arises because, in systematic reviews, only a small fraction of the screened papers is usually relevant, while the majority is not. The strict criterion of identifying all (or *nearly all*) relevant papers further amplifies this challenge. Moreover, it is essential to recognise that current search strategies start with a pool of potentially relevant eligible studies from a broad search. This assumption may soon be outdated with the advent of generative AI technologies and LLMs, potentially revolutionising the criteria and methodologies for evaluating these approaches. To address RQ2, we formulated three sub-questions, outlined below.

***RQ2.1**: What are the shortcomings of the common evaluation measures used in automated citation screening?*

We assess specific metrics currently used for the evaluation of the citation screening task, such as the Work Saved over Sampling (WSS) and the True Negative Rate (TNR) measures. We enumerate their drawbacks and explain what makes some of them unsuitable for assessing automated citation screening quality. While a detailed discussion is available in subsequent chapters, it's essential to understand that these measures, although popular, might not capture the nuances of the citation screening process.

***RQ2.2**: Which properties should evaluation measures have for appropriate assessment of citation screening algorithms?*

In order to assess the performance of citation screening algorithms accurately, it is essential to identify the properties that evaluation measures should possess. For instance, a metric should be sensitive to the inherent class imbalance and the critical need to retrieve nearly all relevant papers, ensuring a comprehensive and reliable evaluation. Fulfilling these properties should ensure a comprehensive and reliable evaluation of the algorithms.

***RQ2.3**: How could automated citation screening be evaluated differently to consider outcomes of systematic literature reviews?*

Traditionally, the evaluation of automated citation screening has focused on binary classification performance measures, such as Precision, Recall, and F1 score, to assess the algorithms' effectiveness in identifying relevant documents. However, to consider the outcomes of systematic literature reviews, the evaluation could be conducted differently, incorporating additional dimensions and measures. One approach is to evaluate the algorithms based on the impact their predictions have on the changes in review outcomes. This could involve assessing the algorithms' ability to prioritise relevant documents more likely to influence the review conclusions. It is crucial to recognise that this is a challenging attempt given the subjective nature of measuring a single paper's contribution to an overall systematic review's conclusions.

RQ3: How to use machine learning models for automated citation screening with highly imbalanced datasets so the results could be generalised to other reviews?

Automatic classification of documents is a well-explored problem in NLP, with high importance and multiple real-world applications [3]. Two major obstacles to developing effective text classifiers in practice are the lack of labelled data and class imbalance [104]. Both of these problems exist in automated citation screening. For automated citation screening, models should ensure consistent performance across multiple datasets, each with its unique characteristics. In the domain of citation screening, datasets often exhibit a skew towards one class, notably the class of documents that are not relevant or the "negative class" (excluded documents). Systematic literature review datasets vary by the total number of documents and the class balance, meaning that some of them can be 'easier' to learn by machine learning models than others. Here, we focus on four sub-questions: RQs 3.1 and 3.2 concerning model performance, while RQs 3.3 and 3.4 target the broader applications of automated citation screening.

***RQ3.1**: How do neural classification methods perform in the citation screening step, particularly when compared to traditional methods?*

We measure the performance of classic deep learning architectures (Recurrent and Convolutional Neural Network-based) on benchmark datasets of systematic literature reviews in

medicine. We found that none of the three tested architectures could produce consistent gains over established baseline models on benchmark datasets. Moreover, on average, the highest scores were obtained with the statistical Support Vector Machine (SVM) based method, highlighting the strength and potential versatility of more traditional methods in this domain.

***RQ3.2**: Which external knowledge sources can be used to improve the quality of automated citation screening?*

Previous work on classification for citation screening mainly focused on using only the paper's title and abstract information as the models' input. This information was sometimes extended by extracting additional data from medical taxonomies or predefined study characteristics (like population and outcomes) and treating this as additional input. Approaches focusing on retrieval and ranking also utilised search queries or the systematic review title to improve their search. Other automation approaches use information like citation graphs and the similarity of a paper to previously labelled publications. However, these methods potentially introduce biases as they only approximate actual SLR protocol.

The eligibility criteria specified in the study protocol contain a list of reasons for including and excluding a paper from the review. Manual screeners use this information as the critical component for deciding if a paper should be included or excluded. However, until now, approaches have not used eligibility criteria (list of criteria for inclusion and exclusion) as the ultimate decision-making criterion focusing on the similarity of already labelled papers. We test how the eligibility criteria section can be used as an input to neural network approaches, specifically in a zero-shot setting.

***RQ3.3**: How can recent advancements in language models be applied for automatic eligibility screening of full text publications?*

Another foreseeable application, mirroring manual workflows, could be during the full text eligibility screening. As full texts of publications exceed the maximum input size of many language models, they have been out of the scope of the previous research. However, recent developments in LLMs made it possible to process documents containing more than 10,000 words [81]. We evaluate several Transformer-based models and the GPT-x LLMs on the full text publication screening dataset, finding that fine-tuned Transformers still achieve the best results.

***RQ3.4**: Can these machine learning approaches generalise to systematic literature reviews conducted in a domain other than medicine?*

Systematic literature reviews in medicine follow a structured process. The eligibility criteria are detailed, as they adhere to strict guidelines. In other scientific disciplines, the criteria can also be written in less detail or using less strict vocabulary and structure. This can lead to lower generalisation of models trained using this information as an input feature. We propose a tool that performs automated search and screening using eligibility criteria. It can be used to measure the quality of automation approaches in scientific

disciplines beyond medicine. Researchers can also use the tool for conducting academic exploratory literature reviews.

RQ4: What techniques can be used to improve eligibility screening of patients to clinical trials?

The existing methods for the automatic selection of patients for clinical trials primarily focus on topical relevance, but there is a growing need to explore and develop techniques that can enhance the eligibility screening process [216]. This research question aims to investigate novel approaches and strategies that can improve the accuracy and efficiency of patient eligibility screening.

***RQ4.1**: What is the impact of individual sections of clinical trial text on the performance of a lexical retrieval approach?*

Clinical trials are multi-fielded semi-structured documents. Understanding the contribution of individual sections of clinical trial text on the performance of a lexical retrieval approach can help decide on the limitation of these approaches and make more informed decisions on how the clinical trials matching process can be improved. We analyse the impact of each section separately, such as the trial description, tested condition and inclusion and exclusion criteria, showing the limitations of lexical retrieval models in this task.

***RQ4.2**: How can information extraction techniques improve the retrieval of eligible clinical trials?*

Information extraction techniques have the potential to enhance the retrieval of eligible trials by extracting specific data elements from clinical trial documents, shifting the focus more on the inclusion and exclusion criteria. These techniques can involve using rule-based systems, statistical models, or machine learning algorithms to identify and extract relevant information. They can also be used to structure the patient description text, written in a free text. We show how the number of retrieved relevant trials can be improved by employing drug and disease entity recognition and negation detection both on patient and clinical trial documents.

1.4 Published Research

The subsequent chapters of this thesis are based on the following published research. The software developed to run the experiments presented in this thesis is open-source, governed by the Apache-2.0 license, and available at the following website: `https://github.com/WojciechKusa`.

Chapter 2 on eligibility screening for patients to clinical trials matching:

Wojciech Kusa, Óscar E. Mendoza, Petr Knoth, Gabriella Pasi, Allan Hanbury. “Effective

Matching of Patients to Clinical Trials using Entity Extraction and Neural Re-ranking".[3] In: *Journal of Biomedical Informatics*. JBI 2023. Journal paper. [136]

Chapter 4 on new citation screening datasets:

Wojciech Kusa, Óscar E. Mendoza, Matthias Samwald, Petr Knoth, and Allan Hanbury. "CSMeD: Bridging the Dataset Gap in Automated Citation Screening for Systematic Literature Reviews." In *37th Conference on Neural Information Processing Systems Track on Datasets and Benchmarks*. NeurIPS 2023. New Orleans, United States. Long paper. [137]

Chapter 5 on binary evaluation of citation screening:

Wojciech Kusa, Aldo Lipani, Petr Knoth, Allan Hanbury. "An Analysis of Work Saved over Sampling in the Evaluation of Automated Citation Screening in Systematic Literature Reviews". In: *Intelligent Systems with Applications*, pp. 200193, 2023, ISSN: 2667-3053. ISWA 2023. Journal paper; [134]

Wojciech Kusa, Aldo Lipani, Petr Knoth, Allan Hanbury. "VoMBaT: A Tool for Visualising Evaluation Measure Behaviour in High-Recall Search Tasks". In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR 2023. Taipei, Taiwan. System demonstration paper. [135]

Chapter 6 on outcome-based evaluation of citation screening:

Wojciech Kusa, Guido Zuccon, Petr Knoth, Allan Hanbury. "Outcome-based Evaluation of Systematic Review Automation". In: *Proceedings of the 13th International Conference on the Theory of Information Retrieval*. ICTIR 2023. Taipei, Taiwan. Long paper. [140]

Chapter 7 on conducting citation screening using binary classification algorithms:

Wojciech Kusa, Allan Hanbury, Petr Knoth. "Automation of Citation Screening for Systematic Literature Reviews using Neural Networks: A Replicability Study". In: *Proceedings of the 44th European Conference on Information Retrieval*. ECIR 2022. Stavanger, Norway. Reproducibility paper. [130]

Chapter 8 on citation screening using eligibility criteria:

Wojciech Kusa, Petr Knoth, and Allan Hanbury. "CRUISE–Screening: Living Literature Reviews Toolbox." In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. CIKM 2023. Birmingham, United Kingdom. System demonstration paper; [133]

Wojciech Kusa. "Rapid Systematic Reviews: Zero-shot Citation Screening with the Usage of Eligibility Criteria". In: *The 17th Conference of the European Chapter of the Association for Computational Linguistics Student Research Workshop*. EACL SRW 2023. Dubrovnik, Croatia, Workshop paper. [128]

[3] Except for contributions made in Sections 3.4 and 5.4 which are part of the PhD thesis by Óscar E. Mendoza from the University of Milano-Bicocca: "Adaptation of neural-enhanced retrieval model to domain-specific tasks".

Other publications related to the biomedical NLP and scientific document processing:

Jason Alan Fries, Leon Weber, Natasha Seelam, Gabriel Altay, Debajyoti Datta, Samuele Garda, Myungsun Kang, Ruisi Su, Wojciech Kusa, Samuel Cahyawijaya, Fabio Barth, Simon Ott, Matthias Samwald, Stephen Bach, Stella Biderman, Mario Sänger, Bo Wang, Alison Callahan, Daniel León Periñán, Théo Gigant, Patrick Haller, Jenny Chim, Jose David Posada, John Michael Giorgi, Karthik Rangasai Sivaraman, Marc Pàmies, Marianna Nezhurina, Robert Martin, Michael Cullan, Moritz Freidank, Nathan Dahlberg, Shubhanshu Mishra, Shamik Bose, Nicholas Michio Broad, Yanis Labrak, Shlok S. Deshmukh, Sid Kiblawi, Ayush Singh, Minh Chien Vu, Trishala Neeraj, Jonas Golde, Albert Villanova del Moral, Benjamin Beilharz. "BigBIO: A Framework for Data-Centric Biomedical Natural Language Processing". In: *Thirty-sixth Conference on Neural Information Processing Systems Track on Datasets and Benchmarks*. NeurIPS 2022. Long paper; [69]

Wojciech Kusa, Georgios Peikos, Óscar E. Mendoza, Allan Hanbury, Gabriella Pasi. "DoSSIER at MedVidQA 2022: Text-based Approaches to Medical Video Answer Localization Problem". In: *The 21st Biomedical Natural Language Processing at ACL 2022*. BioNLP 2022. Workshop paper; [132]

Wojciech Kusa, Edoardo Mosca, and Aldo Lipani. "*Dr LLM, what do I have?*: The Impact of User Beliefs and Prompt Formulation on Health Diagnoses". In: *3rd Workshop on NLP for Medical Conversations at IJCNLP-AACL 2023*. NLPMC 2023. Workshop paper; [138]

Óscar E. Mendoza, Wojciech Kusa, Alaa El-Ebshihy, Ronin Wu, David Pride, Petr Knoth, Drahomira Herrmannova, Florina Piroi, Gabriella Pasi, Allan Hanbury. "Benchmark for Research Theme Classification of Scholarly Documents". In: *Third Workshop on Scholarly Document Processing at COLING 2022*. SDP 2022. Workshop paper; [59]

Anjani Dhrangadhariya, Wojciech Kusa, Henning Müller, Allan Hanbury. "HEVS-TUW at SemEval-2023 Task 8: Ensemble of Language Models and Rule-based Classifiers for Claims Identification and PICO Extraction". In: *The 17th International Workshop on Semantic Evaluation*. SemEval 2023. Workshop paper; [55]

Jason Alan Fries, Natasha Seelam, Gabriel Altay, Leon Weber, Myungsun Kang, Debajyoti Datta, Ruisi Su, Samuele Garda, Bo Wang, Simon Ott, Matthias Samwald, Wojciech Kusa. "Dataset Debt in Biomedical Language Modeling". In: *Workshop on Challenges & Perspectives in Creating Large Language Models at ACL 2022*. Workshop paper. [68]

Wojciech Kusa, Yasin Ghafourian. "DoSSIER at TREC 2021 Clinical Trials Track." In: *Proceedings of the Thirtieth Text REtrieval Conference*. TREC 2021. Virtual. Shared task paper. [129]

Wojciech Kusa, Patrick Styll, Maximilian Seeliger, Oscar E. Mendoza, Allan Hanbury. "DoSSIER at TREC 2023 Clinical Trials Track." In: *Proceedings of the Thirty-Second Text REtrieval Conference*. TREC 2023. Virtual. Shared task paper. [139]

1.5 Thesis structure

This thesis is structured as follows. Chapter 2 introduces clinical trials eligibility screening and presents our work on eligibility screening in the patient-to-trials paradigm. Related work for clinical trials screening is included in Chapter 2. Chapter 3 introduces systematic literature reviews and state of the art concerning citation screening datasets, models and their evaluation. In Chapter 4 we introduce CSMED: the largest to date citation screening meta-dataset, and CSMED-FT: the first dataset specifically designed to evaluate citation screening of full text publications. Our contributions to the evaluation of citation screening for systematic literature reviews are showcased in Chapters 5 and 6. These chapters focus on relevance-based and impact-based evaluation techniques, respectively. In Chapter 7, we present the work on citation screening using binary classification algorithms. In contrast, in Chapter 8, we focus on screening using external information such as eligibility criteria, with extensions for the case of full text screening and literature reviews beyond medicine. We conclude the thesis in Chapter 9 by summarising the main findings and discussing future research opportunities for eligibility screening. Figure 1.4 presents the contributions with respect to each chapter.

Figure 1.4: Contributions with respect to each chapter.

CHAPTER 2

Matching Patients to Clinical Trials with Eligibility Criteria; A Pilot Study

In the previous chapter, we introduced two examples of eligibility screening in the medical setting: citation screening for systematic literature reviews and eligibility screening for matching patients to clinical trials. In this chapter, we will focus on the problem of matching patients to clinical trials, focusing on improving the performance of lexical methods.

Clinical trials (CTs) are crucial to the progress of medical science, specifically in developing new treatments, drugs, or medical devices [204]. Awareness and access to these studies are still challenging both for patients and physicians, making the recruitment of patients a significant obstacle to the success of trials [181, 204].

Even if a sufficient number of patients is found, the recruitment process requires screening the patients for eligibility, which is a labour-intensive task [63]. Automated identification of eligible participants not only promises great benefits for translational science [181] but also aids patients by allowing them to be included in specific trials [125].

In recent years, several initiatives have been proposed to build automatic systems for matching patients to CTs [125, 216, 215, 235]. The task has been defined as an information retrieval problem under the patient-to-trials evaluation paradigm [215] (Figure 1.3c). Here, the query is constituted by patient-related information, either in the form of electronic health records (EHRs) or ad-hoc queries, and the documents are the CTs currently recruiting patients [125].

This retrieval task involves the semantic complexity of matching the patients' information with heterogenous, multi-fielded CT documents [219]. In addition to this, the eligibility

criteria often use complex language structures (e.g. concepts can be negated twice) and medical jargon given in either semi-structured or unstructured ways [49].

To date, the existing approaches have revealed a significant lack of balance between efficiency and effectiveness. While pipeline-based models showcase promising performance, the substantial model sizes required to achieve competitive results raise concerns regarding costly deployment and limitations on reproducibility.

We evaluate the utility of individual sections of CT text on the performance of lexical retrieval system. We also develop a data enrichment process for both queries and documents for supporting CT search with a probabilistic lexical model as a first-stage retriever. It consists of entity recognition and negation detection modules applied to both the patient description and the eligibility section of CTs. The data enrichment process also provides the classification of both patient descriptions and CT eligibility criteria into current, past and family-medical conditions. The extracted information boosts the importance of affirmative and negative mentions of diseases and drugs for both the documents and queries in the traditional retrieval scenario. Finally, we compare several filtering techniques. We evaluate our work on the TREC Clinical Trials track 2021 and 2022 collections.

2.1 Related Work

This section describes previous work on CT matching with various paradigms, approaches to extract information from clinical data and from patient-related information, and neural re-ranking for CT retrieval.

2.1.1 Clinical trials matching

The TREC Clinical Trials track concerns the task of matching single patients to clinical trials. Other tasks concerning CT matching mentioned in the literature are cohort-based retrieval [126] and trial-to-trial retrieval [279].

In the context of the TREC CT track, patient-related information is written as free-text, whereas the document collection consists of a snapshot of *ClinicalTrials.gov* database.[1] Each clinical trial contains multiple fields, including two titles (brief and official one), condition, summary, detailed description, and eligibility criteria. The content of these fields can range from structured (e.g. gender and age of eligible patients) through semi-structured (e.g. eligibility criteria section) to unstructured (e.g. description and summary). The eligibility criteria field contains inclusion and exclusion criteria, a core aspect of the CT matching task. Trials were judged using a graded relevance scale of three points: 0 if the patient is not relevant to the CT, 1 if the patient is topically relevant but excluded based on the eligibility criteria, and 2 when the patient fulfils the eligibility criteria.

[1] https://clinicaltrials.gov

The TREC CTs track differs from previous medical TREC tracks in several aspects. TREC Precision Medicine 2017–2020 [214] is concerned with matching CTs to a patient summary consisting of only the patient's disease, relevant genetic variants, and basic demographic information. On the other hand, TREC CT topics consist of an unstructured patient summary. TREC Clinical Decision Support 2014–2016 [213] used topics written similarly (free-text patient descriptions), but the task was to match patients to PubMed publications, instead of CT documents. Moreover, none of the previous tracks used a graded relevance scale focused on eligibility.

Figure 2.1 provides an example of a patient's EHR description and of sections from a relevant CT. Using a bag-of-words approach to tackle the patient-to-trial matching problem may pose difficulties as both the patient's description and the CTs contain many irrelevant terms, thereby introducing noise. Moreover, both can contain negated key terms (for instance, the exclusion criteria), the handling of which is essential for deciding eligibility but may not be trivial even when using n-grams or neural network-based models [77, 245]. Additionally, the sections of queries and documents may have different importance because of their time dependency (i.e., past or present conditions) and because they can refer to either patients or patients' family medical history. One can try to overcome these issues by structuring both the query and documents, and extracting relevant items first.

As illustrated in Figure 2.1, both patient and clinical trial description share drug and disease keywords (e.g., salmeterol, fluticasone, asthma). Notably, these drug mentions also appear in the exclusion criteria of the clinical trial. Existing approaches that directly incorporate exclusion criteria, either for filtering purposes or to assign negative weights based on keyword co-occurrence, may erroneously exclude such patients. However, it is essential to note that these drug entities in exclusion criteria are considered relevant only in the context of allergic reactions. Therefore, a comprehensive understanding of the intricate semantics associated with this task is essential for effectively addressing this problem.

Previous work attempted to solve a CT matching task using various lexical and neural models. Leveling [148] annotated a corpus with terms from medical dictionaries and with negations for retrieving trials for the TREC Precision Medicine track. A large number of systems reported in the TREC CT used variants of the Okapi BM25 model [109] or the Divergence from Randomness (DFR) model [7] that has demonstrated potential in the biomedical information retrieval field.

2.1.2 Information extraction from clinical data

Information extraction from clinical data has been an active area of research in recent years. Previous work has focused on automatic extraction of eligibility criteria from clinical trial protocols. For instance, Dasgupta et al. [48] presented a method for identifying and segmenting eligibility criteria into five semantic categories, including demographic information, health status, treatment history, laboratory test reports, and lifestyle. The

Patient Description - #23

A 39-year-old man came to the clinic with cough and shortness of breath that was not relieved by his inhaler. He had these symptoms for 5 days during the past 2 weeks. He doubled his oral corticosteroids in the past week. He is a chef with a history of asthma for 3 years, suffering from frequent cough, wheezing, and shortness of breath and chest tightness. The symptoms become more bothersome within 1-2 hours of starting work every day and worsen throughout the work week. His symptoms improve within 1-2 hours outside the workplace. Spirometry was performed revealing a forced expiratory volume in the first second (FEV1) of 63% of the predicted. His past medical history is significant for seasonal allergic rhinitis in the summer. He doesn't smoke or use illicit drugs. His family history is significant for asthma in his father and sister. He currently uses inhaled corticosteroid (ICS) and fluticasone 500 mcg/salmeterol 50 mcg, one puff twice daily.

Clinical Trial - NCT03755544

Title: Salmeterol/Fluticasone Easyhaler in the Treatment of Asthma and COPD

Eligibility:

Main Inclusion Criteria:

• Male or female patients with asthma or COPD who have been using salmeterol/fluticasone propionate combination treatment for at least 3 monthsbefore the study

• Age ≥ 18 years

• Written informed consent obtained.

Main Exclusion Criteria:

• Pregnant or lactating female patients

• Participation in other clinical studies during the study.

• Known hypersensitivity (allergy) to salmeterol, fluticasone propionate or the excipient lactose.

Description:

A prospective, open-label, non-interventional, multicentre study in adult patients with asthma or COPD who are treated with Salme-terol/fluticasone Easyhaler. During the study the Salmeterol/fluticasone Easyhaler will be used according to the Summary of Product Characteristics. Clinical effectiveness of the treatment will be evaluated with change in asthma or COPD symptoms during 12 weeks treatment.

Legend: Disease | Drug | Negation | Negation relation | Current MC | Past MC | Family MH

Figure 2.1: An example of a clinical trial and a description of a patient eligible for this trial. Highlighted items are described in detail in Section 2.2. Example adapted from Pradeep et al. [203].

EliIE system [111] was proposed for converting free-text eligibility criteria for clinical research into a structured and formalised format using a 4-step process including entity and attribute recognition, negation detection, relation extraction, normalisation of concepts and output structuring.

Other studies aimed to extract information from patients' health records. The development and uptake of NLP methods for processing free-text clinical notes has been growing in recent years. A systematic review of the literature by Sheikhalishahi et al. [232] showed that there is a significant increase in the use of machine learning methods for NLP in clinical notes related to chronic diseases, and that deep learning is an emerging methodology. The ConText algorithm aims to determine whether conditions mentioned in clinical reports are negated, hypothetical, historical, or experienced by someone other than the patient [85]. The n2c2 n2c2/OHNL 2019 shared task [234] focused on extracting family history information from clinical notes. Garcelon et al. [71] utilised heuristics to detect medical history and negated terms in patients' records.

Despite these efforts, there has been a lack of approaches that integrate information extraction techniques to enhance both query and document representation. Specifically, there is a lack of methods that effectively combine the extracted terms to determine

eligibility ranking. This presents an opportunity for further exploration in the field.

2.1.3 Neural approaches for CT

Several approaches using Transformer-based architectures and pre-trained models, such as BERT [53], have achieved state-of-the-art effectiveness in some of the biomedical information processing applications. In CT retrieval, there have been multiple attempts to use BERT embeddings in both dual-encoder and cross-encoder retrieval setups with different pre-trained models such as BioBERT or ClinicalBERT [107, 220, 219]. These results correspond to implementations of methods applied to traditional ad-hoc retrieval tasks and have not outperformed multiple experiments under traditional retrieval models [216, 217]. On the other hand, Pradeep et al. [203] proposed a multi-stage neural ranking system for the CTs matching problem, relying on T5-based models, currently with state-of-the-art results in multiple retrieval tasks, including CT. According to the findings presented in TREC CT 2021 [216], large, T5-based models outperform smaller transformers models in CT retrieval.

Kusa et al. [136][2] proposed the TCRR (Topical and Criteria Re-Ranking) neural method. This method draws inspiration from curriculum learning principles, where concepts are progressively learned from simple to complex. The model is initially trained on identifying documents of topical relevance where both eligible and excluded documents are relevant. Subsequently, it targets the more challenging eligibility classification, distinguishing only eligible documents as relevant. This dual-objective training uses a pre-trained BERT language model as a cross-encoder, enhanced with a linear combination layer for fine-tuning through a pairwise loss function. The inference process mirrors this training by employing the first stage retrieval rank to perform a two-fold re-ranking of top trials, leveraging the distinct models stemming from the dual training objectives. Usage of this training method shows that BERT-based models can provide a viable alternative to T5-based models in clinical trial retrieval.

2.2 Methodology

This section describes the steps for processing CT documents and patient descriptions. We used these processed text as input to probabilistic lexical retrieval models. We conduct our experiments on the TREC Clinical Trials 2021-22 track collections.

2.2.1 Clinical trials processing

Clinical trials are *fielded documents* with the following fields: brief summary, brief title, identifier, detailed description, drug name, drug keywords, eligibility criteria, gender, general keywords, intervention type, maximum age, minimum age, official title, and

[2]As described in Chapter 1, this part of the work was developed by Óscar E. Mendoza from the University of Milano-Bicocca and is described in his PhD thesis "Adaptation of neural-enhanced retrieval model to domain-specific tasks".

primary outcome. Intervention type, gender, and primary outcome refer to controlled vocabularies; age-related fields are numeric. All other fields except clinical trial ID are textual.

We parse the content of a clinical trial document to split it into specific sections. The *eligibility criteria* section is crucial, as it outlines the specific requirements and conditions a patient must meet or the characteristics they must possess to participate in the trial. This can range from age, gender, and health status to previous treatments and current medications. Ensuring a patient meets these criteria is essential for the safety and integrity of the research and to demonstrate that the results are scientifically valid. Our CT processing is focused on making the eligibility criteria as fine-grained as possible so we can easily discriminate aspects referring to medical history and drugs. We start by using a parser based on heuristics to separate the eligibility criteria section of clinical trials into inclusion and exclusion criteria. Then, we extract single criterion items from the two criteria sections.

We further classify each item from inclusion and exclusion as concerning one of the three sections: '*current medical condition*', '*past medical condition*' and '*family medical history*'. Our motivation is that admission notes (which the topics simulate), consist of several sections that do not have equal impact on the patients' relevance to the trial and may even be irrelevant to their eligibility. Similarly, clinical trials can have different types of information stored in their eligibility section.

We then use a pre-trained entity extraction model together with an algorithm for determining negation to detect *affirmative* and *negative* drug and disease entities in both inclusion and exclusion sections. In the next step, we remove double negations coming from negated exclusion criteria. For every entity in the exclusion criteria, we swap their modifier (from affirmative to negative and vice versa). It allows us to create a single list of eligibility criteria keywords by concatenating extracted inclusion criteria keywords with exclusion criteria keywords for which we swapped the affirmative/negative modifiers. For instance, the exclusion criterion 'Patients who are *not smoking*' becomes the inclusion criterion 'Patients who are *smoking*'. This step may not always be correct; nevertheless, we found it to be a good approximation, allowing us to collapse these two sections into one.

The final result is a single list of extracted entities, classified by their section and modifier. All extracted keywords from a document D_i can be described by the set $K_{D_i} = \{A_i^{cmc}, A_i^{pmc}, A_i^{fmh}, N_i^{cmc}, N_i^{pmc}, N_i^{fmh}\}$, where A stands for a list with affirmative entities, N for negative entities, and cmc, pmc and fmh for current medical conditions, past medical conditions, and family medical history, respectively.

We can then enrich the CT document representation by expanding them with the extracted keywords. Traditionally, this would have been done by boosting the importance of extracted terms. However, in order to preserve the semantic information about each extracted entity (section and modifier information), we use prefixing to create special tokens describing each entity. Furthermore, as many of these entities are

multi-word expressions, we concatenate the tokens using the underscore character '_' to create a single token. Specifically, we create new tokens by adding them the prefixes '*cmc*', '*pmc*' and 'fmh' for each respective section and additionally '*no*' when an entity is negated (e.g. N_i^{pmc} = ['*myasthenia gravis*','*shortness of breath*'] becomes ['*pmc_no_myasthenia_gravis*', '*pmc_no_shortness_of_breath*']). We append the new tokens to the pre-processed document and use the enriched document to create an index for the lexical retrieval models. An example gold output for the CT document NCT03755544 from Figure 2.1 is: `['cmc_asthma', 'cmc_copd', 'cmc_salmeterol', 'cmc_fluticasone', 'pmh_salmeterol', 'pmh_fluticasone', 'pmh_propionate']`

2.2.2 Patient description processing

As we are essentially aiming to match the patient to the CT criteria, we follow a similar approach to enrich the input query. A patient's description is split into the sections of current medical conditions, past medical conditions, and family medical history. As for the trials, we run an entity and negation detection algorithm for each section. Extracted keywords for query Q_j can be represented as $K_{Q_j} = \{A_j^{cmc}, A_j^{pmc}, A_j^{fmh}, N_j^{cmc}, N_j^{pmc}, N_j^{fmh}\}$, where each element contains a list of extracted entities. We follow the same procedure as in Section 2.2.1 to create special tokens by adding the section and negation prefixes. Finally, the query for lexical models containing the original patient description is enriched by appending the extracted entities after tokenisation. An example gold output for Patient #23 from Figure 2.1 is: `['cmc_oral_corticosteroids', 'cmc_asthma', 'cmc_cough', 'cmc_no_smoke', 'cmc_no_illicit_drugs', 'cmc_fluticasone', 'cmc_inhaled_corticosteroids', 'cmc_salmeterol', 'pmc_seasonal_allergic_rhinitis', 'fmh_asthma']`

2.2.3 Ranking

In our experiments we consider the following sections (fields) of clinical trials as input when creating an index: brief title, official title, description, summary, conditions and criteria. For experiments in Section 2.4.2, we also append the enhanced eligibility criteria representation as an additional text.

2.2.4 Filtering

Following approaches from previous work [148, 129, 219], we employ filtering on the age and gender to eliminate trials for which patients would be excluded based on the demographic criteria. We parse the age and gender information from patient descriptions for all patients. In clinical trials, this corresponds to '*minimum_age*', '*maximum_age*' and '*gender*' fields. In this step, we remove the trials for which the patient is ineligible based on these two criteria.

Furthermore, we try rule-based parsing to extract information about smoking and alcohol consumption from both patients and clinical trials. Similarly to the demographic filters, we use this information to filter out ineligible patients.

2.3 Experiment Setup

This section details the datasets and baselines we have employed as well as the evaluation procedure. Additionally, we discuss the implementation of the methods described in the paper. The code is available under the following URL.[3]

2.3.1 Dataset

The corpus released by TREC contain 375,580 clinical trials. In 2021, 75 topics (patient notes) were used, and 50 more were created in 2022. There are 35,832 relevance judgements in 2021 and 35,394 in 2022. More details of the datasets can be found in Table 2.1. Clinical trial documents released by TREC are in XML format and consist of several sections. For our experiments, we use the sets of topics as they were provided.

Table 2.1: Statistics of TREC CT datasets from 2021 and 2022.

	2021	2022
Documents	375,580	375,580
Topics	75	50
Avg. topic length (tokens)	133.4	105.9
Avg. topic length (sentences)	11.2	9.4
Total judgements	35,832	35,394
– Eligible (2)	5,570	3,939
– Excluded (1)	6,019	3,036
– Not relevant (0)	24,243	28,419
Unique Trials judged	26,162	26,585

2.3.2 Evaluation

We follow the evaluation procedure from the TREC Clinical Trials track, which is the standard evaluation procedure for ad-hoc retrieval tasks. As the relevance assessment is given using graded relevance scale (eligible, excluded, or not relevant), the performance of the models is measured using normalised discounted cumulative gain (nDCG). We present results as reported by TREC, using nDCG@5 and nDCG@10, Precision at 10 (P@10), and Reciprocal Rank (RR).

We treat unjudged documents as non-relevant, ensuring that our results are not biased towards models that retrieve a large number of unjudged documents. Furthermore, we

[3] https://github.com/ProjectDossier/patient-trial-matching

focus on Precision as the primary metric for optimising retrieval models. Our goal is to identify trials for which patients are eligible, and Precision provides strict feedback to achieve this aim.

2.3.3 Implementation details

We use ScispaCy [180] and medspaCy [64] to implement our entity extraction experiments. We apply the spaCy NER model trained on the BC5CDR dataset[4] to detect disease and drug mentions.

We have decided to use the ConText algorithm [85] to determine whether extracted entities were negative or affirmative. While more recent algorithms are available for identifying assertions in clinical text [259], we opted for the ConText algorithm due to its established track record and availability inside the medspaCy library. Moreover, as our approach is focused on enriching not only 125 queries but also 375,000 clinical trial documents, an additional criterion for selecting the ConText model was its scalability.

Text is lowercased, and tokenised with the spaCy's `en_core_sci_lg` model[5]; punctuation and stopwords are removed. As a main lexical retrieval model, we use the BM25+ [253] "out-of-the-box", i.e. without parameter optimisation, implemented in the Rank-BM25[6] Python package. Furthermore, for the first two experiments, we also test two other lexical models, namely TF-IDF [242] and DFR model based on inverse document frequency with Bernoulli after-effect and H2 normalisation (In_expB2) [7], both implemented in the Terrier search engine[7]. Our Clinical Trials parsing script is available as a separate open-source package[8].

2.4 Results

We begin by assessing the effectiveness of using clinical trial sections. Subsequently, we examine the influence of extracted entities and filtering techniques.

2.4.1 Clinical trials sections

We first evaluate the utility of different sections of CTs. For this experiment, we use the raw, unmodified version of the clinical trial and patient description text. We successfully extracted the 'inclusion' and 'exclusion' sections from the eligibility criteria for 91% of the clinical trials. For the remaining 9% of trials where extraction did not work, we attributed the empty text to both the 'inclusion' and 'exclusion' sections. We create

[4] https://s3-us-west-2.amazonaws.com/ai2-s2-scispacy/releases/v0.5.3/en_ner_bc5cdr_md-0.5.3.tar.gz

[5] https://s3-us-west-2.amazonaws.com/ai2-s2-scispacy/releases/v0.5.3/en_core_sci_lg-0.5.3.tar.gz

[6] https://github.com/dorianbrown/rank_bm25

[7] http://terrier.org

[8] https://github.com/WojciechKusa/clinical-trials

Table 2.2: Impact of CTs' sections on the performance of the BM25+ retrieval model. The first group contains results using only a single section as a document representation, and the second group represents results using several concatenated sections. The 'criteria' section represents the original text from the *eligibility criteria* section of a clinical trial. The 'inclusion' and 'exclusion' sections are derived from the 'criteria' section using heuristic methods. Underlined values indicate highest score within the group, **bold** values indicate highest score overall. The identifier of each run is in the first column.

		TREC CT 2021				TREC CT 2022			
#	Input sections	nDCG@5	nDCG@10	P@10	RR	nDCG@5	nDCG@10	P@10	RR
1.	brief title	.218	.205	.131	.298	.216	.189	.128	.297
2.	official title	.245	.215	.137	.293	.237	.205	.140	.370
3.	description (*desc.*)	.354	.317	.195	.408	.324	.277	.168	.381
4.	summary (*sum.*)	.332	.315	.192	.376	.346	.305	.220	.480
5.	conditions (*cond.*)	.168	.164	.109	.245	.165	.155	.102	.224
6.	inclusion	.405	.391	.252	.478	.373	.337	.230	.459
7.	exclusion	.120	.117	.048	.114	.169	.137	.068	.173
8.	criteria	.397	.367	.199	.411	.363	.338	.216	.437
9.	brief title + official title (*tit.*)	.270	.256	.172	.322	.261	.220	.150	.369
10.	sum. + criteria + tit.	.470	.445	.255	.467	.450	.427	.292	**.542**
11.	desc. + criteria + tit.	.490	.448	.259	.470	.426	.394	.258	.446
12.	sum. + desc. + tit.	.402	.386	.243	.443	.414	.381	.272	.491
13.	sum. + desc. + tit. + cond. (*all*)	.400	.380	.228	.437	.407	.379	.272	.473
14.	all + inclusion	**.508**	.462	**.276**	**.505**	.464	**.437**	**.312**	.520
15.	all + exclusion	.398	.367	.203	.395	.386	.363	.238	.451
16.	all + criteria	.491	**.464**	.272	.492	**.465**	.426	.290	.506

several indexes and retrieval models with different combinations of sections as input features. The results for the BM25+ model are presented in Table 2.2. The first eight rows represent results when only one CT section was used to create an index, whereas the remaining rows present runs conducted on the concatenations of selected sections. Meanwhile, the results for In_expB2 and TF-IDF retrieval models are presented in Table 2.3.

Among single section runs, the usage of the *inclusion* field alone (run 6) yields the highest Precision@10 and nDCG@5 scores for the BM25+ model, both for 2021 and 2022 data. Moreover, for 2021 topics, the *inclusion* section also achieves the highest nDCG@10 and RR from all single input sections (runs 1–8). Run 6 is also on par with run 13, that concatenates all sections except eligibility criteria.

Notably, for the 2022 collection, for all single-field runs, the *summary* field consistently achieves the highest on-average Reciprocal Rank (RR) across the three evaluated retrieval models. This distinction can be attributed to the nature of RR, which is the multiplicative inverse of the rank of the first relevant trial. The *summary* field, often offering a broader overview of CT, might match more queries due to its generic terms, thus potentially positioning the first relevant document higher. However, metrics like Precision@k and nDCG@k emphasise the presence of multiple relevant items within the top k results. Hence, sections with specific details, such as the *inclusion* criteria, might find more eligible trials for a patient, explaining the observed variances across different evaluation

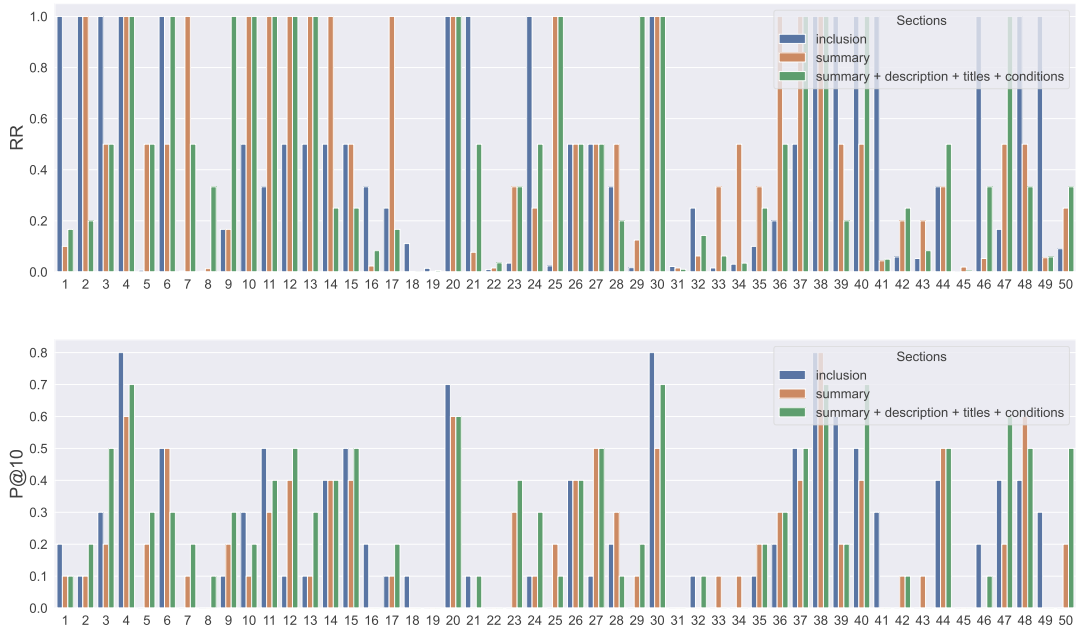


Figure 2.2: Topic-by-topic Reciprocal Rank (top) and P@10 (bottom) scores comparison for a BM25+ model with different document representations for TREC CT 2022 data.

metrics. This observation is further supported by a topic-by-topic comparison of RR and P@10 for the BM25+ model, as illustrated in Figure 2.2. Nevertheless, there are patients for which the *inclusion* section run outperforms the *summary* in terms of RR.

Concatenating more sections to create an index improves the on-average nDCG scores. However, this does not always hold for the metrics that consider the distinction between eligible and ineligible (P@10 and RR).

The *exclusion* section achieves the worst results from all single section runs (run 7), even when compared to runs using only the title of a clinical trial. Moreover, simply adding the text from the *exclusion* section for the bag-of-words approaches decreases the retrieval performance when compared to using the *inclusion* section only (run 16 versus 14). These outcomes motivate our subsequent experiments and document enrichment techniques described in Section 2.2.1, where we try to structure the knowledge contained in the eligibility section to take advantage of the available data.

The results for In_expB2 and TF-IDF retrieval models follow a similar trend, with the differences for 2022 data even higher than for the BM25+ model. The only noticeable difference is the nDCG scores for the TF-IDF model being higher when using all sections with eligibility criteria compared to using the *inclusion* section (run 16 versus 14). This outcome shows that our findings can be generalised to other lexical models.

Table 2.3: Impact of CT sections on the performance of In_expB2 and TF-IDF retrieval models. For each model, the first group contains results using only a single section as a document representation, and the second group represents results using several concatenated sections. The 'criteria' section represents the original text from the *eligibility criteria* section of a clinical trial. The 'inclusion' and 'exclusion' sections are derived from the 'criteria' section using heuristic methods. Underlined values indicate highest score within the group, **bold** values indicate highest score overall for each model. The identifier of each run is in the first column.

		TREC CT 2021				TREC CT 2022			
#	Input sections	nDCG@5	nDCG@10	P@10	RR	nDCG@5	nDCG@10	P@10	RR
	In_expB2								
1.	brief title	.174	.167	.112	.221	.222	.193	.134	.296
2.	official title	.194	.179	.109	.252	.245	.209	.146	.380
3.	description (*desc.*)	.354	.327	.200	.458	.341	.296	.186	.380
4.	summary (*sum.*)	.299	.288	.172	.313	.373	.322	.222	.511
5.	conditions (*cond.*)	.119	.119	.081	.172	.141	.135	.094	.196
6.	inclusion	.398	.370	.225	.445	.389	.345	.238	.485
7.	exclusion	.147	.131	.051	.134	.138	.133	.070	.142
8.	criteria	.386	.360	.192	.409	.347	.322	.224	.409
9.	brief title + official title (*tit.*)	.252	.235	.149	.325	.300	.248	.162	.423
10.	sum. + criteria + tit.	.454	.426	.227	.467	.474	.435	.286	**.574**
11.	desc. + criteria + tit.	.462	.437	.248	.419	.437	.417	.292	.441
12.	sum. + desc. + tit.	.441	.405	.252	.517	.455	.415	.282	.488
13.	sum. + desc. + tit. + cond. (*all*)	.440	.411	.252	.533	.463	.420	.282	.493
14.	all + inclusion	**.518**	**.482**	**.281**	**.553**	**.506**	**.480**	**.346**	.539
15.	all + exclusion	.395	.365	.203	.377	.425	.388	.254	.473
16.	all + criteria	.480	.455	.267	.441	.490	.449	.312	.508
	TF-IDF								
1.	brief title	.196	.172	.107	.253	.221	.193	.130	.305
2.	official title	.203	.181	.109	.256	.238	.200	.138	.353
3.	description (*desc.*)	.313	.280	.160	.396	.309	.272	.162	.387
4.	summary (*sum.*)	.281	.263	.147	.327	.336	.288	.196	.496
5.	conditions (*cond.*)	.124	.127	.087	.180	.152	.144	.094	.201
6.	inclusion	.411	.377	.229	.466	.383	.333	.232	.444
7.	exclusion	.145	.132	.053	.146	.139	.129	.072	.125
8.	criteria	.383	.364	.199	.421	.338	.316	.220	.405
9.	brief title + official title (*tit.*)	.235	.213	.129	.300	.276	.223	.146	.397
10.	sum. + criteria + tit.	.444	.411	.214	.436	.416	.389	.260	.497
11.	desc. + criteria + tit.	.458	.429	.232	.435	.403	.385	.264	.438
12.	sum. + desc. + tit.	.364	.335	.195	.429	.392	.354	.236	.480
13.	sum. + desc. + tit. + cond. (*all*)	.362	.332	.184	.435	.405	.358	.234	**.505**
14.	all + inclusion	.478	.446	**.260**	**.481**	.430	.406	**.282**	.474
15.	all + exclusion	.380	.345	.183	.381	.380	.342	.222	.454
16.	all + criteria	**.482**	**.450**	.248	.454	**.437**	**.407**	.274	.478

2.4.2 Impact of extracted entities

To determine the impact of the extracted entities, we selected the optimal configuration of input sections from the previous step, which used the summary, description, titles, conditions, and inclusion criteria (run 14). We use these sections as a base document representation and enriched it with different combinations of extracted entities: *c* – only current medical conditions, *cf* – current and family medical history, *cp* – current and

Table 2.4: Experimental results for runs with index and query expanded with extracted entities. Letters describe usage of extracted affirmative and negative medical entities for (c) current conditions, (p) past conditions, and (f) family history. **Bold** values indicate highest score overall. The identifier of each run is in the first column.

		TREC CT 2021				TREC CT 2022			
#	Model	nDCG@5	nDCG@10	P@10	RR	nDCG@5	nDCG@10	P@10	RR
14.	all + inclusion	.508	.462	.276	.505	.464	.437	.312	.520
16.	all + criteria	.491	.464	.272	.492	.465	.426	.290	.506
14a.	+ c	.524	.480	.292	.542	.500	.459	.328	**.528**
14b.	+ cf	.524	**.481**	**.293**	.542	.500	.460	**.330**	**.528**
14c.	+ cp	.532	.478	.287	**.555**	.501	.460	.328	.521
14d.	+ cfp	**.532**	.480	.288	**.555**	**.502**	**.460**	.328	.521

past medical conditions, *cfp* – current, family and past medical conditions.

The results for the BM25+ model are presented in Table 2.4. Using extracted items from patients positively impacts the final score. The highest Precision scores are achieved with extracted affirmative and negated entities for the current and family medical history. The low impact of past medical condition can be explained by an infrequent occurrence of this data in patient descriptions in the TREC dataset and the quality of the ConText algorithm. Extracted entities contribute more positively to the measures where judgements distinguish between eligible and ineligible patients. The best-performing model (14d) comprises all available extracted data (affirmative and negative entities for current, past and family medical history) to enrich the index. This tells us that our proposed method can potentially improve the retrieval with complex negated sentences. However, the relative performance gain is low, and a detailed analysis is needed to understand how it can be further improved.

Topic #48 of TREC CT 2021 and entities extracted from it using our approach are presented in Table 2.5. The table shows drug and disease mentions with information if these mentions are negated. These entities are classified into current, past and family medical conditions. Upon examination, our entity extraction and section classification models may produce false negatives. For instance, they failed to recognise *'MI'* (Myocardial Infarction) as a disease in the family history or *'fungal hyphae'* as a current condition. In our exploration of other topics and clinical trial documents, we also encountered false positives, where entities were extracted erroneously and did not correlate with drug or disease mentions. These inaccuracies affect the final retrieval result. By further fine-tuning our models on domain-specific data, we anticipate an enhancement in retrieval quality.

Results for In_expB2 and TF-IDF retrieval models are presented in Table 2.6. The In_expB2 model on TREC CT 2021 data is the only one for which our query and document enrichment techniques do not improve results. We hypothesise that this is the case as the starting model (run 14) was already a very strong model compared to other

Table 2.5: Example entities extracted for Topic #48 from TREC CT 2021.

Topic #48: Fernandez is a 41 year man who is a professional soccer player. He came to the clinic with itchy foot. Physical exam revealed localized scaling and maceration between the third and fourth of his right toe. It became inflamed and sore, with mild fissuring. The dorsum and sole of the foot was unaffected. There is no pus or tearing in the affected area. He didn't use ant topical ointment on the lesion and has no positive history for any underlying disease such as DM. He smokes 15 cigarettes per day and drinks a beer per day. His family history is positive for hyperlipidemia in her mother and MI in her father. He is in relation with several partners and use condom during the intercourse. His physical exam and lab studies were normal otherwise. Tinea pedis infection confirmed as his diagnosis by the observation of segmented fungal hyphae during a microscopic KOH wet mount examination.

Section	Entity	Is negated
Current MC	itchy	—
	sore	—
	fissuring	—
	tearing	✓
	Tinea pedis infection	—
	KOH	—
Past MC	DM	✓
Family MH	hyperlipidemia	—

baselines. For the TF-IDF model, we can observe that the enrichment with current and past medical entities yields the best results both for 2021 and 2022 data.

2.4.3 Effectiveness of filtering

Next, we test several filtering methods as described in Section 2.2.4. As a base run, we take our best configuration from the previous experiment: BM25+ run enriching data with current medical conditions and medical history of the patient and family (run 14d). Results for TREC CT 2021 are presented in Table 2.7.

Our filtering results align with other researchers' results, confirming that utilising age and gender fields can improve the quality of the final matching. The usage of both filters (run e) removes, on average, 26.3% trials from the top 1000 retrieved documents for all topics of the 2021 collection, improving the P@10 score by 4.9 percentage points over the unfiltered run. Out of these two fields, the contribution of the age filter has more impact and is significantly better than the base run.

On the other hand, smoking and alcohol related-filtering does not help to improve the results further (runs f and g). We grouped these filters together as our algorithm did not identify any smoker, and only nine drinking patients in the TREC CT 2021 topics. Despite only these few mentions, we observe deterioration of the results.

Figure 2.3 presents a topic-by-topic analysis of the results in terms of the number of relevant trials ranked in top 20 using lexical models for the three best runs from each

Table 2.6: Experimental results for runs with index and query expanded with extracted entities for In_expB2 and TF-IDF retrieval models. Letters describe usage of extracted affirmative and negative medical entities for (c) current conditions, (p) past conditions, and (f) family history. **Bold** values indicate highest score overall for each model. The identifier of each run is in the first column.

		TREC CT 2021				TREC CT 2022			
#	Model	nDCG@5	nDCG@10	P@10	RR	nDCG@5	nDCG@10	P@10	RR
	In_expB2								
14.	all + inclusion	**.518**	**.482**	**.281**	**.553**	.506	.480	.346	.539
16.	all + criteria	.480	.455	.267	.441	.490	.449	.312	.508
14a.	+ c	.499	.457	.272	.483	.515	**.484**	.340	.555
14b.	+ cf	.491	.455	.272	.479	**.524**	.482	**.342**	.554
14c.	+ cp	.494	.461	.267	.494	.521	.479	.336	**.559**
14d.	+ cfp	.492	.457	.267	.490	.521	.475	.332	.547
	TF-IDF								
14.	all + inclusion	.478	.446	.260	.481	.430	.406	.282	.474
16.	all + criteria	.482	.450	.248	.454	.437	.407	.274	.478
14a.	+ c	**.484**	.452	.259	.463	**.496**	**.439**	.302	**.536**
14b.	+ cf	.481	.446	.259	.457	.493	.437	.302	.528
14c.	+ cp	.483	**.459**	**.261**	**.515**	.475	**.439**	**.306**	.511
14d.	+ cfp	.477	.453	.259	.508	.477	**.439**	.304	.517

Table 2.7: Filtering results on TREC CT 2021 data. Letters describe the used filters: (A) Age, (G) Gender, (S) Smoking and (D) Drinking. **Bold** values indicate highest score overall. Superscripts denote significant differences in paired Student's t-test with $p \leq 0.05$. The identifier of each run is in the first column.

#	Model	nDCG@5	nDCG@10	P@10	RR	% filtered trials
a	14.	0.508	0.462	0.276	0.505	—
b	14d.	0.532	0.480	0.288	0.555	—
c	14d. + A	0.554^{af}	0.509^{abdf}	0.335^{abdf}	0.603^{abdf}	23.4%
d	14d. + G	0.537^{f}	0.483^{f}	0.288	0.556^{b}	5.7%
e	14d. + AG	$\mathbf{0.561}^{abdf}$	$\mathbf{0.513}^{abcdf}$	$\mathbf{0.337}^{abdf}$	$\mathbf{0.604}^{abdf}$	26.3%
f	14d. + SD	0.526	0.475	0.284	0.546	0.7%
g	14d. + AGSD	0.555^{af}	0.509^{abdf}	0.335^{abdf}	0.595^{abdf}	26.7%

experiment. We can observe an incremental gain both from extracted entities and filtering.

2.4.4 Finding eligible trials

Figure 2.4 presents two plots with an averaged per patient count of relevant and excluded trials depending on a cutoff point for TREC 2022 collection. We compare our result to

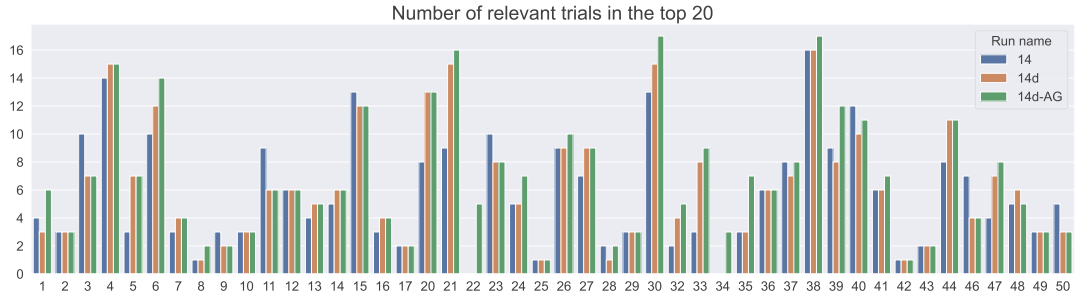


Figure 2.3: Topic-by-topic number of relevant trials in the top 20 for the three best BM25+ runs from each experiment: 14 – baseline, 14d – further query and index enriched with extracted entities, and 14d+AG – further filtered for age and gender.

the TCRR neural re-ranking model [136][9]. Both applied techniques, namely extracting drug and disease entities and filtering by age and gender, impacts finding more eligible trials. However, only the run with filtering is able to retrieve consistently fewer ineligible trials than the baseline run. We can also see that, on average, our best run (14d-AG), retrieves twice as many trials for which a patient is eligible than ineligible.

Upon comparison with a neural TCRR method, it can be observed that our approach yields less favourable outcomes for retrieving eligible trials (although similar in terms of excluded trials as shown in the lower part of Figure 2.4). However, it is important to note that methods like TCRR and other neural re-rankers need a solid first-stage retrieval model to work well. Our query and document enrichment approach builds a strong foundation for first-stage retrieval models, which helps improve the performance of TCRR. Moreover, when directly compared to TCRR, our method offers lower latency, an important factor to consider.

There are several limitations of this study, both related to the dataset and the models. Usage of the TREC CT collection implies that the patient descriptions are relatively short, i.e., EHR admission note-style documents. We acknowledge that our approaches could have problems handling longer sequences.

Furthermore, the topics are written only in English. This does not concern clinical trials, for which the *ClinicalTrials.gov* database is the leading international source. Nevertheless, multilingual medical retrieval may present challenges for both lexical and neural models, as the nuances and complexities of medical terminology can vary significantly across languages. Addressing these limitations and developing strategies for multilingual medical retrieval is an essential area for future research.

[9]In this paper, the TCRR model was developed by Óscar E. Mendoza from the University of Milano-Bicocca and is described in his PhD thesis "Adaptation of neural-enhanced retrieval model to domain-specific tasks".

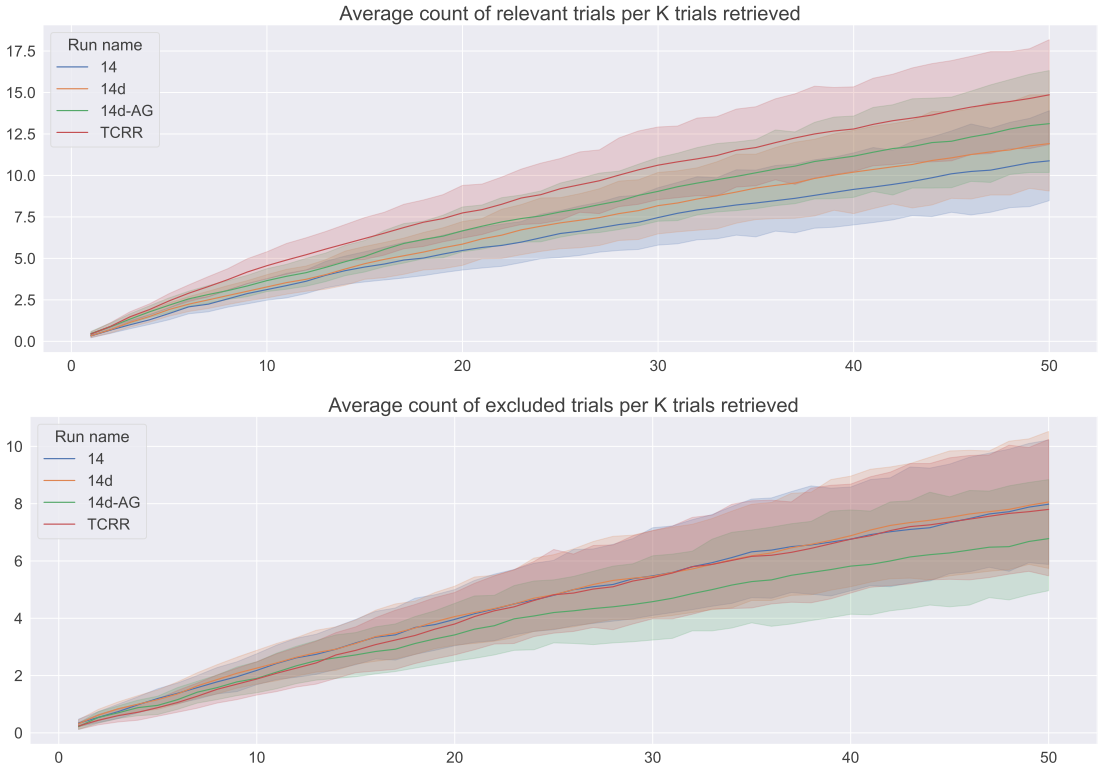


Figure 2.4: Averaged per patient count of relevant (top) and excluded (bottom) trials depending on a cut-off of K trials retrieved (x-axis) for TREC CT 2022 collection.

2.5 Summary

This chapter presents an approach for clinical trial retrieval under the patient-to-trial paradigm. We investigate the impact of individual clinical trial sections showing that the 'inclusion' section alone contributes the most to the final retrieval score. Moreover, we evaluate the handling of complex eligibility criteria for matching patients to clinical trials by combining input from information extraction modules into a lexical retrieval model. The extracted drug and disease entities and their negations positively impact the retrieval of eligible trials. Filtering based on gender and age proved to be successful in eliminating ineligible trials. Even though our proposed system involves many single components, it showcases an alternative approach to the clinical trial matching problem, emphasising the importance of eligibility criteria. In future work, we plan to measure the impact of extracted entities on neural re-ranking models.

CHAPTER 3

Background and Literature Review

In the remaining part of the thesis, we focus on the task of citation screening for systematic literature reviews. This chapter introduces the related work and state-of-the-art for this task. Specifically, in Section 3.2, we present the process of systematic literature reviews in detail. In Section 3.3, we overview the task of citation screening, presenting current automation approaches. Next, in Sections 3.4 and 3.5, we focus on available datasets and evaluation approaches for citation screening, respectively. In Section 3.6, we outline tools available to researchers for reviewing the literature, primarily in the context of systematic reviews.

We base our literature review about the automation of systematic literature reviews and citation screening on three recent surveys of this topic:

- Systematic review conducted by O'Mara-Eves et al. [190] in 2015;
- Update to the review above, completed by Norman [182] in 2020;
- Systematic review conducted by van Dinter et al. [261] in 2021.

We extend these reviews to cover published research until May 2023.

3.1 Notation

In Table 3.1, we present the definitions, standard evaluation measures, and notation used throughout the thesis. Items are ordered by importance, starting with the most general terms. While concepts related to systematic literature reviews adhere to standard definitions, some notation connected to evaluation is specific to this thesis.

Table 3.1: Definitions of concepts and terms used in the thesis.

Concept	Definition
Systematic Literature Review, SLR	A research method that involves reviewing existing literature in a structured and comprehensive way to answer a specific research question.
Review Protocol	A detailed plan that describes the rationale, hypothesis, and planned methods of the systrmatic literature review.
Cochrane	An international network dedicated to creating and disseminating SLRs on the effects of healthcare interventions. The Cochrane Library is a key resource for systematic reviews in health care.
Publication	Any piece of academic or scholarly work, such as articles, reports, or papers, that is disseminated to a wider audience.
Study	An organised and detailed piece of research conducted with the objective of answering a specific question or testing a particular hypothesis. It involves the systematic gathering and interpretation of data to derive meaningful insights on a targeted subject.
Eligibility Criteria	Guidelines or standards predefined in the review protocol, used to determine which studies or publications should be included or excluded from a systematic review. These criteria ensure that only relevant and appropriate studies are considered, thereby maintaining the review's validity and reliability.
Includes	Refers to the set of publications included in an SLR based on the eligibility criteria.
Excludes	Refers to the set of publications that don't meet the set eligibility criteria and are omitted from the SLR.
Meta-Analysis	A statistical technique that combines the findings from independent studies to synthesise evidence on a particular topic or research question. It provides a quantitative estimate of the overall effect of a particular intervention or treatment based on pooled data from multiple studies.
SLR Outcome	A comprehensive conclusion or set of findings that arise from analysing the literature reviewed within a systematic literature review. For instance, in a systematic review exploring the efficacy of a certain drug on reducing blood pressure, the SLR outcome might state: "After analyzing 50 studies, the drug was found to reduce systolic blood pressure by an average of 10 mmHg more than the placebo."
Forest Plots	Graphical representation of the estimated results from individual studies in a meta-analysis, along with the overall results.
Group Size	The total count of individuals enrolled in the experimental or control group.

Table 3.1: Definitions of concepts and terms used in the thesis.

Number of Events	The count of specific occurrences or outcomes observed within the study groups, such as number of individuals exhibiting particular symptoms or adverse events.
Weight of a Study	The importance or influence given to a study when calculating an average or combined effect in a meta-analysis.
Effect Size	Measure of the strength of the relationship between two variables in a statistical population (e.g., risk ratio or standardised mean difference) in events between the experimental and control group.
RevMan	A Cochrane tool available for conducting meta-analysis and managing SLR.
CLEF TAR	The "Conference and Labs of the Evaluation Forum – Technology Assisted Reviews". CLEF is an initiative that focuses on the evaluation of information access systems, and the TAR track specifically dealt with technologies aimed at assisting systematic reviews.
Run	A specific execution of a machine learning system, using a particular model, parameters, and data. In the case of citation screening, runs can be retrieval (publications returned for a query), classification (publications relevant to the SLR topic) or ranking (publications sorted by their relevance to SLR topic).

Measure	**Definition**
Recall	Also known as sensitivity, measures the proportion of relevant items retrieved compared to the total number of relevant items.
TNR	True Negative Rate, also known as specificity, a traditional measure for evaluating citation screening.
$nDCG$	Normalised Discounted Cumulative Gain, a measure of ranking quality.
WSS	Work Saved over Sampling, a traditional measure for evaluating citation screening.

Notation	**Definition**
$\mathcal{I}$	set of relevant documents that should be included in the review, *includes*
$\mathcal{E}$	set of irrelevant documents that should be excluded in the review, *excludes*
$\lvert\mathcal{I}\rvert$	number of *includes*
$\lvert\mathcal{E}\rvert$	number of *excludes*
N	total number of documents $\lvert\mathcal{I}\rvert + \lvert\mathcal{E}\rvert$
TP	number of *true positives*, i.e., includes classified correctly
TN	number of *true negatives*, i.e., excludes classified correctly
FP	number of *false positives*, i.e., excludes classified incorrectly

Table 3.1: Definitions of concepts and terms used in the thesis.

FN	number of *false negatives*, i.e., includes classified incorrectly
$r\%$	a recall value of $r\%$
$n_{r\%}$	rank of a document for which the recall level of $r\%$ is achieved
$X@r\%$	evaluation measure X calculated at a fixed recall value of $r\%$
$X@d$	evaluation measure X calculated at a fixed cut-off of d documents
$X@d\%$	evaluation measure X calculated at a fixed cut-off of $d\%{\cdot}N$ documents
O_o	Original outcome of the review.
O_p	Predicted outcome based on evaluated system.
CI_{lower}	Lower bound of the confidence interval of the original outcome.
CI_{upper}	Upper bound of the confidence interval of the original outcome.
MoD	Magnitude of difference in the outcome effect size: $\frac{\|O_o - O_p\|}{\|O_o\|}$.
Δ_{CI}	Distance between the predicted outcome and the closest bound of the CI: $\min(\|O_p - CI_{lower}\|, \|O_p - CI_{upper}\|)$.
$\mathcal{I}_{P_i}$	Publication *Influence*, it quantifies the extent to which a publication affects the outcomes of an SLR, as defined by the formula: $\sum_j^J \left(\sum_k^K \frac{MoD_{j,k}}{n_{ps_k}} \right)$.

3.2 Systematic Literature Reviews

What is perceived to be the first analytical approach to aggregate the outcomes of several clinical studies was published in 1904 by Karl Pearson [196]. Synthesis of research findings, systematically and critically appraised, emerged in the 1970s under the term 'meta analysis' [240]. The Cochrane Collaboration[1] was established in 1993 and has created the ground for evidence-based medicine. Now, Cochrane is an international network of researchers, academics and practitioners dedicated to the principles of managing healthcare knowledge in such a manner that ensures their high quality, availability and completeness [94]. There are more than 220,000 records published between 2000 and 2022 tagged as SLRs in PubMed[2]. Under the assumption that the number of publications was constant throughout the years, there was, on average, 10,000 SLRs published per year. PROSPERO, international prospective register of systematic reviews, had 15,667 new registrations in 2018 alone, and this data comes from the times before the COVID-19 pandemic [182].

3.2.1 How is a systematic literature review conducted?

Systematic literature reviews consist of multiple steps which do not necessarily follow the linear order. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) is the standard methodology followed by most authors for conducting SLRs in

[1] https://www.cochrane.org
[2] https://pubmed.ncbi.nlm.nih.gov/

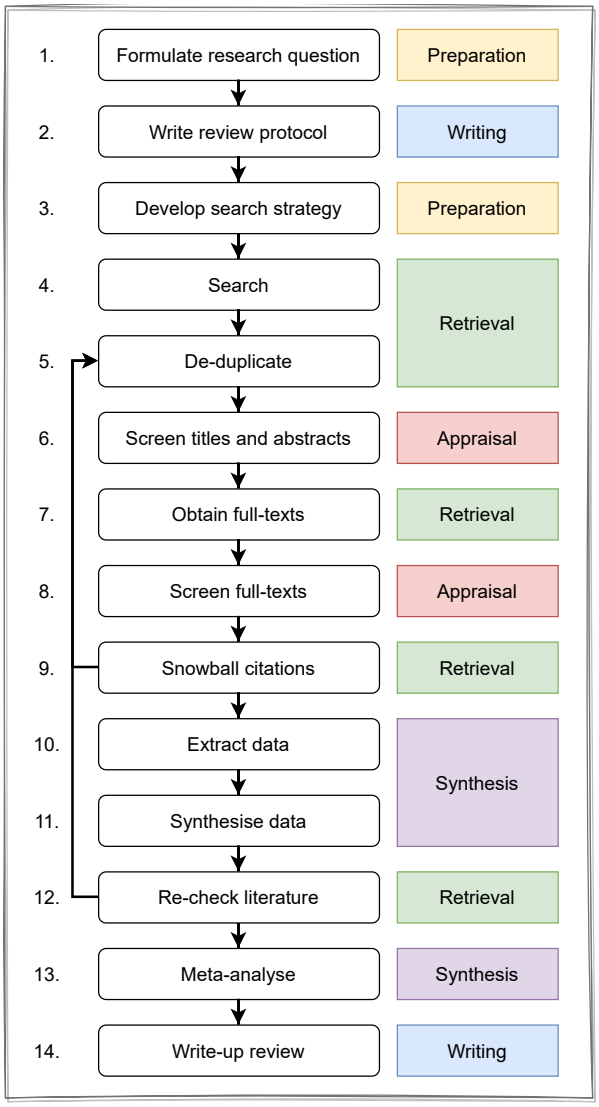


Figure 3.1: Fourteen steps of the systematic literature review process clustered into five high-level categories. Steps and process description according to the Cochrane [94], figure adapted based on Tsafnat et al. [254].

medicine [174]. Depending on the granularity, previous studies enumerated between four and up to 15 tasks that might be included in the SLR process [254]. A representation of the ordered fourteen SLR steps is presented in Figure 3.1. On a high level, the task can be categorised into five different phases: Preparation, Writing, Retrieval, Appraisal, and Synthesis.

Preparation: The preparation stage is foundational, laying the groundwork for the systematic review. The first step, formulating the research question, involves delineating

the scope and objectives of the review. This is followed by the development of a search strategy, which sets out the plan for identifying relevant literature sources. It includes selecting keywords, databases, and other sources to be searched.

Writing: In the writing phase, the review protocol is created. This protocol outlines the methods and processes that will be followed throughout the systematic review. It ensures transparency and replicability of the review process. Later in the process, after data extraction and synthesis, the results and insights are consolidated and presented in a comprehensive review.

Retrieval: Retrieval encompasses a set of steps focused on obtaining the literature. The first step is searching databases and other sources based on the pre-defined search strategy. Once the initial pool of articles is obtained, de-duplication is done to remove any redundant articles. After the appraisal stages, there is often a need to obtain more literature, either by obtaining full texts not available initially or by 'snowballing' citations–looking at the references and citations of obtained articles to find more relevant articles. Re-checking the literature ensures that no recent or crucial articles have been missed.

Appraisal: Once articles have been retrieved, they must be appraised for relevance and quality. This process begins by screening titles and abstracts to filter out unrelated or low-quality articles. For those articles that pass this initial screening, full texts are obtained and screened. This deeper dive allows reviewers to ascertain the relevance and rigour of each article, ensuring that only the most pertinent and trustworthy sources inform the review's findings. In the medical SLRs, the standard is to have two reviewers appraise and agree on the decision on each paper.

Synthesis: The synthesis phase is where insights start to emerge. Data from the included articles is extracted, capturing the vital information required to answer the research question. This data is then synthesised, often combining quantitative or qualitative findings from different articles to derive more comprehensive insights. If the data allows, a meta-analysis might be performed in some cases. This statistical technique combines results from multiple studies to provide a more robust estimate of an effect or phenomenon.

The SLR process, while complex, provides a high level of clarity and depth, making it invaluable in informing research, policy, and practice across various fields. Among the most time-consuming steps of the SLR process, the study selection and data extraction repeatedly come first [186, 34, 11, 231]. Surprisingly, Haddaway and Westgate [82] report that the most significant proportion of time was needed for administration and planning, which are beyond the reach of current automation techniques. Automated citation screening methods could also help in updating the timeliness of reviews as it would allow for the creation of "living reviews" (Section 3.2.3).

One crucial framework often used within the SLRs of clinical research is the *PICO framework*. PICO stands for Patient/Population, Intervention, Comparison, and Outcome. This framework facilitates the structuring of clinical questions and subsequently aids in the systematic search of literature [94]:

- Patient/Population – Defines the patient or population group in question. It is crucial for pinpointing the exact demographic or condition under investigation.
- Intervention – Describes the treatment, exposure, or management strategy that is being considered for the patient or population group.
- Comparison – Refers to the alternative against which the intervention is compared. It can be a placebo, another treatment strategy, or no intervention at all.
- Outcome – Enumerates the effects or endpoints that are being measured. This could range from physiological measures to patient satisfaction.

The PICO framework provides a clear method to translate clinical questions into components that can be used to search databases in an organised and efficient manner. However, it is important to clarify that the PICO framework is particularly applicable and predominantly used within Randomised Clinical Trials (RCTs). Within the SLR process, the PICO framework is used for instance in the retrieval phase to help formulate search strategies. Moreover, when specifying inclusion and exclusion criteria for studies in the review, the PICO criteria offer a systematic approach to deciding which studies are relevant to the posed research question.

3.2.2 Challenges in conducting systematic literature reviews

Conducting systematic literature reviews comes with several challenges that can affect the process and results [83].

Bias and objectivity concerns in selection SLRs aim to provide unbiased insights, but inadvertent biases in study selection, weighting, or emphasis can skew the review's results [10, 247]. Ensuring consistent criteria and transparency throughout the selection process is vital to maintaining the review's credibility and integrity.

Database limitations and coverage Relying on databases for literature sourcing presents a challenge: no single database captures every relevant publication [192, 160]. If a database omits essential journals or articles, it can inadvertently introduce gaps in the SLR. Hence, multi-database searches and alternative sourcing methods are vital to ensuring a review's comprehensiveness.

Heterogeneity in study designs Researchers working on SLRs must navigate through many sources when dealing with different study designs, from RCTs to observational studies [156]. The inherent diversity complicates the synthesis process, necessitating robust strategies to compare and consolidate results uniformly.

Language and translation barriers The global nature of research means that crucial studies might be published in languages unfamiliar to the reviewers [179]. Overlooking or misinterpreting these due to translation inaccuracies can skew SLR results. As such, investing in accurate translation services and multilingual review teams is crucial to maintaining an SLR's comprehensiveness and reliability.

Redundancy in review efforts Systematic literature reviews (SLRs) often involve repetitive tasks, such as screening similar articles or using overlapping search terms, leading to inefficiencies in the review process [103, 10]. By developing better tools and protocols, there exists a significant potential for streamlining the SLR process, enhancing productivity, and effectively reducing redundant efforts.

Escalating workload The exponential growth in the volume of published literature means reviewers face the daunting task of comprehensively reviewing and synthesising vast amounts of information [182, 16]. As this workload increases, it is essential to adapt methodologies, leverage technology, and develop new tools to cope with the surge, ensuring reviews remain thorough and reliable.

Extended duration of systematic review completion By their nature, SLRs are thorough and meticulous, demanding significant time and effort to ensure every piece of relevant literature is considered [94, 143, 26]. This extended duration can be challenging for researchers and stakeholders awaiting results. Hence, there is a pressing need for innovative strategies that can expedite the review process without sacrificing its rigour and quality. Moreover, certain research domains evolve at a very fast pace, making some SLR findings obsolete soon after they're published [236, 19]. In such dynamic fields of science, periodic reviews, rapid updates, and agile methodologies become essential to keep the literature review current and relevant.

3.2.3 Living reviews

The ever-evolving landscape of knowledge requires a more agile approach to literature reviews. As described in the previous section, traditional systematic reviews, while comprehensive and rigorous, can quickly become outdated as new studies and findings emerge. To address this limitation, there has been a growing interest in the concept of living literature reviews in the medical community [61, 282]. Unlike their static counterparts, living reviews are dynamic and iterative, ensuring that they are always up-to-date and in tune with the latest evidence. This continuous evolution is instrumental in providing clinicians, researchers, and policymakers with the most relevant and current insights.

The process of conducting a living review involves consistent monitoring of the research landscape for new evidence that can potentially impact the review's conclusions. The living review is promptly updated to incorporate these findings when significant new research is identified. These constant updates ensure that the review remains a reliable

source of information, reducing the time lag between the emergence of new evidence and its assimilation into clinical or policy recommendations. Living reviews represent a paradigm shift in evidence synthesis, ensuring that medical decisions are informed by the most recent and robust evidence available. To ensure timely updates, living review creators often leverage advanced tools and technologies to automate and scale up the search and appraisal of relevant literature [164].

3.2.4 Systematic literature reviews outside medicine

The process applied in systematic literature reviews in medicine was also transferred into other scientific domains, such as environmental sciences [21], software engineering [119], social sciences [201] and engineering [31]. The Campbell Collaboration[3], a sibling organisation to Cochrane, emerged in 1999 and adapted Cochrane methodology to produce systematic reviews on broader public policy issues. Searching the Web of Science for the query: "systematic literature review" shows clear dominance in medical sciences, followed by business economics, psychology, computer science, and engineering sciences with a total of almost 130,000 publications [31]. However, outside the medical domain, there exist fewer procedures, standards and guidelines. Moreover, there are fewer dedicated tools that help researchers. The need for procedures, guidelines and tools offering some automation is and will be manifested in these other disciplines.

3.2.5 Exploratory literature reviews

Finally, academic literature reviews (usually more exploratory than "systematic") are also conducted in the academic setting by myriads of PhD and Masters students [76]. This process enables students to familiarise themselves with the current state of the art, theory and methods in their field. They can also identify gaps that could be addressed by their research. Overlapping reviews are very often repeated by different groups as there is no data sharing and exchange format that could enable reusing past reviews as it is in systematic reviews in medicine [239, 103]. Guidelines and methodologies also aim to improve this process but do not mention any automated approaches, and the search process itself is not very structured but, instead, exploratory [202]. Students conduct their literature review searches using multiple tools [241]. However, this area is still underexplored compared to the systematic literature reviews. Further studies are needed to assess the adoption of various tools for conducting exploratory literature reviews by early-career academics. There is substantial potential in developing standards and tools that academics could adopt for the purpose of literature reviews.

3.3 Citation Screening

All the documents retrieved from the search step constitute the input to the citation screening (CS) step. In a manual screening scenario, annotators (also called reviewers

[3] https://www.campbellcollaboration.org

or screeners) need to screen (appraise) all these documents to select only the fraction relevant to the systematic review study, which should be included in the final review (also known as *includes*). The remaining documents are irrelevant to the review topic and should be excluded (also known as *excludes*).

Previous publications estimated that reviewers can screen from 0.13 to 2.88 abstracts per minute (20 seconds to 7.7 minutes per publication), while screening one full text article takes them from 4.3 to 5 minutes [186, 233]. Traditionally, for SLRs in medicine, every publication needs to be appraised by at least two reviewers. Because the total number of retrieved documents can go into tens of thousands, it is essential to find a way of accelerating this process [190].

3.3.1 Task formulation

We start by introducing the task of citation screening for SLRs and presenting the notations used for its formulation. An SLR is characterised by various attributes, including the title, abstract, research question $\mathcal{RQ}$, and eligibility criteria $\mathcal{C}$. We refer to all these attributes collectively as the SLR protocol. Eligibility criteria comprise a set of rules and conditions that a document must meet for inclusion in the SLR. Given a pool of all documents denoted as $\mathcal{D}$, the main goal of automated citation screening is to assist researchers in identifying relevant publications for inclusion in an SLR. Each document $d \in \mathcal{D}$ has attributes such as its title, abstract, main content, authors, and publication year.

Document retrieval The initial step involves document retrieval, which aims to generate a set of potentially relevant documents $\mathcal{D}' \subseteq \mathcal{D}$ based on the research question $\mathcal{RQ}$. This step commonly involves querying bibliographic databases with specific keywords and Boolean expressions. We can formulate this step as a retrieval function r, such that $r(\mathcal{RQ}, \mathcal{C}) = \mathcal{D}'$.

However, the retrieved set $\mathcal{D}'$ may contain a large number of false positives (irrelevant documents). Therefore, the following citation screening step is needed to exclude irrelevant documents from the SLR. The task of citation screening for SLRs can be formally defined as follows:

Definition 3.3.1 (**Citation screening**).
Given a set of documents $\mathcal{D}'$ and a set of eligibility criteria $\mathcal{C}$, the task of citation screening for SLR is to systematically determine for each document $d \in \mathcal{D}'$ whether it satisfies the criteria $\mathcal{C}$. This decision can be represented as a binary label $y_d \in \{0, 1\}$, where $y_d = 1$ if document d satisfies the criteria $\mathcal{C}$, and $y_d = 0$ otherwise.

It is important to emphasise that strict adherence to traditional Boolean search methods as the initial step is not obligatory. Theoretically, the entire literature collection $\mathcal{D}$, could serve as the initial "pool" from which screening techniques are then applied to

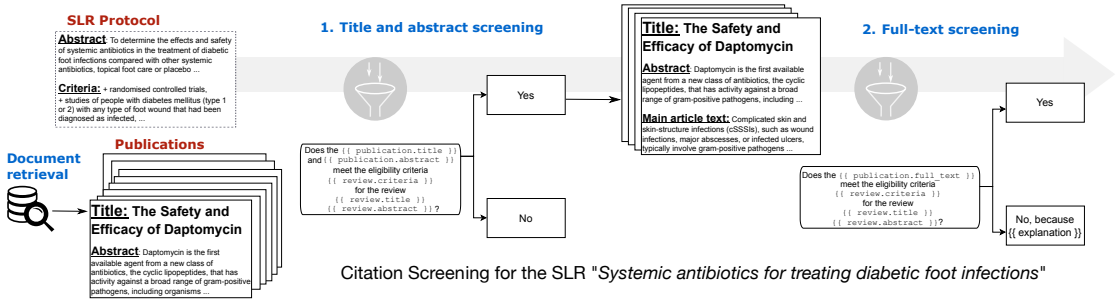


Figure 3.2: Illustration of the citation screening process, separated into two tasks (1) title and abstract screening and (2) full text screening. Tasks are represented as a specific example of question-answering when a single question asks for a fulfilment of all eligibility criteria $\mathcal{C}$ at once.

isolate documents meeting specific criteria $\mathcal{C}$. This perspective suggests a more flexible and potentially more inclusive approach to literature review, leveraging the strengths of modern computational methods. Nevertheless, this method raises another concerns regarding trust and transparency [160].

The manual citation screening is conducted in two steps, differing in which attributes of documents are considered (as shown in Figure 3.2): (1) title and abstract screening and (2) full text screening. In the first step, the relevance of each document is evaluated based on its title and abstract. In the second step, a more thorough assessment is performed by examining the full text of the documents included in the previous step.

Binary classification and document ranking are central screening task formulations for screening methods. Both directly assess documents against eligibility criteria in a straightforward, binary fashion. Binary classification functions as the foundational method, determining the relevance of each document by mapping its features to a binary outcome. Document ranking extends this assessment by ordering the documents in a manner that reflects their likelihood of meeting the set criteria, hence prioritising the review process. In contrast, question answering and natural language inference represent alternative formulations. Both offer a more fine-grained analysis of documents, often revealing insights not immediately apparent through binary classification or ranking methods.

Binary classification When citation screening is treated as a binary classification problem, each document $d \in \mathcal{D}'$ is assigned a binary label $y_d \in \{0, 1\}$ to indicate its relevance ($y_d = 1$) or irrelevance ($y_d = 0$) to the SLR per the criteria $\mathcal{C}$.

We denote the feature representation of document d as $\mathbf{x}_d \in \mathbb{R}^n$, where n is the dimensionality of the feature space. The goal is to learn a classifier $f(\mathbf{x}_d; \boldsymbol{\theta})$ parameterised by $\boldsymbol{\theta}$ that maps the feature representation to the predicted relevance score:

$$y_d \approx f(\mathbf{x}_d; \boldsymbol{\theta}) \tag{3.1}$$

The classifier can also leverage decisions made on previously screened papers. This historical decision data acts as an additional source of information for the classifier. It can either modify the feature representation $\mathbf{x}_d$ or directly influence the classification function f via parameter updates to θ.

In this case, let $\mathcal{H}$ represent the set of previously screened documents with their associated decisions. The classifier may incorporate these decisions into its learning process, possibly leading to a refined parameter set $\boldsymbol{\theta'}$ and an updated classification function $f'(\mathbf{x}_d, \mathcal{H}; \boldsymbol{\theta'})$.

$$y_d \approx f'(\mathbf{x}_d, \mathcal{H}; \boldsymbol{\theta'}) \tag{3.2}$$

Document ranking In the context of citation screening, document ranking involves ordering documents by their likelihood of meeting the eligibility criteria $\mathcal{C}$. The ranking function ρ, parameterised by $\boldsymbol{\phi}$, assigns a score to each document $d \in \mathcal{D}'$ such that documents more likely to satisfy $\mathcal{C}$ are positioned higher in the ranking:

$$\forall d, d' \in \mathcal{D}', \mathcal{C}(d) > \mathcal{C}(d') \implies \rho(d; \boldsymbol{\phi}) > \rho(d'; \boldsymbol{\phi}) \tag{3.3}$$

The document ranking approach can be converted to a binary classification task using a threshold τ in the ranking scores. Documents with a ranking score above τ are classified as relevant ($y_d = 1$), and those below as irrelevant ($y_d = 0$). For a given document d, if $\rho(d; \boldsymbol{\phi}) > \tau$, then y_d is predicted to be 1, else 0. The threshold τ can be determined based on the specific requirements of the SLR or through empirical methods such as maximising a particular evaluation metric.

Question answering An alternative formulation of the citation screening task is to frame it as a question-answering (QA) problem. In this approach, we transform the eligibility criteria $\mathcal{C}$ into a set of questions $\mathcal{Q} = \{q_1, \cdots, q_{|C|}\}$, where each question q_k corresponds to a specific criterion in $\mathcal{C}$.

For each document $d \in \mathcal{D}'$, we obtain a set of predicted answers $\hat{A}^d = \{\hat{a}_k^d | \text{meets}(q_k, \hat{a}_k^d)\}$, where $\text{meets}(q_k, \hat{a}_k^d)$ denotes that the document d should meet the criterion expressed by the question q_k. The final relevance label $\hat{y}_d$ of a document d can be determined by aggregating the predicted answers $\hat{A}^d$ using a logical combination function, such as the logical AND operation. This question-answering formulation offers a more fine-grained assessment of a document's relevance concerning various aspects of the eligibility criteria $\mathcal{C}$.

Natural Language Inference The task of natural language inference (NLI) involves determining the logical relationship between a premise and a hypothesis. In the context of automated document screening, the premise represents the eligibility criteria $\mathcal{C}$, and the hypothesis corresponds to the content of a document d. The NLI model, denoted as

α, is trained to predict whether the hypothesis is true (entailment), false (contradiction), or undetermined (neutral) given the premise:

$$\hat{y}_d = \alpha(\mathcal{C}, \text{Content}(d); \boldsymbol{\xi}) \tag{3.4}$$

Where $\boldsymbol{\xi}$ are the parameters of the NLI model. The entailment score can then be used to infer the relevance of the document to the SLR. If the predicted result is entailment, then the document satisfies the eligibility criteria and should be included in the SLR. NLI explores the subtleties of logical relationships between document content and the SLR criteria, adding a layer of inferential judgment to the screening process.

These four formulations of the citation screening task–binary classification, document ranking, QA, and NLI–represent distinct but complementary approaches to the citation screening process. Each contributes to a comprehensive strategy for efficiently identifying relevant literature in an SLR. These methods can be integrated to leverage their strengths, using classification as an initial filter, question answering and NLI for detailed analysis, and ranking to organise the screening workflow.

3.3.2 Automated citation screening

Automated citation screening is an umbrella term for using natural language processing (NLP), machine learning (ML) and information retrieval (IR) techniques with the goal of decreasing the time spent on manual screening. According to a survey on the topic of automation of systematic literature reviews by van Dinter et al. [263], 25 out of 41 analysed primary studies published between 2006 and 2020 addressed (semi-)automation of the citation screening process. Another, older systematic review from 2014 found in total 44 studies dealing implicitly or explicitly with the problem of screening workload [190]. Both of these surveys highlight that citation screening was automated the most often in the past among other SLR steps. This can be attributed primarily to three factors: (1) the importance and cost involved when conducting this step manually, (2) the relatively low entrance level for researchers working on this topic as it can be represented as a binary document classification problem, and (3) the availability of SLR datasets.

Screening automation is a general term for various approaches aimed at reducing workload during the screening stage of systematic reviews [190]. These approaches can be divided into four different categories [182]:

1. *screening reduction* – classification or ranking algorithms to automatically exclude non-relevant publications;

2. *screening prioritisation* – ranking relevant records earlier in the process of screening;

3. *automation as a second screener* – classification or ranking methods to include relevant publications instead of a second annotator;

4. *visual text mining* – using NLP algorithms to present similarities allowing workers to locate relevant documents faster.

Screening reduction approaches traditionally train a supervised model on an annotated dataset sample to determine whether a paper should be included or excluded from the review. They have the downside of requiring gathering manual annotations before making any predictions for every new systematic review. In particular, many previous studies used half of the dataset as a training dataset [99, 124, 130] which might limit their applicability to larger systematic reviews.

Although innovative, the prioritisation approach in automated citation screening is not without limitations, especially in context of medical systematic literature reviews where comprehensiveness is essential. In a study by Tsou et al. [257], the prioritisation accuracy varied greatly across datasets and approaches tested. Furthermore, the prioritisation method, by design, does not reduce the total number of studies that need to be reviewed. It merely rearranges the order in which they are assessed. While this might streamline the initial phase of the review, enabling the reviewers to focus on later stages of the process, it does not lessen the overall workload. Reviewers still need to assess each study, making the time savings marginal at best. Furthermore, reliance on automated prioritisation might inadvertently lead to overconfidence in the initial results, potentially causing reviewers to give less attention to studies ranked lower, creating additional bias in the review process.

Previous (semi-)automation approaches ranged from statistical models like naïve Bayes classification [17, 166], support vector machine (SVM) [37, 165, 99, 36], voting perceptron [35] and random forest [122] to neural networks [124, 262, 130]. Martinez et al. [165] proposed a system which combines both prioritisation via ranking and filtering via classification. A significant limitation of all these approaches is the need for a large number of manual annotations that must be completed before developing a reliable model for every new systematic review [255]. Moreover, the majority of the classification models are evaluated only retrospectively which might raise questions of data leakage when considering large amounts of data used for pretraining language models.

Researchers looked at utilising external information in automated citation screening. Timsina et al. [249] used UMLS tokens in order to improve the classification quality. Tsafnat et al. [255] extracted four critical characteristics of observational studies (population, exposure, confounders and outcomes) and used them to filter studies showing significant performance gains over standard approaches. Brockmeier et al. [27] trained a named entity recognition model to extract PICO phrases which were used as additional features for a relevance classification model.

A question-answering approach has been explored to enhance the efficiency of systematic reviews, building upon the advancements in automated citation screening [295]. This approach, detailed by Zou and Kanoulas [294], introduces an interactive method for high-recall information retrieval. The methodology revolves around formulating direct

questions about specific entities or information within documents that are likely to be relevant but have not yet been identified. This approach aims to rapidly find the remaining crucial documents in a collection by engaging reviewers in a question-and-answer format. The system, utilising a Sequential Bayesian Search-based method, optimises the sequence of questions to maximise efficiency in document retrieval. This technique is particularly beneficial when identifying the last few relevant documents in a collection, which are often the most challenging to locate using traditional methods. Notably, this approach has been demonstrated to be effective even in scenarios where reviewers provide noisy or imperfect answers.

A big part of the research focusing on reducing the screening workload has investigated the use of active learning [35, 172, 269, 87]. Active learning approaches for document screening were also introduced into some commercial software [100]. However, in systematic literature reviews, active learning for screening is not as popular as in the legal domain, where it has received considerable attention, with continuous active learning being able to significantly reduce the burden of screening [44]. Despite the growing number of papers in this domain highlighting potential work savings with certain algorithms or systems, these savings are largely theoretical and evaluated a posteriori. Realizing these efficiencies in practice hinges on developing a method to effectively determine the appropriate time to stop the screening process.

Statistical stopping criteria have emerged as a vital component in automated screening for systematic reviews, addressing the challenge of determining when to cease screening while ensuring relevant studies are not missed [30, 151]. These criteria, based on statistical inference, estimate the probability of encountering additional relevant studies and suggest stopping when this probability falls below a predefined threshold. Callaghan and Müller-Hansen [30] introduced a statistical stopping criterion for active learning in document screening, where screening continues until the recall surpasses a set target, assessed through a hypothesis test using random sampling from unseen documents. The process involves estimating the likelihood of missing relevant documents based on the hypergeometric distribution and stops screening when this likelihood falls below a certain confidence level, ensuring both efficiency and reliability in the screening process. Stevenson and Bin-Hezam [243] have developed a novel stopping method for screening based on point processes, demonstrating its effectiveness in achieving high recall with minimal document screening and outperforming several alternative methods across various datasets. Implementing these criteria requires balancing the risk of missing relevant studies against the resources required for screening, making their integration into automated systems a nuanced yet critical aspect of the screening process. This approach coincides with active learning methods, offering an additional layer of efficiency by dynamically adjusting the screening process based on real-time data analysis [35, 172].

Using automation as a second reviewer in systematic reviews suggests that while single-reviewer screening can be efficient, it risks missing a significant number of studies. Waffenschmidt et al. [266] found that single screenings by experienced reviewers missed a median of 5% of studies, with varying impacts on meta-analysis findings. Similarly,

Gartlehner et al. [73] reported a 13% miss rate in single-reviewer screenings. They suggested that while single-reviewer screening may not meet the high standards expected in systematic reviews, it could be a viable option for rapid reviews where methodological rigour is balanced against the need for speed. These findings point towards the potential of automated systems to act as a second reviewer, aiming to enhance the speed and reduce the likelihood of overlooking relevant studies.

A different strategy for automating systematic reviews was presented during the CLEF eHealth Lab Technology Assisted Reviews (CLEF TAR) in Empirical Medicine task [112, 113, 114] running between 2017 and 2019. The organisers curated a benchmark collection of 129 systematic literature reviews with citations, eligibility decisions and review protocols. Specifically, in the CLEF TAR 2019 edition, the challenge was to find all relevant documents from a set of PubMed articles given a Boolean query. It overcomes the need to create an annotated dataset but makes it harder to incorporate reviewers' feedback. Participants experimented with IR and NLP approaches during this task, using active learning and relevance feedback.

The introduction of the Transformer architecture [264] was the giant leap forward in deep learning for NLP. The BERT model [53] and its variants, which are based on the Transformer architecture, have pushed state of the art for many NLP tasks. It involves massive pre-training of a language model on unsupervised corpora and a detailed supervised fine-tuning on a usually much smaller corpora of a downstream task. Pretrained Transformer models have already replaced other architectures in multiple fields, including text classification, information retrieval, and ranking [290]. So far, deep neural network-based models have not managed to consistently outperform other approaches for citation screening in medical systematic reviews [130]. This might be primarily due to the very high class imbalance and lack of positive examples to train the model. Ioannidis [102] used BERT-based models to work on document screening within the CLEF TAR task achieving better results than the traditional IR models. Yang et al. [289] showed an approach to successfully training the BERT model on a legal e-discovery dataset, and this approach has a potential to also be applicable to systematic literature reviews. To our knowledge, these were the first usages of a generative neural network models in a document screening task.

3.3.3 Fixing the search

The search for documents is an essential step preceding the screening. It involves developing a search strategy (the search query) using a research question and eligibility criteria. This query is then issued to one or more search engines that index published literature (e.g. PubMed[4] or Embase[5]). Search strategies are commonly represented as Boolean queries and are developed by information specialists in an iterative process [225]. These queries are large and complex as they need to cover multiple synonyms, acronyms

[4] https://pubmed.ncbi.nlm.nih.gov
[5] https://embase.com

and spelling variations peculiar to the medical domain. The key motivation for using Boolean queries instead of best-match retrieval systems is the need for reproducibility enforced with a deterministic result set and static collection statistics.

As manual query formulation might take several weeks or even months [211, 117], another strategy for improving the process focuses on the automated formulation of a search query. This could fix many flaws coming from the currently used search process. It could provide reviewers with seed studies or information specialists with an initial query to begin the formulation process [227]. For instance, Karimi et al. [118] investigated the early provision of the quality of search strategies by returning ranked results during the strategy formulation process.

There are two main approaches to query formulation: conceptual formulation [33] and objective formulation [89]. Objectively derived Boolean queries compared to the conceptual approach have been found to typically yield retrieval effectiveness results of higher sensitivity [90]. However, the objective approach can only be applied to meta-reviews and not to typical systematic review query formulation.

The general approach consists of composing query logic using a systematic review protocol and seed studies. Several previous studies proposed this approach using techniques like extracting PICO terms and linking to medical taxonomies (like MeSH[6] and UMLS[7]), followed by postprocessing techniques to increase the precision [227, 228]. AutoFormulate is a tool that, based on seed studies, can generate a search string [153]. Other researchers proposed visual text mining techniques to support building the search query [170].

A comparison of automatic Boolean query formulation techniques found that they are still only somewhat effective compared to the original, manually formulated queries [228]. Automatic query formulation approaches generate outcomes that are less "natural" for doctors and reviewers. Therefore the queries might be harder to understand, and the adoption might be even slower than it is for citation classification approaches.

3.3.4 Technology-assisted reviews

High-recall search tasks or High-Recall Information Retrieval (HRIR) are terms describing the objective of locating all or nearly all relevant documents in a collection. Technology-Assisted Review (TAR), an approach arising from HRIR, combines information retrieval and machine learning to refine the process of reviewing extensive volumes of documents. TAR systems aim to support human reviewers by automating repetitive tasks and highlighting the most relevant documents for review, thus helping organisations save time and resources.

Citation screening for systematic literature reviews can be seen as an example of TAR task [190]. Other examples of high-recall search tasks are legal electronic discovery [288],

[6]The Medical Subject Headings (MeSH) ontology is an hierarchically organised index of biomedical concepts.

[7]The Unified Medical Language System (UMLS) is an integration of a number of key medical and biomedical terminologies, including MeSH.

construction of evaluation collections [144], and responses to the freedom of information laws requests [168]. Workshops such as LegalAIIA [41] and ALTARS [56] have popularised HRIR applications among the research community. The TREC Legal [46, 251, 12, 188], TREC Total Recall [78], and the CLEF eHealth Lab Technology-Assisted Review [112, 113, 114] have provided researchers with access to datasets and standardised evaluation methods.

One of the most critical aspects of HRIR systems is recall, which measures the proportion of relevant documents that are retrieved by the system. One of the key goals of TAR systems is to detect as many relevant documents (True Positives, TPs) as possible while excluding as many irrelevant documents (True Negatives, TNs) as possible. By reducing the number of TNs, TAR systems can save time and resources for human reviewers. However, caution must be exercised in implementing TAR systems as poor performance could result in legal sanctions, personal liability, and economic costs, as demonstrated in legal discovery scenarios [58].

3.4 Citation Screening Datasets

In this section, we present the review of available citation screening datasets and discuss the limitations of these datasets. We then outline the approach to standardising biomedical datasets using the BigBio framework. Finally, we summarise datasets for other SLR steps.

To find relevant citation screening datasets, we searched Google Scholar[8] and Semantic Scholar[9] for publications introducing new datasets for the citation screening task. We then searched for the forward citations of the original publication to find usages of the datasets. From our list we excluded private datasets used in only one publication. We further excluded datasets with only one SLR. As a results, we found 12 datasets fulfilling the criteria. Table 3.2 presents a summary of these datasets.

A dataset created by Cohen et al. [35] containing 15 SLRs is the first and, to the best of our knowledge, up until today, the most commonly used to evaluate the effectiveness of machine learning models for the CS task. They constructed a test collection of 15 different systematic reviews produced by the Oregon Evidence-based Practice Centre (EPC) related to the efficacy of medications in several drug classes. Wallace et al. [269] released another datasets consisting of three systematic reviews related to the clinical outcomes of various treatments. Both these datasets contain SLRs with a small number of citations (varying from 310 to 4751). Howard et al. [99] introduced new collection of five substantially larger reviews (from 4479 to 48 638 citations) that have been used to assess the performance of the SWIFT-review tool. These datasets were created using broader search strategies which justifies a higher number of citations.

[8] https://www.scholar.google.com
[9] https://www.semanticscholar.org

Table 3.2: A comparison of publicly available benchmark datasets used in the experiments on automated citation screening for systematic literature reviews, sorted by the publication year. We included all publicly available datasets and private datasets which were used in more than one publication. The "Avg. size" refers to the average number of citations contained in each review within the dataset. The "Avg. ratio of included" indicates the average percentage of those citations that were included in the final review. The "Additional data" column describes if the review contains metadata other than coming from the citation list.

	Introduced in	#reviews	Domain	Avg. size	Avg. ratio of included	Additional data	Publicly available
1.	Cohen et al. (2006) [35]	15	Drug	1,249	7.7%	—	✓
2.	Wallace et al. (2010) [269]	3	Clinical	3,456	7.9%	—	✓
3.	Miwa et al. (2014) [172]	4	Social science	8,933	6.4%	—	—
4.	Howard et al. (2016) [99]	5	Mixed	19,271	4.6%	—	✓
5.	Scells et al. (2017) [225]	93	Clinical	1,159	1.2%	Search queries	✓
6.	CLEF TAR 2017 [112]	50	DTA	5,339	4.4%	Review protocol	✓
7.	CLEF TAR 2018 [113]	30	DTA	7,283	4.7%	Review protocol	✓
8.	CLEF TAR 2019 [114]	49	Mixed	2,659	8.9%	Review protocol	✓
9.	Alharbi et al. (2019) [5]	25	Clinical	4,402	0.4%	Review updates	✓
10.	Parmar (2021) [195]	6	Biomedical	3,019	21.6%	—	—
11.	Hannousse et al. (2022) [84]	7	Computer science	340	11.7%	Review protocol	✓
12.	Wang et al. (2022) [275]	40	Clinical	1,326	—	Review protocol and seed studies	✓

Starting in 2016, a new dataset (*test collection*) was released almost every year. All these 12 datasets differ in the total number of reviews, subdomain, average review size, and percentage of included studies. However, the overall tendency shows a very high-class imbalance towards the negative class (i.e., irrelevant publications). Datasets introduced by Miwa et al. [172] and Parmar [195] are not publicly available, yet they were used in three and two research papers, respectively, so we included them in our comparison.

Until 2017 all of the datasets contained only the citation list with eligibility decisions [182]. More recently, datasets started to include titles of SLRs and search queries used for finding publications. Additional metadata is limited to search queries [225], review protocols (three datasets released as a part of the CLEF TAR shared-task by Kanoulas

et al. [112, 113, 114]), review updates [5] and seed studies [275].

So far, there was a little attention to review automation outside of the medical domain. The only available datasets are four social science reviews by Miwa et al. [172], and seven computer science reviews by Hannousse and Yahiouche [84]. Compared to the general interest and rate of production of SLRs in other domains, this overall underrepresentation of benchmark datasets could be improved. We also found a dataset containing one large SLR of environmental policies [98], which has different scope and format than other screening datasets.

Overall, we found, that in the citation screening domain, there is a lack of a standardised benchmark on which other researchers evaluate their approaches. Many previous works compare their models only on their (often private) dataset without showing the performance gain over previous work. Papers from the ML and NLP domains, very often evaluate their approaches on datasets introduced by Cohen et al. [35]. This dataset is, at the moment of writing, 17 years old and contains reviews characterised by a short generic title (e.g. 'ADHD' or 'Statins') and a list of PubMed IDs with their binary eligibility decisions. On the other hand, IR focused papers present their evaluation on CLEF TAR task datasets.

We have also evaluated a comprehensive catalogue of medical artificial intelligence datasets and benchmarks by Blagec et al. [22], only three citation screening datasets are mentioned: Cohen et al. [35], Wallace et al. [268], and Miwa et al. [172]. Of these three datasets, only two are publicly available. Additionally, another five private SLRs used in only one publication [237] are mentioned.

3.4.1 Limitations of existing datasets

Through our review, we identified twelve CS datasets reported in research papers, of which ten have been publicly released. During this analysis, we identified several datasets' shortcomings; some are also prevalent in other machine learning problems. Below, we summarise our findings, highlighting the key issues.

Poor documentation One major concern with previous datasets is the lack of documentation. None of the datasets we examined implement a datasheet [75] or data card [207], which are essential tools for ensuring transparency and reproducibility. Additionally, seven datasets do not provide clear licenses or terms of use. Inconsistencies were also found for one of the datasets [225], in terms of the number of the available content: the paper states 93 SLRs, but we found a list of 176 reviews on the corresponding GitHub repository.

Limited applicability Previous datasets are often small and lack crucial metadata like SLR research question or eligibility criteria, limiting their use to only evaluation of classification tasks. Older datasets typically provide only the title of the review, which limits their applicability for the comprehensive evaluation of neural language

understanding models. The most widely used dataset to date [35] was released in 2006. As ML and NLP techniques continue to advance rapidly, it is crucial to have up-to-date datasets that reflect the complexities and nuances of the current research landscape.

The datasets also do not contain the information about why a particular paper was excluded from the review. Without this data, the automated citation screening problem cannot be tackled in any other way than a binary decision. This is not the case in real life, as a typical SLR contains at least several exclusion and inclusion criteria, and the decision about every paper can be presented as a multi-dimensional relevance problem.

Lack of canonical splits Another significant challenge of previous datasets is the absence of canonical train-test splits. Depending on the field of research, practices may vary. As discussed before, In the ML and NLP domains, the prevailing practice is to use inter-review splits, where each review is treated as an individual dataset, and a set of citations is selected for training and testing. Conversely, IR publications often report intra-review splits, treating each review as a "topic" or query, and averaging the results across multiple queries.

In this sense, only the three TAR[10] datasets contain pre-defined canonical splits, yet, only at the intra-review level. For three other datasets [35, 269, 99], previous works have demonstrated significant variability in model evaluation based on the selection of cross-validation splits, particularly for the smallest datasets that contain a limited number of relevant documents [262, 130]. The lack of standardised splits makes it challenging to compare different approaches and hinders the fair evaluation of models' performance.

Dataset overlap We also evaluated the overlap between datasets at the level of entire systematic reviews This analysis aimed to understand the potential duplication of information and data leakage across different datasets.

Table 3.3 presents the extent of overlap observed between the train and test splits of the datasets. We discovered that at least 11 SLRs were present in multiple collections [225, 112, 113, 114]. SLRs released as part of the SIGIR 2017 collection [225] are also present among the test splits in CLEF TAR 2017 and 2019 collections. The TAR 2019 collection is most severely affected, with 3 SLRs present both in its training and test splits, accounting for approximately 6% of the test partition [114]. While this overlap is not a significant concern when evaluating unsupervised methods like BM25 [218], it poses a potential threat to conducting fair comparisons with large language models (LLMs). Machine learning models, and especially LLMs, have the capability to memorise their training data, making it critical to address dataset overlap to ensure unbiased evaluations [96].

It is worth noting that we did not explicitly report the overlap between different CLEF TAR datasets [112, 113, 114]. The creators of the dataset have already acknowledged

[10]TAR stands for Technology-Assisted Reviews and was a shared task organised at CLEF between 2017 and 2019 by Kanoulas et al. [112, 113, 114].

Table 3.3: List of overlapping Cochrane systematic literature reviews between datasets.

Cochrane review ID	First collection	Other collections
CD011145	sigir2017 (train)	tar2017 (test)
CD010633	sigir2017 (train)	tar2017 (test), tar2018 (train), tar2019 (train)
CD010653	sigir2017 (train)	tar2017 (test), tar2018 (train), tar2019 (train)
CD010542	sigir2017 (train)	tar2017 (test), tar2018 (train), tar2019 (train)
CD009185	sigir2017 (train)	tar2017 (test), tar2018 (train), tar2019 (train)
CD008081	sigir2017 (train)	tar2017 (test), tar2018 (train), tar2019 (train)
CD002143	sigir2017 (train)	sigir2017 (train)
CD001261	sigir2017 (train)	tar2019 (test)
CD011571	tar2019 (train)	tar2019 (test)
CD012164	tar2019 (train)	tar2019 (test)
CD011686	tar2019 (train)	tar2019 (test)

that each new edition of the dataset includes SLRs from the previous editions as part of the training data.

Finally, as some other datasets did not share metadata about the considered reviews (except for the very high-level title of the systematic review like ADHD or COPD), we were not able to map them to specific systematic reviews. An alternative method would involve checking the overlap between the included and excluded documents using for instance their Pubmed IDs.

Lack of common evaluation Another notable deficiency among the previous datasets is the absence of a common set of evaluation measures. For example, the most widely used dataset by Cohen et al. [35] was evaluated using several disparate evaluation measures such as WSS [35], AUC or $Precision@r\%$. However, recent research has exposed limitations and problems with both WSS and AUC as metrics for this task [134]. Only the three TAR datasets provide scripts for evaluating submissions. We delve deeper into the problem of evalaution measures in Section 3.5 of this chapter and in Chapters 5 and 6.

Availability in biomedical benchmarks Recent efforts have focused on creating larger collections of more diverse datasets to evaluate the performance of biomedical NLP models. These efforts include benchmarks like BLUE [198], HunFlair [280], BLURB [79], and BigBio [69], which provide datasets and tasks for evaluating biomedical language understanding and reasoning. Additionally, there are biomedical datasets geared towards prompt-based learning and evaluation of few and zero-shot classification, such as Super-NaturalInstructions [278] and BoX [194]. Out of all benchmarks mentioned above, only BoX contains one CS dataset covering five SLRs, however, this dataset is private. Coverage for other SLR tasks is also limited.

To summarise, previous datasets exhibit certain drawbacks that limit their suitability for comprehensive and standardised evaluation. While the TAR 2017-19 collections stand out as the only ones containing canonical splits and a set of evaluation measures, some of their topics overlap with another dataset [225], and we also identified data leakage in the newest TAR 2019 dataset. Consequently, we believe that developing a new collection is necessary to address these issues and establish a robust foundation for SLR automation evaluation. Moreover, as the topic of systematic review automation has direct commercial applications, it also would be beneficial to the broader research community to have a resource for objective verification of commercial products.

3.4.2 Standardising biomedical datasets

The creation of benchmarks such as ImageNet [52], SQuAD [209] or GLUE [272] were one of the critical components of growing success in the machine and deep learning in many domains. It should also be noted that while the progress made on these benchmarks is unquestionable, all benchmarks and datasets are only proxies for real-world tasks and can exhibit significant biases [256].

Furthermore, a recent trend in deep learning shifted the attention from model-centric machine learning (proposing novel architectures) to data-centric (improving training datasets) [68]. It is inspired by the observation that the performance gains provided by using better training data and commodified model architectures are higher than gains from new architectures. This shift requires a curation of appropriate datasets. For instance, Lee et al. [147] showed that data deduplication leads to more accurate and more robust models.

Unfortunately, implementing these successes in specialised areas such as biomedicine faces substantial obstacles, partly due to the current *dataset debt* in biomedical NLP. The review of available biomedical datasets showed that only 13% of them are available via programmatic interface [68]. There are currently no zero-shot evaluation frameworks for biomedical data similar to BIG-Bench,[11] which currently contains little-to-no biomedical tasks.

The standardisation of datasets will be a critical aspect in the realm of biomedical natural language processing. BigBio, a comprehensive framework, has been instrumental in addressing the challenges inherent in biomedical NLP data [69]. BigBio offers programmatic access to over 120 biomedical NLP datasets, covering a wide range of tasks and languages. This accessibility facilitates the creation of meta-datasets, which are crucial for training and evaluating language models. The framework supports reproducibility by allowing consistent and standardised dataset access, emphasising data-centric machine learning principles. This approach aligns with the trend towards improving training datasets rather than focusing solely on novel model architectures.

A key feature of BigBio is its dual-view dataset loaders: the ‘source’ view preserves the original dataset format, while the ‘BigBio’ view harmonises the dataset into standardised

[11] https://github.com/google/BIG-bench

schemas. Harmonisation in BigBio involves adapting datasets to a set of lightweight, task-specific schemas to standardise biomedical datasets, catering to various NLP tasks. The knowledge base (KB) schema is versatile, encompassing entity-based tasks like named entity recognition and relation extraction. The question-answering (QA) schema supports various QA formats, including multiple-choice and factoid questions. The textual entailment (TE) schema addresses tasks involving the relationship between text spans, such as entailment or contradiction. The Text (TEXT) schema is employed for classification tasks and for handling tasks with single text spans and associated labels. The textual pairs classification (PAIRS) schema is designed for tasks involving relationships between two text spans, like semantic similarity. Finally, the text-to-text (T2T) schema is used for sequence-to-sequence tasks, including translation and summarisation. Each schema is constructed to maximise coverage of relevant task features while maintaining simplicity and flexibility for diverse dataset integration.

3.4.3 Datasets for other SLR steps

Several other datasets have also been introduced that covers other steps of SLR creation. Marshall et al. [162] introduced a large dataset with Cochrane reviews for the task of assessing the risk of bias – a procedure aiming at establishing the quality of input studies. Nye et al. [187] proposed a PICO (Population, Intervention, Comparison and Outcome) extraction dataset containing 5,000 annotated abstracts of biomedical publications. In the query formulation, often the models evaluate their performance on the CLEF TAR 2017-2018 datasets [112, 113]. For the task of systematic review summarisation, an MSLR2022 shared task was introduced [273] consisting of two datasets: [271, 54].

There is poor coverage of SLR datasets among biomedical benchmarks, especially for the task of citation screening. None of the existing benchmarks contains any publicly available citation screening dataset. Only the BoX [194] benchmark incorporates five SLRs. However, these datasets remain inaccessible, not even available through a Data Use Agreement (DUA). This limitation renders the benchmark ineffective for broader research and application purposes.

From other SLR automation tasks, BLURB [79] and BigBio [69] benchmarks contain only the information extraction dataset by Nye et al. [187]. BLUE [198] and CBLUE [293] benchmarks do not contain any SLR-related task. Therefore, there is a clear need to develop and include publicly available SLR datasets in biomedical benchmarks, to facilitate further research and progress in this field.

The latest advances in large language models (LLMs) offer significant potential for aiding in SLR automation but simultaneously raise several concerns. A user study by Yun et al. [291] mentions that SLR practitioners acknowledged the potential utility of LLMs in various tasks, such as generating the first draft of a review, writing plain language summaries, and extracting information from longer texts. On the other hand, domain experts have highlighted several crucial issues, including concerns about hallucinations, the untraceable origins of generated content, and proliferation of bad quality reviews.

As the evaluation of LLMs is often based on benchmark datasets, ensuring that more SLR-related datasets will be included in future benchmarks is essential.

3.5 Evaluation of Citation Screening Automation

Evaluation measures can be classified according to the conceptual class into: performance measures (how successful a system/user is in accomplishing a search task), process measures (describing interaction between the user and the system) and usability measures (what is the user's perception of and experience with the system) [121].

When it comes to evaluating interactive information retrieval, precision and recall are the most commonly used performance measures, as noted by Kelly and Sugimoto [121]. However, assessing precision in system-centered evaluation is a more straightforward process as it only involves determining whether a document has been retrieved or not. In contrast, user-centered evaluation requires documents to be retrieved, viewed, and marked as relevant by a human subject, which can be a more complex process. In addition to performance measures, usability measures are frequently used to evaluate the effectiveness of interactive information retrieval systems. Such measures often focus on aspects such as user satisfaction with the search results, ease of use, usefulness, understandability, ease of learning, general satisfaction, and the amount of time required to conduct the search. In the context of CS automation, most evaluation was conducted using performance measures.

In the remaining parts of this section, we first describe evaluation measure axioms. Then, we detail the measures used in evaluating citation screening automation.

3.5.1 Evaluation measure axioms

Busin and Mizzaro [29] introduced a novel axiomatic approach, termed axiometrics, for analysing evaluation metrics in information retrieval systems. They proposed eight axioms that a metric should satisfy to be considered robust and reliable. The first two axioms address the consistency of relevance measurements across documents and queries. Axiom #1, named "Similarity comparison per document", posits that equal similarity in relevance assessments for a particular document should remain consistent when additional documents are included in the analysis. Axiom #2, "Similarity comparison per query", extends this principle to queries, asserting that equal similarity in relevance assessments for a specific query should be maintained across various query sets.

The following four axioms, Axioms #3 to #6, focus on the intrinsic properties of effectiveness metrics. Axiom #3, "Zero and maximum", mandates that an effectiveness metric must define a true zero and a maximum value, reflecting the extremities of worst and best performances. Axiom #4, "Similarity", requires that the metric should align with the similarity order between system-generated and user-defined relevance measurements. Axiom #5, "System Relevance", and Axiom #6, "User Relevance",

emphasise that the metric should prioritise documents based on system relevance and user relevance, respectively, when their correctness is equivalent.

Axiom #7 and #8 address the stability of effectiveness metrics concerning the expansion of document and query sets. Axiom #7, "Document Set Stability", argues that adding a document to a set should not adversely impact the performance comparison between two IR systems if no such effect was observed in a smaller subset. Axiom #8, "Query Set Stability", similarly posits that enlarging the query set should not reverse the performance ranking between two sysems unless such inversion was evident in a reduced query set.

Complementary research in this field has been conducted by other scholars. Bollmann [24] examined evaluation measures for IR systems, introducing the principles of monotonicity and the Archimedean axiom. They demonstrated that measures adhering to these axioms can be expressed as a function of the count of relevant retrieved and nonrelevant not retrieved documents. Contrasting this approach, Ferrante et al. [67] focused on developing a formal framework for utility-oriented measurements of retrieval effectiveness. Moffat [173] contested the exclusive use of uniform-step interval scales in IR evaluation. They advocated for the current IR metrics such as reciprocal rank or normalised discounted cumulative gain, which translate categorical and ordinal data into real numbers, arguing for their foundational robustness and enhanced interpretability compared to proposed intervalised alternatives.

3.5.2 Screening evaluation metrics

A successful automated citation screening algorithm should miss as few relevant papers as possible and also save reviewers time by removing irrelevant papers. In a more strict scenario for medical systematic literature reviews, the requirement might be not to miss *any* relevant paper. Metrics used in past studies consisted of the traditional metrics used in the NLP and IR, like accuracy, recall, precision, specificity, F-score or AUC and a set of custom metrics like WSS, count of relevant references found, utility or coverage [263]. Evaluation of automatic approaches traditionally relies on binary relevance ratings, very often obtained from the title and abstract screening step [190, 114].

When the screening task is treated as a classification task, measures based on the confusion matrix and the notion of Precision and Recall are commonly used [263, 190]. Aside from Precision and Recall, measures include variations of the harmonised mean between the two, i.e., F_β–score, Yield, Burden [269], $Utility_\beta$ [268], sensitivity-maximising thresholds [47], and AUC [38]. Another measure, Work Saved over Sampling (WSS), measures the amount of work saved when using machine learning models to screen irrelevant publications [35]. Another popular evaluation approach measures system's precision at a fixed recall level (Precision at r% recall, $Precision@r\%$), representing the percentage of relevant retrieved documents [124].

When the screening task is treated as a ranking task (e.g., for the sub-task of screening prioritisation or predicting when to stop screening), then rank-based measures and measures at a fixed cut-off are commonly used, e.g., $nDCG@n$, $Precision@n$, $Recall@n$,

R-Precision [78], and *last relevant found* [225, 100]. Recall versus effort plots using the *knee method* [43] have been also used, plotting the 'effort' scores over the full range of values of recall. Cost- and economic-based metrics are also popular, for instance, CLEF TAR shared task [114] used total cost (TC) and total cost with weighted penalty (TCW) The TREC Total Recall track [78] employed another cut-off-based metric, $recall@aR+b$, which measures the recall achieved when $aR+b$ documents have been identified, where R is the number of relevant documents in the collection and a and b are parameters. When $a=1$ and $b=0$, $recall@aR+b$ is equivalent to R-precision. However, retrospectively evaluating models at different levels of recall might better suit the screening task, because it takes into account the number of relevant documents found and the trade-off between reviewing more documents and potentially finding more relevant ones, versus stopping the review and potentially missing some relevant documents. One challenge arising from these two distinct approaches (classification versus ranking) is the difficulty in going beyond simple effectiveness measures and comparing the real-world savings for users.

Another practical issue for evaluation arises from the fact that the screening is typically conducted in two stages. During the initial phase of screening titles and abstracts, the limited information available often makes it challenging to definitively classify papers as either eligible or ineligible. Papers are usually marked as 'maybe eligible' or discarded as 'not eligible'. Only at the full-text stage can the label be confirmed as 'definitely eligible'. In practice, missing a paper at both of these stages leads to evidence being overlooked. However, there are several other factors to consider. The costs of manual screening are significantly higher at the full-text level. Conversely, the ability to recover a study at the full-text level is greater, as the pool of potentially screened studies is smaller. Finally, it is also important to consider that the majority (if not all) of the machine learning-based tools used to date work with the titles and abstracts of publications, so their assessments should only be compared to the initial screening stage label.

Evaluation of models using active learning was conducted with different metrics that account for labelled and unlabelled samples. Yield and burden were the most common metrics introduced for evaluating active learning models [269]. Yield represents the fraction of positive instances identified by an algorithm, whereas burden represents the fraction of positive instances annotated manually by reviewers. However, the formula for calculating Yield is identical to Recall. Wallace et al. [268] introduces Utility, a weighted sum of recall and (1-Burden) using a β parameter, similar to Fβ-measure. Another evaluation measure, Coverage, was proposed by Miwa et al. [172], which indicates the ratio of positive instances in the data pool annotated during active learning. Hashimoto et al. [87] proposed to estimate WSS in an active learning scenario as: $WSS@95\% = (1 - Burden)$, over a Yield performance of 95%. Despite the number of proposed bespoke metrics, one of the most commonly used ones to evaluate active learning models is the Area Under the ROC Curve (AUC).

A problem that echoes the lack of a common benchmark is a lack of standardised metrics that could be used to compare different approaches. Publications from the field of NLP usually use metrics like F1-score or AUC, whereas publications from IR focus on WSS

or precision-recall curves. Some metrics, like specificity, positive likelihood relation, net reclassification index, coverage, Matthews correlation coefficient (MCC), normalised Discounted Cumulative Gain (nDCG), were used only in one or two studies, making it hard to compare the scores between runs [263]. Another confusion in comparing approaches arises when different papers use different names to introduce the same metric: for instance, recall, sensitivity and yield effectively measure the same thing, and WSS and specificity are strongly correlated.

Among evaluation approaches for citation screening, the most common is cross-validation (van Dinter et al. [263] reports 13 papers), followed by train/test split (appeared in 9 publications) and train/validation/test split (2 studies). When using cross-validation, 11 studies used the 10-fold or 5×2-fold approach [263]. 5×2 cross-validation splits the dataset into two equally sized subsets: train and test, with an even distribution of label classes which are subsequently used to train and test the model. The whole process is then repeated five times [131]. Using half of the dataset for training machine learning models is convenient, mainly if the dataset contains few relevant documents. However, effectively, this assumption gives lower gains in a real-life scenario, especially for large SLRs where half of the dataset that needs to be manually annotated can mean several thousand publications. When there are no predefined test datasets it is also difficult to make comparisons between different models.

3.5.3 Usage of metrics across datasets

Finally, we were interested in checking how recently each dataset was used, where that usage was published, and what kind of evaluation measures were applied to that data. For each dataset from Table 3.2 we searched for the most recent publication using that dataset for an experimental evaluation. We then checked which evaluation measures were used in that publication and where it was published. Table 3.4 presents the summary of our findings. We can see that to this date, most datasets were used in the past two years and simultaneously used by different publications. There is also a disparity in used evaluation measures, yet the basic Precision, Recall and F1-score prevail.

3.6 Digital Tools and Resources in Literature Reviewing

This section delves into the various tools and platforms that facilitate the process of systematic literature reviews, beginning with academic search engines.

3.6.1 Academic search engines

Private academic search engines, citation indices and paywalled collections such as ScienceDirect, Scopus and Web of Science are one source of finding relevant publications. However, the restricted access and associated costs of these platforms might pose challenges for some researchers. Public search engines and publication aggregators such as

Table 3.4: Usage statistics of the SLR datasets, including the latest publication year, venue and evaluation measure. We report two usages in case there was a more recent pre-print published.

	Introduced in	Latest usage in	Latest evaluation measures	Latest venue
1.	2006 [35]	2023 [146, 134]	TNR [134], AUC [146]	ECIR
2.	2010 [269]	2022 [130]	WSS, Precision@95% [130]	ECIR
3.	2014 [172]	2016 [87]	Yield, Burden, WSS [87]	JBI
4.	2016 [99]	2022 [130], 2023 [145]	WSS, Precision@95% [130], AUC [145]	ECIR
5.	2017 [225]	2018 [224]	Precision, Recall, WSS [224]	SIGIR
6.	2017 [112]	2023 [276]	Precision, F1, Recall [276]	WSDM
7.	2018 [113]	2023 [276]	Precision, F1, Recall [276]	WSDM
8.	2019 [114]	2022 [274], 2023 [145]	MAP, Precision, nDCG [274], AUC [145]	ECIR
9.	2019 [5]	2020 [6]	Recall, Precision [6]	JAMIA
10.	2021 [195]	2022 [194]	F1-Score [194]	NAACL
11.	2022 [275]	2023 [277]	Precision, F1, F3, Recall [277]	SIGIR
12.	2022 [84]	2022 [84]	Recall, Precision, Macro F1, Accuracy [84]	MedPRAI

Google Scholar[12], Semantic Scholar [8], CORE [123], OpenAlex [205] and PubMed[13] are becoming increasingly popular for allowing researchers to access the latest publications freely. While one of their functionalities is creating a citation network, their overarching goal is to facilitate academic research. Their support for conducting systematic literature reviews, however, is often minimal.

Several systems were introduced for academics, offering more reviewing-related capabilities than simply searching for papers relevant to the query. For instance, ResearchRabbit[14] provides a graph-based visual interface for finding relevant publications based on citations and document similarity. ZetaAlpha[15] provides personalised recommendations of papers, but the application covers only the domain of Artificial Intelligence. The primary focus of these applications is on exploratory search. Moreover, only a few of the abovementioned tools provide an API, and none of them allow for a traditional systematic literature review workflow.

3.6.2 Data sources

Transitioning from general academic search engines, there are dedicated data sources that cater to specialised fields, providing researchers with detailed and often curated information. In the domain of medical systematic literature reviews, there are specific requirements on the number of databases researchers need to use to ensure the comprehensiveness of the search. For instance, Cochrane suggests using, at minimum, the following three data sources [94]:

[12] https://www.scholar.google.com/
[13] https://www.ncbi.nlm.nih.gov/pubmed/
[14] https://www.researchrabbit.ai
[15] https://www.zeta-alpha.com

1. MEDLINE,[16] which can be accessed through PubMed, is a primary source for biomedical literature.

2. EMBASE[17] is another critical database of biomedical and healthcare publications. However, it is noteworthy that certain institutions might not have a direct subscription to EMBASE. In such cases, Scopus[18] is a viable alternative to retrieve the content present in EMBASE.

3. The Cochrane Central Register of Controlled Trials (CENTRAL),[19] accessible through the Cochrane Library, is an essential resource for bibliographic reports of RCTs.

However, to further minimise potential biases, Cochrane researchers often supplement their searches with databases like *ClinicalTrials.gov* and the International Clinical Trials Registry Platform (ICTRP).[20] This is because the register records in CENTRAL are found to be less comprehensive than the original register entry, leading to the risk of missing out on crucial studies during a search, as highlighted by Hunter et al. [101].

3.6.3 Systematic review toolboxes

There are already a number of tools that help researchers conduct systematic literature reviews. The abundance of tools aiding systematic reviewers led to the creation of a page which aggregates all the available tools[21] [161]. Although there are 241 listed tools, researchers should be cautious as this index is not comprehensive and misses some commercial tools. Actually, not only would the toolbox be needed to categorise these tools, but a dedicated "graveyard" would be required to catalogue all the unsupported, not maintained applications, which makes it harder to create a common standard for this discipline.

Dedicated commercial tools exist to support medical researchers in conducting systematic literature reviews. However, these tools may further widen the knowledge gap as they are usually customised to medical reviews and require purchasing a subscription, which can be a bottleneck to academic researchers from lower- and lower-middle income countries [176, 184].

Colandr [32] is an open-access, machine-learning-assisted tool for finding relevant citations and extracting desired data from PDF articles. Elicit[22] uses large language models to find relevant studies and answers to the review questions. L · OVE[23] (Living OVerview of

[16] https://www.nlm.nih.gov/medline/index.html
[17] https://www.embase.com
[18] https://www.scopus.com
[19] https://www.cochranelibrary.com/central
[20] https://www.who.int/clinical-trials-registry-platform
[21] http://systematicreviewtools.com
[22] https://www.elicit.org
[23] https://www.iloveevidence.com

Evidence) provides a freely available database of living systematic reviews. However, the database covers only medical and health-related reviews. Given this tool's specialisation, the field would benefit from a comprehensive solution that leverages NLP techniques to streamline the living literature review process across domains.

Conducting systematic reviews without dedicated tools is possible but can be long and tedious, leading to an accumulation of errors and making the review difficult to update and reproduce. It was previously shown that with a set of good tools, the systematic review could be conducted within two weeks [34]. This outstanding result was achieved by using ten different tools that helped automate different stages of the whole process. Furthermore, using automation tools for screening can reduce the costs and time of the process by 50% when a machine learning model replaces one human screener [233]. Despite these advancements, a surprising revelation is that one of the most common tools used for extracting evidence from publications is Microsoft Excel [189]. However, in many disciplines outside of medicine, the adoption of any of these tools is still minimal. Therefore, a compelling case exists for broader education and advocacy to increase adoption, ensuring systematic reviews are efficient, accurate, and replicable [230].

3.6.4 Screening reduction tools

As of June 2023, the systematic review toolbox includes 46 applications targeting the screening phase, particularly title and abstract screening. Harrison et al. [86] found 15 of these tools as both accessible and available without requiring specific computing infrastructure. Among the popular commercial offerings are DistillerSR[24], Covidence[25], Evidence Prime[26] or Sciome[27] each presenting varied modules covering different aspects or the entirety of the systematic literature review (SLR) process.

Except for the commercial tools, a plethora of free or even open-source tools is available, usually created by academics. These tools, such as Abstrakr [270], Rayyan [62], or ASReview [260] usually support only one of the systematic review stages. They all offer a user interface for selecting relevant studies, and even some of them provide screening automation or prioritisation models. The APS tool has been proposed as a systematic review search system using continuous active learning on the PubMed collection [152].

The evaluation by Harrison et al. [86] highlighted that these tools vary significantly in cost, scope, and user community. Covidence and Rayyan, scoring over 75% in feature analysis, emerged as particularly notable in the user survey. They were lauded for their usability, aligning well with user requirements, and were favoured by researchers for future use, despite Covidence's associated costs potentially being a drawback. In contrast, tools like Abstrackr, Colandr, and EPPI-Reviewer, while providing valuable functionalities, were noted for certain limitations. For instance, Colandr's processing speed and user interface issues with EPPI-Reviewer were points of concern.

[24] https://www.evidencepartners.com/products/distillersr-systematic-review-software
[25] https://www.covidence.org
[26] https://www.evidenceprime.com/products
[27] https://www.sciome.com/swift-review/

Tsou et al. [257] compared Abstrackr and EPPI-Reviewer, finding varying performance in citation screening prioritisation across different datasets. This variation in performance, especially in the context of large and small reviews, highlights the need for careful consideration of the tool's capabilities in relation to the heterogeneity and complexity of the research topics addressed in systematic reviews.

The landscape of screening tools is complex and varied, with each offering unique features and functionalities. While tools like Covidence and Rayyan stand out for their general applicability and user-friendliness in healthcare research, the choice of tool should be guided by the specific requirements of the systematic review and the resources available to the researchers.

CHAPTER 4

Citation Screening Datasets

In this chapter, we introduce two novel datasets designed for improving robustness, accessibility and standardisation of resources in citation screening. The recent advancements and paradigm shifts in NLP and ML; with the extensive use of pre-trained models and transfer learning [149, 57], and the more recent prompt-based learning [155, 28]; have significantly transformed the field and enhanced the predictive capabilities of models across various tasks. Based on our review of available citation screening datasets and benchmarks (Section 3.4 of Chapter 3), we identified the most representative datasets for the task of citation screening, finding several issues with existing datasets. The available datasets still primarily cater to training supervised algorithms, lacking the scale and granularity necessary to evaluate state-of-the-art models.

To address these limitations and provide a more comprehensive resource for training and evaluating data-centric methods in SLR automation, we create CSMED, consolidating nine previously released collections of SLRs. We further extend a subset of SLRs in CSMED with additional metadata coming from the review protocol. We present how we achieved our goal of curating a meta-dataset that captures the diversity and challenges present in real-world SLRs and allows for easy extensions. Furthermore, we introduce CSMED-FT: a first dataset explicitly designed for evaluation of the full text publication screening task. The chapter details these datasets' construction, characteristics, and potential applications, aiming to facilitate more nuanced and efficient citation screening processes. Finally, we discuss how CSMED could be utilised in the future for a continuous prospective evaluation of systematic review automation with minimal human annotations—a task which could assist in the evaluation of large language models.

4.1 CSMeD: Citation Screening Meta-Dataset

CSMED contributes to the field of citation screening by addressing the scarcity of representative datasets in this area. Its creation, encompassing a wide array of systematic

Table 4.1: A list of source citation screening datasets included in the CSMED. First four datasets contain non-Cochrane SLRs, whereas the other five are based on Cochrane reviews. 'Avg. ratio of included' column present ratio of included publication from the title and abstract screening stage, 'Avg. size' refers to averaged across SLRs document count in the dataset. The 'Additional data' column describes if the review contains metadata other than coming from the citation list: (A): Search queries, (B): Review protocol containing review title, abstract and search strategy, (C): Review updates consisting of changes to included papers. Total values do not account for duplicated reviews. 'DTA' stands for diagnostic test accuracy reviews. ‡ – different number of reviews in the paper versus the GitHub repository; 'Total' counts the higher value and doesn't account for duplicates.

Source	# reviews	Domain	Avg. size	Avg. ratio of included	Additional data	Cochrane reviews
[35]	15	Drug	1,249	7.7%	—	—
[269]	3	Clinical	3,456	7.9%	—	—
[99]	5	Mixed	19,271	4.6%	—	—
[84]	7	Comp. Science	340	11.7%	B	—
[225]	93/176‡	Clinical	1,159	1.2%	A	✓
[112]	50	DTA	5,339	4.4%	B	✓
[113]	30	DTA	7,283	4.7%	B	✓
[114]	49	Mixed	2,659	8.9%	B	✓
[5]	25	Clinical	4,402	0.4%	C	✓
Total	360		3,471	4.4%		

literature reviews (SLRs), marks a step towards a more nuanced understanding and improvement of citation screening processes. This dataset not only addresses the limitations of existing collections, but also expands the scope for evaluating contemporary models. The inclusion of diverse SLRs, enriched with additional metadata, makes CSMED an invaluable resource for developing and testing data-centric methods in systematic literature review automation, offering a comprehensive platform for future research advancements.

4.1.1 Dataset construction details

We search for the original dataset sources, trying to identify the list of publications with eligibility decisions and as much meta-data about systematic reviews as possible. We decided not to host these datasets and to create dataset loaders that rely on that dataset's original (external) location. Currently, nine out of ten public CS datasets identified by us have been included in CSMED. We provide a summary of the datasets in Table 4.1. In total, CSMED consists of 360 SLRs, making it the largest publicly available collection in this domain and the only one providing access to the datasets via a harmonised API.

To ensure interoperability and facilitate the ease of use, we designed data loaders for the datasets in accordance with the BigBio text classification (`TEXT`) schema [69]. This choice offers several advantages. BigBio has the largest coverage of biomedical datasets

and supports access to the datasets via API. Moreover, it is compatible with popular libraries such as Hugging Face's Datasets [150] and the EleutherAI Language Model Evaluation Harness [70], thereby reducing the experimental costs.

Taking advantage of the lists of publications that most of the sources of datasets share as PubMed IDs, we extend the BigBio data loaders to enable the download of PubMed articles. Our harmonised data loaders selectively load the PubMed articles that are a part of each dataset. The single exception is the dataset by Hannousse and Yahiouche [84], which is the only publicly available collection of non-medical SLRs. For this dataset, we extract the content using the SemanticScholar API.[1] As a result, CSMED serves also as the first resource that gathers SLRs from diverse domains.

Adding new citation screening datasets to CSMED is a straightforward process, requiring the implementation of a custom dataloader. This dataloader is responsible for parsing input data, and it should include mechanisms for data splitting and example generation. An illustrative example of a dataloader is available under the following URL.[2] Additionally, CSMED offers utility methods for the automatic downloading of publications from PubMed and SemanticScholar, which are particularly useful when the dataset depends on external resources. It is important to ensure that the dataset's usage complies with its licensing terms and that the dataset authors have obtained all necessary permissions. Subject to these conditions, the dataset can be integrated into the HuggingFace Dataset Hub.[3]

4.1.2 Extending metadata

We present the possibilities of extending the subset of Cochrane SLRs to experiment with screening beyond classification or ranking. The extension of metadata in CSMED is an enhancement that increases the dataset's utility and relevance.

We categorise CSMED datasets into two groups:

1. Datasets containing Cochrane medical SLRs (datasets #5-9 in Table 4.1),
2. Datasets comprising other SLRs (datasets #1-4 in Table 4.1).

This distinction is made because from following the Cochrane protocol, more extensive information on the review is provided. We use the additional data available on reviews websites to extend CSMED. Among the new information, we find the eligibility criteria most valuable—the inclusion of eligibility criteria no longer limits the data to the evaluation of supervised binary classification but opens its application to question-answering or language inference tasks.

[1] https://www.semanticscholar.org/product/api

[2] https://github.com/WojciechKusa/systematic-review-datasets/blob/main/csmed/datasets/datasets/swift/swift.py

[3] https://huggingface.co/datasets

We carefully examine the subset of SLRs produced by Cochrane, aiming to identify potential enhancements and extensions that would help mitigating the existing limitations of previous datasets. Every Cochrane SLR first registers and publishes the protocol containing the review title, abstract, search strategy and the eligibility criteria. This information is all that human experts need to produce the final review, i.e., they first find the relevant studies and then conduct the meta-analysis of their results. As described in Section 3.3.1, the screening process can be also modelled as question-answering, where every publication is compared against the eligibility criteria in order to make the decision about the inclusion, similar to the task of matching patients to clinical trials [214, 216]. In the current approach, we consider only binary relevance (included versus excluded). However, in practice, more categories can be defined by reviewers (e.g. a study can be assigned as a background publication or meta-analysis).

To expand CSMeD, we searched the Cochrane Library[4] for all SLRs from the meta-dataset based on the Cochrane review ID and take their latest open-access version. We extract available information about the review: review title and abstract, eligibility criteria, search strategy and references. Cochrane reports a list of publications that were included and excluded at the full text screening stage. These lists can be treated as representative of all included from the title and abstract screening stage. This assumption is based on practical review processes, where reviewers, especially in cases of large volumes of papers, might not report every exclusion at the full text level due to time and workload constraints. Hence, these publications are considered an approximate but comprehensive representation of those included during the earlier screening phase. Each excluded publication in these reports is also accompanied by a specific reason for exclusion, as determined by the reviewer.

As the previous research on citation screening for medical SLRs evaluated their approaches on the PubMed database, we assign PubMed IDs to these publications. We also define appropriate BigBio data loaders (Section 3.4.2) so the task can be seen as question-answering (QA) or textual pairs classification task (PAIRS).

Table 4.2: Details of the CSMeD expanded meta-dataset. Column '#docs' refers to the total number of documents included in all SLRS within the dataset, '#included' mentions number of included documents on the title and abstract screening stage and 'Avg. %included' the percentage of included publications averaged from all reviews.

Split name	#reviews	#docs	#included	Avg. #docs	Avg. % included	Avg. #words in document
CSMeD-train-basic	30	128,438	7,958	4,281	9.6%	229
CSMeD-train-cochrane	195	372,422	7,589	1,910	21.9%	180
CSMeD-dev-cochrane	100	229,376	4,365	2,294	20.8%	201
CSMeD-all	325	730,236	19,912	2,247	20.5%	195

[4] https://www.cochranelibrary.com/cdsr/reviews

Details of the new expanded CSMED are provided in Table 4.2. We were not able to find meta-data for all Cochrane SLRs, hence the expanded CSMED is smaller than the original meta-dataset. In total, the new expanded dataset consists of 295 unique Cochrane SLRs and 30 non-Cochrane SLRs.

We then establish the canonical splits for our dataset. We design the splits at the intra-review level, which assumes a zero-shot application of models to "held-out" SLRs (models trained on one set of reviews are applied to another set without further training). The entire set of basic SLRs is designated for training. From the Cochrane subset of systematic reviews, we randomly selected 195 of them for the training split and the remaining 100 for the development split. We abstain from designating a test split because CSMED aggregates existing datasets, all of which have been heavily used in prior research. Given the concerns raised about the overlap in these datasets, creating a new, unbiased test collection is recommended.

We decided against introducing inter-review splits due to the dataset's diverse potential applications. For instance, researchers interested in ranking and prioritisation might run the model in a zero-shot setting, while those focusing on active learning-based models would find dynamic definitions of 'train' and 'test' documents more relevant, making such static designations impractical. However, this could easily be extended in future work, especially for the potential introduction of a new test collection.

With this new information, one can envision a modelling benchmark which steps away from simple classification of publications to simulating a complete process of searching and generation of review. In this way we establish a more reliable benchmark for evaluating the capabilities of language models.

4.1.3 Visualisations

Our data analysis methodology involved creating visualisations and summary tables based on curated datasets. We analyse dataset statistics like available data splits, licensing information, dataset and reviews size as well as dataset overlap. This allows us to provide both a detailed view of individual reviews and an overview of datasets containing multiple reviews.

We leverage Streamlit[5] to create interactive visualisations for our meta-dataset. We present essential details for every dataset, such as the number of training samples, character and word counts, and labels and token lengths distribution across dataset splits (example in Figure 4.1). We build upon the existing BigBio schemas and visualisations, extending them to incorporate citation screening-specific details.

We use TF-IDF-based document vectoriser with UMAP [169] to plot two-dimensional representations of the datasets. This approach allows us to effectively capture and display the structural patterns and similarities within a single systematic literature review, aiding

[5] https://streamlit.io

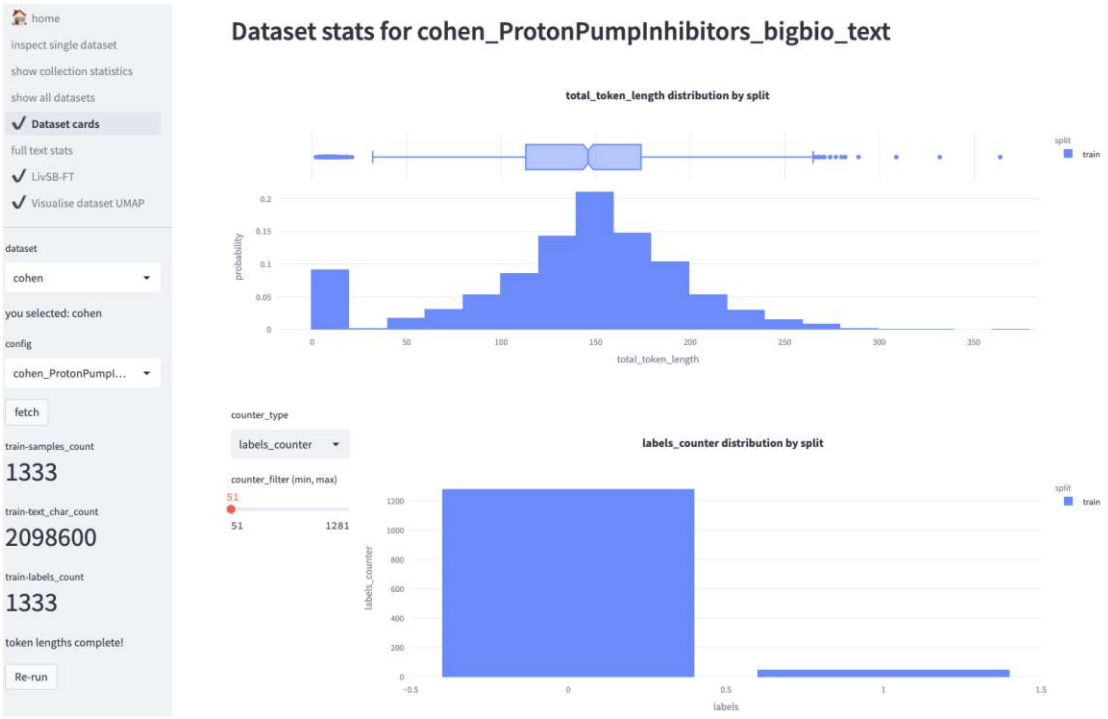


Figure 4.1: Example visualidation with statistics for a "Proton Pump Inhibitors" SLR dataset.

researchers in identifying clusters, outliers, and potential data correlations. An example of UMAP clustering of publications is presented in Figure 4.2.

A live demo of the visualisation interface is available.[6] Some features of the visualisation interface require data preprocessing; they are unavailable in the demo but can be run locally using the code from the GitHub repository.

4.1.4 CSMeD Data Card

Dataset Description: CSMeD is a meta-dataset consisting of nine different citation screening datasets containing 300+ systematic literature reviews (SLRs). Each systematic review consists of a list of publications that need to be classified as either *relevant* or *irrelevant*. All datasets have data loader scripts providing programmatic access aligned with the BigBio framework and HuggingFace datasets library. We preserve the original splits of the datasets. We also generate data cards for every dataset which is part of the CSMeD. CSMeD allows for accessing independent datasets and single systematic reviews, which are part of each dataset.

train-cochrane and dev-cochrane splits contain expanded metadata about systematic reviews such as systematic review title, abstract, eligibility criteria and search

[6] https://systematic-review-datasets.streamlit.app/

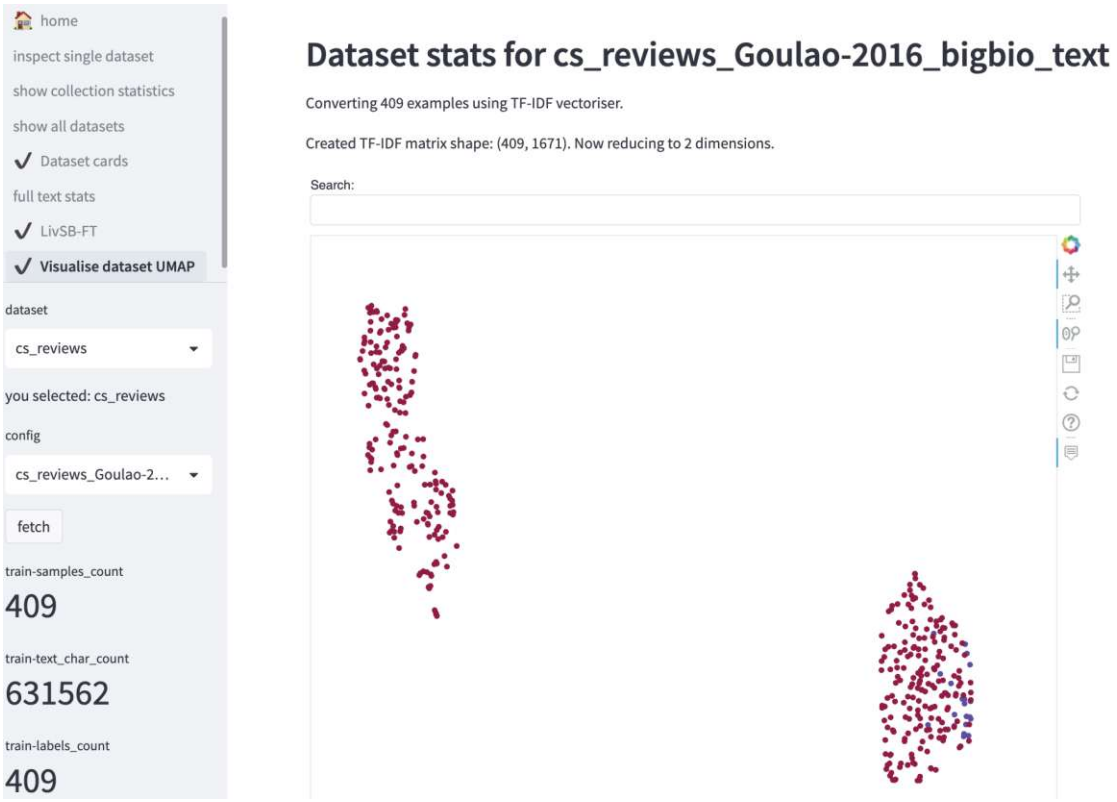


Figure 4.2: Example visualisations with TF-IDF and UMAP representation of documents for a "CS-Goulao-2016" SLR. Based on the plot, one can see that the retrieved documents are grouped in two clusters with all relevant publications belonging to one of them (bottom-right part of the plot). This can be an indicator that any model will likely remove the other "non-relevant" cluster of documents and hence achieve good score in detecting true negatives.

strategy. TRAIN-BASIC is a set of SLRs for which such meta-data was unavailable and it is characterised by the systematic literature review title.

TRAIN-COCHRANE and DEV-COCHRANE splits are suitable for the tasks of question answering, natural language inference, and text pair classification. TRAIN-BASIC is suitable only for the text classification task.

Homepage: `https://github.com/WojciechKusa/systematic-review-datasets`

URL: `https://github.com/WojciechKusa/systematic-review-datasets`

Licensing: CC BY 4.0

Languages: English

Tasks: text classification (`TXTCLASS`), question answering (`QA`), natural language inference (`NLI`), text pairs classification (`PAIRS`).

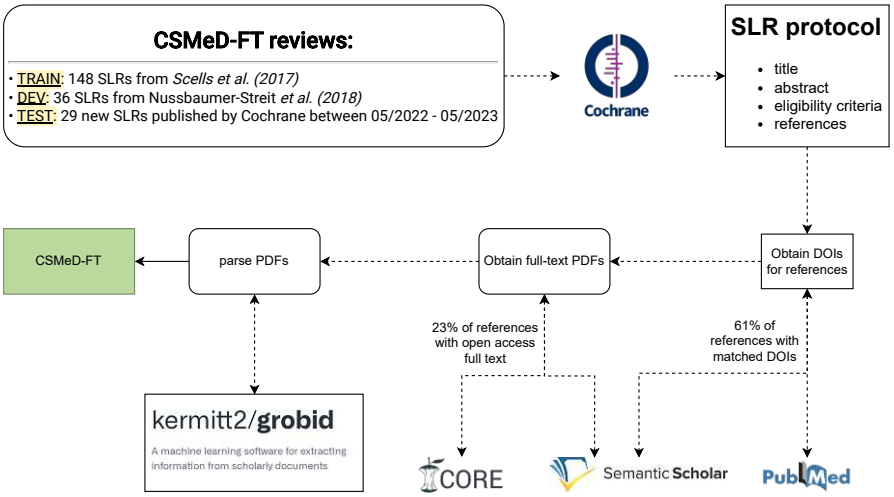


Figure 4.3: CSMeD-ft construction steps.

Schemas: Text (`TEXT`), Text pairs classification (`PAIRS`). Question Answering (`QA`), source (`source`).

Splits: train-basic, train-cochrane, dev-cochrane, *all*

4.2 CSMeD-ft: Full Text Classification Dataset

LLM advancements have enabled processing long context windows [20, 292, 175, 81], with the commercial tools claiming to support 32k [191] or even 100k tokens [9]. Exploiting this capability, we propose CSMeD-ft: a dataset designed explicitly for evaluating the screening of document full texts, to research questions associated with the comprehensive understanding of very long documents

CSMeD-ft is an extension of the CSMeD meta-dataset that specifically focuses on the full text screening step in SLRs. CSMeD-ft is, to the best of our knowledge, the first dataset explicitly targeted at the screening of the publication full text. While previously researchers already used full text screening labels from other datasets to evaluate their models, the input to these models constituted only the titles and abstracts of publications [100].

4.2.1 Dataset construction details

Figure 4.3 depicts CSMeD-ft construction steps. To construct CSMeD-ft, we collected various elements of SLRs from the Cochrane Library website, including the title, abstract and eligibility criteria sections of the SLR and SLRs' appendix and references. The appendix of published SLR contains a search strategy, while the references list papers categorised as: "studies included in the review", "studies excluded from the review", and "additional references". We decided to focus solely on the "included" and "excluded"

categories as there is no definitive way to determine the intended meaning when researchers added papers as additional references. However, in future work, we plan to explore the possibility of extending the dataset to encompass publications from the "additional references" category.

To obtain the full texts of references, we used the DOI (Digital Object Identifier) of each publication. While some references directly provided the DOI, for others, we initially attempted to match them to PubMed IDs and then extracted the DOIs from PubMed and Semantic Scholar. To assign PubMed IDs to the publications parsed from the Cochrane website, we followed a four-step process:

- We check if the PubMed ID information is provided on the Cochrane references webpage.
- We conduct search in PubMed using ENTREZ[7] by searching for the same title and authors.
- We search for the PubMed ID in SemanticScholar using publication DOI from Cochrane references webpage.
- We search again in PubMed, this time with a relaxed requirement by searching for an exact match in the title only.

We then use the PubMed ID to resolve the DOI of the publication. We could match the DOI for more than 61% of references. We take the publications with matched DOIs and use SemanticScholar and CORE[8] APIs to find URLs to their open access full text documents. This process successfully finds URLs to 27% of publications on average. We then download the PDFs and use GROBID [1] to parse the content of these documents into an XML format.

We adopted a time-wise construction approach for CSMED-FT canonical splits, putting the newest reviews in the test set to ensure the integrity and avoid data contamination. Therefore, we select 31 open access Cochrane reviews published in the last year (between 01/06/2022 and 31/05/2023) to create our test set. We used data from previous publications to construct a testing and development set: we selected 60 reviews mentioned in Nussbaumer-Streit et al. [185] for the development set and 176 reviews listed by Scells et al. [225] for the training dataset.

At the moment of constructing this dataset, creating a prompt for LLMs with an input of a few thousand tokens is feasible, albeit costly. According to the OpenAI model pricing[9] as of June 2023, screening 500 full texts with the GPT-4-32k model would cost more than 400 USD. Therefore, we also release a subset of randomly selected 50 documents

[7] https://www.ncbi.nlm.nih.gov/search/
[8] https://core.ac.uk
[9] https://openai.com/pricing

from the test set as CSMeD-ft-test-small. Details of the dataset are presented in Table 4.3.

It should be noted that newer SLRs tend to have more comprehensive metadata and more open access full text publications available. This resulted in token length and label frequency differences across the dataset splits (Figure 4.4). Despite these variations, we decided to retain these splits as they present a more realistic and challenging scenario, closely reflecting real-life circumstances. We release the source code for the entire dataset construction process, enabling transparency and reproducibility. We have also built a dedicated page to explore CSMeD-ft dataset containing full text documents on the CSMeD visualisations dashboard.

Table 4.3: Details of the CSMeD-ft dataset. Column '#included' mentions number of included documents on the full text step. CSMeD-ft-test-small is a subset of CSMeD-ft-test.

Dataset name	#reviews	#docs.	#included	Avg. % included	Avg. #words in document	Avg. #words in review
CSMeD-ft-train	148	2,053	904	44.0%	4,535	1,493
CSMeD-ft-dev	36	644	202	31.4%	4,419	1,402
CSMeD-ft-test	29	636	278	43.7%	4,957	2,318
CSMeD-ft-test-small	16	50	22	44.0%	5,042	2,354

Despite its small size, all labels in the dataset have been created by medical experts. It evaluates the eligibility of publications to the often very complex criteria, both written in a domain-specific medical language.

4.2.2 CSMeD-ft Data Card

Dataset Description The dataset focuses on the task of full text screening for systematic literature review creation. It contains 3,333 systematic literature review and publication pairs with decisions if the publication was included in the systematic literature review. Every excluded publication also contains a textual justification for exclusion. Systematic literature reviews are formatted in the JSON format, whereas publications are stored as CSV files. Token frequency distribution by split and frequency of different kind of instances is presented in Figure 4.4. Newer SLRs (in validation and test splits) have more text, than the older reviews in the training splits. CSMeD-ft-sample is a subset of CSMeD-ft-test dataset. We intend to store the dataset on the TU Wien Research Data repository,[10] currently the dataset is available on the project GitHub repository.

Homepage: `github.com/WojciechKusa/systematic-review-datasets`

URL: CSMeD-ft.zip

Licensing: CC BY 4.0

[10] `https://researchdata.tuwien.ac.at`

Figure 4.4: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Languages: English

Tasks: text pairs classification, natural language entailment

Schemas: TEXT, PAIRS, source.

Splits: TRAIN, DEV, TEST, SAMPLE

Dataset size (document pairs): TRAIN: 2,053, DEV: 644, TEST: 636, SAMPLE: 50

Size of downloaded dataset files: 33.5 MB

Size of the generated dataset files: 112.2 MB

4.3 Discussion

In this section, we delve into the broader implications, challenges, and future directions presented by the introduction of CSMED and CSMED-FT. We discuss the underlying motivations for their creation, their potential for transforming SLR automation, and the challenges of evaluating large language models in this context. This discussion aims to contextualise our contributions within the broader landscape of information retrieval and machine learning, highlighting their relevance and potential impact on the field.

4.3.1 Rationale behind CSMeD

The introduction of CSMeD represents a step forward in the standardisation of systematic literature review automation. This meta-dataset addresses the acute need for comprehensive and diverse datasets in citation screening, a task that has been historically underrepresented. By consolidating nine distinct datasets and incorporating a broad range of SLRs, CSMeD offers more robust training and evaluation of models but also provides insights into the nuanced challenges faced in real-world SLR scenarios.

Despite the potential value in revising existing collections, we recognise that such efforts might not substantially alter prevailing research practices, particularly the reliance on outdated dataset versions. For instance, Lagopoulos and Tsoumakas [141] identified some issues with the CLEF TAR 2019 collection. They subsequently released an updated version containing extended meta-data and addressing concerns related to duplicates and overlap[11]. However, subsequent research has largely ignored this updated dataset. Therefore, we decided to introduce the CSMeD as a new collection.

The development of CSMeD is not an attempt to establish another gold standard dataset but rather to refine and enhance the quality of existing data. Our aim is to shed light on potential research directions that can further the automation of SLRs. In this regard, to ensure impartiality, quality, and continuous improvement, we advocate for the governance of citation screening datasets by independent collaborations, such as the International Collaboration for the Automation of Systematic Reviews (ICASR) [18].[12]

4.3.2 Prospective evaluation of systematic review automation

The evaluation of large language models presents several challenges, prominently including prompt sensitivity [178], construct validity, and data contamination [115, 116]. Prompt sensitivity refers to the model's performance variability based on the input prompt's structure or wording, complicating consistent assessment across different prompts. Construct validity concerns the extent to which the evaluation metrics and tasks genuinely reflect the model's ability to perform intended functions. Contamination, often resulting from data leakage, involves models inadvertently being trained on parts of the test set, leading to inflated performance metrics. These challenges necessitate careful design and implementation of evaluation protocols to ensure the reliability and validity of LLM assessments. CSMeD can offer a platform for this purpose by providing an up-to-date collection of SLRs, coupled with detailed metadata, enabling a more dynamic and realistic evaluation of LLMs.

One of the most promising aspects of our approach and using CSMeD is the potential for prospective evaluation. New Cochrane reviews are continually conducted and published in the Cochrane Library. The approach proposed for the construction of CSMeD can be used to gather the SLR protocols as soon as they are registered in the Cochrane Library.

[11] https://github.com/sakrifor/tar
[12] https://icasr.github.io

The data from the protocols (namely SLR description, search strategy, and eligibility criteria) is sufficient for reviewers to conduct the SLR manually. Thus, the automation approaches can be tested inside such 'sandboxes', and the gold data will be available as soon as the manual review is completed. This strategy ensures that predictions are made before the publication of a review, ensuring no data contamination. Even if LLMs contained some information about underlying PubMed documents in their training data, this was without any labels relevant to screening or meta-analysis results, further preventing contamination. Using Cochrane reviews minimises the data requirements and enhances the real-world applicability of the evaluation. It assesses current model capabilities and their adaptability and scalability in evolving SLR datasets. A notable limitation is the approximate two-year delay from SLR registration to publication, which can be mitigated to some extent by selecting older SLR protocols.

4.4 Summary

This chapter has presented the development and potential applications of two novel datasets for citation screening in systematic literature reviews, CSMED and CSMED-FT. CSMED, with its extensive collection of diverse SLRs and enhanced metadata, addresses significant gaps in existing datasets and sets a new standard for evaluating citation screening methods. CSMED-FT, focuses on full-text screening, further expanding the scope for evaluation of advanced language models in this field. We also discuss how our work can be extended to facilitate evolving collections of systematic reviews. These datasets not only facilitate more efficient and accurate screening processes but also open avenues for future research in applying advanced language models to the complexities of systematic literature reviews.

CHAPTER 5

Relevance-based Evaluation Measures for Citation Screening

As discussed in Chapter 3, multiple evaluation measures have been proposed for assessing the efficiency and effectiveness of citation screening algorithms. Among these measures, one that has garnered significant attention is the Work Saved over Sampling (WSS). This chapter starts by examining Work Saved over Sampling. We investigate the properties of WSS, and assess its terms and their influence on the final WSS score. Similarly to the Discounted Cumulative Gain (DCG) metric [105], we propose to normalise the WSS in order to be able to compare the scores between multiple models and datasets. This representation preserves all the features of the WSS and simultaneously removes some constants from the equation. Furthermore, we show that the normalised WSS is equivalent to the True Negative Rate (TNR, also known as specificity).

Using the derived equation, we calculate and provide benchmark scores for fifteen systematic review datasets with the TNR@95% Recall measure. We show that the incorrect usage of WSS@95% to compare averaged performance across several datasets proved to yield erroneous order of models. Based on our findings, we recommend using TNR at $r\%$ Recall as the evaluation measure for citation screening automation and technology-assisted reviews.

We also examine the behaviour of Precision at $r\%$ Recall cutoff – a measure analogous to WSS, also used in previous research. Similarly to the WSS, we propose its min-max normalised versions.

We introduce a variation of the True Negative Rate, which we call *rectified True Negative Rate* ($reTNR@r\%$). The $reTNR$ penalises models that perform worse than a random ordering of the documents, following the original rationale of WSS.

Finally, we introduce *VoMBaT* – a new visual analytics tool for analysing behaviour of evaluation measures for high-recall tasks. Our visual analytics tool addresses the current

limitations in understanding evaluation measures, especially at different Recall levels, by bridging the gap between technical experts and non-experts in the field of High-Recall IR. Our tool allows researchers and practitioners to gain deeper insights into the behaviour of evaluation measures. Additionally, it offers the ability to simulate potential savings in time and money, specifically in the process of manual versus automatic citation screening across different datasets. Using *VoMBaT*, we present several analyses of evaluation measures in the context of citation screening. We discuss and compare WSS and TNR with Precision and AUC, two other commonly used evaluation measures in this setting.

5.1 Screening Evaluation Measures Based on the Confusion Matrix

In this section, we introduce and discuss a set of evaluation measures derived from the confusion matrix, which forms the basis for assessing the performance of citation screening models. Most of these measures have been previously applied to the evaluation of automated citation screening or TAR models.

Accuracy and Balanced Accuracy We begin with accuracy, a primary measure that represents the overall correctness of the model in classifying both relevant and irrelevant documents. It is calculated as the proportion of true positive and true negative predictions among all predictions, given by:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{5.1}$$

While accuracy provides a general indication of model performance, it can be misleading in imbalanced datasets. Particularly in citation screening, where average percentage of relevant documents in 300 systematic reviews is 7.1%. In such cases, the accuracy metric may not provide a comprehensive insight into the model's performance. To mitigate this issue, Balanced Accuracy (BA) is introduced, offering a more nuanced evaluation by considering the model's effectiveness separately for each class and then averaging these accuracies. BA is defined as:

$$BA = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \tag{5.2}$$

Precision, Recall, and F-score Precision and Recall are pivotal in contexts where either false positives or false negatives have significant implications. Precision calculates the proportion of true positive predictions among all positive predictions, indicating the model's ability to avoid false positives. Recall, on the other hand, assesses the model's

capability to identify all relevant documents. They are formalised as:

$$Precision = \frac{TP}{TP + FP} \tag{5.3}$$

$$Recall = \frac{TP}{TP + FN} \tag{5.4}$$

$$\tag{5.5}$$

The F_β-score harmonises these metrics, offering a single measure that balances Precision and Recall. The β parameter is chosen such that Recall is considered β times as important as Precision:

$$F_\beta = \frac{(1 + \beta^2) \cdot TP}{(1 + \beta^2) \cdot TP + \beta^2 \cdot FN + FP} \tag{5.6}$$

Two commonly used values for β are 2, which weighs Recall higher than Precision, and 0.5, which weighs Recall lower than Precision:

$$F_2 = \frac{5 \cdot TP}{5 \cdot TP + 4 \cdot FN + FP} \tag{5.7}$$

$$F_{0.5} = \frac{1.25 \cdot TP}{1.25 \cdot TP + 0.25 \cdot FN + FP} \tag{5.8}$$

In the context of citation screening, a Recall-oriented task, users might be more interested in minimising false negatives, hence using the β values higher than 1.

Depth for Recall, Burden, Utility and Coverage Depth for Recall (DFR) quantifies the effort required to achieve a certain level of Recall. It integrates the prevalence of relevant documents and Precision to provide insights into the workload associated with manual citation screening:

$$DFR = \frac{Prevalence \cdot r}{Precision} \tag{5.9}$$

where the Prevalence is described as a ratio of relevant documents to all documents in the collection:

$$Prevalence = \frac{\mathcal{I}}{\mathcal{N}} \tag{5.10}$$

Burden, Utility and Coverage were proposed to assess the screening model performance in the active learning setting [269, 172]:

$$Burden = \frac{TP_L + TN_L + FP_L + TP_U + FP_U}{N}, \tag{5.11}$$

$$Utility = \frac{\beta \cdot Recall + (1 - Burden)}{\beta + 1}, \tag{5.12}$$

$$Coverage = \frac{TP_L}{TP_L + FN_L + TP_U + FN_U}, \tag{5.13}$$

given true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for labelled data (TP_L, TN_L, FP_L, FN_L) and unlabelled data (TP_U, TN_U, FP_U, FN_U).

Other measures The Matthews Correlation Coefficient (MCC) is a robust metric that considers all four terms of the confusion matrix, delivering a balanced evaluation even in imbalanced datasets:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5.14}$$

False Discovery Rate (FDR) quantifies the proportion of false positives among all positive predictions. The Diagnostic Odds Ratio (DOR) combines these rates to offer insights into the model's discriminative power:

$$FDR = \frac{FP}{TP + FP} \tag{5.15}$$

$$LR+ = \frac{TPR}{FPR} \tag{5.16}$$

$$LR- = \frac{FNR}{TNR} \tag{5.17}$$

$$DOR = \frac{LR+@r\%}{LR-@r\%} \tag{5.18}$$

Additional measures like Negative Predictive Value (NPV) and False Omission Rate (FOR) can also be considered to provide a comprehensive understanding of the model's performance across various dimensions:

$$NPV = \frac{TN}{TN + FN} \tag{5.19}$$

$$FOR = \frac{FN}{TN + FN} \tag{5.20}$$

5.2 Work Saved over Sampling Measure

The *Work Saved over Sampling* (WSS) is a custom evaluation measure, also based on a confusion matrix, used specifically in the context of automated citation screening evaluation. It was introduced and described by Cohen et al. [35] as "the percentage of papers that meet the original search criteria that the reviewers do not have to read (because they have been screened out by the classifier)." It estimates a reduction in the workload of human screening by using automation tools, assuming a fixed r% Recall level. WSS, given a Recall set at r%, is defined as follows:

$$WSS@r\% = \frac{TN + FN}{N} - (1 - r)\,, \tag{5.21}$$

where TN is the number of true negatives (excludes that were correctly removed), FN is the number of false negatives (includes that were incorrectly marked as irrelevant documents), and N is the total number of documents.

The choice of Recall level is influenced by the domain and characteristics of the review. Past studies on the automation of citation screening typically used 95% Recall as the threshold to preserve a satisfactory quality of the systematic literature review in medicine [35]. In other technology-assisted review tasks, Recall level might be lower, and sometimes this choice is influenced by time or money limitations. For instance, in e-discovery, a commonly used Recall is 80% [289].

WSS has been used in almost 40 previous publications to evaluate the effectiveness of a supervised machine learning system for citation screening [166, 99, 226, 124, 130]. This makes it the de-facto standard evaluation metric in this field. It was also used as one of evaluation measures for the Technology Assisted Review shared task at CLEF by Kanoulas et al. [112, 113].

O'Mara-Eves et al. [190] mention that there is a subjective component for metrics like Fβ-score and WSS. Evaluators determine thresholds and parameters, making it difficult to compare across studies. It is also not always transparent or justified how the thresholds/weights are chosen. Cohen [36] in their later study abandoned WSS in favour of Area Under the ROC Curve (AUC) as they argue that the former metric fails to capture different Recall-precision trade-offs in different reviews. On the other hand, Cormack and Grossman [45] mention that cumulative measures like area under the cumulative Recall curve and average Precision yield very little insight into the actual or hypothetical effectiveness of the models.

Norman [182] notices that despite WSS being relatively easy to interpret in the context of automation of systematic reviews, it is also strongly influenced by random effects and tends to have a large variance. Recall versus effort plots using the *knee method* [43] can be used as a more generalised extension of the WSS metric, plotting the scores over the full range of values of Recall.

5.3 Analysis of the Work Saved over Sampling Measure

We first present an example of the evaluation of automated citation screening with WSS. Later, we examine WSS properties and its terms and their influence on the final score.

5.3.1 Citation screening example

Let us assume an example systematic review with a citation list containing the total number of documents $N = 2000$. Out of them, only 200 (10%) are relevant to the systematic review study and should be included in the final review (also known as *includes*, $\mathcal{I}$). The remaining 1800 documents are irrelevant to the review topic and should be excluded (also known as *excludes*, $\mathcal{E}$). In a manual screening scenario, annotators need to screen all 2000 documents to select only the 200 relevant ones.

Fixing the level of Recall also assumes that the number of true positives and false negatives is static. A Recall of 95% is achieved when the model correctly predicts 190 relevant documents (TP). The remaining 10 includes are treated as false negatives (FN). In practice, different models vary from each other by how many excludes they can screen out automatically (i.e., good models maximise the number of $\mathcal{E}$ classified as true negatives (TN) while minimising the number of false positives (FP)). The WSS measure can be applied both to ranking (where the rank of r% relevant documents is used) and classification (where we a posteriori assume that the model used a specific prediction threshold to achieve the Recall level of r%.)

5.3.2 The $(1 - r)$ term

The $(1 - r)$ term was introduced to measure the advantage of a model when compared to the work saved with respect to a simple random sampling. A Recall level of 95% is on average achieved when 95% of a dataset is randomly sampled, and this provides a 5% saving for reviewers. With the $(1 - r)$ term, the WSS@r% score above 0 means that a model performs better than the random sampling. If the WSS score is below 0, the model performs worse than random.

We argue that the $(1 - r)$ term does not impact the WSS score as it was originally assumed, as it is just a constant value that is being subtracted from all scores from the same level r% of Recall. In particular, for $r = 0.95$, this term will always subtract 0.05 from the final WSS score, which can be seen as redundant if we want to compare multiple results.

5.3.3 The FN term

WSS at a specific r% Recall assumes that exactly $(1 - r)$% of documents that should be included will be misclassified. For a specific r% Recall, the number of False Negatives (FN) is always equal to $\lfloor |\mathcal{I}| \cdot (1 - r) \rfloor$, where with $\lfloor \cdot \rfloor$ we indicate the floor operator. This means that the FN term will also be a constant for every model for the same dataset. Consequently, for a fixed level of Recall, true positives (TP) are equal to $r \cdot |\mathcal{I}|$.

Furthermore, the usage of the FN term in the WSS formula complicates its understanding. In the numerator (which should be maximised since the formula measures work saved), there is a sum of true negatives (a factor that should be maximised) and false negatives (a factor which should instead be minimised). A single evaluation measure should not maximise the sum of correct and wrong decisions simultaneously.

5.3.4 The maximum and minimum WSS value

For every dataset, we can calculate the maximum and minimum values of the WSS score as follows:

$$max(WSS@r\%) = \frac{|\mathcal{E}| + \lfloor |\mathcal{I}| \cdot (1-r) \rfloor}{N} - (1-r), \tag{5.22}$$

$$min(WSS@r\%) = \frac{0 + \lfloor |\mathcal{I}| \cdot (1-r) \rfloor}{N} - (1-r). \tag{5.23}$$

The maximum value of WSS is achieved when at least $r\%$ of included documents are presented first, before any irrelevant document (or in the classification nomenclature $TN = |\mathcal{E}|$). On the other hand, the minimum WSS value is obtained when all excluded documents are ranked before at least one relevant document ($TN = 0$).

The absolute maximum and minimum values of WSS depend on the dataset, and its excludes/includes ratio. $max(WSS)$ approaches 0 in datasets significantly imbalanced towards the positive class (includes):

$$\lim_{|\mathcal{E}|\to 0} max(WSS@r\%) = \lim_{|\mathcal{E}|\to 0} \frac{|\mathcal{E}| + |\mathcal{I}| \cdot (1-r)}{|\mathcal{E}| + |\mathcal{I}|} - (1-r) = 0. \tag{5.24}$$

On the other hand, as the ratio of irrelevant to relevant documents ($|\mathcal{E}|/|\mathcal{I}|$) gets higher, the maximum achievable score by WSS also gets higher (impact of includes in both nominator and denominator gets smaller, and the final score depends more on the excludes). Therefore, $max(WSS)$ approaches r in datasets heavily imbalanced towards the negative class (excludes):

$$\lim_{|\mathcal{I}|\to 0} max(WSS@r\%) = \lim_{|\mathcal{I}|\to 0} \frac{|\mathcal{E}| + |\mathcal{I}| \cdot (1-r)}{|\mathcal{E}| + |\mathcal{I}|} - (1-r) = r. \tag{5.25}$$

Similar considerations can be applied to $min(WSS)$, and its upper and lower bound also depends on the excludes/includes ratio:

$$\lim_{|\mathcal{E}|\to 0} min(WSS@r\%) = \lim_{|\mathcal{E}|\to 0} \frac{0 + |\mathcal{I}| \cdot (1-r)}{|\mathcal{E}| + |\mathcal{I}|} - (1-r) = 0, \tag{5.26}$$

$$\lim_{|\mathcal{I}|\to 0} min(WSS@r\%) = \lim_{|\mathcal{I}|\to 0} \frac{0 + |\mathcal{I}| \cdot (1-r)}{|\mathcal{E}| + |\mathcal{I}|} - (1-r) = r - 1. \tag{5.27}$$

Moreover, $min(WSS)$ will not be negative only in the case when the dataset contains only documents that should be included ($|\mathcal{E}| = 0$). These properties of maximum and

minimum values of WSS mean that this measure does not fulfil the zero and maximum Axiom #3 proposed by Busin and Mizzaro [29] (see Section 3.5.1 of Chapter 3 for an overview of the evaluation metric axioms).

5.3.5 Evaluation with cross-validation

Most of the automated citation screening models require some seed of manually labelled documents to train the machine learning model, which can rank or predict the category of remaining documents. This assumes preparation of the training set, i.e., manually annotating documents for their eligibility. In previous work, evaluation was usually done using stratified 5×2-fold cross-validation that splits the dataset into two equally sized subsets with an even distribution of label classes which are subsequently used to train and test the model [166, 37, 99, 124, 262, 130]. The actual work saved would be measured on the second half of the initial dataset. Effectively, in the example dataset and when using 5×2-fold cross-validation, there would be total of $|\mathcal{N}| = 1000$ documents for the evaluation with WSS, out of which 100 includes $\mathcal{I}$ and 900 excludes $\mathcal{E}$.

This approach implies another practical consideration with the $(1 - r)$ term in the WSS measure. If in the dataset the total number of includes $\mathcal{I}$ is small, such that for a specific level of Recall r, $(1 - r)\%$ of relevant items would be fewer than one document (i.e., $|\mathcal{I}| \cdot (1 - r) < 1$), the number of false negatives will be equal to 0 for all Recalls $\geq r$. Thus, the following equation holds:

$$WSS@r\% = WSS@100\% - (1 - r). \tag{5.28}$$

This means that even when comparing WSS scores for different levels of Recall r, they will differ only by the constant $(1 - r)$ term, and it does not depend on the total number of documents N. If a dataset contains 20 relevant documents, a Recall = 95% means that 19 of them were successfully identified and only one document is a False Negative. Therefore, for WSS@95%, the equation above is true for all datasets where the total number of relevant documents used in the evaluation is fewer than 20 ($|\mathcal{I}| < 20$).

Moreover, when a common practice of using stratified 5×2-fold cross-validation is applied to evaluating a model, and one only calculates the scores on half of the dataset, this, in practice, means that the total size of includes in the dataset for which this equation holds is twice as high (40 relevant examples in the case of $r = 95\%$). From our analysis of 23 commonly used benchmark datasets [130], five have less than 40 includes in total (three of these datasets have even less than 20 includes). This means that there is no difference if one evaluates the same model at 95% or 100% Recall, as these two scores will always only differ by 0.05 for the dataset considered.

5.4 The Normalised WSS

We first present two formulations of the min-max normalisation of the WSS measure. Then we present benchmark results with normalised WSS and highlight the importance of this normalisation.

5.4.1 Min-max normalisation of WSS

As was done in the case of the DCG metric [105], we propose to normalise the WSS metric. As for the nDCG, the normalised WSS will allow for comparison across multiple models and benchmark systematic review datasets. The approach is presented below:

$$nWSS@r\% = \frac{WSS@r\% - min(WSS@r\%)}{max(WSS@r\%) - min(WSS@r\%)} \tag{5.29}$$

With the assumptions from the previous section, we further formulate the equation as:

$$\begin{aligned} nWSS@r\% &= \frac{(TN + \lfloor |\mathcal{I}| \cdot (1-r) \rfloor)/N - \cancel{(1-r)} - \lfloor |\mathcal{I}| \cdot (1-r) \rfloor/N + \cancel{(1-r)}}{(|\mathcal{E}| + \lfloor |\mathcal{I}| \cdot (1-r) \rfloor)/N - \cancel{(1-r)} - \lfloor |\mathcal{I}| \cdot (1-r) \rfloor/N + \cancel{(1-r)}} \\ &= \frac{(TN + \lfloor |\mathcal{I}| \cdot (1-r) \rfloor)/\cancel{\mathcal{N}} - \lfloor |\mathcal{I}| \cdot (1-r) \rfloor/\cancel{\mathcal{N}}}{(|\mathcal{E}| + \lfloor |\mathcal{I}| \cdot (1-r) \rfloor)/\cancel{\mathcal{N}} - \lfloor |\mathcal{I}| \cdot (1-r) \rfloor/\cancel{\mathcal{N}}} \\ &= \frac{TN + \cancel{\lfloor |\mathcal{I}| \cdot (1-r) \rfloor} - \cancel{\lfloor |\mathcal{I}| \cdot (1-r) \rfloor}}{|\mathcal{E}| + \cancel{\lfloor |\mathcal{I}| \cdot (1-r) \rfloor} - \cancel{\lfloor |\mathcal{I}| \cdot (1-r) \rfloor}} \\ &= \frac{TN}{|\mathcal{E}|} \end{aligned} \tag{5.30}$$

Applying this normalisation makes all the constant terms of WSS (FN and $(1-r)$) cancel themselves. The nWSS score for every dataset is always in the range $[0, 1]$. An ideal score is achieved when all the excluded documents are classified as true negatives, and then the nWSS is equal to 1. Conversely, when all the documents that should be excluded are classified incorrectly, $TN = 0$ and thus nWSS $= 0$.

In the case of a Recall threshold at 95%, the nWSS equation is:

$$nWSS@95\% = \frac{TN@95\%}{|\mathcal{E}|}, \tag{5.31}$$

meaning that we only need to estimate the number of true negatives produced by a ranking/classification model when it achieves 95% Recall.

Furthermore, as $|\mathcal{E}|$ is equal to all the negatives that should be excluded, i.e., $|\mathcal{E}| = TN + FP$, this allows us to produce another version of the nWSS:

$$nWSS = \frac{TN}{TN + FP}, \tag{5.32}$$

which is equal to the True Negative Rate (TNR), also known as specificity. This means that nWSS@r% is equal to specificity at a Recall rate of r% (S@r%).

$$nWSS@r\% = TNR@r\% = \frac{TN@r\%}{|\mathcal{E}|}, \tag{5.33}$$

5.4.2 Alternative demonstration for rank-based evaluation

Here we propose an alternative demonstration that uses rank-based evaluation terms. We assume that $n_{r\%}$ is the rank of the last manually screened document in the ordered dataset so as to achieve $r\%$ of Recall. $TN + FN$ is thus equal to $N - n_{r\%}$, and we can then re-write the WSS equations as follows:

$$WSS@r\% = \frac{TN + FN}{N} - (1 - r) = \frac{N - n_{r\%}}{N} - (1 - r). \tag{5.34}$$

In this equation, both N and r are fixed, and the only model and dataset-dependent parameter is $n_{r\%}$. The minimum value of WSS is when the rank is the lowest possible (only $(1 - r)$ of relevant documents were still not seen): $n_{r\%} = N - (1 - r) \cdot |\mathcal{I}|$. The maximum value of WSS is when the rank is equal to $r\%$ of relevant documents: $n_{r\%} = r \cdot |\mathcal{I}|$. We can write the minimum as:

$$\begin{aligned} min(WSS@r\%) &= \frac{N - (N - (1 - r) \cdot |\mathcal{I}|)}{N} - (1 - r) \\ min(WSS@r\%) &= \frac{(1 - r) \cdot |\mathcal{I}|}{N} - (1 - r), \end{aligned} \tag{5.35}$$

and the maximum as:

$$\begin{aligned} max(WSS@r\%) &= \frac{N - r \cdot |\mathcal{I}|}{N} - (1 - r) \\ max(WSS@r\%) &= \frac{(|\mathcal{E}| + |\mathcal{I}|) - r \cdot |\mathcal{I}|}{N} - (1 - r) \\ max(WSS@r\%) &= \frac{|\mathcal{E}| + (1 - r) \cdot |\mathcal{I}|}{N} - (1 - r). \end{aligned} \tag{5.36}$$

We can then write the formula for normalised WSS@r% using document ranking terms:

$$\begin{aligned} nWSS@r\% &= \frac{\frac{N - n_{r\%}}{N} - \cancel{(1 - r)} - \frac{(1-r)\cdot|\mathcal{I}|}{N} + \cancel{(1 - r)}}{\frac{|\mathcal{E}|+(1-r)\cdot|\mathcal{I}|}{N} - \cancel{(1 - r)} - \frac{(1-r)\cdot|\mathcal{I}|}{N} + \cancel{(1 - r)}} \\ nWSS@r\% &= \frac{\frac{N - n_{r\%}}{\cancel{N}} - \frac{(1-r)\cdot|\mathcal{I}|}{\cancel{N}}}{\frac{|\mathcal{E}|+(1-r)\cdot|\mathcal{I}|}{\cancel{N}} - \frac{(1-r)\cdot|\mathcal{I}|}{\cancel{N}}} \\ nWSS@r\% &= \frac{N - n_{r\%} - (1 - r) \cdot |\mathcal{I}|}{|\mathcal{E}| + \cancel{(1 - r) \cdot |\mathcal{I}|} - \cancel{(1 - r) \cdot |\mathcal{I}|}} \\ nWSS@r\% &= \frac{N - n_{r\%} - (1 - r) \cdot |\mathcal{I}|}{|\mathcal{E}|} \\ nWSS@r\% &= \frac{|\mathcal{E}| + r \cdot |\mathcal{I}| - n_{r\%}}{|\mathcal{E}|}. \end{aligned} \tag{5.37}$$

Equation 5.37 is the rank-based version of the nWSS equation. Furthermore, if we substitute the rank-based terms with confusion matrix terms ($n_{r\%} = TP + FP$), we can show that this formula is identical to Equation 5.32:

$$
\begin{aligned}
nWSS@r\% &= \frac{|\mathcal{E}| + r \cdot |\mathcal{I}| - n_{r\%}}{|\mathcal{E}|} \\
nWSS@r\% &= \frac{(TN + \cancel{FP}) + r \cdot |\mathcal{I}| - (TP + \cancel{FP})}{|\mathcal{E}|} \\
nWSS@r\% &= \frac{TN + r \cdot |\mathcal{I}| - TP}{|\mathcal{E}|} \\
nWSS@r\% &= \frac{TN + \cancel{TP} - \cancel{TP}}{|\mathcal{E}|} \\
nWSS@r\% &= \frac{TN}{TN + FP}.
\end{aligned} \tag{5.38}
$$

5.4.3 Benchmark results with $TNR@95\%$

In this section, we compare the WSS and TNR scores in order to demonstrate the problems with using WSS for evaluation. We compare the results of seven different models on fifteen citation screening datasets from Cohen et al. [35]. We are interested in establishing the rank of each model based on the average WSS and TNR scores. We take previous benchmark results from the literature for seven distinct models. We used Equation 5.29 to convert WSS@95% scores reported by previous studies to the TNR@95% scores. The performance of past models evaluated with TNR@95% and WSS@95% is presented in Table 5.1.

When comparing the model's performance using averaged TNR against the average WSS from these 15 datasets, we observe changes in the model's ranking. When ordered by their average WSS score, models from best to lowest score are D, E, C, G, **F**, **B** and A. However, when evaluated with TNR, the order is the following: D, E, C, G, **B**, **F** and A. Hence, with only seven models, we have already noticed that the incorrect usage of WSS to compare averaged performance across several datasets proved to yield an erroneous order of models.

5.5 Additional Evaluation Measures at Specific Recall Levels

In this section, we present normalised versions of standard evaluation measures at specific Recall cutoffs and a variation of true negative rate, penalising models performing worse than random sampling. These refinements are required for accurately assessing model performance in citation screening tasks over several datasets. These metrics can be important for researchers and practitioners in information retrieval and machine learning as they offer more accurate and context-sensitive tools for assessing model performance.

Table 5.1: Evaluation results with WSS and TNR at 95% Recall on systematic review datasets from Cohen et al. [35] described in Chapter 3. The following models were used: A: Cohen et al. [35], B: Matwin et al. [166], C: Cohen [36], D: Howard et al. [99], E: Kontonatsios et al. [124], F: van Dinter et al. [262], G: Kusa et al. [130]. **Bold** indicates highest score.

No	Datset name	Dataset size	Percentage of includes	Models						
WSS@95%				A	B	C	D	E	F	G
1	ACEInhibitors	2,544	1.6%	0.566	0.523	0.733	**0.801**	0.787	0.783	0.783
2	ADHD	851	2.4%	0.680	0.622	0.526	**0.793**	0.665	0.698	0.424
3	Antihistamines	310	5.2%	0.000	0.149	0.236	0.137	**0.310**	0.168	0.047
4	Atypical Antipsychotics	1,120	13.0%	0.141	0.206	0.170	0.251	**0.329**	0.212	0.218
5	Beta Blockers	2,072	2.0%	0.284	0.367	0.465	0.428	**0.587**	0.504	0.419
6	Calcium Channel Blockers	1,218	8.2%	0.122	0.234	0.430	**0.448**	0.424	0.159	0.178
7	Estrogens	368	21.7%	0.183	0.375	0.414	**0.471**	0.397	0.119	0.306
8	NSAIDs	393	10.4%	0.497	0.528	0.672	**0.730**	0.723	0.571	0.620
9	Opioids	1,915	0.8%	0.133	0.554	0.364	**0.826**	0.533	0.295	0.559
10	Oral Hypoglycemics	503	27.0%	0.090	0.085	**0.136**	0.117	0.095	0.065	0.098
11	Proton PumpInhibitors	1,333	3.8%	0.277	0.229	0.328	0.378	**0.400**	0.243	0.283
12	Skeletal Muscle Relaxants	1,643	0.5%	0.000	0.265	0.374	**0.556**	0.286	0.229	0.090
13	Statins	3,465	2.5%	0.247	0.315	0.491	0.435	**0.566**	0.443	0.409
14	Triptans	671	3.6%	0.034	0.274	0.346	0.412	**0.434**	0.266	0.210
15	Urinary Incontinence	327	12.2%	0.261	0.296	0.432	0.531	**0.531**	0.272	0.439
	Average WSS@95% score			0.2343	0.3348	0.4078	**0.4876**	0.4711	0.3351	0.3388
	Rank based on average WSS@95% score			7	6	3	1	2	5	4

TNR@95%		A	B	C	D	E	F	G
1	ACEInhibitors	0.625	0.582	0.795	**0.864**	0.850	0.846	0.846
2	ADHD	0.746	0.687	0.589	**0.862**	0.731	0.765	0.484
3	Antihistamines	0.053	0.210	0.302	0.197	**0.380**	0.230	0.102
4	Atypical Antipsychotics	0.212	0.287	0.246	0.339	**0.429**	0.294	0.301
5	Beta Blockers	0.340	0.425	0.525	0.487	**0.649**	0.564	0.478
6	Calcium Channel Blockers	0.183	0.305	0.518	**0.538**	0.512	0.223	0.244
7	Estrogens	0.284	0.529	0.579	**0.652**	0.557	0.202	0.441
8	NSAIDs	0.605	0.640	0.800	**0.865**	0.857	0.688	0.742
9	Opioids	0.184	0.609	0.417	**0.883**	0.588	0.348	0.614
10	Oral Hypoglycemics	0.176	0.169	**0.239**	0.213	0.182	0.141	0.186
11	Proton PumpInhibitors	0.338	0.289	0.391	0.443	**0.466**	0.303	0.345
12	Skeletal Muscle Relaxants	0.050	0.317	0.426	**0.609**	0.338	0.281	0.141
13	Statins	0.303	0.373	0.553	0.496	**0.630**	0.504	0.469
14	Triptans	0.086	0.334	0.409	0.478	**0.500**	0.326	0.268
15	Urinary Incontinence	0.347	0.387	0.542	**0.655**	**0.655**	0.360	0.550
	Average TNR@95% score	0.3022	0.4094	0.4888	**0.5721**	0.5550	0.4050	0.4141
	Rank based on average TNR@95% score	7	5	3	1	2	6	4

5.5.1 Precision and F_{beta} at a Recall cutoff

The problem with evaluating models at a specific Recall value is that most measures will not be bounded between [0, 1]. Therefore, their min and max values depend on the class imbalance, and these measures do not fulfil the Axiom #3 proposed by Busin and

Mizzaro [29]. This leads to problems in comparing performance across different datasets and makes it difficult to assess the true performance of models.

Besides WSS, previous studies also used Precision at a fixed Recall cutoff ($P@r\%$) to evaluate screening prioritisation algorithms [124]. Using the analysis from Section 5.3.3 that for a specific r% Recall, TP will be equal to $r \cdot |\mathcal{I}|$, we can define a minimum and maximum Precision values as follows:

$$max(Precision@r\%) = \frac{r \cdot |\mathcal{I}|}{r \cdot |\mathcal{I}| + 0} = 1, \tag{5.39}$$

$$min(Precision@r\%) = \frac{r \cdot |\mathcal{I}|}{r \cdot |\mathcal{I}| + \mathcal{E}}. \tag{5.40}$$

We can observe that the maximum Precision@r% value will always be equal to 1. However, the minimum Precision value, similarly to WSS, depends on the $\mathcal{I}/\mathcal{E}$ ratio of the dataset:

$$\lim_{|\mathcal{E}|\to 0} min(Precision@r\%) = \lim_{|\mathcal{E}|\to 0} \frac{r \cdot |\mathcal{I}|}{r \cdot |\mathcal{I}| + \mathcal{E}} = 1, \tag{5.41}$$

$$\lim_{|\mathcal{I}|\to 0} min(Precision@r\%) = \lim_{|\mathcal{I}|\to 0} \frac{r \cdot |\mathcal{I}|}{r \cdot |\mathcal{I}| + \mathcal{E}} = 0. \tag{5.42}$$

Therefore, we define min-max normalised version of $Precision@r\%$ ($nP@r\%$) as:

$$\begin{aligned}
nP@r\% &= \frac{\frac{TP}{TP+FP} - \frac{TP}{TP+|\mathcal{E}|}}{1 - \frac{TP}{TP+\mathcal{E}}} \\
nP@r\% &= \frac{\Big(TP \cdot (TP + |\mathcal{E}|) - TP \cdot (TP + FP)\Big) / \Big((TP + FP) \cdot (TP + |\mathcal{E}|)\Big)}{\Big(\cancel{TP} + |\mathcal{E}| - \cancel{TP}\Big) / \Big(TP + |\mathcal{E}|\Big)} \\
nP@r\% &= \frac{TP \cdot |\mathcal{E}| - TP \cdot FP}{(TP + FP) \cdot \cancel{(TP + |\mathcal{E}|)}} \cdot \frac{\cancel{(TP + |\mathcal{E}|)}}{|\mathcal{E}|} \\
nP@r\% &= \frac{TP \cdot (|\mathcal{E}| - FP)}{(TP + FP) \cdot |\mathcal{E}|} \\
nP@r\% &= \frac{TP \cdot TN}{|\mathcal{E}| \cdot (TN + FP)}.
\end{aligned} \tag{5.43}$$

Similar generalised equation of its normalised version can be proposed for $F_{beta}@r\%$ measure:

$$nF_{beta}@r\% = \frac{(r+\beta^2)\cdot|\mathcal{I}|\cdot TN}{|\mathcal{E}|\cdot(r\cdot|\mathcal{I}|+\beta^2\cdot|\mathcal{I}|+FP)}, \tag{5.44}$$

and specific examples for F_1-score, $F_{0.5}$-score and F_3-score are presented in equations below:

$$nF_1@r\% = \frac{(r+1)\cdot|\mathcal{I}|\cdot TN}{|\mathcal{E}|\cdot(r\cdot|\mathcal{I}|+|\mathcal{I}|+FP)} \tag{5.45}$$

$$nF_{0.5}@r\% = \frac{(r+0.25)\cdot|\mathcal{I}|\cdot TN}{|\mathcal{E}|\cdot((r+0.25)\cdot|\mathcal{I}|+FP)} \tag{5.46}$$

$$nF_3@r\% = \frac{(r+9)\cdot|\mathcal{I}|\cdot TN}{|\mathcal{E}|\cdot((r+9)\cdot|\mathcal{I}|+FP)} \tag{5.47}$$

Using these normalised equations ensures that the process of averaging scores from multiple reviews (topics) maintains mathematical rigour. This approach contrasts with the methodologically inadequate practice of averaging non-normalised scores, which results in difficulties in model comparisons, misleading interpretations and inconsistencies in evaluation.

5.5.2 Rectified TNR

Additionally, we introduce a new evaluation measure for analysis: rectified True Negative Rate ($reTNR$) and its min-max normalised version ($nreTNR$):

$$reTNR@r\% = \begin{cases} TNR@r\%, & \text{if } \frac{FP@r\%}{\mathcal{E}} < r\% \\ \frac{TN@r\%}{\mathcal{E}}, & \text{otherwise} \end{cases} \tag{5.48}$$

$$nreTNR@r\% = \frac{reTNR - \min(reTNR)}{\max(reTNR) - \min(reTNR)} \tag{5.49}$$

$reTNR$ penalise models which perform worse than a random ordering of the documents, i.e., when, for a given $r\%$ of Recall, the true negative rate is lower than $(1-r)$, $reTNR$ score is equal to the $(1-r)$. This threshold is equal to a simple random sampling, as savings of $(1-r)\%$ are achieved when, on average, $r\%$ of the dataset is randomly sampled. An intuition for this measure is that all models performing worse than random sorting are equally bad, and, especially when averaging scores, they should not influence the actual work saved (see Figure 5.1). This follows up on the original intuition of the WSS metric introduced by Cohen et al. [35], which obtains zero value for a model performance equal to a random sampling.

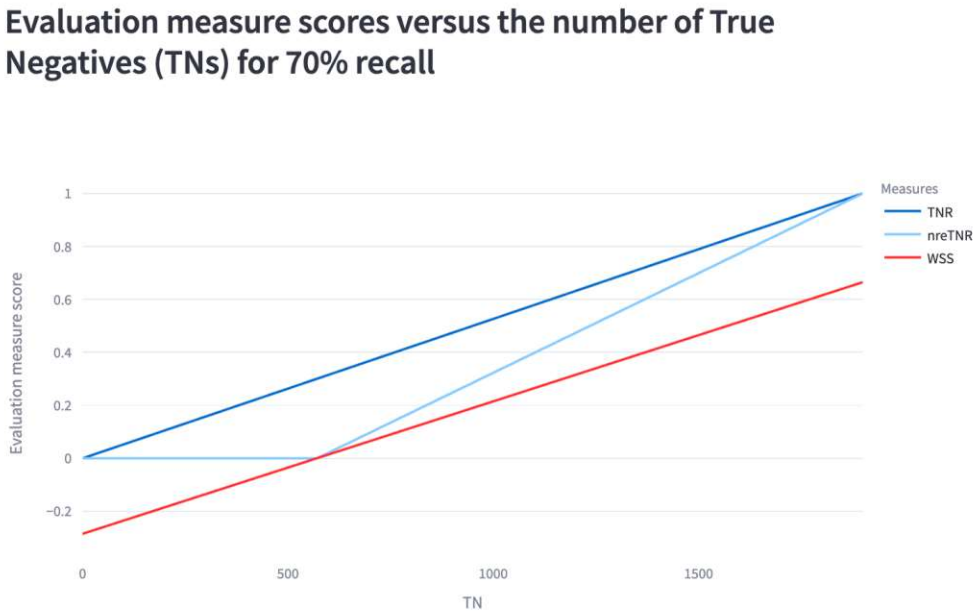


Figure 5.1: Plot presenting a TNR, nreTNR and WSS behaviour for a custom dataset containing $N = 2000$ documents out of which 5% are relevant ($|\mathcal{I}| = 100$). Visualisation shows how the number of detected true negatives (TNs) influences the score of each evaluation measure. Evaluations conducted at a Recall cutoff = 70%.

5.6 VoMBaT Visual Analytics Tool

The dynamics between the values of true negatives and Recall can be difficult to comprehend when using measures such as $WSS@r\%$, $TNR@r\%$ or $Precision@r\%$. These measures do not provide a clear understanding of the number of true negatives found by the model and the time saved as a result. Additionally, it can be challenging to translate a particular score of these measures into real time and money benefits. The complexity of these measures can pose a challenge for practitioners in effectively utilising TAR systems and accurately evaluating their performance.

To this end, we developed VoMBaT (**V**isualisation **o**f **M**easure **B**eh**a**viour for **T**AR) – a visual analytics toolbox focusing on high-recall evaluation scenario. We implemented fifteen evaluation measures based on the confusion matrix terms described in the previous Sections: $Precision$, $Accuracy$, $Balanced\ Accuracy$, $F_1 - score$, $F_3 - score$, $F_{0.5} - score$, TNR, WSS, MCC, FDR, NPV, FOR, DFR and DOR (Equations 5.1–5.20).

The tool we developed offers an interface to compare different evaluation measures, providing insights into their impact. The tool does not compare scores from actual runs but instead takes only two dataset parameters: dataset size and a percentage of relevant documents in the dataset, making it domain agnostic and applicable for many TAR applications. The target users for the tool are researchers, practitioners and other

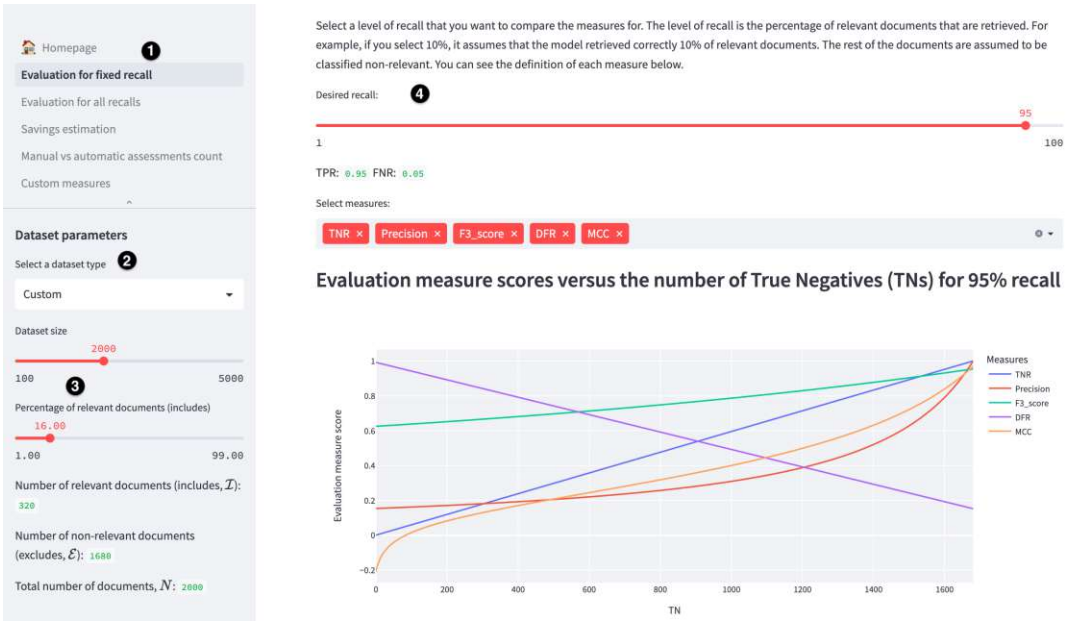


Figure 5.2: The VoMBaT page for comparing several evaluation measures at a fixed Recall level. Users can navigate to other pages and select dataset parameters using the sidebar on the left, while Recall level and evaluation measure selection are on the top.

stakeholders involved in high-recall search tasks who want to understand the evaluation measures better. These users may include data scientists, machine learning engineers, legal professionals, and academic researchers. Additionally, the tool can be used to help users in their decision-making process about the quality of TAR models and to evaluate the potential savings in time and resources in a variety of settings.

Our tool is made using Python 3.10, Plotly and Streamlit. VoMBaT is available as an open-source package[1] under the Apache-2.0 license and the demo is available under the following URL.[2] The interface consists of five subpages described in detail in this section.

5.6.1 Interface

The navigation between five pages is implemented using a sidebar on the left-hand side of the screen (❶ on Figure 5.2). A set of predefined dataset parameters (the total number of documents N and a percentage of relevant documents $\mathcal{I}$) was prepared for each of these pages ❷. Users can also define custom dataset size and a percentage of relevant documents ❸. There are two types of predefined datasets:

- Three synthetic examples of dataset parameters showing extreme options for the distribution of relevant documents ($\mathcal{I}$) in the dataset: balanced, heavily unbalanced

[1] https://github.com/WojciechKusa/VoMBaT
[2] https://vombat.streamlit.app

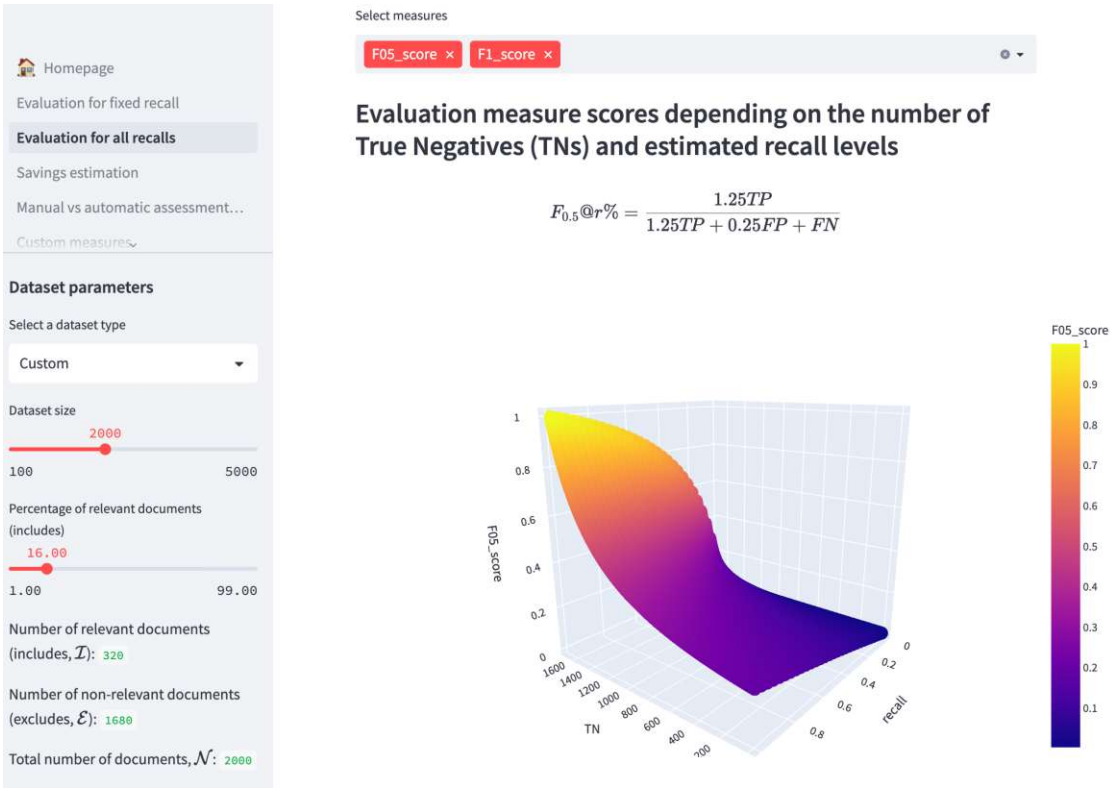


Figure 5.3: The page for comparing one evaluation measure across all Recall levels. Users can select dataset parameters using the sidebar on the left.

towards positive class (example of a very good search query), and heavily unbalanced towards negative class (very typical in systematic reviews).

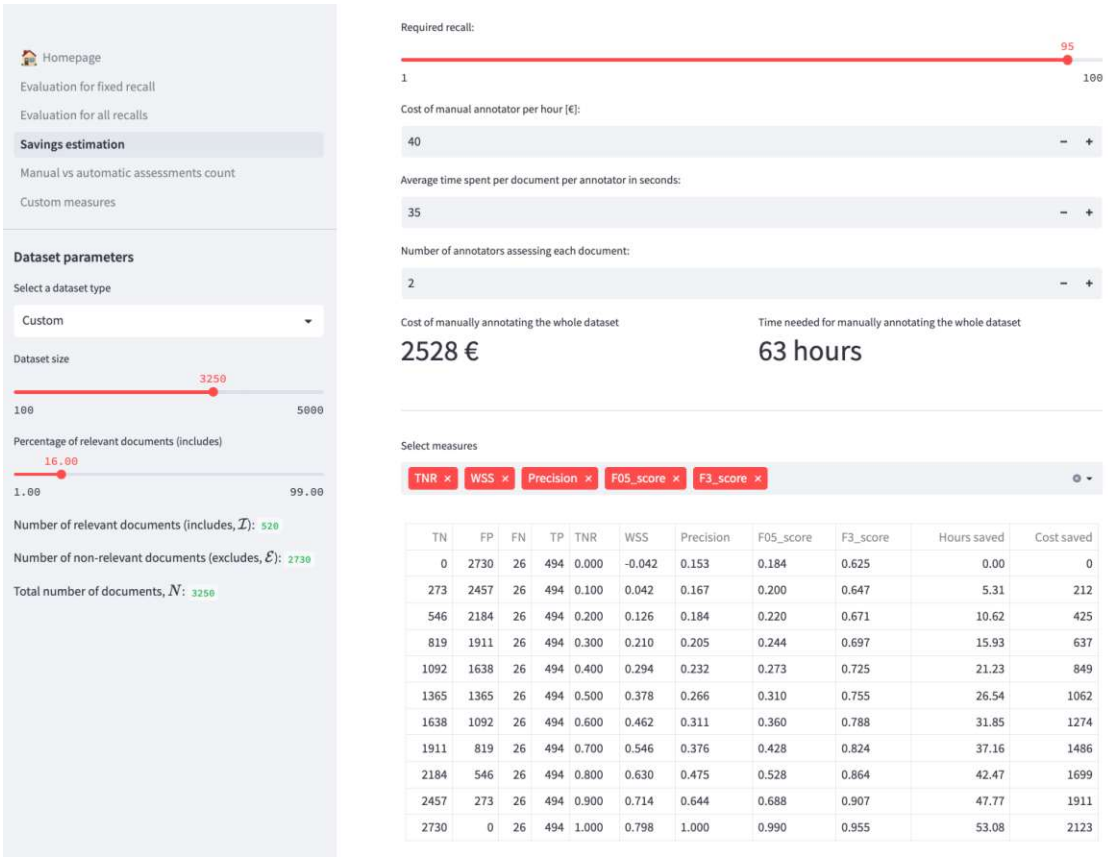
- Fifteen datasets which use the N and $\mathcal{I}$ values from systematic reviews in the field of medicine introduced by Cohen et al. [35].

Evaluation for a fixed Recall level

This page presents a comparison of evaluation measures for a fixed level of Recall (❹ on Figure 5.2). Users need to select a level of Recall to compare the measures first. The level of Recall is the percentage of relevant documents that are retrieved. For example, if one selects 10%, it assumes that the model retrieved 10% of relevant documents correctly. The rest of the documents are assumed to be classified as non-relevant.

Evaluation for all Recall levels

This page presents 3D plots of possible evaluation measure scores for all Recall and TN levels (Figure 5.3). First, up to four measures from the set of predefined ones can be selected. Each measure is plotted in a separate interactive 3D plot, and the x-axis and



TN	FP	FN	TP	TNR	WSS	Precision	F05_score	F3_score	Hours saved	Cost saved
0	2730	26	494	0.000	-0.042	0.153	0.184	0.625	0.00	0
273	2457	26	494	0.100	0.042	0.167	0.200	0.647	5.31	212
546	2184	26	494	0.200	0.126	0.184	0.220	0.671	10.62	425
819	1911	26	494	0.300	0.210	0.205	0.244	0.697	15.93	637
1092	1638	26	494	0.400	0.294	0.232	0.273	0.725	21.23	849
1365	1365	26	494	0.500	0.378	0.266	0.310	0.755	26.54	1062
1638	1092	26	494	0.600	0.462	0.311	0.360	0.788	31.85	1274
1911	819	26	494	0.700	0.546	0.376	0.428	0.824	37.16	1486
2184	546	26	494	0.800	0.630	0.475	0.528	0.864	42.47	1699
2457	273	26	494	0.900	0.714	0.644	0.688	0.907	47.77	1911
2730	0	26	494	1.000	0.798	1.000	0.990	0.955	53.08	2123

Figure 5.4: The page for estimating time and money savings that can be achieved depending on the value of evaluation measures. Users can select dataset parameters using the sidebar on the left.

y-axis represent the number of TNs and the estimated Recall level, respectively. The score of the selected measure is presented on the z-axis.

Savings estimation

This page presents the simulation of time and money savings that can be achieved depending on the value of evaluation measures (Figure 5.4). Users can use this simulation to determine the minimum threshold for the evaluation measures that can be accepted in order to reduce the manual screening time and the cost of the evaluation. Users can adjust factors such as the average time per document, the number of manual assessments per document, and the cost of annotators. During the manual document review, each document is assessed by an annotator. Furthermore, in the case of systematic literature reviews in medicine, each document is screened by at least two people. Savings can be achieved when the model's automatic assessments are accurate enough to replace manual checks for certain documents, effectively eliminating True Negatives. The more TNs the

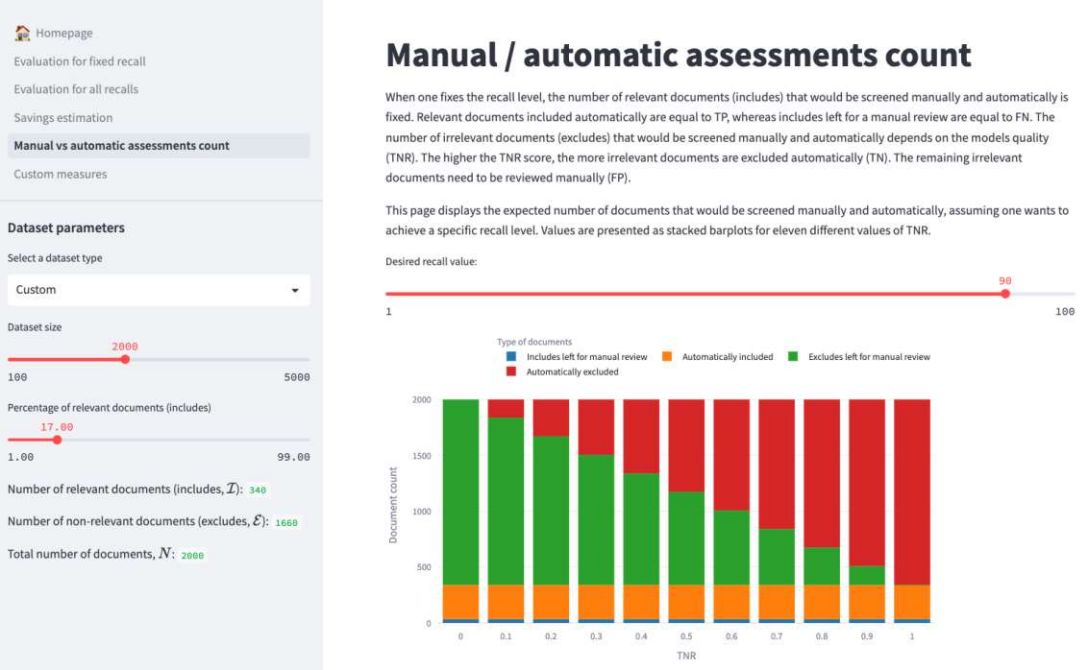


Figure 5.5: The page for comparing manual and automatic assessments count depending on the TNR score. Users can select dataset parameters using the sidebar on the left.

model can remove, the greater the potential for cost and time savings.

Manual vs automatic assessment comparison

This page assumes that the number of relevant and irrelevant documents to be reviewed manually or automatically is fixed once the desired Recall level is established (Figure 5.5). The relevant documents included in the automatic assessment are equal to the true positives. In contrast, the remaining relevant documents that need to be reviewed manually are the false negatives. The number of irrelevant documents that will be reviewed automatically or manually depends on the model's TNR score. The higher the TNR score, the more irrelevant documents will be automatically excluded, representing the true negatives. The remaining irrelevant documents will need to be reviewed manually, which are the false positives (FP). This page provides a visual representation of the expected number of documents that will be reviewed automatically or manually based on a specified Recall level. The values are presented as stacked bar plots for eleven different TNR scores.

Custom measures

Finally, we allow users to write and test custom evaluation measures using confusion matrix terms as building blocks (Figure 5.6). The equation written by user in the text box is converted to Python code using reverse polish notation to support basic mathematical

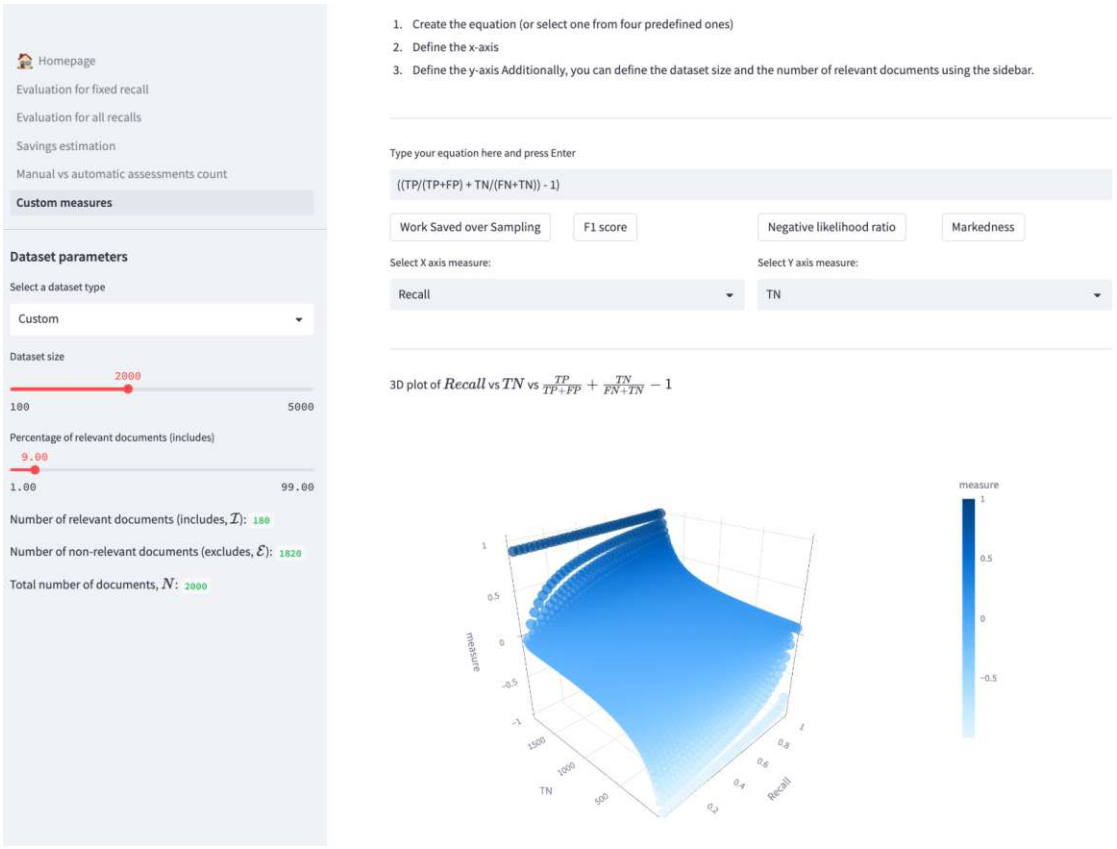


Figure 5.6: The page for comparing custom evaluation measures. Users can select dataset parameters using the sidebar on the left.

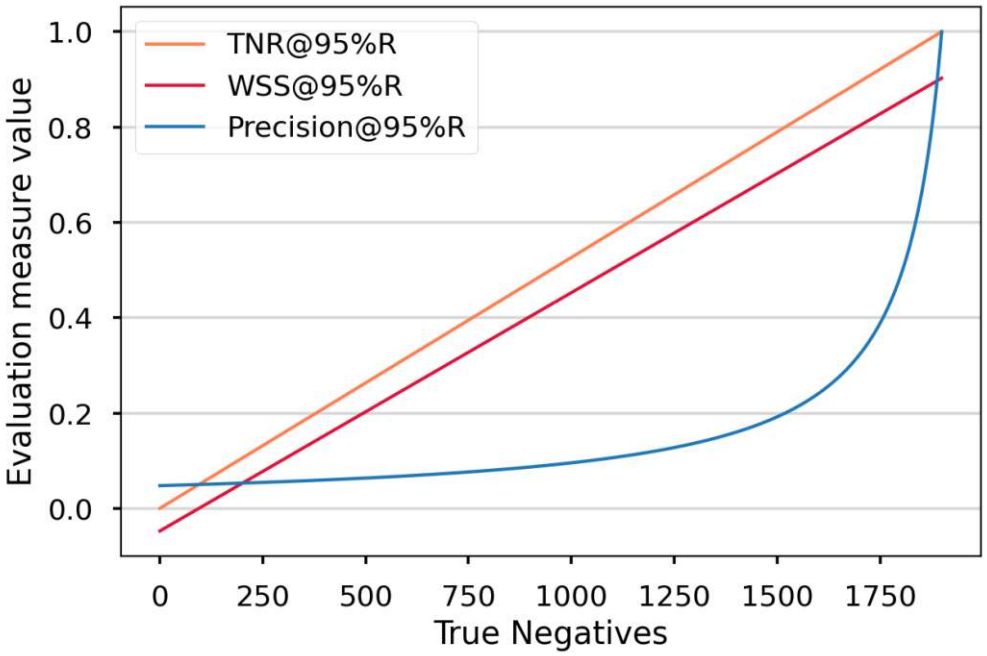
operations. The user has the option to select two variables (by default, it is Recall and TN, as it is on other pages) which will be plotted for comparison with the evaluation measure. The interface is similar to the *Evaluation for all Recalls* page.

5.7 Discussion

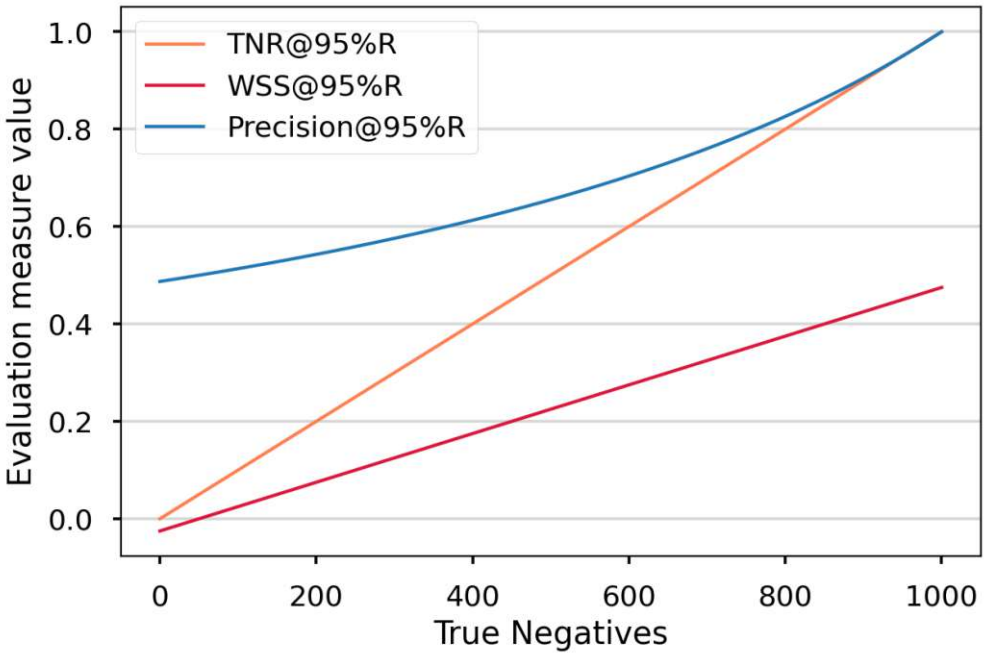
We first compare WSS and TNR to Precision and AUC, two other commonly used evaluation measures in citation screening automation. We then discuss the ability to model cost savings when using these measures. Finally, we present additional limitations of current evaluation measures.

5.7.1 Comparison with Precision

Figure 5.7 presents the dynamic of evaluation measures' scores as a function of the number of true negatives detected by an algorithm for a fixed Recall level of 95%. We consider two types of datasets having the same total number of documents $N = 2000$ but



(a) Evaluation measures' scores versus the number of True Negatives for an imbalanced dataset with 5% of positive examples ($|\mathcal{I}| = 100$, $|\mathcal{E}| = 1900$).



(b) Evaluation measures' scores versus the number of True Negatives for a perfectly balanced dataset ($|\mathcal{I}| = |\mathcal{E}| = 1000$).

Figure 5.7: Dynamics of evaluation measures (WSS, TNR (nWSS) and Precision) scores as a function of the number of True Negatives (TN) at 95% Recall for two sample datasets.

differing in the $|\mathcal{I}|/|\mathcal{E}|$ ratio: heavily imbalanced towards the negative class with only 5% of positive examples (Figure 5.7a), and perfectly balanced dataset (Figure 5.7b). On both datasets, WSS and TNR scores rise linearly with the rising number of true negatives detected by the algorithm, but a change in the Precision scores is not linear, and its derivative depends on the class imbalance. In addition, out of these three measures, only TNR is always bounded by 0 and 1. Again, minimum Precision value depends on the class imbalance, which for WSS is the case for both minimum and maximum values.

TNR score can also be directly translated to the number of documents reviewers do not need to screen manually. Furthermore, when used with appropriate multipliers, assuming all documents are equally time-consuming to screen, one can convert the TNR score into the time and money saved by using automation tools.

5.7.2 Comparison with AUC

As already mentioned, measures like ROC or Precision-Recall curve are more suitable for comparing a model's effectiveness across multiple Recall levels. However, they do not allow for automatic comparisons across multiple models and are not suitable for score aggregations across several datasets. Fawcett [65] mentions that even though ROC curves may be used to evaluate classifiers, care should be taken when using them to conclude classifier superiority.

Figure 5.8 presents ROC curves and corresponding AUC scores for two hypothetical models on the same dataset. Model A, which obtains a higher AUC score, quickly achieves >60% Recall, but its score plateaus and only manages to exceed Recall of 80% at the very end. On the other hand, model B, which "struggles" initially but reaches perfect Recall at an FPR level of 0.35, obtains a lower AUC score. For the general search task, model A might be more suitable. However, for technology-assisted reviews where we want to ensure that the model achieves very high Recall (and even in the case of rapid reviews or e-discovery, this should very rarely be lower than 70%), model B is the only one which delivers some gain to the user.

Hence, we believe that compared to TNR, AUC scores can favour models that achieve good Recall scores at low values of FPR, which are of no value for citation screening tasks. An alternative can be to calculate partial AUC score (pAUC), a practice for highly sensitive diagnostic tests [167, 106]. Similarly to the TNR@r% and WSS@r% calculations, one could parameterise AUC by the desired minimum Recall (TPR) level. Then, the pAUC is computed in the part of the ROC space where the Recall is greater than a given threshold r.

5.7.3 Cost savings

Based on the analysis of the plots, it is apparent that there are two main types of evaluation measures relevant to this task. Measures such as WSS, DFR, and TNR are linearly correlated with the number of TN and FP predicted by the algorithm. On the other hand, $Precision$ (and analogically, F_β–score) have different, non-linear characteristics.

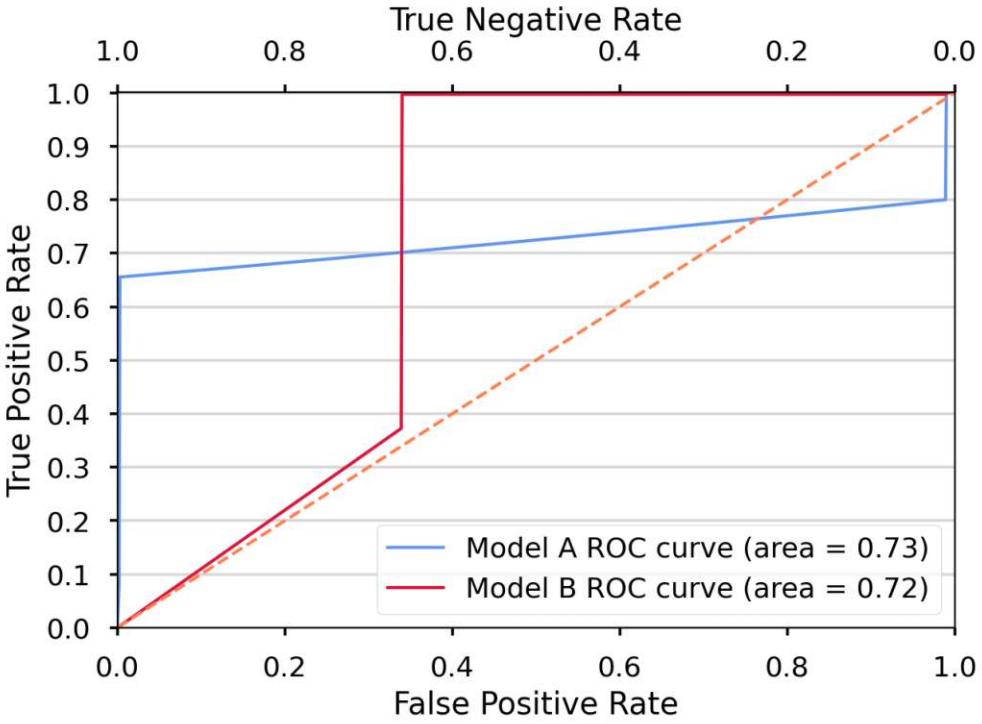


Figure 5.8: Receiver Operating Characteristic (ROC) curves for two hypothetical models with their corresponding AUC scores. Model A achieves a higher value of AUC, despite the fact that its TPR performance reaches 80% only at the FPR level almost equal to 100%, and model B achieves maximum Recall at FPR level of 35%.

For datasets that consist of a significant number of non-relevant documents, *Precision* values only start to increase as the number of TN increases (due to the constant value of TP and the $(\mathcal{E} - TN)$ term in the denominator).

This leads us to conclude that TNR–style measures would be more directly transferable to cost savings. However, measures focusing on *Precision* can be more useful when evaluating models for a fully automated final stage of the screening process, for instance during full text screening in systematic reviews. In this case every document successfully screened by the TAR system is of high importance. This is because, in practice, filtering the last few percent of documents can bring the most significant gains to users, as the remaining, not relevant documents can be easily screened using other techniques. This behaviour can be observed from the analysis of VoMBaT pages described in Section 5.6.1, where the TNR score grows linearly with decreasing time and cost of conducting the review[3].

[3]Under the assumption that every document takes the same amount of time for screening.

5.7.4 Limitations of current measures

WSS, TNR or Precision cannot account for the amount of manual work required to kick-start the automated screening. Current classification approaches use some type of cross-validation to train and evaluate their models. Usage of different train/test splits provides another challenge as TNR@r% (unlike *Burden* or *Utility*) does not measure the amount of data that needs to be labelled manually before training the classifier. To overcome this problem, plotting the learning curves for TNR@r% could be one way to compare the performance of these models.

5.8 Summary

This chapter analyses Work Saved over Sampling (WSS), a measure commonly used to evaluate automated citation screening models. We inspect the terms and properties of WSS and show drawbacks of the measure.

We propose min-max normalisation of Work Saved over Sampling at $r\%$ Recall (nWSS@r%). It improves on WSS as it normalises possible scores into the $[0, 1]$ range. This enables fair comparison between different models and score aggregations from multiple datasets. nWSS also simplifies over WSS as it does not contain two WSS terms that were shown to be constants by our analysis. Moreover, we show that nWSS is equal to True Negative Rate (TNR), further simplifying the understanding of the measure.

TNR has a linear correlation with the number of documents that a manual reviewer does not need to screen and can be directly translated to the time (and money) saved when using automation tools. We suggest the usage of TNR at $r\%$ of Recall as an evaluation measure for the citation screening task if the score is to be compared between multiple models across several datasets. We propose a variant of TNR: nreTNR (normalised rectified TNR), which penalises models which perform worse than a random ordering of the documents.

Furthermore, we introduce an interface to analyse and understand behaviours of evaluation measures used in a high-recall setting. We implemented a dashboard with fifteen evaluation measures, focusing on the ones used in technology-assisted review tasks. The interface enables a comparison of how these measures behave depending on specific values of Recall and true negatives. For TNR, it also provides the estimate of saving when using automatic models and a count of documents that need to be screened automatically versus manually. Our tool helps to increase the understanding of evaluation measures used in high-recall search tasks and especially TAR systems.

CHAPTER 6

Impact-based Evaluation Measures for Citation Screening

As shown in Chapters 3 and 5, current evaluation measures for automated citation screening methods in systematic literature reviews are limited to binary relevance assessment, where each publication is considered either relevant or irrelevant. These evaluation measures do not account for the influence of each publication on the review outcome. This is a vital issue, as the assumption that all relevant publications are equally important to the final outcome of the systematic review is not necessarily valid. Without an accurate assessment of the importance of each document, the conclusions of a systematic review may be biased or incomplete. To address this issue, in this chapter, we propose a novel methodology for assessing citation screening based on evaluating outcome differences, which enables us to determine the influence of each publication on the systematic review.

To understand the effectiveness of automated citation screening methods, practitioners have relied on metrics based on the notions of Recall, Precision and cost – and of a binary assessment of relevance[1] [130, 263, 190]. This practice assigns to every publication to be included in the review the same importance. So, for example, if method M_1 identifies $\{A, B, C\}$ as potentially relevant publications while method M_2 identifies publications $\{A, D, E\}$, and the ground truth is that the relevant publications are $\{A, B, D\}$, then M_1 and M_2 achieve the same Recall, Precision and cost. However, we argue, that the two sets $\{A, B, C\}$ and $\{A, D, E\}$ may not be equally important, and thus identifying either of B or D may not be equivalent if the outcomes of the review were considered. In fact, if excluded, some publications can significantly change a review's conclusion to the extent that the conclusion might be the opposite (e.g., from favouring a drug to favouring a placebo) [185, 183]. On the other hand, not including other publications might have only a small quantitative impact on the outcomes of the review.

[1] Every publication to be included in the review is labelled as relevant, while every excluded publication is non-relevant.

Nussbaumer-Streit et al. [185] compared repeated literature searches using 14 abbreviated approaches (combinations of various databases with and without searches of reference lists) on a sample of 60 Cochrane systematic reviews of clinical interventions. They re-calculated the main summary-of-findings table of each Cochrane review and asked original review authors whether the conclusions changed compared to the original review. They found that in only 2% of cases (95% CI: 0%–9%), combining one database with another or with searches of reference lists was falsely reaching an opposite conclusion compared to comprehensive searches. This outcome shows that identifying *all* relevant studies is not always crucial for obtaining the same review results.

Marshall et al. [163] presented a study exploring the potential changes in systematic review outcomes when rapid review methods are applied. By recalculating meta-analyses for the first dichotomous outcome in 2,512 Cochrane systematic reviews, they simulated the effects of using rapid review strategies, such as searching only PubMed, excluding older studies, and limiting the review to larger trials or the single largest trial. Their results highlight the variability and potential risks associated with these methods, demonstrating that changes in pooled odds ratios and statistical significance can occur frequently, depending on the rapid review approach used. Notably, the study finds that searching only PubMed presents the least risk of significant changes to the outcomes, suggesting that this method might be suitable under certain conditions, such as scoping reviews or situations requiring urgent synthesis. This study underlines the importance of considering the impact of different literature search and selection strategies on the reliability of systematic review outcomes.

Building on these insights, we propose a new evaluation framework that considers inclusion and exclusion information and meta-analysis data from reviews created by Cochrane to estimate outcomes and weights of included publications. This information can be used to assess the quality of ranking and classification methods. This framework allows for assessing automatic approaches from the angle of how closely their *outcomes* – not just their set of included publications – are to the outcomes of the original review. By comparing the outcomes of the automated model to those of the original review, we can gain a better understanding of the quality of the automated approach and its effect on the final outcome of the review.

We propose two different evaluation types: *outcome-based evaluation* and *review-based evaluation* and present experimental results for both types on the CLEF TAR 2019 dataset [114]. *Outcome-based evaluation* measures the extent to which it is possible to obtain different study outcomes from those obtained by the original study when using documents retrieved by the search results. This evaluation is expressed by five aspects of analysis focusing on different features of review outcomes. We explore initial experiments on the CLEF TAR 2019 dataset [114]. Our simulation results show that by randomly removing one publication per review (average Recall of 92% publications), 95% of outcomes remain unchanged. However, after removing five publications (average Recall of 63%), 76% of the outcomes are still the same, showing that the relationship between Recall and achieved outcomes is not linear. We also show that the outcome-based

evaluation emphasises different aspects of the models' performance than the traditional IR evaluation measures. We finally propose multi-objective optimisation to handle the problem of non-estimable outcomes.

Review-based evaluation weighs each publication retrospectively based on their independent influence on the review outcome. Such weighting can be included in traditional measures like nDCG or TNR. We present one example of how these weights can be used in the nDCG measure showing that the ordering of runs changes compared to using binary relevance judgements.

Finally, we discuss the essential limitations of each of the methodologies and suggest directions that can be explored in the future to fully operationalise our proposal. We believe that this new evaluation approach will provide a better understanding of the impact of automatic literature screening methods on the outcome of systematic literature reviews and help identify areas in which these methods can be improved.

6.1 Evaluation Framework

Our evaluation framework for automated citation screening involves four parts which are detailed in the following subsections (a graphical description is presented in Figure 6.1). The first step is data extraction, where we extract statistics of the studies included in the review and match studies to publications. The statistics contain information about outcomes and effect sizes reported in the systematic review. The second step is model evaluation, where we use the extracted data to estimate the review's outcomes for rankings or classifications of the citation list. The third step is the analysis of the results, where we compare the outcomes obtained from the alternative rankings to the outcomes of the original review. Finally, the last part involves estimating the influence of each publication based on its contribution to the outcomes of the review by considering factors like the difference in outcome when a publication is missing and the number of studies and outcomes reported by the publication. Our proposed framework allows for a more nuanced evaluation of automated citation screening methods. By considering the influence of each publication on the review's outcomes, we can identify which publications are most important to retrieve and prioritise them accordingly. Our framework can be used to weigh publications for the traditional evaluation methods.

6.1.1 Data extraction

Cochrane systematic reviews distinguish between *study* and *publication*. When conducting eligibility screening for systematic reviews, reviewers evaluate the documents on the level of publications. Each study can be reported by several publications. Each publication may present different aspects or findings of the same study, but they are all derived from the same underlying research. We assume that a study has been found if at least one publication that systematic review creators classified as reporting that study was successfully retrieved.

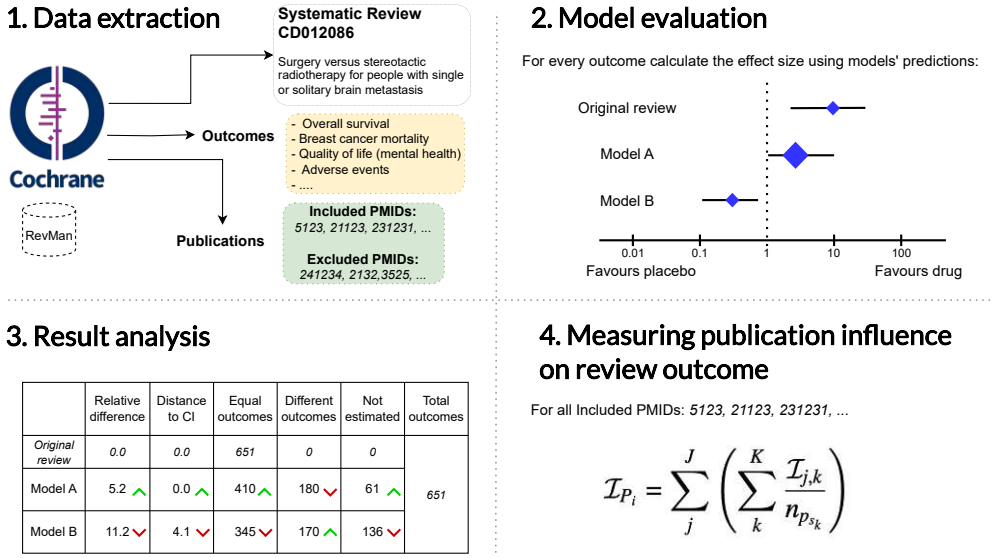


Figure 6.1: Four steps of the proposed evaluation framework.

For every review, based on its Cochrane review ID, we identify its corresponding RevMan file and list of included publications. A RevMan file is the format used by Cochrane containing all statistical data about studies and outcomes included in the review. Outcomes of Cochrane reviews are reported in the following hierarchy: one comparison can have several outcomes, and one outcome can consist of a few subgroups. We extract all metadata from the RevMan files, such as the comparisons, outcomes and subgroups and the results of every included study. Note that the use of RevMan files is for experimental convenience, but is not a requirement of the framework: the required data could be provided in other formats. Furthermore, Cochrane recently announced that future systematic literature reviews would contain statistical data in more common CSV and RIS formats.[2]

Cochrane reports a list of included publications and studies which correspond to them. Traditionally, retrieval was conducted at the level of publications [112, 113, 114]. In order to be able to re-use previous relevance judgements, we need to assign PubMed IDs to these publications. We follow the same four-step process for matching PubMed IDs to publications as described in the CSMED-FT creation procedure (Section 4.2 of Chapter 4).

6.1.2 Model evaluation

When conducting a meta-analysis, for every outcome, each study has its weight and effect size calculated first (respectively columns 6 and 7 on example forest plots in Figure 6.2). Effect size is an essential statistical concept in the analysis of research data [95]. It is

[2] https://www.cochrane.org/news/cochrane-improving-...-our-reviews

a measure that quantifies the magnitude of difference between two groups in a study. Researchers use a variety of effect measures to compare outcome data between two intervention groups, including odds ratios and mean differences.

For instance, in ratio effect measures, a value of 1 represents no difference between the groups [51, 50]. On the other hand, in difference measures, a value of 0 represents no difference between the groups. Values higher or lower than these "null" values may indicate either benefit or harm of an experimental intervention, depending on the order of the interventions in the comparison and the nature of the outcome. Every estimate is expressed with a measure of uncertainty, such as a confidence interval (CI) or standard error (SE).

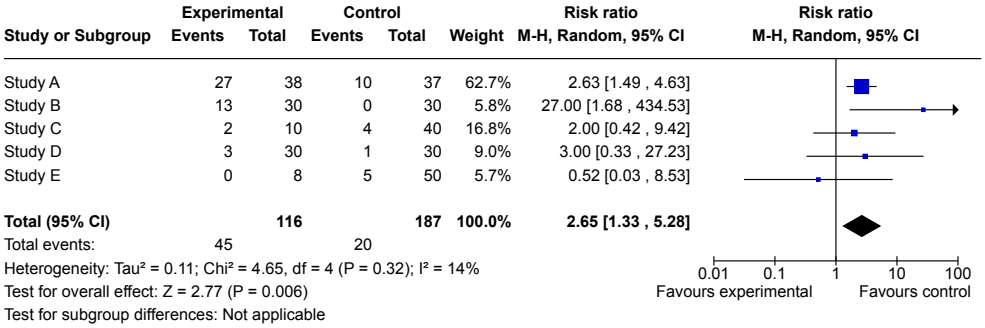
Effects depend on the number of events reported by that study, whereas weights assigned to each study are influenced by other studies included in this outcome. So when removing one study from the meta-analysis, only the weights of the remaining studies will change, but their effect sizes will stay the same (compare Figures 6.2a and 6.2c). There are several types of outcomes reported by Cochrane, in our work, we focus on the dichotomous and continuous outcomes only and calculate them following the approach by Deeks and Higgins [50].

Our framework supports evaluating arbitrary classification and retrieval runs, and calculates the final outcomes of the review based on publications included in the run. When evaluating classification or retrieval runs, we take all publications predicted as relevant. When evaluating ranking runs, we need to assume a cut-off point. Previous studies working on systematic review automation used either cut-off at $r\%$ of Recall [35, 134], or at $d\%$ of total dataset size [112, 113].

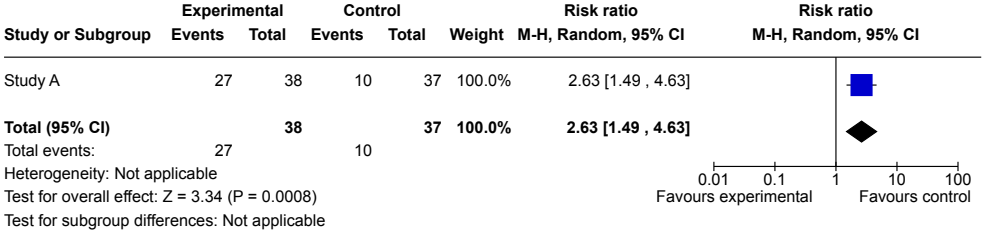
6.1.3 Outcome stability assessment

We examine the outcomes generated by the run and compare them with the outcomes obtained by the original review (Figure 6.2). We extend the analysis done by Nussbaumer-Streit et al. [185], who proposed two categories of "changed conclusions": (1) if the new review drew the opposite conclusion, (2) if it is not possible to draw a conclusion or the new conclusion has less certainty. We distinguish five aspects of analysis for review outcomes against the original review (Figure 6.2a). The first two of these aspects are real-valued, whereas the remaining three are categorical:

1. *Magnitude of difference* — By how much are the outcomes different in their effect size (Figure 6.2a versus 6.2b)? In other words, what is the numerical *influence* on the review outcome when certain studies are not included? This is measured by calculating the relative difference in effect size between the original outcome O_o and predicted outcome O_p: $MoD = \frac{\|O_o - O_p\|}{\|O_o\|}$. When $O_o = 0$ and $O_p \neq 0$, we assume $MoD = 100\%$; otherwise, when $O_o = O_p = 0$, we set $MoD = 0\%$. Similarly, when the predicted outcome cannot be estimated, we assume $MoD = 100\%$.



(a) Hypothetical review outcome with 5 included studies. Recall = 100%.



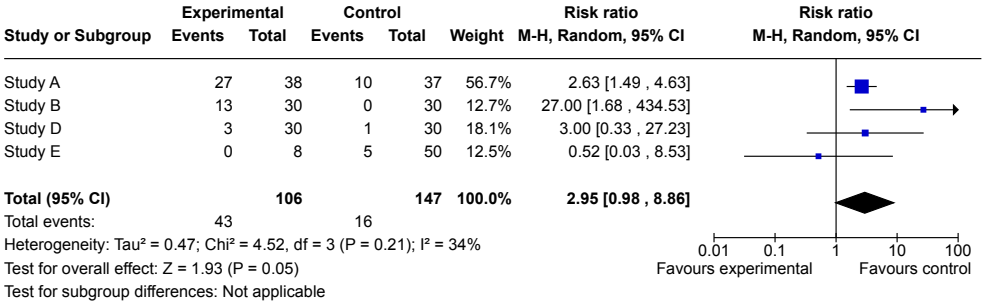
(b) Not including studies B, C, D and E still keep the review outcome approximately the same (absolute difference: 0.02, relative difference: 0.0076). Recall = 20%.

Figure 6.2: Different review outcomes represented as forest plots. Each row is a single study. Columns from the right represent, respectively: (1) the study identifier, (2) number of events in the experimental group (e.g., patients with specific symptoms or adverse events), (3) experimental group size, (4) number of events in the control group, (5) control group size, (6) the weight of a study, and (7) effect size of a study: a difference (e.g., risk ratio or standardised mean difference) in events between experimental or control group. Simulations and figures done using RevMan Web, available at `http://revman.cochrane.org`. The figure is continued on the next two pages.

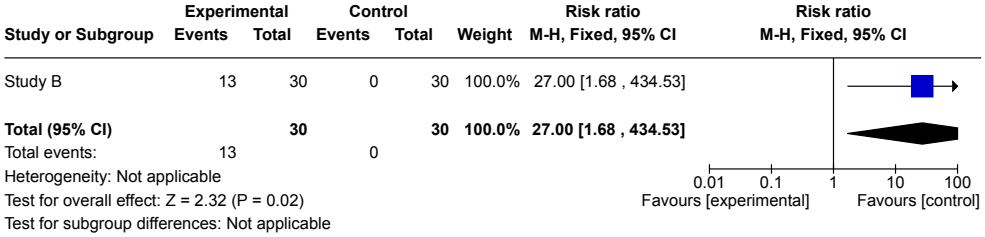
2. *Distance from CI* — Is the new outcome within the Confidence Interval (CI) of the original outcome (Figure 6.2c)? The answer is a distance between the predicted outcome O_p and the closest of the pair $(CI_{\text{lower}}, CI_{\text{upper}})$:

$$\Delta_{CI} = \begin{cases} \|O_p - CI_{\text{lower}}\| & \text{if } O_p < CI_{\text{lower}}, \\ \|O_p - CI_{\text{upper}}\| & \text{if } O_p > CI_{\text{upper}}, \\ 0 & \text{otherwise.} \end{cases}$$

3. *Overestimation/underestimation* — Is the outcome overestimated or underestimated compared to the original one (Figure 6.2d)? We first check if the calculated outcome is equal (due to the limits of precision of data reported in RevMan files, we use the relative and absolute tolerance of 10^{-5} and 10^{-6} respectively). Then, if the outcome is different, we check if the result is higher than the original (overestima-



(c) Not including study C will overestimate the review outcome, yet it will be within the 95% CI range. Recall = 80%.



(d) Not including studies A, C, D and E will overestimate the review outcome, and it will be above the 95% CI range of the original outcome. Recall = 20%.

Figure 6.2: (cont.) Different versions of review outcomes continued.

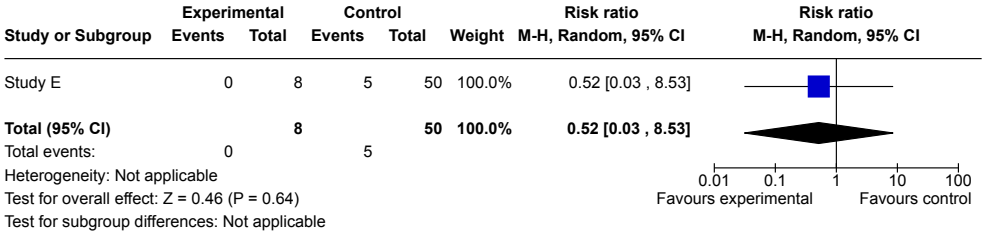
tion) or lower (underestimation). The answer has three options: "overestimated", "underestimated", and "equal".

4. *Sign* — Does the outcome have the same sign as the original one (Figure 6.2e)? In other words, are the new conclusions opposite to the original ones? The answer is binary: "yes"/"no". This aspect corresponds to the first category from Nussbaumer-Streit et al. [185].

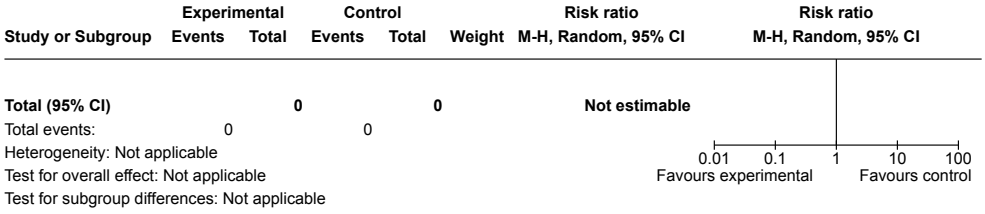
5. *Estimability* — Is it possible to calculate the outcome (Figure 6.2f)? An outcome cannot be calculated if there are no included studies concerning it. The answer is binary: "yes"/"no".

6.1.4 Measuring publication *influence* on the review outcome

So far, our analyses were conducted at the level of single outcomes. However, from the perspective of a retrieval algorithm, we are interested in primarily retrieving publications reporting the most critical outcomes for a given systematic review. To bridge this gap, we propose a method for measuring the influence of each publication. This *Influence* value can be incorporated into traditional evaluation measures like *nDCG* or *Precision* to weight publications.



(e) Not including studies A, B, C and D will change the study outcome – from 'favours control' to 'favours experimental'. Recall = 20%.



(f) Not including any study makes the outcome not estimable. Recall = 0%.

Figure 6.2: (cont.) Different versions of review outcomes continued.

First, we count the number of publications reporting the same study. If several publications report a study, not including one of these publications will not mean that this study will not be reported. Therefore, we assume that this could be a factor which is negatively correlated with the weight of the publication. For a study k, we define the number of publications reporting this study as $n_{p_{s_k}}$.

The *Influence* for a publication P_i can be defined as:

$$\mathcal{I}_{P_i} = \sum_{j}^{J} \left(\sum_{k}^{K} \frac{MoD_{j,k}}{n_{p_{s_k}}} \right), \tag{6.1}$$

where $MoD_{j,k}$ is the magnitude of difference (analysis aspect 1) of a study k on outcome j, when this study is missing from the predictions. When an outcome reports effects within subgroups, we sum the *Influences* from every subgroup first into $MoD_{j,k}$. The inner sum of Equation 6.1 measures the *Influences* of all K studies reported by that publication for a given outcome, whereas the outer sum adds all J outcomes which are reported by that publication. The calculated *Influence* score can be incorporated into the relevance assessments. The higher the $\mathcal{I}_{P_i}$ score is, the more influential a publication is on the final systematic review outcomes.

6.2 Experiment Setup

Two types of evaluation can be conducted using the proposed evaluation framework:

1. **Outcome-based evaluation** – in this case, we do not treat a dataset as a collection of systematic reviews but rather a collection of outcomes. The problem of conducting a systematic review is multi-dimensional. One can think of it as having several outcomes reporting different dimensions of the review, and the evaluation of the user's needs is conducted independently from the perspective of each outcome. We do not want to average across reviews, each containing a different number of outcomes. We add or average these outcome-level results instead.

2. **Review-based evaluation** – similar to the traditional evaluation measures, one can calculate the results per review by averaging or adding all the single outcomes results for a given review. This can be further averaged across the collection of systematic reviews. However, here we focus on using calculated publication *Influence* to evaluate the scores with traditional evaluation measures.

Before we present the results, we first discuss the dataset and runs used.

6.2.1 Dataset

We used 38 systematic reviews of interventions from the CLEF TAR 2019 training and test datasets [114]. We selected this collection as this was the only dataset inside CSMED which contained Cochrane SLRs of interventions. The inclusion of this dataset was further motivated by its use in an evaluation campaign, providing a diverse array of participant runs which are invaluable for the appraisal of various evaluation metrics.

Each review consists of a Cochrane ID, a protocol, and a list of publications described by their PubMedIDs with qrels both on the title and abstract level and a full text level. We enhanced the dataset by collecting RevMan files and information about the data and analysis as described in Section 6.1.1.

Out of 38 reviews in CLEF TAR 2019, our script found studies and outcomes for 32 reviews (17 in the training subset and 15 in the test subset). We summarise the statistics of the 32 reviews we consider in Table 6.1. There is a significant discrepancy in the number of outcomes reported by the reviews, ranging from as few as 2 or 3 outcomes in small reviews to 128 outcomes in the largest one. Moreover, an additional challenge is that the majority of these outcomes are reported by just one or two studies.

These 32 reviews report 1640 included publications, out of which we managed to find PubMed IDs for 1175 of them (71.6%). Next, we wanted to match publications identified with our script to the CLEF TAR 2019 qrels based on the PubMed ID. There were, in total, 778 relevant documents on the full text level identified in the CLEF TAR for these 32 reviews. We successfully merged 741 publications (95.2% of the total in CLEF TAR); there are only 37 publications in CLEF TAR 2019 qrels which we do not have in our records.

Table 6.1: Statistics of the considered dataset.

Dataset split		CLEF TAR 2019 Training		CLEF TAR 2019 Test	
Reviews' type		— Interventional —			
# Reviews		17		15	
Total # comparisons		54	100%	77	100%
Extracted comparisons		54	100%	69	89.6%
Total # outcomes		272	100%	516	100%
Extracted outcomes		267	98.2%	453	87.8%
— Dichotomous		158	58.1%	261	50.6%
— Continuous		109	40.1%	192	37.2%
Outcomes per review	Min	2		3	
	Median	9		15	
	Max	41		128	
Studies per outcome	Min	1		1	
	Median	2		2	
	Max	55		40	

6.2.2 Runs

We use 34 official CLEF TAR 2019 runs from three teams. The teams used a variety of ranking methods, including traditional BM25, interactive BM25, continuous active learning, relevance feedback, and various stopping criteria. Additionally, we included 40 runs based on the reproducibility of the active learning method by Yang et al. [289]. In total, we evaluated 74 runs. However, in this chapter, we predominantly present the results on a subset of 28 most diverse runs. The selection of the 28 runs was influenced by two primary considerations. Firstly, we aimed to represent a broad spectrum of approaches, given that many runs exhibited substantial similarities. Secondly, we sought to improve the visual clarity of our data presentations. By choosing a non-redundant subset, we prevent our figures from becoming overcrowded, facilitating easier interpretation of the results.

The CLEF TAR collection provides relevance judgements at two distinct screening levels: title and abstract, and full text. In the context of SLRs, only the publications that make it through the full text screening contribute to SLR outcomes. Therefore, our framework requires full text assessments, and we used relevance judgements from the full text level. However, it is important to note that the runs we evaluated were trained solely on the titles and abstracts of the publications. While this might not be fair towards the evaluated systems, our experiments aim not to establish which systems are better. Indeed, we seek to provide an example of the operationalisation of our framework and its implications.

6.3 Outcome-based Evaluation

We first run a simulation study to understand the results of our evaluation framework better in a controlled manner. Then, we discuss the usage of the evaluation framework with retrieval and classification runs on CLEF TAR 2019 collection.

6.3.1 Preliminary simulation

We execute a preliminary simulation to understand the effect our outcome-oriented evaluation has on the analysis of systematic review automation methods. We first perform an analysis based on the notion of a publication – we then turn to consider individual studies.

We simulate the evaluation framework by taking the set of included *publications* for each review and randomly removing $\{1, 2, 3, 4, 5, 10, 15, 20, 30, 50, 100\}$ publications from the set and then re-calculating the outcomes. We repeat the simulation 20 times with different random seeds. In other words, we are interested in exploring the impact of false negatives on the final review outcome. We compare the outcomes with the 'gold' outcomes from the original review. Results from all 32 systematic reviews are reported in Table 6.2. In our analysis, we consider the metrics from all five analysis aspects (Section 6.1.3), as well as the Recall.

Figure 6.3 presents box plots of averaged relative difference (aspect (1)) values from our simulation at a cut-off at 20% of the total number of documents. These results validate our expectations regarding the behaviour of this aspect of analysis as the relative difference grows with the number of removed publications. On the other hand, the distance to confidence intervals (aspect (2), Figure 6.4) does not show any specific trend on the CLEF 2019 reviews.

Out of all the metrics, the one that changes the most when varying the number of removed

Table 6.2: Results of the simulation on the *publication* level. Outcomes are aggregated across 32 systematic reviews and are averaged from 20 different random seeds.

		N relevant **publications** removed from the review											
	Analysis Aspect	gold	1	2	3	4	5	10	15	20	30	50	100
1	Mean relative difference	0.0	0.9	2.5	5.3	7.1	10.0	18.3	26.2	36.5	54.9	65.5	84.5
2	Mean distance from CI	0.000	0.002	0.003	0.004	0.007	0.008	0.013	0.042	0.102	0.018	0.008	0.083
3	Equal outcome	824	786	750	706	657	623	496	410	340	256	164	80
	Different	0	38	73	117	167	200	328	413	483	567	659	743
	- Underestimated	0	17	27	38	57	66	98	103	90	55	58	23
	- Overestimated	0	20	45	79	109	134	229	309	393	512	601	720
4	Have same sign	824	815	800	774	756	735	663	597	516	365	277	121
	Have different sign	0	9	24	49	67	88	160	227	307	458	546	702
5	Reported outcomes	824	816	804	781	767	743	675	610	529	371	284	128
	Missing outcomes	0	7	20	43	56	80	148	213	294	452	539	695
Average *Recall* for publications		1.00	0.92	0.84	0.75	0.70	0.63	0.45	0.35	0.28	0.22	0.14	0.05
Average *Recall* for studies		1.00	0.97	0.91	0.80	0.77	0.68	0.53	0.43	0.37	0.31	0.22	0.12

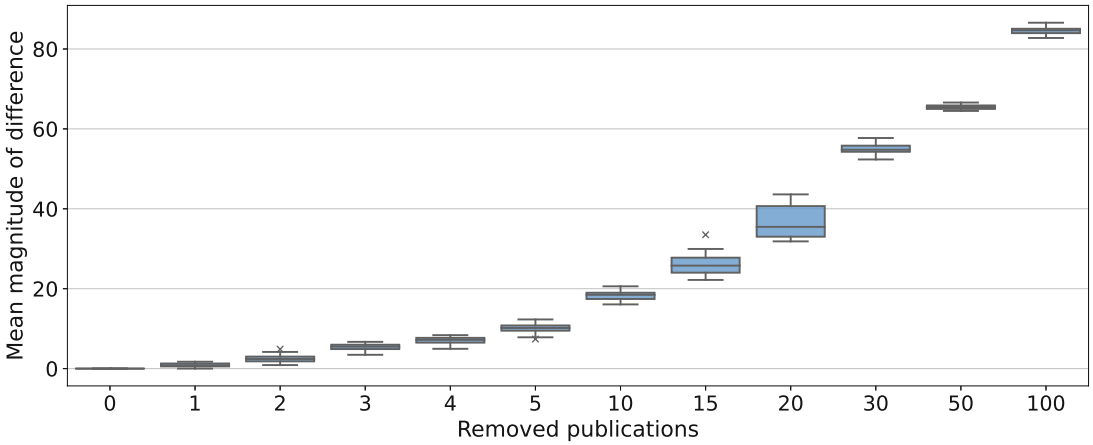


Figure 6.3: Box plots presenting relative difference values from 20 simulations on the publication level. Note that the intervals on the x-axis are not uniform.

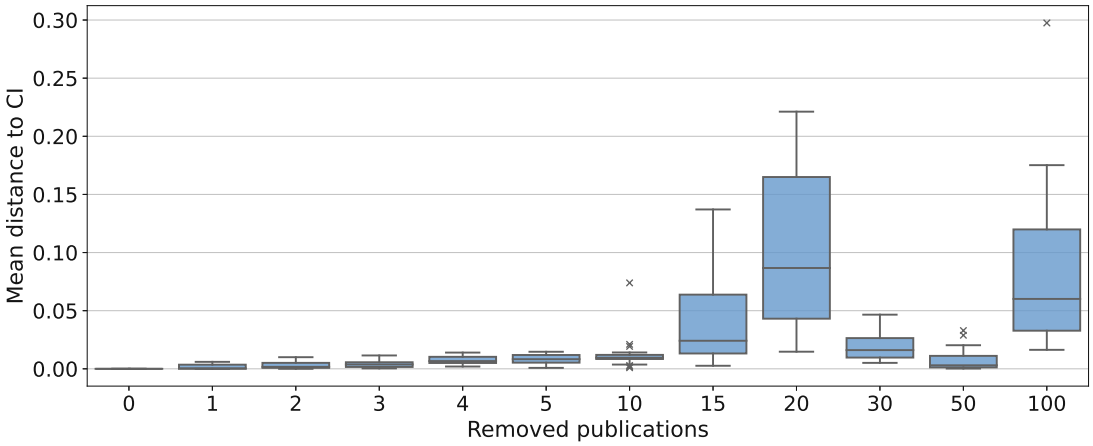


Figure 6.4: Box plots presenting distance to confidence intervals values from 20 simulations on the publication level. Note that the intervals on the x-axis are not uniform.

publications is estimability (5). As more publications are removed, it becomes more and more challenging to calculate outcomes, predominantly because half of the original outcomes relied on one or two studies. At the very extreme, when 100 publications are removed from every review, only 15% of outcomes are still estimable.

The measures of overestimation and underestimation (3) show growing trends with more publications being removed. Already not including one publication per review (achieving an average Recall of 92% for publications and 97% for studies) changed 38 outcomes (4.6% of the total number of outcomes). This shows that the commonly used threshold of 95% Recall does not enforce preserving the same outcomes of the review. Finally, the sign (aspect (4)) is not very descriptive across the simulations as it is mainly influenced

Table 6.3: Results of the simulation on the *study* level. Outcomes are aggregated across 32 systematic reviews and are averaged from 20 different random seeds.

		N relevant **studies** removed from the review											
	Analysis Aspect	gold	1	2	3	4	5	10	15	20	30	50	100
1	Mean relative difference	0	3.6	9	12.6	17.1	20.9	46.3	59	65.9	80.2	89.2	97.7
	Maximum relative difference	0	100	204	322.3	477.1	523.7	100	100	100	111.8	100	100
2	Mean distance to CI	0.000	0.016	0.028	0.039	0.053	0.080	0.010	0.023	0.034	0.060	0.168	0.000
3	Equal outcome	824	667	564	483	419	375	234	158	124	76	30	17
	Different	0	156	259	340	404	448	590	665	699	748	794	806
	- Underestimated	0	67	100	122	134	145	104	93	78	44	29	0
	- Overestimated	0	89	159	217	270	302	486	572	621	704	764	806
4	Have same sign	824	785	742	712	674	642	435	330	272	155	83	18
	Have different sign	0	38	81	112	149	181	388	493	551	668	740	805
5	Reported outcomes	824	795	752	724	689	659	443	338	281	164	89	19
	Missing outcomes	0	29	71	99	135	164	380	485	542	659	734	805
Average *Recall* for studies		1.00	0.83	0.73	0.65	0.56	0.51	0.33	0.24	0.18	0.12	0.05	0.02

by non-estimable outcomes.

Next, we perform the simulations using the same methodology but now removing studies, rather than publications – remember that current metrics evaluate at a publication level, not at a study level. Results are presented in Table 6.3. As expected, systematic review outcomes are less robust to missing studies than to missing publications. This is because several publications might report the same study, and under the assumptions in our approach, retrieving one publication is sufficient to classify the study as found.

6.3.2 Evaluation with CLEF TAR 2019 runs

In this section, we use the predictions from runs described in Section 6.2.2 and evaluate them using our framework. We use only the reviews which are part of the test split of the CLEF TAR 2019 dataset.

We further consider two baselines:

gold – the best possible run which returns all relevant studies from the original systematic review first.

max-with-qrels – this run takes into account the limitations of the CLEF TAR collection and our PubMed articles matching process. It uses all relevant studies identified in the CLEF TAR 2019 qrels as relevant and places them first.

We follow the evaluation procedure of CLEF TAR and calculate the following traditional evaluation measures: Mean Average Precision (MAP), last relevant found, Recall@k% of top-ranked publications, with k in {5, 10, 20, 30, 50}, Work Saved over Sampling at r% of Recall with r in {95%, 100%} (WSS@95%, WSS@100%), $nDCG$@20% of top-ranked publications and Area Under Recall Curve ($AURC$). CLEF TAR as their

primary reporting measure used MAP; therefore, we will treat MAP as the reference measure when sorting runs.

We calculate the relative difference in study outcomes (analysis aspect (1) in Section 6.1.3) for every outcome in all reviews. The lower the average score is, the better the runs, as their effect differs less from the original review effect. As considered runs were rankings, we follow the same procedure as for Recall and nDCG, namely we calculate the relative difference at k% of top-ranked publications with k in {5, 10, 20, 30, 50}.

Figure 6.5 presents a box plot of relative difference per outcome calculated at 30% cut-off of dataset size for 15 test CLEF TAR reviews. We do not evaluate baselines with traditional measures, yet for the purpose of sorting runs, we assume that they achieved the highest MAP score. Except for the best run, all other runs changed their rank when ordered using their mean relative difference score compared to the MAP-based ranking. While top runs, according to MAP scores, have low variability, there are runs among the top 10 which show considerable fluctuation. This means there are specific reviews for which these runs will lead to significantly different decisions about the outcome. This behaviour is comparable for relative difference at other cut-offs k.

What is also interesting is that the mean relative difference at 30% cut-off for the *max-with-qrels* baseline run is 6.24. Furthermore, for the relative difference score calculated at 100% of documents, this baseline score is also not equal to 0. This means that the limitations of the CLEF TAR collection and qrels establish a lower bound for the best achievable value of relative difference.

Figure 6.6 presents correlation between relative difference calculated at 20% cut-off of dataset size and evaluation measures used at CLEF TAR 2019. The score correlates positively with the last relevant found, but there is a negative correlation with all other measures. This confirms our intuition that a higher average relative difference score across outcomes means a worse model effectiveness, as the ideal 'best' model should achieve a difference of 0.

6.3.3 Pareto frontier optimisation

Based on the simulation results, we note a problem with non-estimable outcomes. Should these outcomes be assigned a zero score or maybe an infinite value? This raises the issue of handling these values in the evaluation process for calculating relative difference scores. In our study, we assigned a zero value to non-estimable outcomes, which allowed us to assume that the relative difference equals 100%. Nevertheless, this yields the problem of when the actual outcome is equal to the zero value (i.e., the study does not favour the experimental nor the control group), as the difference, in this case, would also be zero. One way to overcome the issue of non-estimable outcomes would be to evaluate both estimability and relative difference implemented, for instance, using the Pareto frontier [158].

Figure 6.7 presents the Pareto frontier evaluated at a cut-off at 5% of the total number of documents. On the x-axis, we show the number of non-estimable outcomes for each

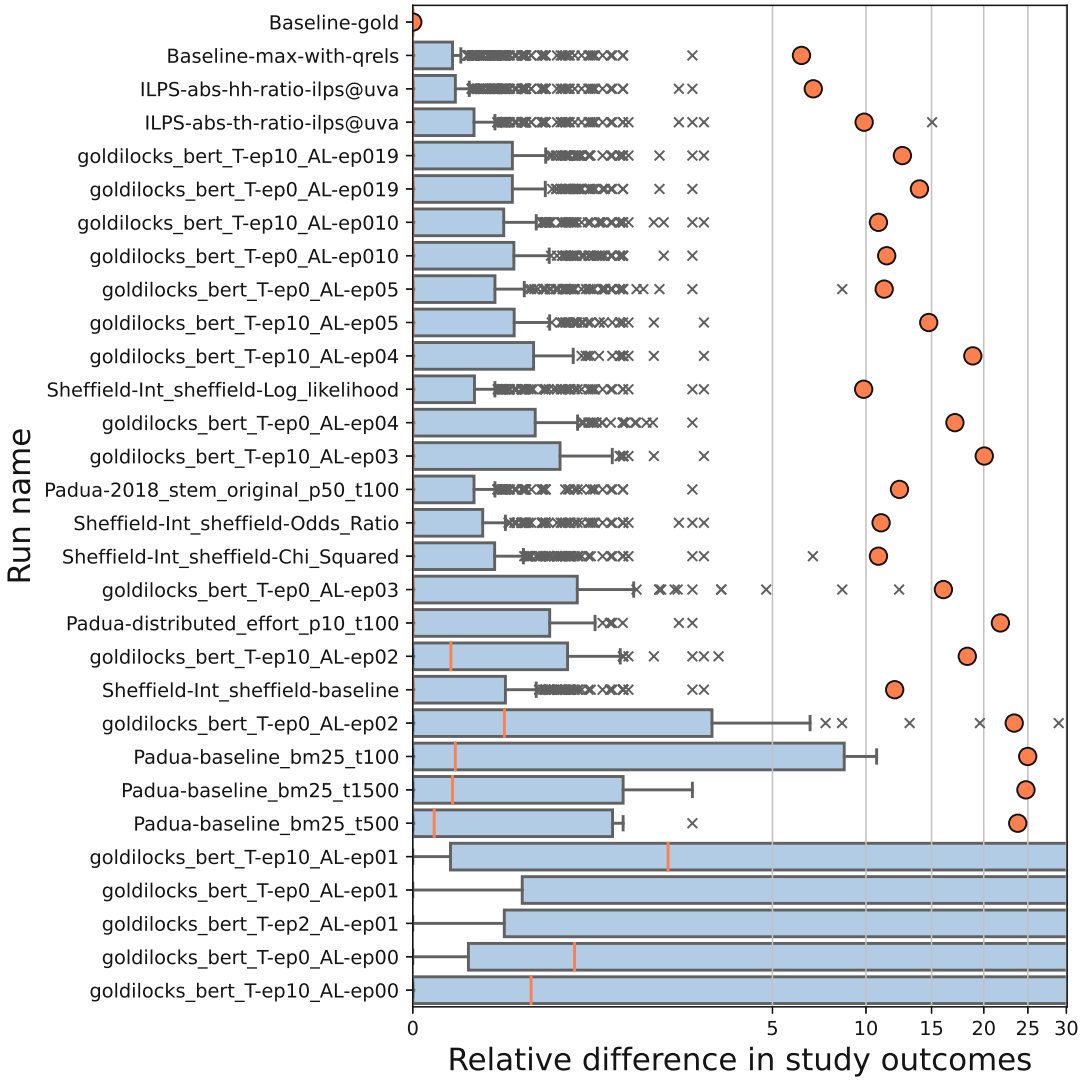


Figure 6.5: Box plot presenting runs with their relative difference in study outcomes for an evaluation with a cut-off at 30% of the total number of documents for each review. Runs are sorted by their MAP score. The orange circle denotes the mean relative difference @30%. The x-axis is cut at 30, while the outliers exist up to the value of 100; we cut for visualisation purposes.

run. On the y-axis, there is a sum of relative difference for estimable outcomes. We min-max normalise the sums including the gold baseline run (gold represents the best achievable score of $(0, 0)$). Both objectives should be minimised, i.e., we want to have as few non-estimable outcomes as possible and for all estimated outcomes, the difference would be as close to zero as possible. Contrary to the previous evaluations, we can notice

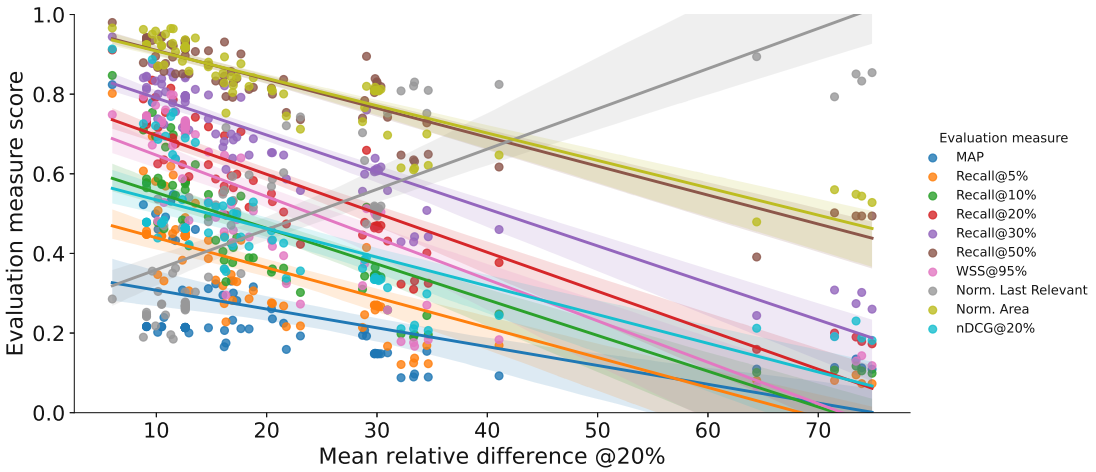


Figure 6.6: Linear regression fits between relative difference at 20% cut-off of documents and other evaluation measures scores. Correlations for relative difference at other cut-offs follow similar trends.

that no single run would dominate on both dimensions.

6.4 Review-based Evaluation

Here, we test our proposal of estimating publication *Influence* (Section 6.1.4). For each publication, we calculate its *Influence* using Equation 6.1.

We investigate if we can use the estimated *Influence* of publications to substitute the binary relevance judgements. We calculate the nDCG score using two versions of document (publication) gains: (1) binary based on the full text screening level qrels, (2) graded based on the publication *Influence*.

We use the same runs as in the outcome-based evaluation. Figure 6.8 presents the scores for two versions of nDCG scores calculated at 20% of dataset size; runs are sorted by mean nDCG. We can observe that the ranking of runs when using gain based on the calculated *Influence* differs from when using the original binary qrels. Moreover, when the best run achieves mean nDCG@20% with binary qrels at a level of 91%, its corresponding nDCG@20% using publication *Influence* is only 50%. What could have been seen as a solved task reveals that the models could still be improved to prioritise the most *influential* relevant publications. Notably, for all runs, at least one review exists, for which a run's nDCG score weighted with *Influence* equals 0.

We only demonstrate the usage of *Influence* with the nDCG measure, but it would also be possible to apply it when evaluating with other evaluation measures like *Precision* or TNR.

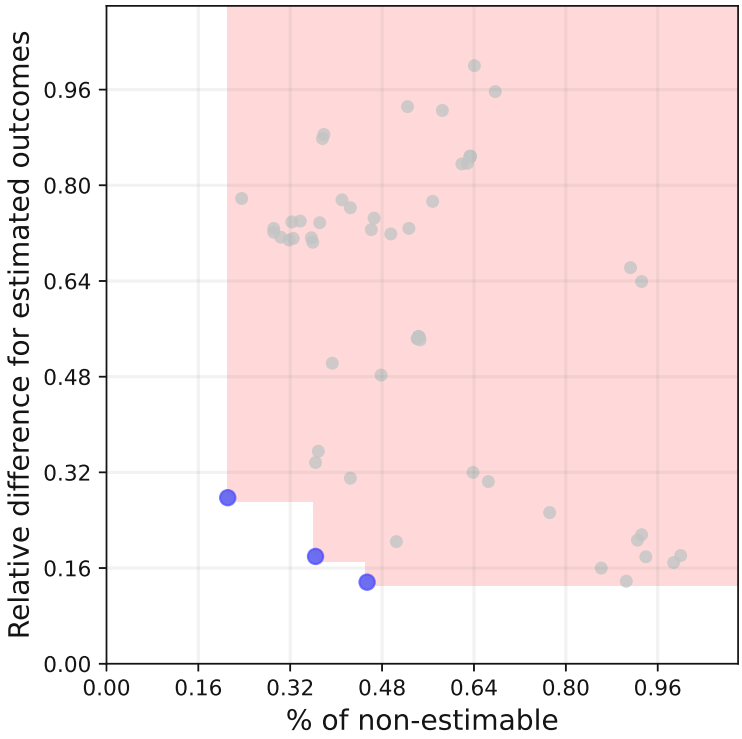


Figure 6.7: Visualisation of the Pareto frontier for two objectives: (1) number of non-estimable outcomes on the x-axis and (2) sum of relative difference for estimable outcomes on the y-axis. Both objectives are to be minimised. Runs are evaluated at a cut-off at 5% of the total number of documents for each review. Non-dominated runs are marked with a blue colour. The Pareto frontier was calculated using the method by Herman and Woodruff [92].

6.5 Discussion

The primary objective of this work was to introduce the concept of evaluating automated methods for systematic reviews based on their influence on review outcomes, rather than relying on binary qrels. In this section, we reflect on the potential challenges and limitations that arise when attempting to fully operationalise our proposed framework.

6.5.1 Improvements in measuring and evaluating publication influence

First, we discuss various challenges that arise when attempting to measure and evaluate the influence of publications in systematic reviews. From assumptions about outcome independence to difficulties with non-estimable outcomes, we detail the hurdles and potential approaches to overcome them.

Improvement 1: Measuring publication influence. In our work, we have presented

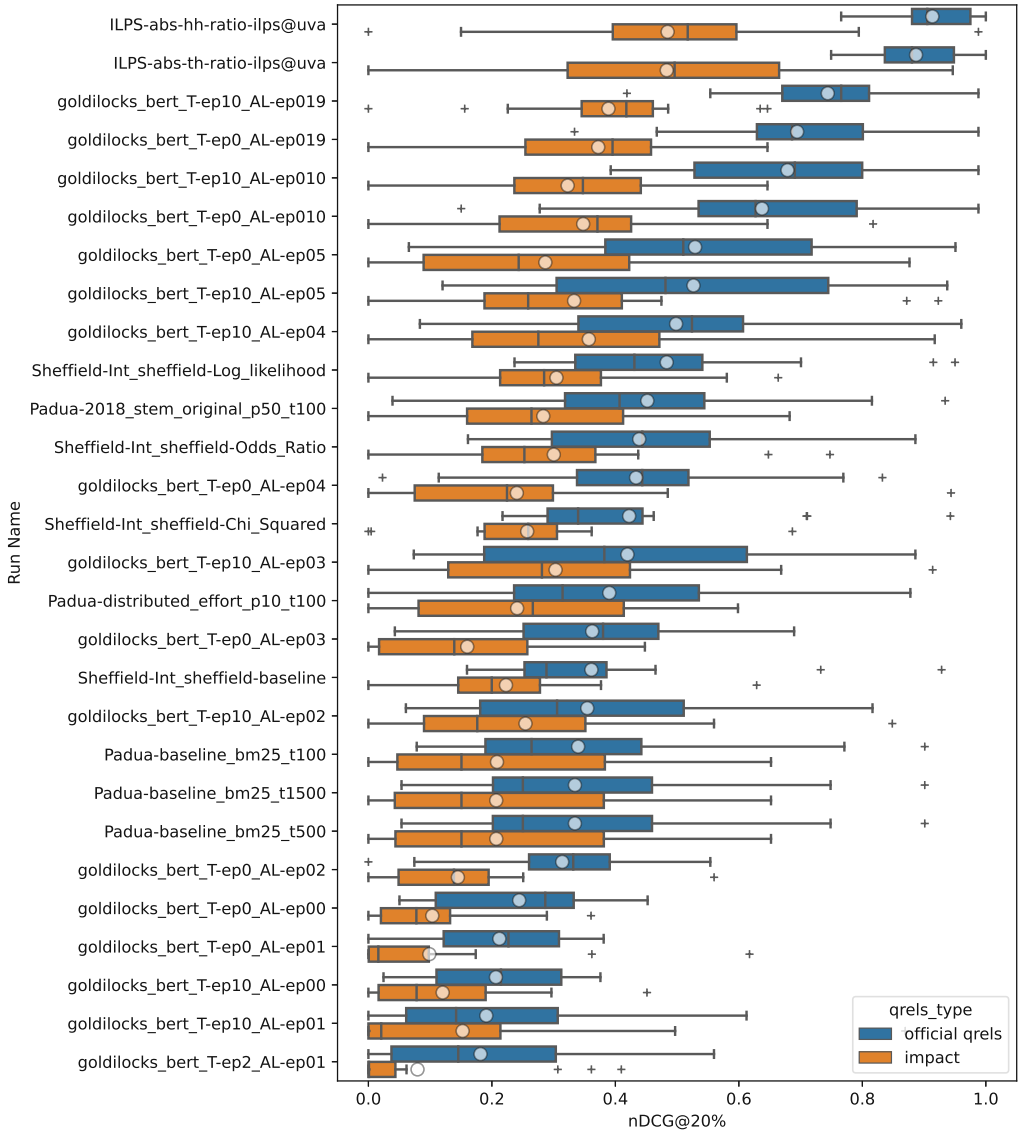


Figure 6.8: Box plot presenting nDCG@20% scores for each run. Orange bars represent original binary qrels, and blue represent weighting based on the *Influence*. Runs are sorted by mean nDCG (white circle) of original binary qrels.

a single approach to measuring study influence and weighting publications, but it is worth considering other potential techniques. We decided to use relative difference as it simplified the calculation of differences in non-estimable effects. An alternative approach would be to calculate the absolute difference in study outcomes. Furthermore, one could also envision other types of weighting based not only on numerical outcomes but on publication metadata or time needed to screen that publication manually.

Improvement 2: Correlations between outcomes. Another significant limitation of the approach proposed in this chapter is the assumption of independence between publications, studies, and outcomes. It is assumed that every outcome is equally important and independent of each other, which may not hold true in many cases. For example, an outcome measuring fatal adverse events in patients may be more relevant than an outcome about mild symptoms. Therefore, it may be necessary to consider the importance of outcomes in a hierarchy to better capture the impact of automated methods on the review. As our experiments are only proof of concept, domain expertise and further research is needed to explore the possibility of creating such a hierarchy of outcomes.

Improvement 3: Non-estimable outcomes. What should be done if an outcome is not estimable? Should it be assigned a zero score or rather an infinite value? This raises the question of how to handle these values in the evaluation process. In our study, we assigned a zero value, which allowed us to assume that the MoD is equal to 100%.

Additionally, there is the issue of what to do when the original outcome is equal to 0 and the prediction is non-estimable. In this case, our methodology would also assign a zero value to the MoD score. While this may seem reasonable since the study was inconclusive, it also means that we are not taking into account any potential reduction in uncertainty that the study might have provided. This can limit the interpretability of the review outcome and hinder our ability to draw meaningful conclusions.

Another important consideration is calculating the confidence intervals in the presence of non-estimable outcomes. One option could be to exclude these outcomes from the calculation, but this can lead to a biased estimate of the standard deviation.

Improvement 4: Publications without numerical outcomes. Additional consideration must be taken for studies (publications) that do not report any numerical outcomes. Such studies do not convey direct comparisons but are still relevant to systematic review. Not including them in our evaluation framework would not change the score, but their contribution could influence the review as these papers might report information such as descriptions of study design, intervention details, participant characteristics, and adverse events. In our work, *Influence* for these publications is equal to 0. If we substituted or multiplied the binary relevance by the *Influence*, this would mean that publications without studies would have a relevance of 0. One alternative option is to add the *Influence* to the original relevance–R_{P_i} would then be:

$$R^{'}_{P_i} = R_{P_i} + \mathcal{I}_{P_i}, \tag{6.2}$$

where non-relevant documents have $\mathcal{I}_{P_i} = 0$. In this way, we make sure that all relevant documents have non-zero $R^{'}_{P_i}$. Such relevance grades could also be used for the title and abstract screening.

Improvement 5: Accounting for several outcomes. Merging multiple outcomes into a single score can be challenging as each outcome may have different meanings and importance for the review. The approach of adding differences used in our methodology

can be problematic in situations where the outcomes are not directly comparable or have different units of measurement. One possible way to deal with this issue is to use weighting to reflect the importance of each outcome for the review. The weighting can be based on factors such as the relevance of the outcome to the research question, the number of studies contributing to the outcome, or the confidence level of the effect estimate. Another alternative could be to use a multi-criteria decision analysis (MCDA) approach to merge the outcomes [197, 193, 296].

Improvement 6: Outcome importance. We decided to consider all outcomes of systematic reviews in our evaluation framework. However, not all outcomes might be of equal importance and weight to the user. For instance, Marshall et al. [163] in their analysis only used the "primary" outcome (numbered 1.1 in the reviews). An alternative would be to use only the first key outcomes mentioned in the "Summary of findings" section of the systematic review. Unfortunately, the "Summary of findings" section is not structured and might require human supervision if used on a large scale.

Improvement 7: Accounting for false positives. This approach does not consider the potential outcome of false positives, i.e., publications that have been identified as relevant by the automatic methods but that are instead excluded by the reviewers. In our analysis, they are treated as if they were having no effect on the final review, and their *Influence* equals 0. However, we believe that meaningful evaluation metrics for systematic review automation should consider both the effect of missing relevant publications and the cost of assessing not-relevant publications, which requires incorporating false positives into the evaluation framework.

Improvement 8: Publication weighting. In our study, we opted for using uniform sampling of publications for the preliminary simulation (Section 6.3.1). Other alternatives (albeit controversial) could be using the impact factor of a journal or the number of citations of a publication.

6.5.2 Broadening the evaluation framework

As the field progresses, our evaluation framework must adapt and extend to include a wider range of outcome types and systematic review methodologies. This section explores such potential extensions.

Extension 1: Different outcome types. While our proposed evaluation framework focuses on continuous and dichotomous outcomes, other types of outcomes may be reported in systematic reviews, including ordinal, count, and time-to-event data. In our analysis, however, we found that continuous and dichotomous outcomes comprised most of the outcomes in the dataset we studied, accounting for 92% of all reported outcomes across 32 CLEF TAR 2019 reviews. We anticipate that our evaluation framework could be generalised to incorporate other types of outcomes.

Extension 2: Other types of systematic reviews. We focus only on systematic reviews of interventions which have a clear structure and evaluate the effectiveness of

specific treatments, programs, or policies by comparing experimental setups with control groups. However, there are several other types of systematic reviews, such as diagnostic test accuracy reviews, prognostic reviews, and qualitative research reviews, each of which presents unique challenges for automation and evaluation [114]. Future work should investigate how this outcome-based evaluation framework can be extended to these other types of reviews.

Extension 3: Title and abstract screening. We work on the outcomes extracted from the full text screening and use relevance judgements from full text screening to judge the runs. However, most models are trained on titles and abstracts, which might make this an unfair comparison. What should be the *Influence* of publications included in the title and abstract screening stage but excluded during full text screening? Should it be zero as it is for other not-relevant publications, or maybe their *Influence* should be negative as screening full texts takes more time and a potential false positive at this stage could be more problematic?

Extension 4: New collection. We argue that the CLEF TAR 2019 collection is not the best for our analysis. It is, however, the only one that provides runs and reviews of clinical interventions. It has several problems with respect to missing documents (see Figure 6.5 where the delta for *maximum-with-data* baseline run is non-zero). A possible approach might be to select good quality reviews for this analysis, similar to Nussbaumer-Streit et al. [185].

Extension 5: Prospective evaluation. Our framework only supports retrospective evaluations of the automatic search strategy creation and screening methods: it does not suit prospective evaluations [248]. That is, our outcome-based evaluation framework requires annotating all publications first, including the studies' effect and weight on the final outcome of the review. This means this evaluation can only occur after the screening (and the review) is completed. However, this limitation is shared with the common practice for evaluating automatic methods in this space [130, 99, 124].

Extension 6: Beyond citation screening. Another challenge is to determine how to incorporate our proposed framework into the broader systematic review process. Automated screening methods are just one aspect of the larger review process, and future research will need to consider how to integrate our approach with other stages of the review, such as PICO identification [187] or review summarisation [273].

Extension 7: Beyond systematic reviews. Previous studies already mentioned that the traditional Cranfield (TREC system evaluation) paradigm is not exhaustive for interactive information retrieval and proposed *usefulness* as an alternative dimension [39]. However, this paradigm is still not commonly used when performing system-centred evaluation [4]. In this chapter, we showed one example of why it is important to start thinking beyond binary relevance judgements and focus more on the utility of the information.

An exploration on how our proposed framework could be applied in other domains beyond systematic reviews can be another future work. The concepts introduced in

this chapter could have broader applications in other areas of information retrieval. Further investigation can help determine the utility and generalisability of our proposed framework.

Extension 8: What does it mean for the users? In the survey by Wagner et al. [267], one-third of respondents were willing to tolerate a risk of over 10% of receiving incorrect answers in exchange for faster evidence synthesis. However, evaluating retrieval effectiveness for systematic reviews solely based on the percentage of relevant documents retrieved, such as aiming to retrieve 95% of relevant documents, does not guarantee correct study outcomes. A user study would be needed to actually assess the quality of our numerical outcomes and measure the impact on users [72, 206].

6.5.3 Potential limitations and biases

Evaluation measures can sometimes be misleading if not used appropriately. In this section, we shed light on potential pitfalls when relying too heavily on certain measures and the ever-present challenge of publication bias in systematic reviews.

Limitation 1: Ensuring the appropriate use of evaluation measure. A practice that can be observed across the field is treating evaluation measures as an optimisation objective. Our evaluation approach should not be used for optimising models. The notion of difference in study outcomes is only known a-posteriori when the review is completed. Using absolute differences in study outcomes as an optimisation objective might lead to overfitting to biases in data.

Limitation 2: Recognising publication bias in systematic reviews. We acknowledge that there is an intrinsic bias in the publication of systematic literature reviews. A significant number of studies that yield negative or null results often remain unpublished, leading to their omission from these reviews. This exclusion can inadvertently create a biased representation of the evidence, emphasising positive outcomes over equally valid negative findings. It is crucial for researchers to consider these unpublished studies to ensure a comprehensive and balanced review of the existing literature.

Limitation 3: Variance in evaluation protocol adherence. Additionally, while we attempted to follow the evaluation protocols from the Cochrane Handbook closely, for 2.4% of outcomes, our effect calculations yielded marginally different results. To mitigate this issue in future work, it would be ideal to have access to RevMan or another official program designed explicitly for calculating study outcomes. This would help ensure that all types of outcomes are accurately and consistently covered, thus enhancing the reliability of the review process.

6.6 Summary

This chapter puts forward a novel, outcome-based evaluation framework for assessing the effectiveness of automatic search strategies and citation screening methods in the context

of systematic literature reviews. Our proposed framework evaluates the quality of these methods based on how closely the outcomes of their included publications match the actual review outcomes. We believe that this approach offers a more accurate reflection of real-world scenarios where not all included publications have the same influence on the final review outcome.

In addition to proposing the approach, we explore five analysis aspects that it enables, including measuring the numerical difference in predicted systematic review outcomes. We run initial experiments to simulate the impact of false negatives on reviews' outcomes showing that five missing publications per review can change 24% of outcomes. We also compare the evaluation results obtained using our framework with those obtained using traditional evaluation methods on CLEF TAR 2019 runs, highlighting the differences in focus between the two approaches. Finally, we propose a method for measuring the *Influence* of each publication and demonstrate its effectiveness in the gain of the nDCG metric.

While our proposed evaluation framework opens a different perspective over traditional methods, we acknowledge that many challenges remain to be addressed for this evaluation to be operationalised, which we outline in the final section. In our assessment, this framework represents a step forward in developing more effective and realistic methods for evaluating automation methods. It is particularly pertinent in the context of systematic literature reviews in medicine and other domains in which the importance of systematic reviews is increasing.

CHAPTER 7

Automated Citation Screening as Binary Classification

This chapter explores the application of deep neural networks in the field of automated citation screening. We present our work on conducting eligibility screening represented as a binary classification using three different models. This chapter provides an empirical foundation on how our work on datasets and evaluation measures from Chapters 4 and 5 can be applied to tackle real problems. By focusing on binary classification, we keep the overall study design simple and propose a more rigid pipeline to multiple implementations of models for this task. We also show the limitations of tackling the citation screening as a binary classification. Therefore, this chapter addresses a specific technical challenge and fits into the larger narrative of advancing information retrieval methods in academic research.

We reproduce two recent papers which proposed using neural networks for citation screening [124, 262]. We chose these studies since, to the best of our knowledge, they were the first ones to successfully address the screening problem using deep neural networks. Both papers represent citation screening as a binary classification task and train an independent model for each dataset. Both papers use deep learning and motivate it by claiming a substantial superiority of deep neural networks over traditional (statistical) models. In the remaining sections of this chapter, we will use the name **Paper A** to refer to the study by Kontonatsios et al. [124] and **Paper B** to indicate work by van Dinter et al. [262].

Kontonatsios et al. [124] (**Paper A**) was the first one to apply deep learning algorithms to automate the citation screening process. They have used three neural network-based denoising autoencoders to create a feature representation of the documents. This representation was fed into a feed-forward network with a linear SVM classifier trained in a supervised manner to re-order the citations. van Dinter et al. [262] (**Paper B**)

presented the first end-to-end solution to citation screening with a deep neural network. They developed a binary text classification model with the usage of a multi-channel convolutional neural network. Both papers claimed to yield significant workload savings of at least 10% on most benchmark review datasets.

We compare these models with traditional approaches and assesses their performance improvement on 23 benchmark datasets. Additionally, we evaluate these models alongside a simpler fastText-based shallow neural network model [23]. We present our challenges regarding replicability in terms of datasets, models and evaluation methodologies.

Moreover, we investigate if the models are invariant to different data features and random initialisations. 18 out of 23 datasets are available as a list of Pubmed IDs of the input papers with assigned decisions (included or excluded). As we needed to recreate data collection scripts for both papers, we wanted to measure if the choice of the document features would influence the final results of the replicated models. We also assess the training time necessary for each model. Our data collection and experiment scripts and detailed results are publicly available on GitHub[1].

Finally, we present the empirical analysis of the true negative rate and normalised Precision measures, evaluated at a fixed recall cutoff. We demonstrate these metrics' utility in assessing different aspects of the screening process.

7.1 Experiment Setup

In this section we present considered models, datasets and the evaluation procedure.

7.1.1 Models

We formulate the task as a binary classification for relevance prediction (see Section 3.3.1 of Chapter 3 for more information). We test the following three neural network-based models:[2]

DAE-FF Paper A presents a neural network-based, supervised feature extraction method combined with a linear Support Vector Machine (SVM) trained to prioritise eligible documents. The data preprocessing pipeline contains stopword removal and stemming with a Porter stemmer. The feature extraction part is implemented as three independent denoising autoencoders (DAE) that learn to reconstruct corrupted Bag-of-Words input vectors. Their concatenated output is used to initialise a supervised

[1] https://github.com/ProjectDoSSIER/CitationScreeningReplicability

[2] The research and models discussed in this chapter were primarily conducted and evaluated during the 2020-2021 period and published at ECIR 2022 [130]. Since then, the fields of NLP and IR have evolved, especially towards the use of larger language models, predominantly based on the Transformers architecture and the BERT model. Despite these advancements, the insights provided in this chapter remain enduring and can be generalized to other models, particularly in aspects relevant to reproducibility and methodological approaches.

feed-forward neural network (FF). These extracted document vectors are subsequently used as an input to an L2-regularised linear SVM classifier. To address the class imbalance, the classifier uses a class-weighting mechanism by adjusting the regularisation parameter $C = 1 \times 10^{-6}$.

Multi-Channel CNN Paper B presents a multi-channel convolutional neural network (CNN) to discriminate between includes and excludes. It uses static, pre-trained GloVe word embeddings [199] to create an input embedding matrix. This embedding is inserted into a series of parallel CNN blocks consisting of a single-dimensional CNN layer followed by global max pooling. Outputs from the layers are concatenated after global pooling and fed into a feed-forward network. The authors experimented with a different number of channels and Conv1D output shapes. Input documents are tokenised and lowercased, punctuation and non-alphabetic tokens are removed. Documents are padded and truncated to a maximum length of 600 tokens. Class imbalance is handled with oversampling. For our replicability study, we have chosen the best performing MODEL_2.

fastText We also test a shallow neural network model which is based on fastText word embeddings [23]. This model is still comparable to more complex deep learning models in many classification tasks. At the same time, it is orders of magnitude faster for training and prediction, making it more suitable for active learning scenarios where reviewers could alter the model's predictions by annotating more documents. To make it even simpler, we do not use pre-trained word embeddings to vectorise documents. Data preprocessing is kept minimal as we only lowercase the text and remove all non-alphanumerical characters.

7.1.2 Hyperparameters

Paper A optimised only the number of training epochs for their DAE model. In order to do so, they used two datasets: Statins and BPA reviews and justified this choice with differences between smaller datasets from Clinical and Drug reviews and SWIFT reviews. Other hyperparameters (including the minibatch size and the number of epochs for the feed-forward model) are constant across all datasets. Paper B used the Statins review dataset to tune a set of hyperparameters, including the number of epochs, batch size, dropout, and dense units. We keep the hyperparameters as set in the original papers.

7.1.3 Data

All 23 systematic literature reviews are summarised in Table 7.1. The statistics include the dataset source, the total number of documents, number and percentage of eligible documents, maximum WSS@95% score and the availability of additional bibliographic metadata. Every document consists of a title, an abstract, and an eligibility decision (included or excluded). Moreover, 18 SLRs contain also bibliographic metadata about documents sourced from PubMed. There is no information about the SLRs beyond their very generic title, like *ADHD* or *Opioids*. The percentage of eligible documents (includes) varies between SLRs, from 0.55% to 27.04%, but on average, it is about 7%, meaning that

Table 7.1: Statistics of 23 systematic literature reviews used in this experiment, a subset of CSMeD-basic containing medical SLRs.

	Datset name	Introduced in	#Citations	Included citations	Excluded citations	Maximum WSS@95%	PubMed ID
1	ACEInhibitors		2,544	41 (1.6%)	2,503 (98.4%)	93.39%	Yes
2	ADHD		851	20 (2.4%)	831 (97.6%)	92.65%	Yes
3	Antihistamines		310	16 (5.2%)	294 (94.8%)	89.84%	Yes
4	Atypical Antipsychotics		1,120	146 (13.0%)	974 (87.0%)	81.96%	Yes
5	Beta Blockers		2,072	42 (2.0%)	2,030 (98.0%)	92.97%	Yes
6	Calcium Channel Blockers		1,218	100 (8.2%)	1,118 (91.8%)	86.79%	Yes
7	Estrogens		368	80 (21.7%)	288 (78.3%)	73.26%	Yes
8	NSAIDs	Drug [35]	393	41 (10.4%)	352 (89.6%)	84.57%	Yes
9	Opioids		1,915	15 (0.8%)	1,900 (99.2%)	94.22%	Yes
10	Oral Hypoglycemics		503	136 (27.0%)	367 (73.0%)	67.96%	Yes
11	Proton PumpInhibitors		1,333	51 (3.8%)	1,282 (96.2%)	91.17%	Yes
12	Skeletal Muscle Relaxants		1,643	9 (0.6%)	1,634 (99.5%)	94.45%	Yes
13	Statins		3,465	85 (2.5%)	3,380 (97.5%)	92.55%	Yes
14	Triptans		671	24 (3.6%)	647 (96.4%)	91.42%	Yes
15	Urinary Incontinence		327	40 (12.2%)	287 (87.8%)	82.77%	Yes
	Average drug		1,249	56 (7.7%)	1,192 (92.3%)	87.33%	15/15
16	COPD		1,606	196 (12.2%)	1,410 (87.8%)	82.80%	No
17	Proton Beam	Clinical [269]	4,751	243 (5.1%)	4,508 (94.9%)	89.89%	No
18	Micro Nutrients		4,010	258 (6.4%)	3,752 (93.6%)	88.57%	No
	Average clinical		3,456	232 (7.9%)	3,223 (92.1%)	87.08%	0/3
19	PFOA/PFOS		6,331	95 (1.5%)	6,236 (98.5%)	93.50%	Yes
20	Bisphenol A (BPA)		7,700	111 (1.4%)	7,589 (98.6%)	93.56%	Yes
21	Transgenerational	SWIFT [99]	48,638	765 (1.6%)	47,873 (98.4%)	93.43%	Yes
22	Fluoride and neurotoxicity		4,479	51 (1.1%)	4,428 (98.9%)	93.86%	No
23	Neuropathic pain | CAMRADES		29,207	5,011 (17.2%)	24,196 (82.8%)	77.84%	No
	Average SWIFT		19,271	1,206 (4.6%)	18,064 (95.4%)	90.44%	3/5
	Average (All datasets)		5,454	329 (7.0%)	5,125 (93.0%)	87.97%	18/23

the datasets are highly imbalanced. These 23 SLRs are from three different collections introduced by Cohen et al. [35], Wallace et al. [269], and Howard et al. [99], respectively (see Section 3.4 for general overview of these datasets).

Paper A trained and evaluated their model on all 23 datasets coming from three categories. Paper B used 20 datasets from the Clinical and SWIFT categories. Paper B states that, on average, 5.2% of abstracts are missing in all 20 datasets, varying between 0% for *Neuropathic Pain* and 20.82% for *Statins*. Compared to previous papers, Paper B reports fewer citations for three datasets (Table 6 in the original paper): *Statins*, *PFOA/PFOS* and *Neuropathic Pain*. This difference is insignificant compared to the dataset size, e.g. 29,207 versus 29,202 for *Neuropathic Pain*, so it should not influence the model evaluation.

The model prepared by Paper B uses also pre-trained 100-dimensional GloVe word embeddings which we downloaded separately from the original authors' website[3] according to the instructions provided by the Paper B GitHub Readme.

[3] `https://nlp.stanford.edu/data/glove.6B.zip`

7.1.4 Evaluation

Original papers used Work Saved over Sampling (WSS) as their primary evaluation measure. In Chapter 5 we have shown the problems with WSS. Therefore, we present the evaluation using the True Negative Rate at 95% Recall (TNR@95%). For comparability to the original papers, in our replicability study we decided to use the implementations of the WSS metric provided by Papers A and B. Then we take these scores and using Equation 5.29 we calculate the TNR values. Furthermore, we also calculate normalised Precision at 95% Recall (nP@95%). Following original papers, we consider the relevance judgements from the title and abstract screening step.

Both papers use a stratified 10×2 cross-validation for evaluation. In this setting, data is randomly split in half: one part is used to train the classifier, and the other is left for testing. This process is then repeated ten times, and the results are accumulated from all ten runs. We also use this approach to evaluate the quality of all three models.

7.2 Results

We first present results of the replicability study. Then we dive into the impact of input features on the model performance. Finally, we present the measurements of training time and the normalised Precision at 95% Recall scores for each model.

7.2.1 Replicability study

TNR@95% scores from older benchmarks and original papers, along with our replicated results, are presented in Table 7.2. For all datasets, both Paper A and B provide only mean score from cross-validation runs. Therefore, we were not able to measure statistical significance between our replicated results and the original ones. To quantify the difference, we decided to calculate the absolute delta between reported and replicated scores: $|x - y|$. Both models report a random seed for the cross-validation splits but not for the model optimisation. Usage of different seeds for model optimisation might be one of the reasons why we were not able to achieve the same results.

For two datasets (*Bisphenol A (BPA)* and *Triptans*), Paper A reports two different results for the DAE-FF model (Tables 5 and 6 in the original paper). We suppose this was only a typing mistake, as we managed to infer the actual values based on the averaged TNR@95% score from all datasets available in the original paper.

The average delta between our replicated results and the original ones from Paper A is 3.9%. Only for three datasets is this value higher than 10%. If we consider different random seeds used for training models, these results confirm the successful replication of Paper A's work.

For Paper B, the average delta is 18.78%. For 11 out of 20 datasets, this delta is more than 10%. For the two largest datasets: *Transgenerational* and *Neuropathic Pain* we

Table 7.2: TNR@95% results for replicated models compared with original results and benchmark models. TNR@95% scores are averages across ten validation runs for each of the 23 review datasets. Underlined scores indicate the highest score within the three tested models, **bold** values indicate the highest score overall.

No	Dataset name	Cohen (2006)	Matwin (2010)	Cohen (2008/2011)	Howard (2016)	**Paper A**	Paper A replicated	Absolute delta	**Paper B**	Paper B replicated	Absolute delta	fastText classifier
1	ACEInhibitors	0.625	0.582	0.795	**0.864**	0.850	0.848	0.16%	0.846	0.423	42.27%	0.846
2	ADHD	0.746	0.687	0.589	**0.862**	0.731	0.705	2.64%	0.765	0.771	0.59%	0.484
3	Antihistamines	0.053	0.210	0.302	0.197	**0.380**	0.343	3.67%	0.230	0.195	3.50%	0.102
4	Atypical Antipsychotics	0.212	0.287	0.246	0.339	**0.429**	0.269	16.00%	0.294	0.143	15.12%	0.301
5	Beta Blockers	0.340	0.425	0.525	0.487	**0.649**	0.521	12.78%	0.564	0.457	10.72%	0.478
6	Calcium Channel Blockers	0.183	0.305	0.518	**0.538**	0.512	0.428	8.35%	0.223	0.125	9.84%	0.244
7	Estrogens	0.284	0.529	0.579	**0.652**	0.557	0.522	3.58%	0.202	0.157	4.55%	0.441
8	NSAIDs	0.605	0.640	0.800	**0.865**	0.857	0.870	1.31%	0.688	0.721	3.33%	0.742
9	Opioids	0.184	0.609	0.417	**0.883**	0.588	0.635	4.74%	0.348	0.302	4.62%	0.614
10	Oral Hypoglycemics	0.176	0.169	**0.239**	0.213	0.182	0.221	3.84%	0.141	0.070	7.14%	0.186
11	Proton Pump Inhibitors	0.338	0.289	0.391	0.443	**0.466**	0.361	10.53%	0.303	0.185	11.83%	0.345
12	Skeletal Muscle Relaxants	0.050	0.317	0.426	**0.609**	0.338	0.338	0.04%	0.281	0.352	7.18%	0.141
13	Statins	0.303	0.373	0.553	0.496	**0.630**	0.549	8.13%	0.504	0.340	16.43%	0.469
14	Triptans	0.086	0.334	0.409	0.478	0.500	0.477	2.32%	0.326	**0.506**	18.03%	0.268
15	Urinary Incontinence	0.347	0.387	0.542	**0.655**	**0.655**	0.600	5.48%	0.360	0.255	10.49%	0.550
	Average drug	0.302	0.409	0.489	**0.572**	0.555	0.513	5.57%	0.405	0.333	11.04%	0.414
16	COPD	—	—	—	—	**0.809**	0.808	0.08%	—	0.196	—	0.680
17	Proton Beam	—	—	—	—	**0.910**	0.906	0.41%	—	0.426	—	0.852
18	Micro Nutrients	—	—	—	—	0.758	**0.759**	0.09%	—	0.263	—	0.693
	Average clinical	—	—	—	—	**0.826**	0.824	0.19%	—	0.295	—	0.742
19	PFOA/PFOS	—	—	—	0.867	**0.911**	0.901	0.98%	0.122	0.360	23.79%	0.841
20	Bisphenol A (BPA)	—	—	—	0.813	**0.855**	0.841	1.36%	0.854	0.424	42.93%	0.696
21	Transgenerational	—	—	—	0.775	0.768	**0.780**	1.16%	0.769	0.050	71.93%	0.424
22	Fluoride and neurotoxicity	—	—	—	0.930	0.858	0.865	0.68%	**0.943**	0.868	7.57%	0.444
23	Neuropathic pain	—	—	—	**0.884**	0.784	0.772	1.24%	0.798	0.160	63.84%	0.790
	Average SWIFT	—	—	—	**0.854**	0.835	0.832	1.09%	0.697	0.372	42.01%	0.639
	Grand average	—	—	—	—	**0.651**	0.623	3.90%	—	0.337	18.78%	0.506

were not able to successfully train the Multi-Channel CNN model. All of these results raise concerns about replicability.

Next, we compare our replicated results and the original ones from Paper A and B to previous benchmark studies. Paper A only compares their model to custom baseline methods and does not mention the previous state of the art results. None of the tested neural network-based models can improve on the results by Howard et al. [99], which uses a log-linear model with word-score and topic-weight features to classify the citations. This means that even though deep neural network models can provide significant gains in TNR@95% scores, they can still be outperformed by classic statistical methods.

7.2.2 Impact of input features

As we encountered memory problems when training the Paper B model on *Transgenerational* and *Neuropathic pain* datasets, we exclude these two datasets from our comparisons in the remaining experiments.

None of the papers provided the original input data used to train the models. We wanted to measure if the results depend on how that input data was gathered. We implemented two independent data gathering scripts using the `biopython` package as suggested by Paper B to obtain 18 out of 23 datasets. One implementation relied on the Medline

Table 7.3: Influence of input document features on the $TNR@95\%$ score for three tested models. "All features" column means a single string concatenating Title, Abstract, Author and Journal information. For each row, **bold** values indicate the highest score for each model, underlined values indicate the highest score across all 3 models.

	DAE-FF				Multi-Channel CNN				fastText classifier			
Datset name	All features	Title and Abstract	Abstract only	Title only	All features	Title and Abstract	Abstract only	Title only	All features	Title and Abstract	Abstract only	Title only
ACEInhibitors	0.785	0.709	0.658	**0.806**	0.367	0.461	0.648	0.525	**0.783**	0.776	0.765	0.441
ADHD	0.639	0.500	0.404	**0.651**	**0.704**	0.528	0.692	0.580	0.424	**0.470**	0.444	0.200
Antihistamines	**0.275**	0.168	0.265	0.016	0.135	**0.204**	0.114	0.105	0.047	0.124	0.175	**0.192**
Atypical Antipsychotics	0.190	0.221	**0.230**	0.046	0.081	**0.086**	0.050	0.013	**0.218**	0.188	0.185	0.095
Beta Blockers	**0.462**	0.451	0.390	0.408	**0.399**	0.243	0.134	0.211	**0.419**	**0.419**	0.407	0.262
Calcium Channel Blockers	**0.347**	0.337	0.297	0.137	0.069	0.083	0.004	**0.117**	0.178	0.139	0.060	**0.244**
Estrogens	**0.369**	0.358	0.331	0.145	0.083	0.076	0.051	**0.092**	**0.306**	0.199	0.108	0.241
NSAIDs	**0.735**	0.679	0.690	0.658	**0.601**	0.443	0.358	0.225	**0.620**	0.506	0.512	0.535
Opioids	**0.580**	0.513	0.499	0.280	0.249	**0.420**	0.413	0.287	**0.559**	0.558	0.534	0.245
Oral Hypoglycemics	0.123	**0.129**	0.107	0.019	0.013	**0.021**	0.004	0.005	**0.098**	0.049	0.042	0.016
Proton PumpInhibitors	**0.299**	0.291	0.153	0.285	**0.129**	0.121	0.059	0.118	0.283	0.228	0.174	**0.360**
Skeletal Muscle Relaxants	0.286	0.327	**0.430**	0.125	0.300	**0.329**	0.242	0.202	0.090	0.142	0.180	**0.210**
Statins	**0.487**	0.434	0.392	0.255	**0.283**	0.231	0.120	0.082	**0.409**	0.376	0.281	0.228
Triptans	**0.412**	0.253	0.320	0.199	**0.440**	0.404	0.407	0.129	0.210	0.205	0.211	**0.075**
Urinary Incontinence	0.483	**0.531**	0.482	0.372	**0.180**	0.161	0.046	0.099	**0.439**	0.310	0.170	0.434
Average drug	**0.431**	0.394	0.373	0.293	**0.269**	0.254	0.223	0.185	**0.339**	0.313	0.283	0.252
COPD	0.665	0.665	0.676	**0.677**	0.128	**0.372**	0.087	0.093	0.312	**0.553**	0.546	0.545
Proton Beam	**0.812**	0.810	0.790	0.799	0.357	0.489	0.408	**0.559**	0.733	0.761	**0.771**	**0.771**
Micro Nutrients	0.663	0.648	0.665	**0.677**	0.199	0.255	0.251	**0.268**	**0.608**	0.602	0.605	0.601
Average clinical	**0.713**	0.708	0.670	**0.718**	0.228	**0.372**	0.249	0.307	0.551	0.638	**0.640**	0.639
PFOA/PFOS	0.713	0.839	**0.847**	0.696	0.305	**0.405**	0.391	0.109	0.779	**0.796**	0.778	0.292
Bisphenol A (BPA)	**0.780**	0.754	0.715	0.631	0.369	0.300	**0.612**	0.182	**0.637**	0.630	0.499	0.079
Fluoride and neurotoxicity	0.806	**0.838**	0.758	0.726	**0.808**	0.688	0.654	0.452	**0.390**	0.375	0.292	0.250
Average SWIFT	0.766	**0.782**	0.774	0.684	0.494	0.464	**0.552**	0.247	**0.602**	0.600	0.523	0.207
Average (All datasets)	**0.520**	0.498	0.481	0.410	0.295	**0.301**	0.274	0.212	**0.407**	0.400	0.368	0.301

module, where a document was represented as a dictionary of all available fields. The second implementation returned all possible fields (title, abstract, author and journal information) concatenated in a single string. Furthermore, we examined how robust the models are, if the input data contained only titles or abstracts of the citations. Results are presented in the Table 7.3.

The best average $TNR@95\%$ results are obtained for all three models when they use all available features (Figure 7.1). All models achieved better results when using just the abstract data compared to the titles alone. This reaffirms our common sense reasoning that titles alone are not sufficient for citation screening. However, there are some specific datasets for which best results were obtained when the input documents contained only titles or abstracts. While this experiment does not indicate why this is the case, we can offer some potential reasons: (1) it could be that eligible citations of these datasets are more similar in terms of titles or abstract; (2) it could be that these models are not able to retrieve relevant information when there is too much noise. Intra- and inter-class dataset similarity need to be further evaluated in future studies.

As presented in Table 7.2, the fastText classifier model was not able to outperform the original results from Paper A and B. However, compared to our replicated results of Paper B, the fastText classifier achieves higher $TNR@95\%$ scores on 18 out of 23 datasets.

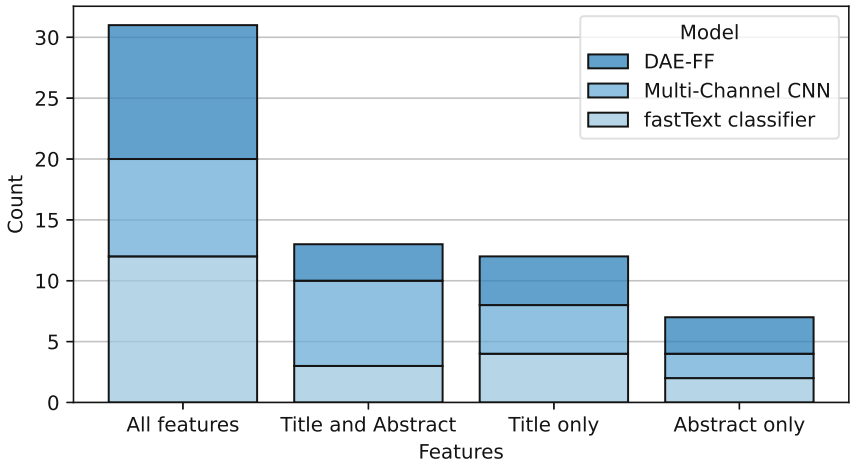


Figure 7.1: A count of experiments in which a model using a specific input feature achieved the best results. Models that use all available features scored the best results 49% of times for a specific (model, dataset) combination.

It is also more robust to random initialisation compared to Multi-Channel CNN.

7.2.3 Training time

We computed the training time for each of the models. The relationship between dataset size and model training time is visualised in Figure 7.2. For the DAE-FF model, we calculated both the training procedure of denoising autoencoder, feed-forward networks, and linear SVM. The DAE component is the most time-absorbing component as it consumes, on average, 93.5% of the total training time. For the fastText and Multi-Channel CNN models, we calculated the training procedure of the binary classifier.

For small datasets containing less than 1,000 documents, one validation fold for fastText took on average 2 seconds, for Multi-Channel CNN 13 seconds, and DAE-FF 82 seconds. Training time difference increases for larger models, where the speed of fastText is even more significant. For the largest dataset, *Transgenerational*, the mean training time for fastText is 78 seconds, for Multi-Channel CNN 894 seconds and for DAE-FF, it is 18,108 seconds. On average, the fastText model is 72 times faster than DAE-FF and more than eight times faster than Multi-Channel CNN, although this dependency is not linear and favours fastText for larger datasets.

7.2.4 Normalised Precision at 95% Recall

In this section, we measure the normalised Precision at a Recall level of 95% ($nP@95\%$). We are inspired by Paper A, which reported Precision at 95% (P@95%) scores in their experiments. However, as P@95% exhibits the same problems as WSS@95%, we decide

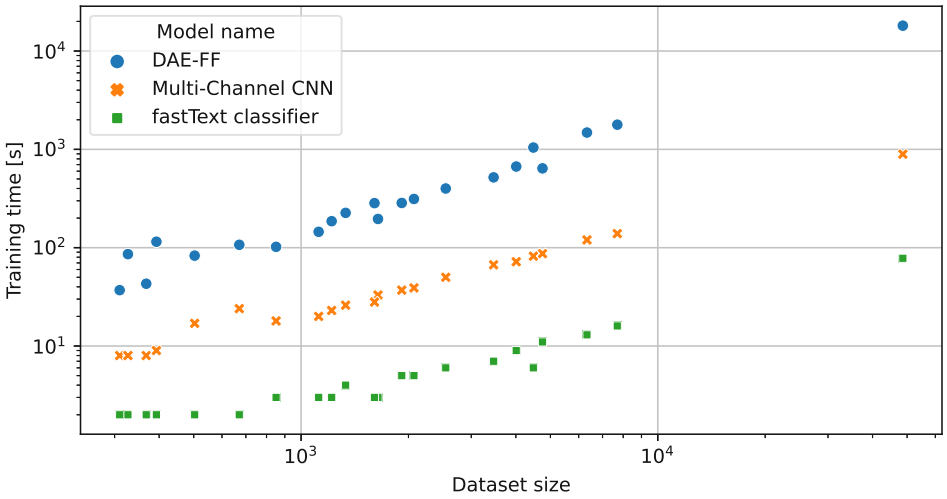


Figure 7.2: The relationship between dataset size and a model training time for the three evaluated models. Both training time and dataset size are shown on a logarithmic scale.

to report the $nP@95\%$ score (see Chapter 5 for a detailed explanation of problems with WSS and Precision when evaluated at a fixed Recall level). We calculate $nP@95\%$ using Equation 5.43 from Section 5.5.1.

Table 7.4 shows mean scores for each model across all three review groups. Similarly to the $TNR@95\%$ metric, the best-performing model is DAE-FF, achieving a mean $nP@95\%$ on 21 datasets equal to 11.1%. This method significantly outperforms Multi-Channel CNN and fastText models by 4.1 percentage points (pp) and 5.0pp, respectively.

Finally, for completeness, we calculate and compare the average P@95% (non-normalised) as reported in Paper A. Paper A presents an average P@95% of 19% aggregated across 23 review datasets. In contrast, our score stands at 16.7% over 21 review datasets, aligning closely with the findings in Paper A.

7.3 Discussion

Our replicability study offers insights into the practical application of the models. In this section, we discuss the challenges with the reproducibility process and the stability of model predictions across datasets and validation splits. We also compare two evaluation measures employed in this study: True Negative Rate and normalised Precision.

Table 7.4: A comparison of Normalised Precision at 95% Recall ($nP@95\%$) for the three models across 21 benchmark datasets. We did not evaluate the two largest SWIFT datasets: *Transgenerational* and *Neuropathic pain*. **Bold** denotes the highest score in each dataset. A † symbol indicates that DAE-FF's average $nP@95\%$ is statistically significantly superior to the other models, based on the Wilcoxon signed-rank test with Bonferroni correction.

Dataset name	DAE-FF	Multi-Channel CNN	fastText
ACEInhibitors	0.063	0.069	**0.075**
ADHD	0.048	**0.069**	0.030
Antihistamines	0.016	**0.021**	0.014
Atypical Antipsychotics	**0.055**	0.045	0.047
Beta Blockers	**0.036**	0.025	0.018
Calcium Channel Blockers	**0.064**	0.014	0.024
Estrogens	**0.168**	0.079	0.089
NSAIDs	**0.331**	0.266	0.161
Opioids	**0.037**	0.016	0.015
Oral Hypoglycemics	**0.072**	0.030	0.034
Proton PumpInhibitors	**0.024**	0.021	0.017
Skeletal Muscle Relaxants	**0.005**	0.004	0.003
Statins	**0.034**	0.026	0.020
Triptans	0.026	**0.032**	0.016
Urinary Incontinence	**0.190**	0.070	0.079
Average Drug	**0.078**	0.052	0.043
COPD	**0.342**	0.101	0.213
Proton Beam	**0.317**	0.249	0.222
Micro Nutrients	**0.152**	0.116	0.069
Average Clinical	**0.270**	0.155	0.168
PFOA/PFOS	**0.125**	0.050	0.092
Bisphenol A (BPA)	**0.059**	0.040	0.037
Fluoride and neurotoxicity	**0.162**	0.119	0.008
Average SWIFT	**0.115**	0.070	0.046
Average (21 datasets)	**0.111**†	0.070	0.061

7.3.1 Challenges with reproducibility

The authors of both papers uploaded their code into public GitHub repositories.[4,5] Both models were written in Python 3 and depend primarily on TensorFlow and Keras deep learning frameworks [2]. The whole implementation was uploaded in four commits for Paper A and one commit for Paper B (excluding commits containing only documentation).

[4] https://github.com/gkontonatsios/DAE-FF
[5] https://github.com/rvdinter/multichannel-cnn-citation-screening

Except for the uploaded Python scripts, there is no explicit information about versions of the packages used to train and evaluate the models. This missing information is crucial for replicability, as, for TensorFlow alone, in 2020, there were 27 different releases related to 6 different MINOR versions[6].

Availability of the code alone does not guarantee a replicable experimental setup. If the project was not documented for the specific software versions, it might be challenging to reconstruct these requirements based exclusively on the code, especially if the experiments were conducted some time ago. In the case of code written in Python, explicitly writing environment version with, for example, pip's `requirements.txt` or conda's `environment.yml` files should be sufficient in most of the cases to save time for researchers trying to replicate the experiments.

Neither paper included the original datasets they used to train and evaluate their models. Paper A provided sample data consisting of 100 documents, which presents the input data format accepted by their model, making it easier to re-run the experiments. Paper B does not include sample data but describes where and how to collect and process the datasets. Thorough descriptions of dataset collection and preparation are crucial, as these steps are time-consuming when inadequately explained.

Finally, paper B also tried to replicate the DAE-FF model from Paper A. They stated that *"(...) we aimed to replicate the model (...) with open-source code via GitHub. However, we could not achieve the same scores using our dataset. After emailing the primary author, we were informed that he does not have access to his datasets anymore, which means their study cannot be fully replicated."*. Our results are contrary to findings by Paper B: we managed to replicate the results of Paper A successfully without having access to their original datasets. Unfortunately, Paper B does not present any quantitative results of their replicability study. Therefore, we cannot draw any conclusions regarding those results as we do not know what Paper B authors meant by *"cannot be fully replicated"*.

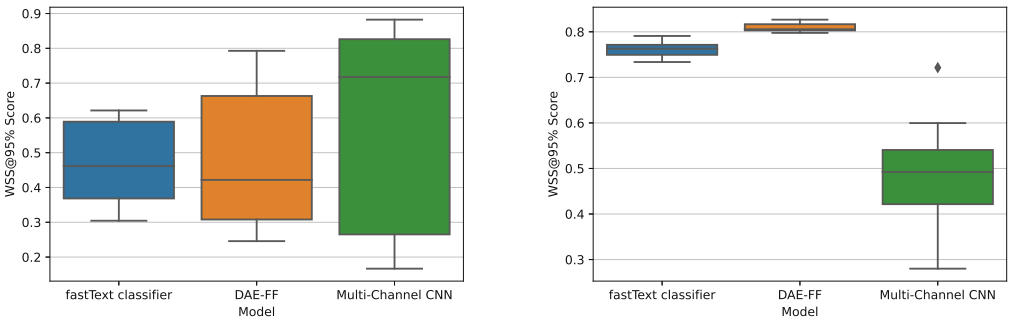
7.3.2 Model stability across validation splits

Figure 7.3 presents results for *ADHD* and *Proton Beam* datasets for all three models. The Multi-Channel CNN model has the widest range of $TNR@95\%$ scores across cross-validation runs. This is especially evident in the datasets from the Clinical group (i.e. *Proton Beam*), for which the DAE-FF and fastText models yield very steady results across every cross-validation fold. This could mean that the Multi-Channel CNN model is less stable, and its good performance is dependent on random initialisation.

7.3.3 Comparison of TNR and nPrecision

In Chapter 5, we proposed and theoretically analysed True Negative Rate ($TNR@r\%$) and normalised Precision ($nP@r\%$) as two distinct measures for evaluation of citation screening at a fixed Recall level. We suggested that the $TNR@r\%$ measure is generally

[6] `https://pypi.org/project/tensorflow/#history`



(a) ADHD review dataset. (b) Proton Beam review dataset.

Figure 7.3: Example boxplots with $TNR@95\%$ scores for three models. Input features are titles and abstracts.

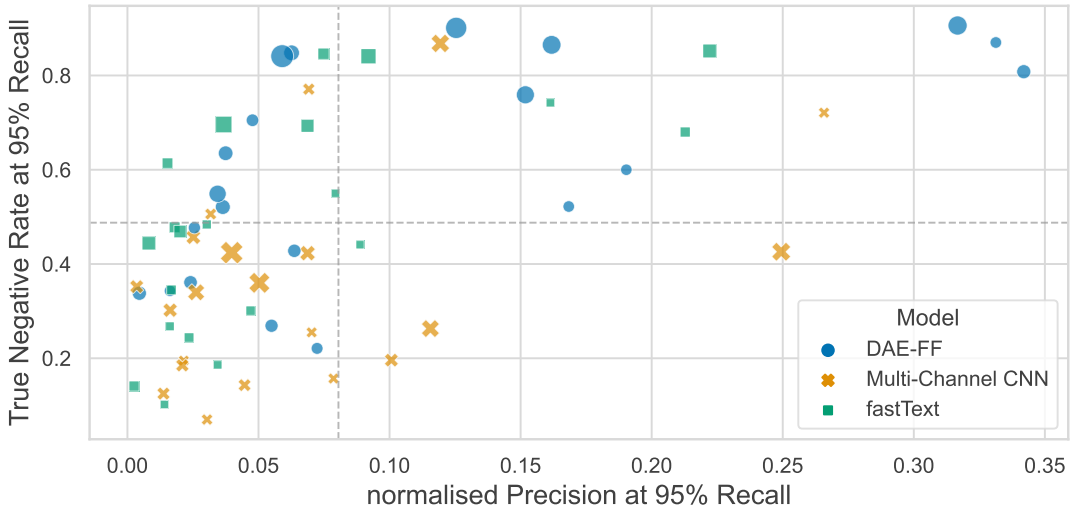


Figure 7.4: Scatter plot of normalised Precision (nP) versus True Negative Rate (TNR) at 95% Recall across three tested models. This figure illustrates the trade-off between $nP@95\%$ and $TNR@95\%$ scores. The size of each marker represents the relative size of the dataset used. Average $nP@95\%$ and $TNR@95\%$ are indicated by dashed grey lines.

better suited for evaluating screening as it is based on the number of correctly identified irrelevant documents (TNs). Furthermore, it can be used to estimate time and money savings when using automation models. On the other hand, nPrecision can be used to evaluate the successful screening of the few last relevant documents, as its score is not linear and disproportionately rewards the latter stages of accurate screening. This section analyses the empirical results obtained on 21 systematic review datasets using the three tested models.

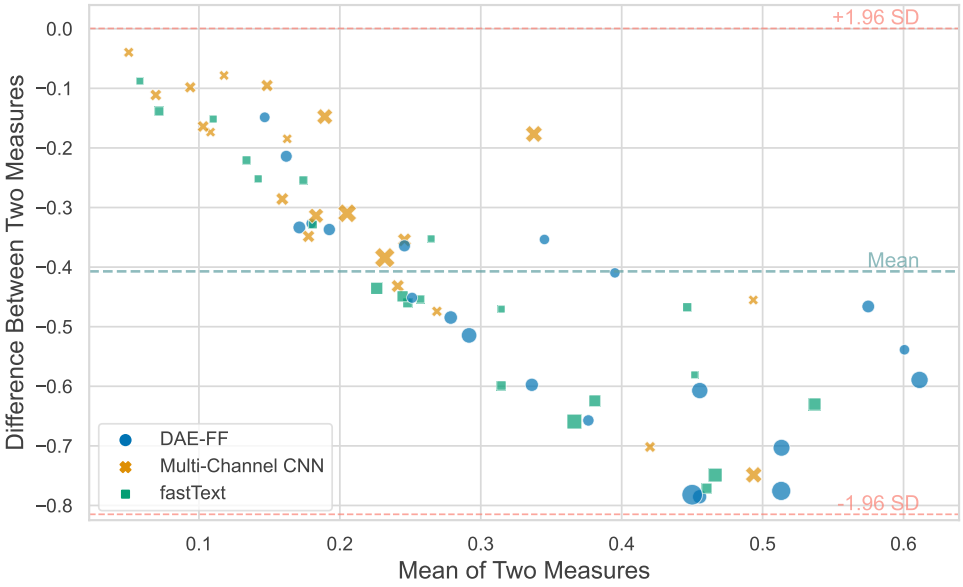


Figure 7.5: Visualisation of agreement between $nP@95\%$ and $TNR@95\%$ across three tested models using a Bland-Altman plot. Each point represents the mean of the two measures plotted against their difference, with the marker size indicating the dataset size. The mean difference and the limits of agreement at ±1.96 standard deviations are marked as dashed horizontal lines.

Figure 7.4 displays a scatter plot contrasting $TNR@95\%$ with $nP@95\%$. The plot reveals a range of values for both metrics across the tested models, indicating variability in performance. The moderate positive correlation suggests that models with higher normalised Precision at 95% Recall tend to have higher True Negative Rates at the same recall level. However, the dispersion of data points highlights that while there is a relationship between the two measures, they do not necessarily increase in tandem. This observation underscores the importance of considering both metrics to gain a comprehensive understanding of a model's performance in citation screening tasks.

A Pearson coefficient of 0.60 between $nP@95\%$ and $TNR@95\%$ indicates a positive, yet not equivalent, relationship. A paired T-Test, yielding a p-value well below 0.05, further confirms the distinctiveness of these metrics.

Figure 7.5 uses a Bland-Altman plot to examine $nP@95\%$ and $TNR@95\%$ agreement. This plot, showing differences against averages, assesses model consistency across datasets. General agreement between $nP@95\%$ and $TNR@95\%$ is observed, though variances in larger datasets for DAE-FF and fastText models indicate some measure inconsistencies.

Correlation analysis shows a weak positive correlation (0.09) between dataset size and $nP@95\%$, suggesting only a minimal influence of dataset size on normalised Precision. In contrast, a moderate positive correlation (0.43) between dataset size and $TNR@95\%$

indicates that larger datasets tend to yield higher TNRs. This difference further highlights that nPrecision and TNR focus on distinct aspects of the screening task. Analysis on a larger number of datasets and models could enhance these findings.

7.4 Summary

In this chapter, we presented the results of our experiments citation screening using three architectures based on deep neural networks. The model proposed by Paper A consists of a denoising autoencoder combined with feed-forward and SVM layers (DAE-FF). Paper B introduces a multi-channel convolutional neural network (Multi-Channel CNN). We used the 23 publicly available datasets to measure the quality of both models. The average delta between our replicated results and the original ones from Paper A is 3.59%. Considering that we do not know the random seed used for the training of original models, we can conclude that the replication of Paper A was successful. The average delta for Paper B is 17.63%. In addition to that, this model is characterised by a significant variance, so we cannot claim successful replication of this method. These observations underscore the importance of stability in model performance, particularly in the context of citation screening, where accuracy is paramount.

Subsequently, we evaluated the fastText classifier and compared its performance to the replicated models. This shallow neural network model based on averaging word embeddings achieved better $TNR@95\%$ results when compared to replicated scores from Paper B and, at the same time, is, on average, 72 and 8 times faster during training than both Paper A and B models. This analysis highlights a trade-off between complexity and performance, emphasising the potential of less complex models in real-world applications where computational efficiency is a concern.

None of the tested models can outperform all the others across all the datasets. DAE-FF achieves the best average results, though it is still worse when compared to a statistical method with the log-linear model. Models using all available features (title, abstract, author and journal information) perform best on the average of 21 datasets when compared to just using a title, abstract or both. We have noted a deficiency across all models as none of them could create stable gains over all datasets, and they had a considerable variation in the scores.

We also presented the empirical analysis of two evaluation measures: $TNR@r\%$ and $nP@r\%$ when evaluated at a fixed recall cutoff. We showed how these two evaluation measures can be used to present model performance on different aspects of the screening process, complementing our findings from the previous Chapters.

Since the completion of this work, new neural models, even larger and more capable according to general benchmarks, have been introduced. However, we believe that insights from this chapter are enduring and can be generalised to other models. Especially if not even stronger, the findings related to the reproducibility in the context of large language models.

CHAPTER 8

Automated Citation Screening with Eligibility Criteria

Natural language prompting has recently demonstrated significant progress for the pretraining of language models, for tasks like large scale multi-task supervision [208, 229, 200] and improving zero-shot classification via explicit, multi-task prompted training data [223, 281, 28]. With appropriate tools and integration of expert-labelled datasets, these performance gains were reported to scale to thousands of prompted training tasks [286]. Such changes hold great promise for improving learning with limited labels which would be very helpful for systematic review researchers.

An initial trend with the improvements in neural network architectures was that they became even more data-greedy, making it harder to train for a problem with minimal gold standard annotated data. A recent trend was to improve models' capacity through intensive research in transfer learning [14] and few-shot learning [287, 74]. Zero-shot learning extends in this direction and aims to make predictions on classes whose instances were not observed during the model's training [285]. Prompt-based learning is a strategy of training large language models to use the same model for different tasks without re-training in a zero-shot setting [155].

Building on experiments from previous chapters, we explore how systematic review metadata can be used to facilitate the screening process. We want to assess the quality of automated screening in a scenario which does not require expensive manual annotations for every new review. Inspired by our work in clinical trial matching, we focus on using external information to improve screening quality. We start by experimenting with the CSMED-COCHRANE dataset. Then, we evaluate several fine-tuned Transformer models and zero-shot prompting of GPT models on the CSMED-FT dataset. We are interested in understanding if and which information is most suitable for screening on title and abstract, and full text levels.

Finally, we present CRUISE–*Screening*, a web application that helps researchers conduct living literature reviews. Compared to other tools, CRUISE–*Screening* combines search and screening stages into one workflow using large language models. Thanks to this, researchers can use the tool as an information system to systematise, manage and record their literature review workflows.

8.1 Citation Screening with External Information

Our proposed approach is based on the recent advancements in language modelling to conduct a zero-shot ranking or classification of papers. We consider metadata in the CSMED-COCHRANE dataset as various query representations for measuring their impact on screening.

8.1.1 Experiment setup

In this experiment, we evaluate the impact of the SLR protocol section on ranking for statistical and neural models in a zero-shot setting. We use two statistical models BM25 and TF-IDF, and three Transformer-based models: MiniLM-L6-v2[1], mpnet-base-v2[2] and BioBert-snli[3] from the SentenceTransformers library [212]. MiniLM model uses 256 tokens, whereas MPNet and BioBERT use 512 tokens.

We test four different SLR meta-data sections from the SLR protocol as input representations: (1) title, (2) abstract, (3) search strategy and (4) eligibility criteria. Predictions are run on the CSMED-COCHRANE-DEV split. We use the `retriv` Python library for implementing the pipeline [15].

8.1.2 Evaluation

We select True Negative Rate at 95% Recall (TNR@95%) and normalised Precision at 95% Recall (nP@95%) as primary evaluation measures. We also evaluate the average position at which the last relevant item is found [112, 113, 114], calculated as a percentage of the dataset size (*Last Rel*). Lower values of *Last Rel* indicate better performance. Additionally, we compute traditional evaluation measures: $nDCG$@10, MAP and Recall at rank k (R@k), with k in {10, 50, 100} following the evaluation from Kanoulas et al. [112].

8.1.3 Results

Table 8.1 presents the results on the CSMED-DEV-COCHRANE dataset. Overall, we find that models using SLR abstracts and eligibility criteria perform the best with the consistent superiority of neural network-based models over traditional retrieval

[1] `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`
[2] `https://huggingface.co/sentence-transformers/all-mpnet-base-v2`
[3] `https://huggingface.co/pritamdeka/S-BioBert-snli-multinli-stsb`

Table 8.1: Results of zero-shot evaluation on CSMED-COCHRANE-DEV dataset. For each measure, **bold** values indicate the highest score for each model across query representation. Underlined values indicate the highest score across all tested models.

Model	Representation	TNR@95%	nP@95%	Last Rel	nDCG@10	MAP	R@10	R@50	R@100
BM25	Title	0.469	0.142	72.2	0.438	0.388	0.349	0.623	0.704
	Abstract	**0.474**	**0.170**	**63.6**	**0.503**	**0.453**	**0.379**	**0.657**	**0.757**
	Search strategy	0.379	0.093	72.1	0.336	0.311	0.268	0.507	0.625
	Criteria	0.430	0.145	67.0	0.452	0.417	0.345	0.629	0.725
TF-IDF	Title	0.439	0.126	75.1	0.334	0.322	0.295	0.575	0.661
	Abstract	**0.490**	**0.147**	**62.8**	**0.417**	**0.404**	**0.348**	**0.640**	**0.728**
	Search strategy	0.372	0.078	72.9	0.271	0.272	0.233	0.500	0.595
	Criteria	0.453	0.139	67.0	0.375	0.372	0.305	0.616	0.704
MiniLM	Title	0.472	0.217	68.1	0.470	0.414	0.379	0.673	0.763
	Abstract	0.492	**0.240**	65.5	**0.517**	0.451	**0.398**	**0.680**	**0.782**
	Search strategy	0.411	0.171	71.4	0.370	0.346	0.320	0.609	0.688
	Criteria	**0.527**	0.198	**60.9**	0.497	**0.456**	0.384	0.657	0.747
MPNet	Title	0.467	0.230	66.6	0.476	0.429	0.376	0.684	0.774
	Abstract	0.516	**0.265**	63.8	**0.556**	0.482	**0.420**	**0.692**	0.777
	Search strategy	0.429	0.181	68.6	0.400	0.372	0.328	0.614	0.699
	Criteria	**0.545**	0.216	**58.5**	0.514	**0.488**	0.393	0.691	**0.784**
BioBERT	Title	0.439	0.141	66.7	0.391	0.369	0.337	0.624	0.717
	Abstract	0.494	0.166	64.4	0.463	0.448	**0.367**	0.655	**0.768**
	Search strategy	0.369	0.098	72.9	0.350	0.335	0.273	0.523	0.635
	Criteria	**0.507**	**0.182**	**62.7**	**0.494**	**0.468**	0.358	**0.681**	0.765

models. The topical similarity between the publications and the SLR abstract suggests an important role of the abstract in the automated screening process.

Across all measures for both statistical models, representing SLR using its abstract consistently outperforms others. This indicates that abstracts, as a source of external knowledge, contain more comprehensive and relevant information for automated citation screening compared to titles or search strategies. This finding is aligned with our analysis of the statistical models for the clinical trials matching task, which also showed the inability of these models to comprehend the eligibility criteria (Chapter 2).

On the other hand, more advanced neural models tend to utilise the eligibility criteria information better. TNR@95% is higher when using the criteria information for all three considered Transformer-based models. Similar considerations can be given about other evaluation measures, where we notice that with growing model size and input window, their performance is getting better when using the criteria section compared to SLR abstract. However, It should be noted that the criteria section is typically more relevant to the full text screening step than title and abstract screening.

The best-performing model, MPNet, using SLR eligibility criteria, achieves TNR@95% equal to 0.545, meaning that this model can remove, on average, more than half of the true negatives when achieving a recall of 95%. We also see that TNR and nP measures are not always aligned between model and representation combination.

Table 8.2 shows the word count statistics of SLR sections for CSMED-COCHRANE-ALL

Table 8.2: Systematic literature review protocol section lengths in number of words for CSMeD-cochrane-all dataset.

Word count	Abstracts	Titles	Search strategy	Eligibility criteria
Mean	720.8	10.8	567.6	852.2
25th Percentile	574.0	7.0	129.5	450.5
50th Percentile	718.0	10.0	273.0	662.0
75th Percentile	878.0	13.0	610.0	1005.0
90th Percentile	976.0	17.0	1196.8	1503.4

dataset. The text is truncated for more than half of the examples in the case of SLR abstracts and eligibility criteria. This also prevents the use of the cross-encoder approach, where the concatenated publication and SLR section would exceed the maximum context window for typical BERT-style models. Using models allowing for longer input sequences could enhance the ranking quality. Exploring large language models or advanced training scenarios like the Topical-Criteria Re-Ranking curriculum learning [136] might also reveal the potential for further improving the results.

8.2 Full Text Screening

We present how CSMeD-FT can be used to evaluate LLMs capabilities in understanding eligibility criteria sections for the purpose of screening very long documents. Specifically, we are interested in evaluating how the models are able to process the inclusion and exclusion criteria. We run experiments both with fine-tuned Transformer models and zero-shot prompting of GPT models.

8.2.1 Experiment setup

As the combined input size of systematic review and publication can be very big (9,246 mean number of tokens on a training split measured with a GPT-4 tokeniser), we select only models that allow for at least 4k tokens context input. For fine-tuning, we choose Longformer [20] and BigBird [292], and their domain-specific models pretrained on clinical data: Clinical-BigBird and ClinicalLongformer [154]. For zero-shot evaluation, we select GPT-3.5-turbo-0301, GPT-4-8k-0314 and GPT-3.5-turbo-16k-0613 accessed via OpenAI API. GPT-4-8k and GPT-3.5-turbo-16k are the only model capable of handling more than 4k input tokens, with context window size of 8k and 16k tokens respectively.

For all models, we concatenate the review text with publication. We truncate the review description text to half of the available context window (2,000 tokens for 4k models, 4,000 tokens for 8k model and 8,000 tokens for 16k model) and fill the rest of available input with a publication.

Table 8.3: Statistics of a review text with respect to the fit within 2,048 tokens context window.

	CSMED-FT-TRAIN	CSMED-FT-DEV	CSMED-FT-TEST	CSMED-FT-SAMPLE
Avg # splits	1.13	1.24	1.83	1.74
Median # splits	1	1	1	1
Max # splits	2	2	4	4
Min # splits	1	1	1	1
More than 1 splits	13%	24%	42%	42%

Transformer models fine-tuning

We select the following model checkpoints from HuggingFace Transformers library:

- Longformer-base[4]
- BigBird-roberta-base[5]
- Clinical-Longformer[6]
- Clinical-BigBird[7]

We want to decide whether a publication fulfils all inclusion criteria and none of the exclusion criteria to include it in the SLR. Specifically, this means matching the eligibility criteria of SLR with the full text of the candidate publication. As input, the model receives the text of the review and publication and is asked to predict a binary category. We concatenate the review title with the eligibility criteria section to create the review text. For publications, we concatenate the title, abstract and the main text.

As available input text (review text + publication text) almost always exceeds the available context window of considered models (4,096 tokens), we use the following approach to allocate available space. We use the `TokenTextSplitter` method from the langchain library[8] with the gpt-3.5-turbo-0301 model to select the review text that would fit the context window. We select at most half of the available context window, so in the context of all Transformer models, review text equals at most 2,048 tokens. This action truncates some part of the eligibility criteria section, i.e. for 13% of items in the trainset and 42% in the test set (Table 8.3). We fill the remaining input sequence with the publication text.

We run our experiments on a single server with 4 Nvidia RTX 3090 GPUs with 24GB of RAM each. We fine-tune the Transformer models on CSMED-FT-TRAIN for four

[4] https://huggingface.co/allenai/longformer-base-4096
[5] https://huggingface.co/google/bigbird-roberta-base
[6] https://huggingface.co/yikuan8/Clinical-Longformer
[7] https://huggingface.co/yikuan8/Clinical-BigBird
[8] https://github.com/hwchase17/langchain

epochs. We use a per-device batch size of 1 with eight gradient accumulation steps. We test several learning rates with the best results on the validation split with 1×10^{-5}. We set weight decay to 0.01. We use AdamW [157] with default values of $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We evaluate models after each epoch on the validation set and select the model with the highest macro F1-score.

Zero-shot prompting of GPT-x models

Similarly, as for the fine-tuned classification models, we reserve at most half of the context window size for the systematic literature review description and fill the remaining tokens with the publication text. We measure the text length using the OpenAI library tiktoken[9], which provides tokenisers for GPT-3.5 and GPT-4 models.

If a whole publication text cannot fit inside a single context window, we run several predictions with non-overlapping sliding windows over the full text document. The final decision D for including a publication in the SLR is determined by the product of the binary predictions $P_1, P_2, \ldots, P_N$ for each of the N context windows.

$$D = P_1 \times P_2 \times \ldots \times P_N, \tag{8.1}$$

where $P_i = 1$ means that the model predicts to include a publication based on the i-th sliding window, and $P_i = 0$ means that the model predicts to exclude a publication. If $D = 1$, the publication is included; if $D = 0$, the publication is excluded. This means a publication cannot be included if there is even a single window with a prediction of 0 (exclude). In case of GPT-3.5-turbo-16k model, only for 4 out of 50 documents the model was unable to accommodate the full text of combined review and publication inside one prompt. We use the following prompt template:

Input Template:

```
Does the following scientific paper fulfill all eligibility \
criteria and should it be included in the systematic review? \
Answer 'Included' or 'Excluded'. \
Systematic review: "{{r.title}}" \n "{{r.criteria}}" \n\n \
Publication: "{{p.title}}" \n "{{p.abstract}}" \n \
"{{p.main_text}}" \n\n \
Answer:
```

Output Template:

```
{{label}}
```

[9] https://github.com/openai/tiktoken

Answer Choices:

```
Included ||| Excluded
```

We set our experimental budget to 50 USD. For the GPT-4 model we conduct the experiments only on the CSMED-FT-TEST-SMALL subset.

8.2.2 Evaluation

In assessing the performance of our models, we employ macro-averaged Precision, Recall, and F1-score as our primary metrics. The choice of macro-averaging is given the nature of full-text screening in systematic literature reviews. Unlike title and abstract screening, where irrelevant publications significantly outnumber relevant ones, the full-text stage often presents a more balanced dataset. This requires an evaluation approach that equally penalises models for over-inclusion or over-exclusion of documents. Therefore, we opted for a measurement approach that treats FP and FN with equal weight, reflecting their comparable impact on the review process. We also report the percentage of documents each model includes, combining TP and FN, to provide a practical perspective on model performance.

8.2.3 Training and inference time

One training epoch took around 30 minutes both for BigBird and Longformer-based models. For inference, Longformer architecture processed, on average, 2.9 samples per second, whereas BigBird models 2.65 samples per second. Making predictions on the entire test split of 636 documents took less than 4 minutes for all models.

For the GPT-3.5-turbo-16k model, making predictions on all 636 examples of the CSMED-FT-TEST split took 44 minutes. However, this value was heavily influenced by the default OpenAI's rate limits of 180,000 tokens per minute for our organisation.

8.2.4 Results

Results of the full text experiment are summarised in Table 8.4. On CSMED-FT-TEST-SMALL, GPT-4-8k strongly outperforms other models. However, this difference is not statistically significant. The GPT-3.5-turbo-16k achieves the highest Precision; this improvement can be attributed to the model's expanded context window and the limitations other GPT-based models have with our simple aggregation rules. However, this might also be caused by overfitting towards the positive class, as this model includes almost twice as many publications as other models. On CSMED-FT-TEST set, Clinical-BigBird, significantly outperforms zero-shot GPT-3.5 model and pre-trained models based on the LongFormer architecture.

Interestingly, both BigBird-based models outperform their counterparts using the Longformer architecture. The typical overall tendency to domain-pre-trained models achieving

Table 8.4: Results of the full text screening experiment averaged over documents. The statistical significance was assessed with a McNemar's t-test ($p < 0.05$) with Bonferroni correction for multiple testing. *Clinical-BigBird* on the CSMED-FT-TEST split showed statistically significant improvements compared to the *stratified random* baseline, *Longformer*, *Clinical-Longformer*, and *GPT-3.5-turbo-16k*, indicated by †. Stratified baseline is averaged from 100 different random seeds. '% incl.' describes the percentage of documents predicted as relevant by models (TP + FN).

	CSMED-FT-TEST-SMALL				CSMED-FT-TEST			
	% incl.	P	R	F1	% incl.	P	R	F1
ORACLE	44%	—	—	—	43.7%	—	—	—
stratified random	50%	0.497	0.498	0.495	—	0.499	0.499	0.498
'include all'	100%	0.220	0.500	0.306	100%	0.219	0.500	0.304
Longformer	40%	0.467	0.468	0.466	40.4%	0.398	0.400	0.398
BigBird-roberta-base	42%	0.572	0.571	0.572	45.1%	0.575	0.575	0.575
Clinical-Longformer	36%	0.547	0.544	0.542	35.1%	0.436	0.441	0.435
Clinical-BigBird	36%	0.590	0.584	0.583	32.8%	**0.623**†	**0.611**†	**0.609**†
GPT-3.5-turbo-0301	54%	0.585	0.586	0.580	—	—	—	—
GPT-4-8k-0314	58%	0.674	**0.672**	**0.660**	—	—	—	—
GPT-3.5-turbo-16k	80%	**0.712**	0.638	0.576	75.9%	0.538	0.528	0.475

higher scores over their open-domain counterparts is also preserved. We believe that fine-tuning the Transformer models first on larger NLI/QA corpora could help improve the results.

8.3 CRUISE–*Screening*

We present CRUISE–*Screening*, a web-based tool for conducting living literature reviews. The tool is developed to improve the efficiency of the literature review process. Its inception is rooted in the earlier discussions on using eligibility criteria for citation screening and the integration of LLMs in this process. This tool attempts to extend the rigorous methodologies of SLRs, typically reserved for professional medical reviews, to broader exploratory reviews in various fields. It aims to address the research question of whether the automation approaches employing LLMs and eligibility criteria are applicable beyond the conventional scope of SLRs in medicine.

CRUISE–*Screening* is connected to several search engines via API and facilitates the process of screening for relevant publications using NLP and machine learning methods. We discuss the development and functionalities of CRUISE–*Screening*, as well as present the challenges in developing such a tool. The system integrates search and screening capabilities into a single application and can connect with several search engines and machine learning models. We foresee two use cases for our system: (1) primarily by

researchers wanting to review the literature to locate the relevant work in their field of expertise; and (2) people developing automation models for literature reviews wanting to compare their approaches with others.

Most tools in this domain are developed specifically for systematic reviews, and using them for general literature reviews requires much overhead in installation and covering the review process. Our system, on the other hand, is among the first to apply systematic review concepts to general literature reviews. This sets our system apart from the rest of the literature review tools, which are primarily recommendation systems for papers. Our system can potentially promote collaboration and facilitate the exchange of ideas among researchers. Specifically, this can be conducted with the possibility of directly sharing reviews with other users and with metadata-reach import/export functionalities of CRUISE–*Screening*.

This section introduces CRUISE–*Screening* and presents the resources used, architecture and the methodologies behind its functionalities. Figure 8.1 shows the architecture of our system. CRUISE–*Screening* is built with Python 3.9, Django 4, Bulma and AlpineJS frameworks. The application is open-source under the Apache-2.0 license[10] and a demo is available under this URL.[11]

8.3.1 Data Resources

Good quality input data covering multiple domains is the crucial ingredient of a successful literature review. Nussbaumer-Streit et al. [185] found that combining two separate databases may suffice to reliably determine the conclusions of a systematic review in medicine. Therefore, CRUISE–*Screening* was designed to use multiple data sources and to allow for extending them when needed. Currently, it supports the following four search engines as data sources: Semantic Scholar API[12], CORE API[13], PubMed via Entrez API[14] and internal document storage.

The first three APIs call search engines that are used as primary data sources when searching for documents. Using three different search engines with contrasting scopes and content enables good search results coverage.

The tool also allows for indexing documents in the internal database. It is implemented using Elasticsearch and communicates with the main application using the API. It can be used, for example, when one wants to store private documents or content not covered by other search engines. For this demo, we index the DBLP-Citation-network Version 13 collection[15] created by Tang et al. [246].

[10] https://github.com/ProjectDoSSIER/CRUISE-Screening
[11] https://citation-screening.ec.tuwien.ac.at
[12] https://www.semanticscholar.org/product/api
[13] https://core.ac.uk/services/api
[14] https://www.ncbi.nlm.nih.gov/search/
[15] https://www.aminer.cn/citation

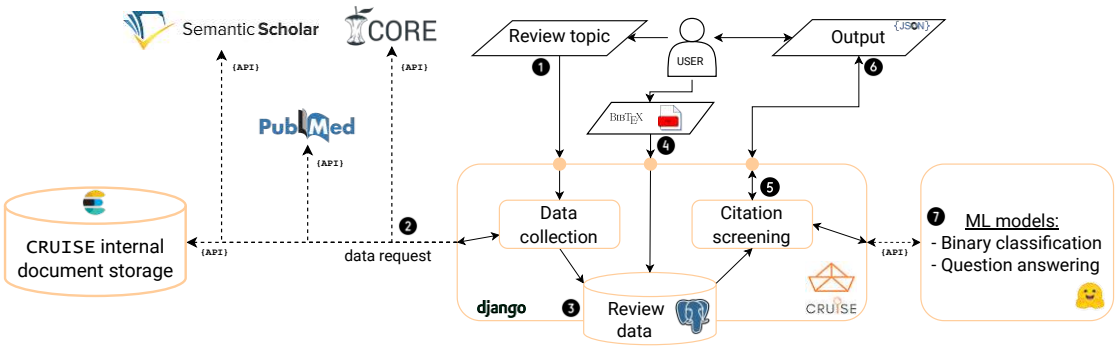


Figure 8.1: Overview of the CRUISE–*Screening* architecture. The numbers are referred to in Sections 8.3.2 and 8.3.3.

The system could easily be expanded to connect to other search engines offering API access. As the system is a meta-search engine, we use a script to deduplicate the search results based on papers' metadata (DOI, title, abstract and authors information).

8.3.2 Screening Workflow

As described in Chapter 3, the typical procedure for finding relevant documents for systematic literature reviews consists of two intertwined stages of search and screening. The first stage corresponds to researchers searching for documents potentially related to the research topic. In the second stage, the documents are screened for their eligibility to the SLR protocol. We have combined and implemented these two stages inside CRUISE–*Screening*.

Search for relevant items

First, the user creates a new literature review by defining the research protocol ❶. The protocol (Figure 8.2) consists of the review's title, description, at least one search query and a set of inclusion and exclusion criteria (eligibility criteria). The tool allows for specifying search engines in which one wants to search for papers, by default selecting all four available sources described in Section 8.3.1. The search can be limited to only the first N results if the reviewer is not interested in a comprehensive literature review.

CRUISE–*Screening* sends API requests to selected search engines and gathers all responses ❷. Merged and deduplicated search results are stored in a PostgreSQL database ❸. In order to support living reviews, the user can re-run the search function periodically to update the list of references. However, since search engines only allow filtering by publication year and not month or day, the tool removes publications older than the year of the previous search during updates. The tool then relies on deduplication to ensure that new publications are not mistakenly added twice.

CRUISE–*Screening* also allows for the additional direct import of data for screening from two sources ❹:

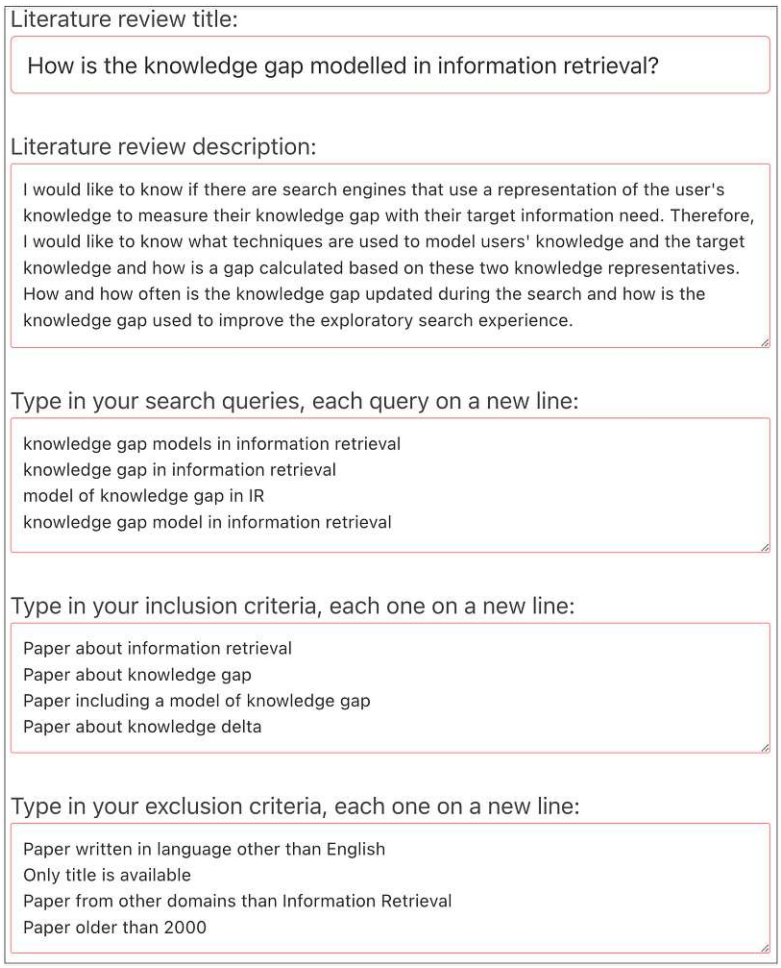


Figure 8.2: Example literature review protocol containing review title, description, search queries and criteria for inclusion and exclusion.

- Bulk upload from reference files – Currently, the tool supports BIB and RIS file extensions. Both of these formats are used by digital libraries (Scopus or IEEE Xplore) and the citation managers like Zotero or Mendeley. These publications are imported into the PostgreSQL database.

- Full text PDF files – Files can be provided either by their URL (assuming they are open access documents) or via a direct upload. These files are processed using GROBID [1] and then added to the database. Documents added this way can also be marked as seed studies. This way, these new documents are labelled as *relevant*, which can speed up the process of automated screening.

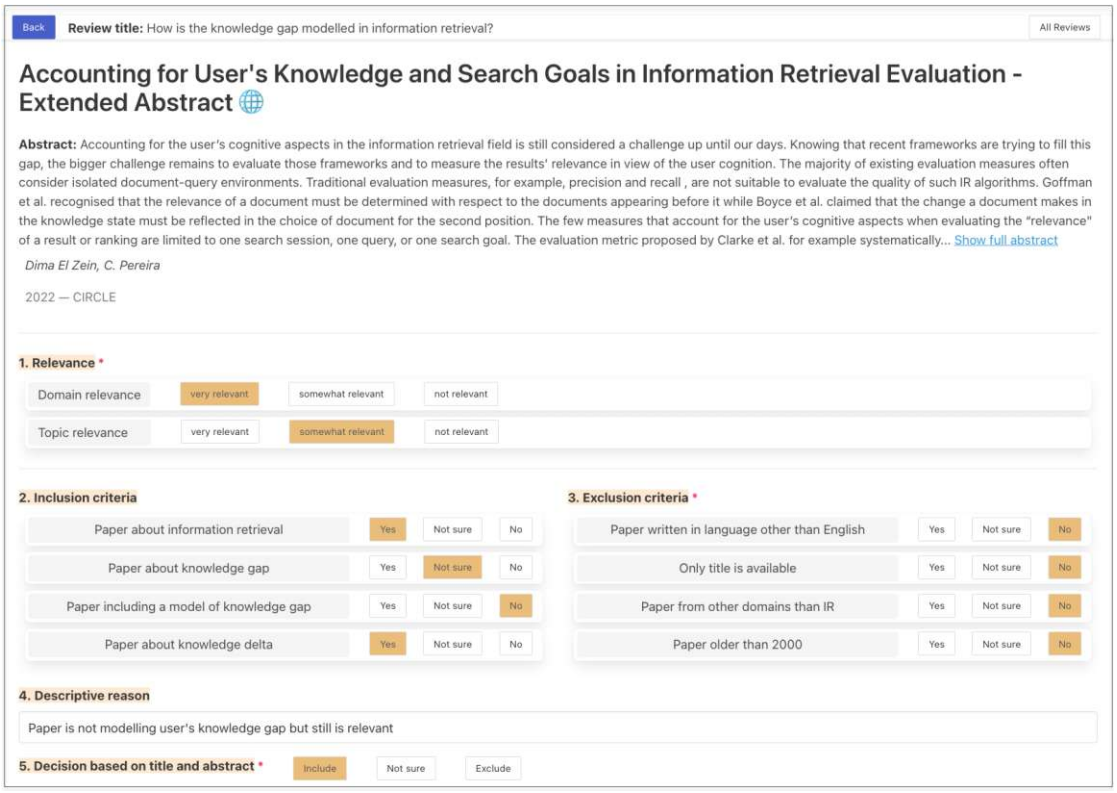


Figure 8.3: Example screening interface in CRUISE–*Screening* presenting single paper with answered questions.

Citation screening

Currently, CRUISE–*Screening* implements the title and abstract screening step ❺ while providing external URLs to full text articles whenever available. Figure 8.3 presents an example screening interface. From the top, it contains the title, abstract, authors, publication venue and year and the link to the full text of the screened paper. Below are two sections with eligibility criteria questions and a main inclusion question.

There are two screening workflows in CRUISE–*Screening*: strict and relaxed. Strict screening requires the annotators to conduct the process by manually answering every eligibility question. It mimics the citation screening process of systematic reviews. This mode could be used for in-depth systematic reviews or gathering manual annotations for training machine learning models.

The relaxed mode does not impose any requirements on which questions the annotator should answer except for the main *include*/*maybe*/*exclude* decision. There are optional questions about the reviewer's prior knowledge of the paper and authors, which reviewers can turn on to control for the selection bias.

8.3.3 Automation Methods

In addition to the fully manual workflow, CRUISE–*Screening* integrates automation methods to increase the speed of the literature review ❼. Using (semi-)automated workflow might also help increase the coverage of review as users can screen more documents in the same time. Implemented approaches include supervised text classification and zero-shot question-answering models. The tool connects to them using an API, which allows for extending the list of supported models.

Text classification

We implemented two examples of supervised classifiers based on previous literature: a logistic regression model using tf-idf text representation and a fastText classifier [110]. These models provide a single *yes*/*no* decision for each paper (corresponding to the main eligibility question from the manual workflow). Reviewers need to annotate a "training set" of at least three included and three excluded papers before using the models. When the reviewer annotates more publications, the models can be retrained to make an improved prediction on the remaining documents.

Question answering

In addition to supervised text classification, CRUISE–*Screening* enables users to conduct automatic screening using prompt-based language models with a question answering approach (see Section 3.3.1 for more details). The advantage of this method is that it does not require pre-labelled data and can make predictions for all inclusion and exclusion criteria. However, it can be computationally intensive and sensitive to the quality of input questions.

Our approach leverages recent language modeling advancements and prompt-based learning for zero-shot classification of papers. We treat this task as an independent inclusion/exclusion criteria analysis. Unlike previous work, our method uses all available information from the SLR protocol, combining eligibility criteria with the title and abstract of each paper in a prompt template. The decision to include a paper is based on whether it meets all inclusion criteria without satisfying any exclusion criteria. The final decision to include a paper in the subsequent review stage is made only when there is no exclusion criterion for which the answer was positive. Moreover, stored for each paper, these decisions can be easily presented to the manual annotator to justify the eligibility decision or simply as a visual hint when conducting the manual screening. An example prompt and workflow are presented in Figure 8.4.

For our demonstration, we used the T0_3B and T0 models [223]. We created a set of prompts for the models, covering all eligibility criteria questions. The example prompt consists of a single eligibility question and the same paper data as available during manual screening (Figure 8.3), namely the title, abstract, authors, journal name and publication year. The API is designed to support any Text2TextGeneration model implemented in the HuggingFace Transformers [283] library.

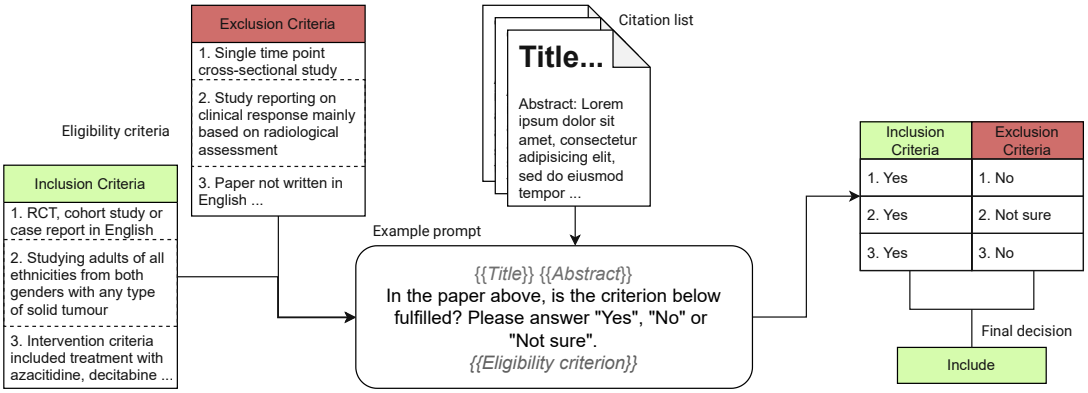


Figure 8.4: Example of citation screening using eligibility criteria and prompt-based learning. For every paper in the citation list, the inclusion and exclusion criteria are compared using the prompt template. This procedure generates two lists with textual answers for each criterion. The final decision is an aggregation of single outputs.

8.3.4 Review management

Once created, the review can be shared with other registered users, and the screening process can be distributed between researchers. There are two different levels of access: administrator and member. As it is often required in medical systematic reviews, the restriction can be imposed that every paper needs to be screened by at least N annotators.

The output of the literature review can be exported in JSON ❻. It contains the literature review protocol and all identified studies with corresponding automatic and manual decisions. CRUISE–*Screening* also enables for exporting completed reviews in a format that is compatible with the CSMED dataloaders, easing the entry level for people willing to provide their data for the purpose of evaluation of systematic review automation.

The evaluation of models is an essential aspect of our tool, as it allows researchers to evaluate their models without the risk of data leakage. Our tool enables researchers to make predictions with several models before starting a new review, storing the results, and evaluating the models after the manual review is conducted, without interfering with the manual workflow.

8.3.5 Deployment

CRUISE–*Screening* is available for on-premises deployment, offering a tailored, local instance for enhanced security and performance. To showcase its functionality, we offer a fully operational demo version accessible online. For local deployment, a Dockerfile and docker-compose file are provided in the GitHub repository,[16] allowing for easy installation and deployment on a local machine or a remote server.

[16] https://github.com/ProjectDoSSIER/CRUISE-Screening

8.3.6 Challenges with results merging

The merging of results from multiple sources can present significant challenges in the context of scientific publications. Although using multiple data sources can lead to better coverage of relevant studies, combining the results from different sources is not a trivial task. The data can be in formats, fields, and identifiers, which require significant effort to reconcile. Additionally, metadata quality can be poor in some cases, which can further complicate the merging process. Therefore, careful consideration must be given to the merging process, and the use of automated tools can help improve the accuracy and efficiency of the process. Nevertheless, human intervention may still be required to resolve any inconsistencies or errors in the data [80].

8.4 Discussion

In this section we discuss challenges with reusing eligibility criteria for screening, creating a meta-search engine for screening and limitations of CRUISE–*Screening*.

8.4.1 Reusing eligibility criteria for screening

In our examination of CSMED, we identified a significant challenge: the eligibility criteria in many past systematic review protocols are not immediately suitable for use in prompts. These criteria are either too complex, combining multiple criteria in one sentence, or do not have a proper sentence structure (e.g. list of drugs used as intervention). While manual annotators typically have access to these details, this information frequently becomes obscured or lost in the SLR publication process.

Addressing this challenge could involve the development of specialised prompt templates designed to accommodate these complex criteria formats. However, a more labour-intensive yet potentially effective approach would involve reformatting these protocols into a more standardised structure.

Another problem is the lack of annotations at the abstract and title level justifying why a particular paper was excluded from the list. Regular expressions and weak supervision [210] could be applied to some of the most trivial cases to get categories or explanations about the labels.

8.4.2 Limitations of CRUISE–*Screening*

This section discusses the limitations that should be considered when using CRUISE–*Screening*.

Data sources: The CRUISE–*Screening* relies on APIs to conduct the search as it acts as a meta-search engine. These APIs could disappear or change over time, affecting the tool's functionality. However, given that we are using multiple resources at once, the risk of this limitation should be mitigated. Moreover, although CRUISE–*Screening* is

connected to several search engines, it may not cover all potential databases or specialised repositories, potentially missing out on some relevant literature.

Search technique: We do not rely on Boolean queries but a set of keyword-based queries, which together, create a pool of retrieved documents. This approach differs from classic systematic reviews. Additionally, we limit the search to the top 500 records for each query by default to speed up the process, which could potentially limit the coverage of relevant studies.

Hallucinations: Large language models can sometimes "hallucinate" and create incorrect predictions or outputs. Users should be aware that the automated screening process could produce false positives or false negatives due to these hallucinations.

Biases: The machine learning models used for screening could have biases in their predictions due to biased training data, which could impact the quality and representativeness of the literature review.

Cost: The deployment and continued use of large language models in the CRUISE–*Screening* can be expensive. The computational requirements for training and deploying these models are substantial, and as models grow in size and complexity, the associated costs may increase. This could potentially impact the scalability and affordability of the tool for researchers with limited resources or budget constraints.

User experience and accessibility: While CRUISE–*Screening* is designed to be user-friendly, there might be a lack of sufficient detail on accessibility features, potentially making it challenging for a wider range of researchers, including those with specific needs or non-technical backgrounds, to use the tool effectively. Furthermore, the current design of the user interface, while functional, may not be optimal for all potential users. We recognise the need for further user studies to assess its intuitive nature and to identify areas of improvement. We aim to improve the tool's usability and accessibility in future iterations.

8.5 Summary

In this chapter, we presented how citation screening can be presented with the eligibility criteria in mind, superseding the binary classification approach used previously. This approach not only streamlines the screening process but also aligns it more closely with the requirements of systematic reviews.

The experiments conducted with various models, including fine-tuned Transformer-based models and GPT variants, demonstrate the feasibility and effectiveness of using these models for both stages of screening. The performance of domain-specific models like

Clinical-Longformer and Clinical-BigBird underscores the importance of contextual and domain-relevant training in achieving higher Precision and Recall.

However, there are areas for improvement and further research. The challenges in leveraging eligibility criteria information for abstract-level screening suggest a need for more sophisticated models or training strategies that can better understand and use this type of data. Additionally, the varying performance of different models indicates the potential for further optimisation, perhaps through ensemble methods or more advanced prompt engineering techniques.

Future research should also focus on refining the input representation for these models, exploring the impact of different sections of SLR protocols, and experimenting with hybrid models that combine the strengths of different architectures. Furthermore, the scalability and generalisability of these models across diverse topics and types of systematic reviews warrant further investigation.

Finally, we introduce CRUISE–*Screening*, a web-based application for conducting living literature reviews. While systematic literature reviews follow strict criteria and are commonly used in healthcare and medical domains, CRUISE–*Screening* was inspired by the techniques used in systematic reviews to bring more structure and rigour into general literature reviews. Using the recent advancements in NLP and prompt engineering techniques, our system supports researchers in conducting living literature reviews.

In future work, we plan to extend the capabilities of CRUISE–*Screening*. We plan to integrate more search engines and implement ranking and screening prioritisation models with active learning into the workflow. We also plan to develop advanced visualisations and analytics to provide more detailed insights into the literature search results.

CHAPTER 9

Conclusion

This thesis proposed novel datasets, evaluation measures and automation approaches to eligibility screening in the medical domain. We demonstrated our methods on two examples from the medical domain: matching patients to clinical trials and citation screening for systematic literature reviews. This chapter concludes the thesis by revisiting the research questions and contributions. Finally, we describe future research opportunities.

9.1 Research Questions and Contributions

Our high-level research question was: *"How can machine learning models help to automate the eligibility screening step in systematic reviews and clinical trial matching?"*. In this section, we summarise our contributions for the research questions introduced in Chapter 1.

RQ1: How should a comprehensive benchmark dataset for systematic literature reviews be constructed?

This research question aimed to understand the essential components and criteria for constructing an effective and comprehensive benchmark dataset for automated citation screening in systematic literature reviews. Our exploration and analysis were segmented into three sub-questions, focusing on the current state of benchmark datasets, the properties of an ideal collection, and the specific construction of a dataset for full text publication screening.

***RQ1.1**: What is the current overview of benchmark datasets for automated citation screening?*

Our investigation into the existing benchmark datasets revealed significant variability in their scope, size, and quality. We identified a lack of diversity and representativeness

in these datasets, which may contribute to biases in model development and evaluation. We also found a dataset overlap and data leakage in the largest available collections. Our contribution includes a comprehensive analysis of these gaps and recommendations for future dataset creation, emphasising the need for diverse and relevant datasets that reflect the evolving research landscape.

***RQ1.2**: Which properties should a benchmark collection have for a valid assessment of citation screening algorithms?*

In response to RQ1.2, we introduced CSMeD, a meta-dataset carefully curated to address the limitations of existing datasets. This dataset is characterised by its diversity, representativeness, and scalability, making it a valuable resource for evaluating various machine learning approaches. CSMeD consists of more than 300 SLRs from medicine, healthcare and computer science disciplines. The comprehensive nature of CSMeD allows for an accurate and nuanced assessment of citation screening algorithms across different paradigms, including classification, ranking or stopping prediction.

***RQ1.3**: How should a dataset for full text publication screening in systematic literature reviews be constructed?*

The development of CSMeD-ft specifically targets the full text publication screening aspect of systematic literature reviews. This dataset was designed with a focus on incorporating a wide range of publication types, extensive metadata, and scalability to ensure its relevance and adaptability to future research needs. It enables the evaluation of language models in processing long documents. Full text screening is a vital part of the systematic literature review process, and until now, rarely been studied by the research community.

RQ1 Conclusion: Our findings underscore the necessity of a governance framework for systematic literature review datasets. This framework should prioritise diversity, representativeness, and adaptability to technological advancements. By establishing such governance, we can ensure that the datasets used for training and evaluating citation screening models are robust, relevant, and reflective of the current and future state of systematic literature review processes.

RQ2: How should citation screening automation approaches for systematic literature reviews be evaluated?

This research question delves into the evaluation of automation approaches in the context of citation screening for systematic literature reviews. We recognise the need for more robust and nuanced evaluation measures. Our study extensively investigated existing evaluation methods and proposed new ones that better align with the practical realities and demands of SLRs.

***RQ2.1**: What are the shortcomings of the common evaluation measures used in automated citation screening?*

Our investigation highlighted the limitations of common evaluation measures in automated citation screening, such as Work Saved over Sampling (WSS). Specifically, we found that when evaluated at a fixed Recall rate, WSS and Precision are not normalised, raising concerns about aggregating scores across several datasets. We addressed these shortcomings by proposing their min-max normalised versions. Moreover, we showed that these measures often fail to account for the intricacies of the citation screening process, such as class imbalance and the need for high recall.

***RQ2.2**: Which properties should evaluation measures have for appropriate assessment of citation screening algorithms?*

Our research underscores the necessity for evaluation measures that recognise the unique challenges of citation screening in systematic reviews. We identified essential properties for these measures, such as the ability to deal with class imbalance and the importance of identifying almost all relevant papers. To this end, we advocated for using measures like TNR, normalised Precision or $nreTNR$ (normalised rectified TNR) that align with these requirements. We showed how TNR can be used to simulate time savings due to its correlation with the number of successfully removed irrelevant documents. These metrics offer a more precise and relevant assessment, ensuring that the algorithms effectively balance the identification of relevant papers with the minimisation of manual screening efforts.

***RQ2.3**: How could automated citation screening be evaluated differently to consider outcomes of systematic literature reviews?*

In a significant shift from traditional binary classification metrics, we explored evaluation methodologies that consider the impact of automated citation screening on the outcomes of systematic literature reviews. This approach led to the development of an outcome-based evaluation framework, assessing not just the screening performance but also the influence of included publications on the final review outcomes. We demonstrated this through initial experiments, showing how missing publications can significantly alter review conclusions. This multidimensional evaluation provides a deeper insight into the real-world effectiveness of citation screening algorithms and a shift in perspective, promising to bridge the gap between theoretical evaluation and practical impacts.

RQ2 Conclusion: The findings from this research represent a significant advancement in the evaluation of automated citation screening for systematic literature reviews. By proposing new evaluation measures and an outcome-based evaluation framework, we have established a more holistic and realistic approach to assessing these technologies. These innovations enhance the accuracy of the evaluation process and align it more closely with the practical needs of systematic review processes. Ultimately, our work paves the way for more effective, efficient, and impactful use of automation in literature review

methodologies, ensuring that the most relevant and influential publications are identified and included in systematic reviews.

RQ3: How to use machine learning models for automated citation screening with highly imbalanced datasets so the results could be generalised to other reviews?

RQ3 explored the intricacies of using machine learning models for effective citation screening in the context of highly imbalanced datasets, aiming to derive generalisable solutions applicable to various systematic reviews.

***RQ3.1**: How do neural classification methods perform in the citation screening step, particularly when compared to traditional methods?*

Our investigation showed that neural network-based classification methods, while technologically advanced, do not consistently surpass the performance of traditional methods like SVM in citation screening tasks. This observation highlights the enduring value of classical algorithms, suggesting a synergistic approach that leverages both traditional and modern methodologies. We also showed the considerable variability in the performance of neural models across validation splits, showing that these models might not provide stable gains. Furthermore, we discussed that the frequently used training and evaluation scenario requiring half of the dataset for training is not very practical, especially for larger SLRs.

***RQ3.2**: Which external knowledge sources can be used to improve the quality of automated citation screening?*

In addressing the role of external knowledge sources, we found that the integration of eligibility criteria into machine learning models, particularly Transformer-based language models, enhances screening accuracy. This approach, especially in zero-shot learning settings, represents a significant improvement in the field of automated citation screening. By harnessing rich external data sources, models can better align with the specific requirements of systematic reviews, thereby increasing their precision and relevance. It also enables the usage of these models with minimal labelled data yet keeps consistent time savings.

***RQ3.3**: How can recent advancements in language models be applied for automatic eligibility screening of full text publications?*

Our study demonstrated the effectiveness of recent advancements in large language models in conducting full-text publication screening. These models' ability to process extensive and complex documents marks a novel and promising direction in automated citation screening. The exploration into larger language models further suggests an expanding horizon for handling comprehensive and intricate full-text analysis.

***RQ3.4**: Can these machine learning approaches generalise to systematic literature reviews conducted in a domain other than medicine?*

Through the development and application of CRUISE–*Screening*, a tool designed for automated search and screening using eligibility criteria, we assessed the generalisability of these models beyond the medical domain. Our findings indicate potential in adapting these methodologies to diverse scientific disciplines, though challenges persist in applying them to less structured domains. This exploration reveals the versatility and adaptability of machine learning approaches in systematic literature reviews across various fields.

RQ3 Conclusion: The key to addressing the class imbalance in automated citation screening lies in the use of eligibility criteria combined with advanced language models. By tailoring machine learning approaches to incorporate specific eligibility criteria, we can enhance the precision and applicability of these models across different domains and review types. This strategy not only addresses the inherent challenges posed by imbalanced datasets but also sets a new standard for efficiency and effectiveness in automated citation screening, paving the way for broader applications in diverse research areas.

RQ4: What techniques can be used to improve eligibility screening of patients to clinical trials?

RQ4 focused on the eligibility screening of patients for clinical trials. Our study explored various techniques to refine the eligibility screening process for clinical trials. We concentrated on dissecting individual sections of clinical trial documents and integrating information extraction methods. The outcome highlights the potential of these techniques in improving the precision and efficiency of matching patients to suitable clinical trials.

***RQ4.1**: What is the impact of individual sections of clinical trial text on the performance of a lexical retrieval approach?*

Our analysis underscored the significance of particular sections of clinical trial texts, notably the *'inclusion'* and *'exclusion'* criteria. By concentrating on these sections, we enhanced the performance of lexical retrieval models, paving the way for more accurate and relevant patient-trial matches. We also showed the problems with using lexical models, which cannot explicitly grasp the content from the exclusion criteria. This insight is crucial for optimising clinical trial matching processes, ensuring patients are efficiently matched with trials that align with their specific health profiles.

***RQ4.2**: How can information extraction techniques improve the retrieval of eligible clinical trials?*

The integration of information extraction techniques marked an advancement in clinical trial retrieval. By incorporating entity recognition and negation detection, we significantly improved the Precision of matching eligible patients to clinical trials. This advancement enhances the accuracy of trial-patient matching and contributes to more effective clinical

trial recruitment. The ability to filter trials based on specific criteria such as gender, age, and medical conditions ensures that patients are presented with the most relevant and potentially beneficial trial options.

RQ4 Conclusion: The outcomes of this research demonstrate a promising direction in clinical trial recruitment and patient matching. By leveraging the detailed analysis of clinical trial texts and incorporating advanced information extraction techniques, we can significantly improve the efficiency and effectiveness of clinical trial matching using a pipeline-based approach. This approach not only benefits patients by connecting them with appropriate trials but also aids researchers in recruiting suitable participants, thereby contributing to the overall advancement of medical research and patient care.

9.2 Future Research

The research presented in this thesis opens avenues for several future explorations, each promising to further optimise and innovate the fields of medical IR, NLP and systematic literature reviews. The following subsections outline targeted areas where such contributions can be made.

End-to-end systematic literature review automation This thesis has focused predominantly on the citation screening aspect of systematic literature reviews. A natural extension of this work is the development of end-to-end solutions that not only identify relevant citations but also assist in data extraction, quality assessment, and synthesis of the included studies. Future research could focus on creating integrated systems based on large language models, specifically the retrieval augmented generation framework. Such models, starting from the research protocol, could create and execute the search query, screen relevant publications and finally extract the outcome data to synthesise it into a coherent review text.

Prospective evaluation of citation screening We have introduced novel evaluation methods and metrics for automated citation screening. Future studies should look into the prospective evaluation of these automated systems in real-world settings. This involves integrating the algorithms into the workflow of ongoing systematic reviews and assessing their performance, usability, and impact on the review quality and efficiency. Additionally, ethical considerations, data privacy, and security should be examined to ensure the responsible application of these technologies. Such prospective evaluation will be crucial for an unbiased assessment of the performance of large language models.

Multilingual medical retrieval This thesis addresses the automation of citation screening and patient-trial matching, where systematic literature reviews, clinical trials and patient descriptions are written in English. The global nature of medicine and healthcare requires systems that can handle a diverse range of languages and dialects. Future research should focus on extending the proposed methods to support multilingual

medical information retrieval. This extension can involve adapting current algorithms to process, understand, and generate information in multiple languages or developing new models tailored to specific linguistic and cultural contexts.

Unified standards for systematic review automation Standardisation in the context of systematic literature review automation is a critical yet largely unexplored area. Future efforts should be directed towards developing international standards for assessing and reporting the performance of automated systems in these domains. This includes establishing common evaluation measures, inter-operable data sharing standards and protocols to enable the consistent evaluation and comparison of different approaches. Research on standardisation should also try to cover systematic literature reviews conducted in disciplines beyond medicine and healthcare. One primary application would be to improve data sharing practices among the largest collaborations creating SLRs like Cochrane and Campbell. Collaboration among stakeholders, including researchers, healthcare professionals, policymakers, and technology providers, will be essential to create and implement these standards globally.

Explainability and interpretability Using complex black-box machine learning models in medical decision-making raises concerns about explainability and interpretability. Future research should address the development of methods that enhance the transparency of the algorithms. This can be achieved by integrating explainable AI techniques into the models, developing visualisation tools to illustrate decision-making processes, and creating platforms for users to interact and engage with the algorithms effectively. Such advancements will contribute to building trust and acceptance among healthcare professionals, patients, and policymakers.

Each research question addressed in this thesis unveils answers and new horizons for exploration. Integrating machine learning, natural language processing, information retrieval, and human expertise is not only a technical challenge. It helps us to see new ideas and possibilities for a future where technology and people work together to make things better.

List of Figures

1.1 The Hierarchy of Evidence in Evidence-Based Medicine (EBM), detailing the quality of evidence from expert opinion to meta-analyses. This pyramid illustrates the increasing reliability and rigor of study designs as one moves up the levels, with meta-analyses representing the pinnacle of evidence quality due to their comprehensive review and analysis of literature. The differentiation between 'filtered' and 'unfiltered' information signifies the degree of critical appraisal and synthesis of evidence. 2
1.2 High-level data flow in Evidence-Based Medicine, depicting how individual patient data contributes to the creation of medical guidelines. Highlighted are sections where the task of eligibility screening is performed. 3
1.3 Three examples of eligibility screening in the medical domain. From the left: (a) citation screening for systematic literature reviews; (b) trial-to-patients clinical trials matching; (c) patient-to-trials clinical trials matching. The funnel in the middle of each diagram represents the eligibility screening process. 4
1.4 Contributions with respect to each chapter. 16

2.1 An example of a clinical trial and a description of a patient eligible for this trial. Highlighted items are described in detail in Section 2.2. Example adapted from Pradeep et al. [203]. 20
2.2 Topic-by-topic Reciprocal Rank (top) and P@10 (bottom) scores comparison for a BM25+ model with different document representations for TREC CT 2022 data. 27
2.3 Topic-by-topic number of relevant trials in the top 20 for the three best BM25+ runs from each experiment: 14 – baseline, 14d – further query and index enriched with extracted entities, and 14d+AG – further filtered for age and gender. 32
2.4 Averaged per patient count of relevant (top) and excluded (bottom) trials depending on a cut-off of K trials retrieved (x-axis) for TREC CT 2022 collection. 33

3.1 Fourteen steps of the systematic literature review process clustered into five high-level categories. Steps and process description according to the Cochrane [94], figure adapted based on Tsafnat et al. [254]. 39

3.2 Illustration of the citation screening process, separated into two tasks (1) title and abstract screening and (2) full text screening. Tasks are represented as a specific example of question-answering when a single question asks for a fulfilment of all eligibility criteria $\mathcal{C}$ at once. 45

4.1 Example visualidation with statistics for a "Proton Pump Inhibitors" SLR dataset. 72
4.2 Example visualisations with TF-IDF and UMAP representation of documents for a "CS-Goulao-2016" SLR. Based on the plot, one can see that the retrieved documents are grouped in two clusters with all relevant publications belonging to one of them (bottom-right part of the plot). This can be an indicator that any model will likely remove the other "non-relevant" cluster of documents and hence achieve good score in detecting true negatives. 73
4.3 CSMED-FT construction steps. 74
4.4 Token frequency distribution by split (top) and frequency of different kind of instances (bottom). 77

5.1 Plot presenting a TNR, nreTNR and WSS behaviour for a custom dataset containing $N = 2000$ documents out of which 5% are relevant ($|\mathcal{I}| = 100$). Visualisation shows how the number of detected true negatives (TNs) influences the score of each evaluation measure. Evaluations conducted at a Recall cutoff = 70%. 95
5.2 The VoMBaT page for comparing several evaluation measures at a fixed Recall level. Users can navigate to other pages and select dataset parameters using the sidebar on the left, while Recall level and evaluation measure selection are on the top. 96
5.3 The page for comparing one evaluation measure across all Recall levels. Users can select dataset parameters using the sidebar on the left. 97
5.4 The page for estimating time and money savings that can be achieved depending on the value of evaluation measures. Users can select dataset parameters using the sidebar on the left. 98
5.5 The page for comparing manual and automatic assessments count depending on the TNR score. Users can select dataset parameters using the sidebar on the left. 99
5.6 The page for comparing custom evaluation measures. Users can select dataset parameters using the sidebar on the left. 100
5.7 Dynamics of evaluation measures (WSS, TNR (nWSS) and Precision) scores as a function of the number of True Negatives (TN) at 95% Recall for two sample datasets. 101
5.8 Receiver Operating Characteristic (ROC) curves for two hypothetical models with their corresponding AUC scores. Model A achieves a higher value of AUC, despite the fact that its TPR performance reaches 80% only at the FPR level almost equal to 100%, and model B achieves maximum Recall at FPR level of 35%. 103

6.1 Four steps of the proposed evaluation framework. 108
6.2 Different review outcomes represented as forest plots. Each row is a single study. Columns from the right represent, respectively: (1) the study identifier, (2) number of events in the experimental group (e.g., patients with specific symptoms or adverse events), (3) experimental group size, (4) number of events in the control group, (5) control group size, (6) the weight of a study, and (7) effect size of a study: a difference (e.g., risk ratio or standardised mean difference) in events between experimental or control group. Simulations and figures done using RevMan Web, available at `http://revman.cochrane.org`. The figure is continued on the next two pages. 110
6.2 (cont.) Different versions of review outcomes continued. 111
6.2 (cont.) Different versions of review outcomes continued. 112
6.3 Box plots presenting relative difference values from 20 simulations on the publication level. Note that the intervals on the x-axis are not uniform. . 116
6.4 Box plots presenting distance to confidence intervals values from 20 simulations on the publication level. Note that the intervals on the x-axis are not uniform. 116
6.5 Box plot presenting runs with their relative difference in study outcomes for an evaluation with a cut-off at 30% of the total number of documents for each review. Runs are sorted by their MAP score. The orange circle denotes the mean relative difference @30%. The x-axis is cut at 30, while the outliers exist up to the value of 100; we cut for visualisation purposes. 119
6.6 Linear regression fits between relative difference at 20% cut-off of documents and other evaluation measures scores. Correlations for relative difference at other cut-offs follow similar trends. 120
6.7 Visualisation of the Pareto frontier for two objectives: (1) number of non-estimable outcomes on the x-axis and (2) sum of relative difference for estimable outcomes on the y-axis. Both objectives are to be minimised. Runs are evaluated at a cut-off at 5% of the total number of documents for each review. Non-dominated runs are marked with a blue colour. The Pareto frontier was calculated using the method by Herman and Woodruff [92]. . 121
6.8 Box plot presenting nDCG@20% scores for each run. Orange bars represent original binary qrels, and blue represent weighting based on the *Influence*. Runs are sorted by mean nDCG (white circle) of original binary qrels. . . 122

7.1 A count of experiments in which a model using a specific input feature achieved the best results. Models that use all available features scored the best results 49% of times for a specific (model, dataset) combination. 136
7.2 The relationship between dataset size and a model training time for the three evaluated models. Both training time and dataset size are shown on a logarithmic scale. 137
7.3 Example boxplots with TNR@95% scores for three models. Input features are titles and abstracts. 140

7.4 Scatter plot of normalised Precision (nP) versus True Negative Rate (TNR) at 95% Recall across three tested models. This figure illustrates the trade-off between nP@95% and TNR@95% scores. The size of each marker represents the relative size of the dataset used. Average nP@95% and TNR@95% are indicated by dashed grey lines. 140
7.5 Visualisation of agreement between nP@95% and TNR@95% across three tested models using a Bland-Altman plot. Each point represents the mean of the two measures plotted against their difference, with the marker size indicating the dataset size. The mean difference and the limits of agreement at ±1.96 standard deviations are marked as dashed horizontal lines. . . . 141

8.1 Overview of the CRUISE–*Screening* architecture. The numbers are referred to in Sections 8.3.2 and 8.3.3. 152
8.2 Example literature review protocol containing review title, description, search queries and criteria for inclusion and exclusion. 153
8.3 Example screening interface in CRUISE–*Screening* presenting single paper with answered questions. 154
8.4 Example of citation screening using eligibility criteria and prompt-based learning. For every paper in the citation list, the inclusion and exclusion criteria are compared using the prompt template. This procedure generates two lists with textual answers for each criterion. The final decision is an aggregation of single outputs. 156

List of Tables

2.1 Statistics of TREC CT datasets from 2021 and 2022. 24

2.2 Impact of CTs' sections on the performance of the BM25+ retrieval model. The first group contains results using only a single section as a document representation, and the second group represents results using several concatenated sections. The 'criteria' section represents the original text from the *eligibility criteria* section of a clinical trial. The 'inclusion' and 'exclusion' sections are derived from the 'criteria' section using heuristic methods. Underlined values indicate highest score within the group, **bold** values indicate highest score overall. The identifier of each run is in the first column. 26

2.3 Impact of CT sections on the performance of In_expB2 and TF-IDF retrieval models. For each model, the first group contains results using only a single section as a document representation, and the second group represents results using several concatenated sections. The 'criteria' section represents the original text from the *eligibility criteria* section of a clinical trial. The 'inclusion' and 'exclusion' sections are derived from the 'criteria' section using heuristic methods. Underlined values indicate highest score within the group, **bold** values indicate highest score overall for each model. The identifier of each run is in the first column. 28

2.4 Experimental results for runs with index and query expanded with extracted entities. Letters describe usage of extracted affirmative and negative medical entities for (c) current conditions, (p) past conditions, and (f) family history. **Bold** values indicate highest score overall. The identifier of each run is in the first column. 29

2.5 Example entities extracted for Topic #48 from TREC CT 2021. 30

2.6 Experimental results for runs with index and query expanded with extracted entities for In_expB2 and TF-IDF retrieval models. Letters describe usage of extracted affirmative and negative medical entities for (c) current conditions, (p) past conditions, and (f) family history. **Bold** values indicate highest score overall for each model. The identifier of each run is in the first column. . 31

2.7 Filtering results on TREC CT 2021 data. Letters describe the used filters: (A) Age, (G) Gender, (S) Smoking and (D) Drinking. **Bold** values indicate highest score overall. Superscripts denote significant differences in paired Student's t-test with $p \leq 0.05$. The identifier of each run is in the first column. 31

3.1 Definitions of concepts and terms used in the thesis. 36
3.1 Definitions of concepts and terms used in the thesis. 37
3.1 Definitions of concepts and terms used in the thesis. 38
3.2 A comparison of publicly available benchmark datasets used in the experiments on automated citation screening for systematic literature reviews, sorted by the publication year. We included all publicly available datasets and private datasets which were used in more than one publication. The "Avg. size" refers to the average number of citations contained in each review within the dataset. The "Avg. ratio of included" indicates the average percentage of those citations that were included in the final review. The "Additional data" column describes if the review contains metadata other than coming from the citation list. 53
3.3 List of overlapping Cochrane systematic literature reviews between datasets. 56
3.4 Usage statistics of the SLR datasets, including the latest publication year, venue and evaluation measure. We report two usages in case there was a more recent pre-print published. 63

4.1 A list of source citation screening datasets included in the CSMED. First four datasets contain non-Cochrane SLRs, whereas the other five are based on Cochrane reviews. 'Avg. ratio of included' column present ratio of included publication from the title and abstract screening stage, 'Avg. size' refers to averaged across SLRs document count in the dataset. The 'Additional data' column describes if the review contains metadata other than coming from the citation list: (A): Search queries, (B): Review protocol containing review title, abstract and search strategy, (C): Review updates consisting of changes to included papers. Total values do not account for duplicated reviews. 'DTA' stands for diagnostic test accuracy reviews. ‡ – different number of reviews in the paper versus the GitHub repository; 'Total' counts the higher value and doesn't account for duplicates. 68
4.2 Details of the CSMED expanded meta-dataset. Column '#docs' refers to the total number of documents included in all SLRS within the dataset, '#included' mentions number of included documents on the title and abstract screening stage and 'Avg. %included' the percentage of included publications averaged from all reviews. 70
4.3 Details of the CSMED-FT dataset. Column '#included' mentions number of included documents on the full text step. CSMED-FT-TEST-SMALL is a subset of CSMED-FT-TEST. 76

5.1 Evaluation results with WSS and TNR at 95% Recall on systematic review datasets from Cohen et al. [35] described in Chapter 3. The following models were used: A: Cohen et al. [35], B: Matwin et al. [166], C: Cohen [36], D: Howard et al. [99], E: Kontonatsios et al. [124], F: van Dinter et al. [262], G: Kusa et al. [130]. **Bold** indicates highest score. 92

6.1 Statistics of the considered dataset. 114
6.2 Results of the simulation on the *publication* level. Outcomes are aggregated across 32 systematic reviews and are averaged from 20 different random seeds. 115
6.3 Results of the simulation on the *study* level. Outcomes are aggregated across 32 systematic reviews and are averaged from 20 different random seeds. . 117

7.1 Statistics of 23 systematic literature reviews used in this experiment, a subset of CSMED-BASIC containing medical SLRs. 132
7.2 TNR@95% results for replicated models compared with original results and benchmark models. TNR@95% scores are averages across ten validation runs for each of the 23 review datasets. Underlined scores indicate the highest score within the three tested models, **bold** values indicate the highest score overall. 134
7.3 Influence of input document features on the TNR@95% score for three tested models. "All features" column means a single string concatenating Title, Abstract, Author and Journal information. For each row, **bold** values indicate the highest score for each model, underlined values indicate the highest score across all 3 models. 135
7.4 A comparison of Normalised Precision at 95% Recall (nP@95%) for the three models across 21 benchmark datasets. We did not evaluate the two largest SWIFT datasets: *Transgenerational* and *Neuropathic pain*. **Bold** denotes the highest score in each dataset. A † symbol indicates that DAE-FF's average nP@95% is statistically significantly superior to the other models, based on the Wilcoxon signed-rank test with Bonferroni correction. 138

8.1 Results of zero-shot evaluation on CSMED-COCHRANE-DEV dataset. For each measure, **bold** values indicate the highest score for each model across query representation. Underlined values indicate the highest score across all tested models. 145
8.2 Systematic literature review protocol section lengths in number of words for CSMED-COCHRANE-ALL dataset. 146
8.3 Statistics of a review text with respect to the fit within 2,048 tokens context window. 147
8.4 Results of the full text screening experiment averaged over documents. The statistical significance was assessed with a McNemar's t-test ($p < 0.05$) with Bonferroni correction for multiple testing. *Clinical-BigBird* on the CSMED-FT-TEST split showed statistically significant improvements compared to the *stratified random* baseline, *Longformer*, *Clinical-Longformer*, and *GPT-3.5-turbo-16k*, indicated by †. Stratified baseline is averaged from 100 different random seeds. '% incl.' describes the percentage of documents predicted as relevant by models (TP + FN). 150

Bibliography

[1] Grobid. https://github.com/kermitt2/grobid, 2008–2022.

[2] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, 2015. URL https://research.google/pubs/pub45166/.

[3] Charu C Aggarwal and ChengXiang Zhai. A survey of text classification algorithms. In *Mining text data*, pages 163–222. Springer, 2012.

[4] Maristella Agosti, Norbert Fuhr, Elaine Toms, and Pertti Vakkari. Evaluation methodologies in information retrieval dagstuhl seminar 13441. In *Acm sigir forum*, volume 48, pages 36–41. ACM New York, NY, USA, 2014.

[5] Amal Alharbi and Mark Stevenson. A dataset of systematic review updates. *SIGIR 2019 - Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1257–1260, 7 2019. doi: 10.1145/3331184.3331358.

[6] Amal Alharbi and Mark Stevenson. Refining Boolean queries to identify relevant studies for systematic review updates. *Journal of the American Medical Informatics Association*, 27(11):1658–1666, 10 2020. ISSN 1527-974X. doi: 10.1093/jamia/ocaa148. URL https://doi.org/10.1093/jamia/ocaa148.

[7] Gianni Amati and Cornelis Joost Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems (TOIS)*, 20(4):357–389, 2002.

[8] Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha,

et al. Construction of the literature graph in semantic scholar. *arXiv preprint arXiv:1805.02262*, 2018.

[9] Anthropic. claude-v1.3-100k, Blog post 'Introducing 100K Context Windows'. https://www.anthropic.com/index/100k-context-windows, 2023.

[10] Madeleine Ballard and Paul Montgomery. Risk of bias in overviews of reviews: a scoping review of methodological guidance and four-item checklist. *Research synthesis methods*, 8(1):92–108, 2017.

[11] Alexandra Bannach-Brown, Piotr Przybyła, James Thomas, Andrew S. C. Rice, Sophia Ananiadou, Jing Liao, and Malcolm Robert Macleod. Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. *Systematic Reviews 2019 8:1*, 8(1):1–12, 1 2019. ISSN 2046-4053. doi: 10.1186/S13643-019-0942-7. URL https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/s13643-019-0942-7.

[12] Jason R Baron, David D Lewis, and Douglas W Oard. Trec 2006 legal track overview. In *TREC*. Citeseer, 2006.

[13] Alexandra Barratt. Evidence based medicine and shared decision making: the challenge of getting both evidence and preferences into health care. *Patient education and counseling*, 73(3):407–412, 2008.

[14] Samar Bashath, Nadeesha Perera, Shailesh Tripathi, Kalifa Manjang, Matthias Dehmer, and Frank Emmert Streib. A data-centric review of deep transfer learning with applications to text data. *Information Sciences*, 585:498–528, 2022.

[15] Elias Bassani. retriv: A Python Search Engine for the Common Man, May 2023. URL https://github.com/AmenRa/retriv.

[16] Hilda Bastian, Paul Glasziou, and Iain Chalmers. Seventy-five trials and eleven systematic reviews a day: How will we ever keep up? *PLoS Medicine*, 7(9), 9 2010. ISSN 15491277. doi: 10.1371/journal.pmed.1000326. URL https://pubmed.ncbi.nlm.nih.gov/20877712/.

[17] Tanja Bekhuis, Eugene Tseytlin, Kevin J Mitchell, and Dina Demner-Fushman. Feature engineering and a proposed decision-support system for systematic reviewers of medical evidence. *PloS one*, 9(1):e86277, 2014.

[18] Elaine Beller, Justin Clark, Guy Tsafnat, Clive Adams, Heinz Diehl, Hans Lund, Mourad Ouzzani, Kristina Thayer, James Thomas, Tari Turner, Jun Xia, Karen Robinson, Paul Glasziou, Olga Ahtirschi, Robin Christensen, Julian Elliott, Sergio Graziosi, Joel Kuiper, Rasmus Moustgaard, Annette O'Connor, Jacob Riis, Karla Soares-Weiser, Camilo Vergara, and Ida Wedel-Heinen. Making progress

with the automation of systematic reviews: Principles of the International Collaboration for the Automation of Systematic Reviews (ICASR), 5 2018. ISSN 20464053. URL https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/s13643-018-0740-7.

[19] Elaine M Beller, Joyce Kee-Hsin Chen, Una Li-Hsiang Wang, and Paul P Glasziou. Are systematic reviews up-to-date at the time of publication? *Systematic reviews*, 2(1):1–6, 2013.

[20] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[21] Gary S Bilotta, Alice M Milner, and Ian Boyd. On the use of systematic reviews to inform environmental policies. *Environmental Science & Policy*, 42:67–77, 2014.

[22] Kathrin Blagec, Jakob Kraiger, Wolfgang Frühwirt, Matthias Samwald, and Assoc Matthias Samwald. Benchmark datasets driving artificial intelligence development fail to capture the needs of medical professionals. 1 2022. URL https://arxiv.org/abs/2201.07040v1.

[23] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 7 2017. ISSN 2307-387X. doi: 10.1162/tacl{_}a{_}00051. URL http://arxiv.org/abs/1607.04606.

[24] Peter Bollmann. Two axioms for evaluation measures in information retrieval. In *SIGIR*, volume 84, pages 233–245. Citeseer, 1984.

[25] Rohit Borah, Andrew W. Brown, Patrice L. Capers, and Kathryn A. Kaiser. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry, 2 2017. ISSN 20446055. URL http://bmjopen.bmj.com/.

[26] Rohit Borah, Andrew W Brown, Patrice L Capers, and Kathryn A Kaiser. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the prospero registry. *BMJ open*, 7(2):e012545, 2017.

[27] Austin J. Brockmeier, Meizhi Ju, Piotr Przybyła, and Sophia Ananiadou. Improving reference prioritisation with PICO recognition. *BMC Medical Informatics and Decision Making*, 19(1):256, 12 2019. ISSN 14726947. doi: 10.1186/s12911-019-0992-8. URL https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0992-8.

[28] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[29] Luca Busin and Stefano Mizzaro. Axiometrics: An axiomatic approach to information retrieval effectiveness metrics. In *Proceedings of the 2013 Conference on the Theory of Information Retrieval*, pages 22–29, 2013.

[30] Max W Callaghan and Finn Müller-Hansen. Statistical stopping criteria for automated screening in systematic reviews. *Systematic Reviews*, 9(1):1–14, 2020.

[31] Stefanie Castillo and Petar Grbovic. The apisser methodology for systematic literature reviews in engineering. *IEEE Access*, 10:23700–23707, 2022.

[32] SH. Cheng, C. Augustin, A. Bethel, D. Gill, S. Anzaroot, J. Brun, B. DeWilde, R. Minnich, R. Garside, Y. Masuda, DC. Miller, D. Wilkie, S. Wongbusarakum, and MC. McKinnon. Using machine learning to advance synthesis and use of conservation and environmental evidence. *Conservation Biology*, 32:762–764, 2018. doi: 10.1111/cobi.1311.

[33] Justin Clark. *Systematic reviewing: Introduction, locating studies and data abstraction*, pages 187–211. Springer Series on Epidemiology and Public Health. Springer, Germany, April 2013. ISBN 978-3-642-37130-1. doi: 10.1007/978-3-642-37131-8{_}12.

[34] Justin Clark, Paul Glasziou, Chris Del Mar, Alexandra Bannach-Brown, Paulina Stehlik, and Anna Mae Scott. A full systematic review was completed in 2 weeks using automation tools: a case study. *Journal of clinical epidemiology*, 121:81–90, 2020.

[35] A. M. Cohen, W. R. Hersh, K. Peterson, and Po Yin Yen. Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association*, 13(2):206–219, 3 2006. ISSN 10675027. doi: 10.1197/jamia.M1929. URL /pmc/articles/PMC1447545//pmc/articles/PMC1447545/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC1447545/.

[36] Aaron M. Cohen. Optimizing Feature Representation for Automated Systematic Review Work Prioritization. *AMIA Annual Symposium Proceedings*, 2008:121, 2008. URL /pmc/articles/PMC2656096//pmc/articles/PMC2656096/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC2656096/.

[37] Aaron M Cohen. Letter: Performance of support-vector-machine-based classification on 15 systematic review topics evaluated with the WSS@95 measure. *Journal of the American Medical Informatics Association : JAMIA*, 18(1):104, 1 2011. doi: 10.1136/JAMIA.2010.008177. URL /pmc/articles/PMC3005879/https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3005879/.

[38] Aaron M Cohen, Kyle Ambert, and Marian McDonagh. A prospective evaluation of an automated classification system to support evidence-based medicine and

systematic review. In *AMIA annual symposium proceedings*, volume 2010, page 121. American Medical Informatics Association, 2010.

[39] M. Cole, J. Liu, N. J. Belkin, Ralf Bierig, Jacek Gwizdka, C. Liu, J. Zhang, and X. Zhang. Usefulness as the Criterion for Evaluation of Interactive Information Retrieval. *Proceedings of the Third Workshop on Human-Computer Interaction and Information Retrieval Cambridge*, pages 1–4, 2009. URL https://mural.maynoothuniversity.ie/15226/. Publisher: HCIR.

[40] Linda Connor, Jennifer Dean, Molly McNett, Donna M. Tydings, Amanda Shrout, Penelope F. Gorsuch, Ashley Hole, Laura Moore, Roy Brown, Bernadette Mazurek Melnyk, and Lynn Gallagher-Ford. Evidence-based practice improves patient outcomes and healthcare system return on investment: Findings from a scoping review. *Worldviews on Evidence-Based Nursing*, 20(1):6–15, 2023. doi: https://doi.org/10.1111/wvn.12621. URL https://sigmapubs.onlinelibrary.wiley.com/doi/abs/10.1111/wvn.12621.

[41] Jack G Conrad and Jeremy Pickens. Second international workshop on ai and intelligent assistance for legal professionals in the digital workplace (legalaiia 2021). In *ASAIL/LegalAIIA@ ICAIL*, page 48, 2021.

[42] Deborah J Cook, Cynthia D Mulrow, and R Brian Haynes. Systematic reviews: synthesis of best evidence for clinical decisions. *Annals of internal medicine*, 126 (5):376–380, 1997.

[43] Gordon V. Cormack and Maura R. Grossman. Engineering quality and reliability in technology-assisted review. *SIGIR 2016 - Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 75–84, 7 2016. doi: 10.1145/2911451.2911510. URL http://dx.doi.org/10.1145/2911451.2911510.

[44] Gordon V Cormack and Maura R Grossman. Scalability of continuous active learning for reliable high-recall text classification. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 1039–1048, 2016.

[45] Gordon V Cormack and Maura R Grossman. Technology-Assisted Review in Empirical Medicine: Waterloo Participation in CLEF eHealth 2017. In *CLEF (working notes)*, 2017.

[46] Gordon V Cormack, Maura R Grossman, Bruce Hedin, and Douglas W Oard. Overview of the trec 2010 legal track. In *TREC*, 2010.

[47] Siddhartha R Dalal, Paul G Shekelle, Susanne Hempel, Sydne J Newberry, Aneesa Motala, and Kanaka D Shetty. A pilot study using machine learning and domain knowledge to facilitate comparative effectiveness review updating. *Medical Decision Making*, 33(3):343–355, 2013.

[48] Tirthankar Dasgupta, Ishani Mondal, Abir Naskar, and Lipika Dey. Extracting semantic aspects for structured representation of clinical trial eligibility criteria. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 243–248, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.clinicalnlp-1.27. URL https://aclanthology.org/2020.clinicalnlp-1.27.

[49] Tirthankar Dasgupta, Ishani Mondal, Abir Naskar, and Lipika Dey. Automatic segregation and classification of inclusion and exclusion criteria of clinical trials to improve patient eligibility matching. In *Explainable AI in Healthcare and Medicine*, pages 291–296. Springer, 2021.

[50] Jonathan J. Deeks and Julian P. T. Higgins. Statistical algorithms in review manager 5. *Statistical Methods Group of The Cochrane Collaboration*, 1(11), 2010.

[51] Jonathan J Deeks, Julian PT Higgins, Douglas G Altman, and Cochrane Statistical Methods Group. Analysing data and undertaking meta-analyses. *Cochrane handbook for systematic reviews of interventions*, pages 241–284, 2019.

[52] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[53] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

[54] Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. MS2: Multi-Document Summarization of Medical Studies. pages 7494–7513, 4 2021. URL https://aclanthology.org/2021.emnlp-main.594http://arxiv.org/abs/2104.06486.

[55] Anjani Dhrangadhariya, Wojciech Kusa, Henning Müller, and Allan Hanbury. HEVS-TUW at SemEval-2023 task 8: Ensemble of language models and rule-based classifiers for claims identification and PICO extraction. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1776–1782, Toronto, Canada, July 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.semeval-1.246.

[56] Giorgio Maria Di Nunzio, Evangelos Kanoulas, and Prasenjit Majumder. Augmented intelligence in technology-assisted review systems (altars 2022): Evaluation metrics

and protocols for ediscovery and systematic review systems. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part II*, pages 557–560, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-030-99738-0. doi: 10.1007/978-3-030-99739-7_69. URL https://doi.org/10.1007/978-3-030-99739-7_69.

[57] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. *Advances in neural information processing systems*, 32, 2019.

[58] David Dowling. Tarpits: The sticky consequences of poorly implementing technology-assisted review. *Berkeley Tech. LJ*, 35:171, 2020.

[59] Óscar E. Mendoza, Wojciech Kusa, Alaa El-Ebshihy, Ronin Wu, David Pride, Petr Knoth, Drahomira Herrmannova, Florina Piroi, Gabriella Pasi, and Allan Hanbury. Benchmark for research theme classification of scholarly documents. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 253–262, Gyeongju, Republic of Korea, October 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.sdp-1.31.

[60] Adrian Edwards and Glyn Elwyn. *Shared decision-making in health care: Achieving evidence-based patient choice*. Oxford University Press, USA, 2009.

[61] Julian H. Elliott, Tari Turner, Ornella Clavisi, James Thomas, Julian P.T. Higgins, Chris Mavergames, and Russell L. Gruen. Living Systematic Reviews: An Emerging Opportunity to Narrow the Evidence-Practice Gap. *PLoS Medicine*, 11(2), 2014. ISSN 15491676. doi: 10.1371/JOURNAL.PMED.1001603. URL /pmc/articles/PMC3928029/https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3928029/.

[62] A Elmagarmid, Z Fedorowicz, H Hammady, I Ilyas, M Khabsa, and M Ouzzani. Rayyan: a systematic reviews web app for exploring and filtering searches for eligible studies for cochrane reviews. In *Evidence-Informed Public Health: Opportunities and Challenges. Abstracts of the 22nd Cochrane Colloquium*, pages 21–26. John Wiley & Sons Hyderabad, India, India, 2014.

[63] Peter J Embi, Anil Jain, and C Martin Harris. Physicians' perceptions of an electronic health record-based clinical trial alert approach to subject recruitment: a survey. *BMC medical informatics and decision making*, 8(1):1–8, 2008.

[64] Hannah Eyre, Alec B Chapman, Kelly S Peterson, Jianlin Shi, Patrick R Alba, Makoto M Jones, Tamara L Box, Scott L DuVall, and Olga V Patterson. Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python. In *AMIA Annual Symposium Proceedings 2021*, (in press, n.d.). URL http://arxiv.org/abs/2106.07799.

[65] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8): 861–874, 2006.

[66] Ashraf Fawzy, Tianshi David Wu, Kunbo Wang, Matthew L Robinson, Jad Farha, Amanda Bradke, Sherita H Golden, Yanxun Xu, and Brian T Garibaldi. Racial and ethnic discrepancy in pulse oximetry and delayed identification of treatment eligibility among patients with covid-19. *JAMA internal medicine*, 182(7):730–738, 2022.

[67] Marco Ferrante, Nicola Ferro, and Maria Maistro. Towards a formal framework for utility-oriented measurements of retrieval effectiveness. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, pages 21–30, 2015.

[68] Jason Fries, Natasha Seelam, Gabriel Altay, Leon Weber, Myungsun Kang, Debajyoti Datta, Ruisi Su, Samuele Garda, Bo Wang, Simon Ott, Matthias Samwald, and Wojciech Kusa. Dataset debt in biomedical language modeling. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 137–145, virtual+Dublin, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.bigscience-1.10. URL https://aclanthology.org/2022.bigscience-1.10.

[69] Jason Fries, Leon Weber, Natasha Seelam, Gabriel Altay, Debajyoti Datta, Samuele Garda, Sunny Kang, Rosaline Su, Wojciech Kusa, Samuel Cahyawijaya, Fabio Barth, Simon Ott, Matthias Samwald, Stephen Bach, Stella Biderman, Mario Sänger, Bo Wang, Alison Callahan, Daniel León Periñán, Théo Gigant, Patrick Haller, Jenny Chim, Jose Posada, John Giorgi, Karthik Rangasai Sivaraman, Marc Pàmies, Marianna Nezhurina, Robert Martin, Michael Cullan, Moritz Freidank, Nathan Dahlberg, Shubhanshu Mishra, Shamik Bose, Nicholas Broad, Yanis Labrak, Shlok Deshmukh, Sid Kiblawi, Ayush Singh, Minh Chien Vu, Trishala Neeraj, Jonas Golde, Albert Villanova del Moral, and Benjamin Beilharz. Bigbio: A framework for data-centric biomedical natural language processing. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 25792–25806. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/a583d2197eafc4afdd41f5b8765555c5-Paper-Datasets_and_Benchmarks.pdf.

[70] Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, September 2021. URL https://doi.org/10.5281/zenodo.5371628.

[71] Nicolas Garcelon, Antoine Neuraz, Vincent Benoit, Rémi Salomon, and Anita Burgun. Improving a full-text search engine: the importance of negation detection

and family history context to identify cases in a biomedical data warehouse. *Journal of the American Medical Informatics Association*, 24(3):607–613, 2017.

[72] Gerald Gartlehner, Gernot Wagner, Linda Lux, Lisa Affengruber, Andreea Dobrescu, Angela Kaminski-Hartenthaler, and Meera Viswanathan. Assessing the accuracy of machine-assisted abstract screening with DistillerAI: A user study. *Systematic Reviews*, 8(1), 11 2019. ISSN 20464053. doi: 10.1186/s13643-019-1221-3. URL https://pubmed.ncbi.nlm.nih.gov/31727159/.

[73] Gerald Gartlehner, Lisa Affengruber, Viktoria Titscher, Anna Noel-Storr, Gordon Dooley, Nicolas Ballarini, and Franz König. Single-reviewer abstract screening missed 13 percent of relevant studies: a crowd-based, randomized controlled trial. *Journal of Clinical Epidemiology*, 121:20–28, 5 2020. ISSN 18785921. doi: 10.1016/j.jclinepi.2020.01.005.

[74] Yao Ge, Yuting Guo, Yuan-Chi Yang, Mohammed Ali Al-Garadi, and Abeed Sarker. Few-shot learning for medical text: A systematic review. *arXiv preprint arXiv:2204.14081*, 2022.

[75] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Commun. ACM*, 64(12):86–92, 2021. doi: 10.1145/3458723. URL https://doi.org/10.1145/3458723.

[76] Rosemary Green. *American and Australian doctoral literature reviewing practices and pedagogies*. Deakin University (Australia), 2009.

[77] Andreas Grivas, Beatrice Alex, Claire Grover, Richard Tobin, and William Whiteley. Not a cute stroke: analysis of rule-and neural network-based information extraction systems for brain radiology reports. In *Proceedings of the 11th international workshop on health text mining and information analysis*, pages 24–37, 2020.

[78] Maura R Grossman, Gordon V Cormack, and Adam Roegiest. TREC 2016 Total Recall Track Overview. In *TREC*, 2016.

[79] 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 Trans. Comput. Heal.*, 3(1):2:1–2:23, 2022. doi: 10.1145/3458754. URL https://doi.org/10.1145/3458754.

[80] Nathalia Sernizon Guimarães, Andrêa JF Ferreira, Rita de Cássia Ribeiro Silva, Adelzon Assis de Paula, Cinthia Soares Lisboa, Laio Magno, Maria Yury Ichiara, and Maurício Lima Barreto. Deduplicating records in systematic reviews: There are free, accurate automated ways to do so. *Journal of Clinical Epidemiology*, 152:110–115, 2022.

[81] Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. Longt5: Efficient text-to-text transformer for long sequences. *arXiv preprint arXiv:2112.07916*, 2021.

[82] Neal R Haddaway and Martin J Westgate. Predicting the time needed for environmental systematic reviews and systematic maps. *Conservation Biology*, 33(2): 434–443, 2019.

[83] Neal R Haddaway, Alison Bethel, Lynn V Dicks, Julia Koricheva, Biljana Macura, Gillian Petrokofsky, Andrew S Pullin, Sini Savilaakso, and Gavin B Stewart. Eight problems with literature reviews and how to fix them. *Nature Ecology & Evolution*, 4(12):1582–1589, 2020.

[84] Abdelhakim Hannousse and Salima Yahiouche. A semi-automatic document screening system for computer science systematic reviews. In *Mediterranean Conference on Pattern Recognition and Artificial Intelligence*, pages 201–215. Springer, 2022.

[85] Henk Harkema, John N Dowling, Tyler Thornblade, and Wendy W Chapman. Context: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of biomedical informatics*, 42(5):839–851, 2009.

[86] Hannah Harrison, Simon J Griffin, Isla Kuhn, and Juliet A Usher-Smith. Software tools to support title and abstract screening for systematic reviews in healthcare: an evaluation. *BMC medical research methodology*, 20(1):1–12, 2020.

[87] Kazuma Hashimoto, Georgios Kontonatsios, Makoto Miwa, and Sophia Ananiadou. Topic detection using paragraph vectors to support active learning in systematic reviews. *Journal of Biomedical Informatics*, 62:59–65, 8 2016.

[88] Zinat Nadia Hatmi. A systematic review of systematic reviews on the covid-19 pandemic. *SN Comprehensive Clinical Medicine*, 3(2):419–436, 2021.

[89] Elke Hausner, Siw Waffenschmidt, Thomas Kaiser, and Michael Simon. Routine development of objectively derived search strategies. *Systematic reviews*, 1(1):1–10, 2012.

[90] Elke Hausner, Charlotte Guddat, Tatjana Hermanns, Ulrike Lampert, and Siw Waffenschmidt. Prospective comparison of search strategies for systematic reviews: an objective approach yielded higher sensitivity than a conceptual one. *Journal of clinical epidemiology*, 77:118–124, 2016.

[91] R Brian Haynes, David L Sackett, W Scott Richardson, William Rosenberg, and G Ross Langley. Evidence-based medicine: How to practice & teach ebm. *Canadian Medical Association. Journal*, 157(6):788, 1997.

[92] Jon Herman and Matthew Woodruff. pareto.py: a *varepsilon−nondomination* sorting routine. https://github.com/jdherman/pareto.py, 2013.

[93] William Hersh. *Information Retrieval: A Biomedical and Health Perspective*. Health Informatics. Springer International Publishing, Cham, health inf edition, 2020. ISBN 978-3-030-47685-4. doi: 10.1007/978-3-030-47686-1. URL http://link.springer.com/10.1007/978-3-030-47686-1http://www.springer.com/series/1114.

[94] JP Higgins, J Thomas, J Chandler, M Cumpston, T Li, MJ Page, and VA Welch. Cochrane handbook for systematic reviews of interventions version 6.2 (updated february 2021). cochrane, 2021. training.cochrane.org/handbook, 2021. Accessed: 2023-09-19.

[95] Julian PT Higgins, Tianjing Li, and Jonathan J Deeks. Choosing effect measures and computing estimates of effect. *Cochrane handbook for systematic reviews of interventions*, pages 143–176, 2019.

[96] hitz-zentroa. LM Contamination Index. GitHub repository, 2023. URL https://github.com/hitz-zentroa/lm-contamination.

[97] Falk Hoffmann, Katharina Allers, Tanja Rombey, Jasmin Helbach, Amrei Hoffmann, Tim Mathes, and Dawid Pieper. Nearly 80 systematic reviews were published each day: observational study on trends in epidemiology and reporting over the years 2000-2019. *Journal of Clinical Epidemiology*, 138:1–11, 2021.

[98] Jingwen Hou, Xiaochen Wang, Jean-Jacques Dubois, R. Byron Rice, Amanda Haddock, and Yue Wang. Extreme systematic reviews: A large literature screening dataset to support environmental policymaking. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, pages 4029–4033, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392365. doi: 10.1145/3511808.3557600. URL https://doi.org/10.1145/3511808.3557600.

[99] Brian E. Howard, Jason Phillips, Kyle Miller, Arpit Tandon, Deepak Mav, Mihir R. Shah, Stephanie Holmgren, Katherine E. Pelch, Vickie Walker, Andrew A. Rooney, Malcolm Macleod, Ruchir R. Shah, and Kristina Thayer. SWIFT-Review: A text-mining workbench for systematic review. *Systematic Reviews*, 5(1):1–16, 5 2016. ISSN 20464053. doi: 10.1186/s13643-016-0263-z. URL https://link.springer.com/articles/10.1186/s13643-016-0263-zhttps://link.springer.com/article/10.1186/s13643-016-0263-z.

[100] Brian E. Howard, Jason Phillips, Arpit Tandon, Adyasha Maharana, Rebecca Elmore, Deepak Mav, Alex Sedykh, Kristina Thayer, B. Alex Merrick, Vickie Walker, Andrew Rooney, and Ruchir R. Shah. SWIFT-Active Screener: Accelerated document screening through active learning and integrated recall estimation. *Environment International*, 138:105623, 5 2020. ISSN 0160-4120. doi: 10.1016/J.ENVINT.2020.105623.

[101] Kylie E Hunter, Angela C Webster, Matthew J Page, Melina Willson, Steve McDonald, Slavica Berber, Peta Skeers, Ava G Tan-Koay, Anne Parkhill, and Anna Lene Seidler. Searching clinical trials registers: guide for systematic reviewers. *Bmj*, 377, 2022.

[102] Alexandros Ioannidis. An Analysis of a BERT Deep Learning Strategy on a Technology Assisted Review Task, 4 2021. URL http://arxiv.org/abs/2104.08340.

[103] John PA Ioannidis. The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *The Milbank Quarterly*, 94(3):485–514, 2016.

[104] Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002.

[105] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 10 2002. ISSN 10468188. doi: 10.1145/582415.582418.

[106] Yulei Jiang, Charles E Metz, and Robert M Nishikawa. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology*, 201 (3):745–750, 1996.

[107] Qiao Jin, Chuanqi Tan, Zhengyun Zhao, Zheng Yuan, and Songfang Huang. Alibaba DAMO Academy at TREC Clinical Trials 2021: Exploring Embedding-based First-stage Retrieval with TrialMatcher. *TREC 2021*, 2021.

[108] Akers Jo, Aguiar-Ibáñez Raquel, Burch Jane, Chambers Duncan, Eastwood Alison, Fayter Debra, Hempel Susanne, Light Kate, Rice Stephen, Rithalia Amber, Stewart Lesley, Stock Christian, Wilson Paul, and Woolacott Nerys. *Systematic Reviews: CRD's guidance for undertaking reviews in health care*. CRD, University of York, York, 1 2009. ISBN 978-1-900991-19-3. URL www.york.ac.uk/inst/crd.

[109] K Sparck Jones, Steve Walker, and Stephen E. Robertson. A probabilistic model of information retrieval: development and comparative experiments: Part 2. *Information processing & management*, 36(6):809–840, 2000.

[110] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference*, volume 2, pages 427–431, 7 2017. ISBN 9781510838604. doi: 10.18653/v1/e17-2068. URL http://arxiv.org/abs/1607.01759.

[111] Tian Kang, Shaodian Zhang, Youlan Tang, Gregory W Hruby, Alexander Rusanov, Noémie Elhadad, and Chunhua Weng. EliIE: An open-source information extraction system for clinical trial eligibility criteria. *Journal of the American Medical Informatics Association*, 24(6):1062–1071, 04 2017. ISSN 1067-5027. doi: 10.1093/jamia/ocx019. URL https://doi.org/10.1093/jamia/ocx019.

[112] Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker. CLEF 2017 Technologically Assisted Reviews in empirical medicine overview. *CEUR Workshop Proceedings*, 1866:1–29, 9 2017. ISSN 1613-0073. URL https://pureportal.strath.ac.uk/en/publications/clef-2017-technologically-assisted-reviews-in-empirical-medicine-.

[113] Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker. CLEF 2018 Technologically Assisted Reviews in empirical medicine overview. *CEUR Workshop Proceedings*, 2125, 7 2018. ISSN 1613-0073. URL https://pureportal.strath.ac.uk/en/publications/clef-2018-technologically-assisted-reviews-in-empirical-medicine-.

[114] Evangelos Kanoulas, Dan Li, Leif Azzopardi, and René Spijker. CLEF 2019 Technology Assisted Reviews in Empirical Medicine Overview. In *CLEF*, 2019.

[115] Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9), 2023.

[116] Sayash Kapoor, Emily Cantrell, Kenny Peng, Thanh Hien Pham, Christopher A Bail, Odd Erik Gundersen, Jake M Hofman, Jessica Hullman, Michael A Lones, Momin M Malik, et al. Reforms: Reporting standards for machine learning based science. *arXiv preprint arXiv:2308.07832*, 2023.

[117] Sarvnaz Karimi, Justin Zobel, Stefan Pohl, and Falk Scholer. The challenge of high recall in biomedical systematic search. In *Proceedings of the third international workshop on Data and text mining in bioinformatics*, pages 89–92, 2009.

[118] Sarvnaz Karimi, Stefan Pohl, Falk Scholer, Lawrence Cavedon, and Justin Zobel. Boolean versus ranked querying for biomedical systematic reviews. *BMC medical informatics and decision making*, 10(1):1–20, 2010.

[119] Staffs Keele et al. Guidelines for performing systematic literature reviews in software engineering. Technical report, Technical report, ver. 2.3 ebse technical report. ebse, 2007.

[120] Matthew D Keller, Brandon Harrison-Smith, Chetan Patil, and Mohammed Shahriar Arefin. Skin colour affects the accuracy of medical oxygen sensors, 2022.

[121] Diane Kelly and Cassidy R. Sugimoto. A systematic review of interactive information retrieval evaluation studies, 1967–2006. *Journal of the American Society for Information Science and Technology*, 64(4):745–770, 2013. ISSN 1532-2890. doi: 10.1002/asi.22799. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.22799. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.22799.

[122] Madian Khabsa, Ahmed Elmagarmid, Ihab Ilyas, Hossam Hammady, and Mourad Ouzzani. Learning to identify relevant studies for systematic reviews using random forest and external information. *Machine Learning*, 102(3):465–482, 3 2016. ISSN 15730565. doi: 10.1007/S10994-015-5535-7. URL https://link.springer.com/article/10.1007/s10994-015-5535-7.

[123] Petr Knoth and Zdenek Zdrahal. CORE: three access levels to underpin open access. *D-Lib Magazine*, 18(11/12):1–13, 2012.

[124] Georgios Kontonatsios, Sally Spencer, Peter Matthew, and Ioannis Korkontzelos. Using a neural network-based feature extraction method to facilitate citation screening for systematic reviews. *Expert Systems with Applications: X*, 6:100030, 7 2020. ISSN 25901885. doi: 10.1016/j.eswax.2020.100030.

[125] Bevan Koopman and Guido Zuccon. A test collection for matching patients to clinical trials. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 669–672, 2016.

[126] Bevan Koopman and Guido Zuccon. Cohort-based clinical trial retrieval. In *Proceedings of the 25th Australasian Document Computing Symposium*, pages 1–9, 2021.

[127] Peter Kranke. Evidence-based practice: how to perform and use systematic reviews for clinical decision-making. *European Journal of Anaesthesiology| EJA*, 27(9): 763–772, 2010.

[128] Wojciech Kusa. Rapid systematic reviews: Zero-shot citation screening with the usage of eligibility criteria. In *The 17th Conference of the European Chapter of the Association for Computational Linguistics Student Research Workshop*, Dubrovnik, Croatia, 2023.

[129] Wojciech Kusa and Yasin Ghafourian. DoSSIER at TREC 2021 Clinical Trials Track. In *Proceedings of the Thirtieth Text REtrieval Conference (TREC 2021)*, 2021.

[130] Wojciech Kusa, Allan Hanbury, and Petr Knoth. Automation of citation screening for systematic literature reviews using neural networks: A replicability study. In Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty, editors, *Advances in Information Retrieval*, pages 584–598, Cham, 2022. Springer International Publishing. ISBN 978-3-030-99736-6. URL https://doi.org/10.1007/978-3-030-99736-6_39.

[131] Wojciech Kusa, Petr Knoth, and Allan Hanbury. Evaluation of Automated Citation Screening with Normalised Work Saved over Sampling: an Analysis. In *1st Workshop on Augmented Intelligence for Technology-Assisted Reviews Systems*, 2022.

[132] Wojciech Kusa, Georgios Peikos, Óscar Espitia, Allan Hanbury, and Gabriella Pasi. DoSSIER at MedVidQA 2022: Text-based approaches to medical video answer localization problem. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 432–440, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.bionlp-1.43. URL https://aclanthology.org/2022.bionlp-1.43.

[133] Wojciech Kusa, Petr Knoth, and Allan Hanbury. CRUISE–Screening: Living Literature Reviews Toolbox. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23)*, page ToBeDetermined, Birmingham, United Kingdom, October 21–25 2023. ACM, New York, NY, USA. doi: https://doi.org/10.1145/3583780.3614736.

[134] Wojciech Kusa, Aldo Lipani, Petr Knoth, and Allan Hanbury. An Analysis of Work Saved over Sampling in the Evaluation of Automated Citation Screening in Systematic Literature Reviews. *Intelligent Systems with Applications*, 18:200193, 2023. ISSN 2667-3053. doi: https://doi.org/10.1016/j.iswa.2023.200193. URL https://www.sciencedirect.com/science/article/pii/S2667305323000182.

[135] Wojciech Kusa, Aldo Lipani, Petr Knoth, and Allan Hanbury. VoMBaT: A Tool for Visualising Evaluation Measure Behaviour in High-Recall Search Tasks. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*, page 5, Taipei, Taiwan, July 23–27 2023. ACM. URL https://doi.org/10.1145/3539618.3591802.

[136] Wojciech Kusa, Oscar E. Mendoza, Petr Knoth, Gabriella Pasi, and Allan Hanbury. Effective Matching of Patients to Clinical Trials using Entity Extraction and Neural Re-ranking. *Journal of Biomedical Informatics*, JBI, 2023.

[137] Wojciech Kusa, Óscar E. Mendoza, Matthias Samwald, Petr Knoth, and Allan Hanbury. CSMeD: Bridging the Dataset Gap in Automated Citation Screening for Systematic Literature Reviews. In *37th Conference on Neural Information Processing Systems Track on Datasets and Benchmarks*, 2023.

[138] Wojciech Kusa, Edoardo Mosca, and Aldo Lipani. "Dr LLM, what do I have?": The impact of user beliefs and prompt formulation on health diagnoses. In *NLPMC 2023: 3rd Workshop on NLP for Medical Conversations at IJCNLP-AACL 2023*, 2023.

[139] Wojciech Kusa, Patrick Styll, Maximilian Seeliger, Oscar E. Mendoza, and Allan Hanbury. DoSSIER at TREC 2023 Clinical Trials Track. In *Proceedings of the Thirty-Second Text REtrieval Conference (TREC 2023)*, 2023.

[140] Wojciech Kusa, Guido Zuccon, Petr Knoth, and Allan Hanbury. Outcome-based evaluation of systematic review automation. In *Proceedings of the 2023 ACM*

SIGIR International Conference on the Theory of Information Retrieval (ICTIR '23), page 9, Taipei, Taiwan, July 2023. ACM. doi: 10.1145/3578337.3605135. URL https://doi.org/10.1145/3578337.3605135.

[141] Athanasios Lagopoulos and Grigorios Tsoumakas. From protocol to screening: A hybrid learning approach for technology-assisted systematic literature reviews. 11 2020. ISSN 23318422. URL http://arxiv.org/abs/2011.09752.

[142] Peder Larsen and Markus Von Ins. The rate of growth in scientific publication and the decline in coverage provided by science citation index. *Scientometrics*, 84(3): 575–603, 2010.

[143] Joseph Lau. Systematic review automation thematic series. *Systematic reviews*, 8: 1–2, 2019.

[144] Dawn Lawrie, James Mayfield, Douglas W Oard, and Eugene Yang. HC4: A new suite of test collections for ad hoc CLIR. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I*, pages 351–366. Springer, 2022.

[145] Eric W Lee and Joyce C Ho. Pgb: A pubmed graph benchmark for heterogeneous network representation learning. *arXiv preprint arXiv:2305.02691*, 2023.

[146] Eric W Lee and Joyce C Ho. Sr-comber: Heterogeneous network embedding using community multi-view enhanced graph convolutional network for automating systematic reviews. In *European Conference on Information Retrieval*, pages 553–568. Springer, 2023.

[147] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*, 2021.

[148] Johannes Leveling. Patient selection for clinical trials based on concept-based retrieval and result filtering and ranking. In *TREC*, 2017.

[149] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703.

[150] Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, et al. Datasets: A community library for natural language processing. *arXiv preprint arXiv:2109.02846*, 2021.

[151] Dan Li and Evangelos Kanoulas. When to Stop Reviewing in Technology-Assisted Reviews. *ACM Transactions on Information Systems (TOIS)*, 38(4), 9 2020. ISSN 15582868. doi: 10.1145/3411755. URL https://dl.acm.org/doi/abs/10.1145/3411755.

[152] Dan Li, Panagiotis Zafeiriadis, and Evangelos Kanoulas. Aps: An active pubmed search system for technology assisted reviews. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2137–2140, 2020.

[153] Hang Li, Harrisen Scells, and Guido Zuccon. Systematic review automation tools for end-to-end query formulation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2141–2144, 2020.

[154] Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *arXiv preprint arXiv:2201.11838*, 2022.

[155] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. 7 2021. URL https://arxiv.org/abs/2107.13586v1.

[156] Theo Lorenc, Lambert Felix, Mark Petticrew, GJ Melendez-Torres, James Thomas, Sian Thomas, Alison O'Mara-Eves, and Michelle Richardson. Meta-analysis, complexity, and heterogeneity: a qualitative interview study of researchers' methodological values and practices. *Systematic Reviews*, 5:1–9, 2016.

[157] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[158] Alexander V Lotov and Kaisa Miettinen. Visualizing the pareto frontier. *Multiobjective optimization*, 5252:213–243, 2008.

[159] Michael Mabe and Mayur Amin. Growth dynamics of scholarly and scientific journals. *Scientometrics*, 51(1):147–162, 2001.

[160] Andrew MacFarlane, Tony Russell-Rose, and Farhad Shokraneh. Search strategy formulation for systematic reviews: Issues, challenges and opportunities. *Intelligent Systems with Applications*, 15:200091, 2022.

[161] Christopher Marshall and Pearl Brereton. Systematic review toolbox: a catalogue of tools to support systematic reviews. In *Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering*, pages 1–6, 2015.

[162] 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.

[163] Iain J Marshall, Rachel Marshall, Byron C Wallace, Jon Brassey, and James Thomas. Rapid reviews may produce different results to systematic reviews: a meta-epidemiological study. *Journal of clinical epidemiology*, 109:30–41, 2019.

[164] Iain J Marshall, Thomas A Trikalinos, Frank Soboczenski, Hye Sun Yun, Gregory Kell, Rachel Marshall, and Byron C Wallace. In a pilot study, automated real-time systematic review updates were feasible, accurate, and work-saving. *Journal of Clinical Epidemiology*, 153:26–33, 2023.

[165] David Martinez, Sarvnaz Karimi, Lawrence Cavedon, and Timothy Baldwin. Facilitating biomedical systematic reviews using ranked text retrieval and classification. In *Australasian Document Computing Symposium (ADCS)*, pages 53–60, 2008.

[166] Stan Matwin, Alexandre Kouznetsov, Diana Inkpen, Oana Frunza, and Peter O'Blenis. A new algorithm for reducing the workload of experts in performing systematic reviews. *Journal of the American Medical Informatics Association*, 17 (4):446–453, 7 2010. doi: 10.1136/JAMIA.2010.004325.

[167] Donna Katzman McClish. Analyzing a portion of the ROC curve. *Medical decision making*, 9(3):190–195, 1989.

[168] Graham Mcdonald, Craig Macdonald, and Iadh Ounis. How the accuracy and confidence of sensitivity classification affects digital sensitivity review. *ACM Transactions on Information Systems (TOIS)*, 39(1):1–34, 2020.

[169] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018.

[170] Germano Duarte Mergel, Milene Selbach Silveira, and Tiago Silva da Silva. A method to support search string building in systematic literature reviews through visual text mining. In *Proceedings of the 30th annual ACM symposium on applied computing*, pages 1594–1601, 2015.

[171] Stéphane M Meystre, Paul M Heider, Andrew Cates, Grace Bastian, Tara Pittman, Stephanie Gentilin, and Teresa J Kelechi. Piloting an automated clinical trial eligibility surveillance and provider alert system based on artificial intelligence and standard data models. *BMC Medical Research Methodology*, 23(1):1–11, 2023.

[172] Makoto Miwa, James Thomas, Alison O'Mara-Eves, and Sophia Ananiadou. Reducing systematic review workload through certainty-based screening. *Journal of Biomedical Informatics*, 51:242–253, 10 2014. ISSN 1532-0464. doi: 10.1016/J.JBI.2014.06.005.

[173] Alistair Moffat. Batch evaluation metrics in information retrieval: Measures, scales, and meaning. *IEEE Access*, 10:105564–105577, 2022.

[174] David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G Altman, and PRISMA Group*. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *Annals of internal medicine*, 151(4):264–269, 2009.

[175] Amirkeivan Mohtashami and Martin Jaggi. Landmark attention: Random-access infinite context length for transformers. *arXiv preprint arXiv:2305.16300*, 2023.

[176] Cristhian D Morales-Plaza, David A Forero-Peña, and Fhabián S Carrión-Nessi. Resource use during systematic review production varies widely: a scoping review: response to nussbaumer-streit et al. *Journal of Clinical Epidemiology*, 142:319–320, 2022.

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

[178] Arvind Narayanan and Sayash Kapoor. Is gpt-4 getting worse over time?, 2023. URL `https://www.aisnakeoil.com/p/is-gpt-4-getting-worse-over-time`. Accessed: 2023-11-28.

[179] Lauge Neimann Rasmussen and Paul Montgomery. The prevalence of and factors associated with inclusion of non-english language studies in campbell systematic reviews: a survey and meta-epidemiological study. *Systematic Reviews*, 7(1):1–12, 2018.

[180] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5034. URL `https://www.aclweb.org/anthology/W19-5034`.

[181] Yizhao Ni, Stephanie Kennebeck, Judith W Dexheimer, Constance M McAneney, Huaxiu Tang, Todd Lingren, Qi Li, Haijun Zhai, and Imre Solti. Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department. *Journal of the American Medical Informatics Association*, 22(1):166–178, 2015.

[182] Christopher Norman. *Systematic review automation methods*. PhD thesis, Université Paris-Saclay ; Universiteit van Amsterdam, 2 2020. URL `https://tel.archives-ouvertes.fr/tel-03060620`.

[183] B Nussbaumer-Streit, I Klerings, AI Dobrescu, E Persad, A Stevens, C Garritty, C Kamel, L Affengruber, VJ King, and G Gartlehner. Excluding non-english publications from evidence-syntheses did not change conclusions: a meta-epidemiological study. *Journal of clinical epidemiology*, 118:42–54, 2020.

[184] B Nussbaumer-Streit, LE Ziganshina, M Mahmić-Kaknjo, G Gartlehner, R Sfetcu, and H Lund. Resource use during systematic review production varies widely: a scoping review: authors' reply. *Journal of Clinical Epidemiology*, 142:321–322, 2022.

[185] Barbara Nussbaumer-Streit, Irma Klerings, Gernot Wagner, Thomas L. Heise, Andreea I. Dobrescu, Susan Armijo-Olivo, Jan M. Stratil, Emma Persad, Stefan K. Lhachimi, Megan G. Van Noord, Tarquin Mittermayr, Hajo Zeeb, Lars Hemkens, and Gerald Gartlehner. Abbreviated literature searches were viable alternatives to comprehensive searches: a meta-epidemiological study. *Journal of Clinical Epidemiology*, 102:1–11, 2018. ISSN 0895-4356. doi: https://doi.org/10.1016/j.jclinepi.2018.05.022. URL `https://www.sciencedirect.com/science/article/pii/S0895435618300179`.

[186] Barbara Nussbaumer-Streit, Moriah Ellen, Irma Klerings, Raluca Sfetcu, Nicoletta Riva, Mersiha Mahmić-Kaknjo, Georgios Poulentzas, P Martinez, Eduard Baladia, Liliya Eugenevna Ziganshina, et al. Resource use during systematic review production varies widely: a scoping review. *Journal of Clinical Epidemiology*, 139: 287–296, 2021.

[187] Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain J Marshall, Ani Nenkova, and Byron C Wallace. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2018, page 197. NIH Public Access, 2018.

[188] Douglas W Oard, Bruce Hedin, Stephen Tomlinson, and Jason R Baron. Overview of the trec 2008 legal track. Technical report, MARYLAND UNIV COLLEGE PARK COLL OF INFORMATION STUDIES, 2008.

[189] Annette M O'Connor, Paul Glasziou, Michele Taylor, James Thomas, René Spijker, and Mary S Wolfe. A focus on cross-purpose tools, automated recognition of study design in multiple disciplines, and evaluation of automation tools: a summary of significant discussions at the fourth meeting of the international collaboration for automation of systematic reviews (icasr). *Systematic reviews*, 9(1):1–6, 2020.

[190] Alison O'Mara-Eves, James Thomas, John McNaught, Makoto Miwa, and Sophia Ananiadou. Using text mining for study identification in systematic reviews: A systematic review of current approaches. *Systematic Reviews*, 4(1):5, 1 2015. ISSN 20464053. doi: 10.1186/2046-4053-4-5.

[191] OpenAI. Gpt-4 technical report, 2023.

[192] Arsenio Paez. Gray literature: An important resource in systematic reviews. *Journal of Evidence-Based Medicine*, 10(3):233–240, 2017.

[193] Joao Palotti, Guido Zuccon, and Allan Hanbury. Mm: a new framework for multidimensional evaluation of search engines. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 1699–1702, 2018.

[194] Mihir Parmar, Swaroop Mishra, Mirali Purohit, Man Luo, Murad Mohammad, and Chitta Baral. In-BoXBART: Get instructions into biomedical multi-task learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 112–128, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.10. URL https://aclanthology.org/2022.findings-naacl.10.

[195] Mihir Prafullsinh Parmar. Automation of title and abstract screening for clinical systematic reviews. Master's thesis, Arizona State University, 2021. URL https://keep.lib.asu.edu/_flysystem/fedora/c7/Parmar_asu_0010N_21179.pdf.

[196] Karl Pearson. Report on certain enteric fever inoculation statistics. *British Medical Journal*, 2:1243 – 1246, 1904.

[197] Georgios Peikos and Gabriella Pasi. Multidimensional relevance in legal and health domains. In *Italian Information Retrieval Workshop*, 2021.

[198] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5006. URL https://aclanthology.org/W19-5006.

[199] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global Vectors for Word Representation. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 1532–1543, 2014. doi: 10.3115/V1/D14-1162. URL https://aclanthology.org/D14-1162.

[200] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019.

[201] Mark Petticrew and Helen Roberts. *Systematic reviews in the social sciences: A practical guide*. John Wiley & Sons, 2008.

[202] Catherine Pickering and Jason Byrne. The benefits of publishing systematic quantitative literature reviews for PhD candidates and other early-career researchers. *Higher Education Research & Development*, 33(3):534–548, 2014.

[203] Ronak Pradeep, Yilin Li, Yuetong Wang, and Jimmy Lin. Neural query synthesis and domain-specific ranking templates for multi-stage clinical trial matching. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)*, 2022.

[204] Taylor R Pressler, Po-Yin Yen, Jing Ding, Jianhua Liu, Peter J Embi, and Philip RO Payne. Computational challenges and human factors influencing the design and use of clinical research participant eligibility pre-screening tools. *BMC medical informatics and decision making*, 12(1):1–11, 2012.

[205] Jason Priem, Heather Piwowar, and Richard Orr. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*, 2022.

[206] Piotr Przybyła, Austin J. Brockmeier, Georgios Kontonatsios, Marie-Annick Le Pogam, John McNaught, Erik von Elm, Kay Nolan, and Sophia Ananiadou. Prioritising references for systematic reviews with robotanalyst: A user study. *Research Synthesis Methods*, 9:470 – 488, 2018.

[207] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1776–1826, 2022.

[208] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140): 1–67, 2020.

[209] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[210] Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data Programming: Creating Large Training Sets, Quickly. *Advances in Neural Information Processing Systems*, pages 3574–3582, 5 2016. ISSN 10495258. URL `https://arxiv.org/abs/1605.07723v3`.

[211] Scott Reeves, Ivan Koppel, Hugh Barr, Della Freeth, and Marilyn Hammick. Twelve tips for undertaking a systematic review. *Medical teacher*, 24(4):358–363, 2002.

[212] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 3982–3992, 8 2019. URL `https://arxiv.org/abs/1908.10084v1`.

[213] Kirk Roberts, Matthew S Simpson, Ellen M Voorhees, and William R Hersh. Overview of the TREC 2015 Clinical Decision Support track. In *TREC*, 2015.

[214] 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 *TREC*, 2017.

[215] 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 *The text REtrieval conference: TREC. Text REtrieval Conference*, 2019.

[216] Kirk Roberts, Dina Demner-Fushman, Ellen M Voorhees, Steven Bedrick, and Willian R Hersh. Overview of the TREC 2021 Clinical Trials Track. In *Proceedings of the Thirtieth Text REtrieval Conference (TREC 2021)*, 2021.

[217] Kirk Roberts, Dina Demner-Fushman, Ellen M Voorhees, Steven Bedrick, and Willian R Hersh. Overview of the TREC 2022 Clinical Trials Track. In *Proceedings of the Thirtieth Text REtrieval Conference (TREC 2022)*, 2022.

[218] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. Okapi at TREC-3. *Nist Special Publication Sp*, 109:109, 1995.

[219] Maciej Rybinski, Jerry Xu, and Sarvnaz Karimi. Clinical trial search: Using biomedical language understanding models for re-ranking. *Journal of Biomedical Informatics*, 109:103530, 2020.

[220] Maciej Rybiński, Vincent Nguyen, and Sarvnaz Karimi. CSIROmed Team Report of TREC 2021 Clinical Trials track: Experiments with BERT Reranking Methods. 2022.

[221] David L Sackett. Evidence-based medicine. In *Seminars in perinatology*, volume 21, pages 3–5. Elsevier, 1997.

[222] David L Sackett, William M C Rosenberg, J A Muir Gray, R Brian Haynes, and W Scott Richardson. Evidence based medicine: what it is and what it isn't. *BMJ*, 312(7023):71–72, 1996. ISSN 0959-8138. doi: 10.1136/bmj.312.7023.71. URL `https://www.bmj.com/content/312/7023/71`.

[223] Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. Multitask

prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=9Vrb9D0WI4.

[224] Harrisen Scells and Guido Zuccon. Generating better queries for systematic reviews. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 475–484, 2018.

[225] Harrisen Scells, Guido Zuccon, Bevan Koopman, Anthony Deacon, Leif Azzopardi, and Shlomo Geva. A test collection for evaluating retrieval of studies for inclusion in systematic reviews. *SIGIR 2017 - Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1237–1240, 8 2017. doi: 10.1145/3077136.3080707.

[226] Harrisen Scells, Guido Zuccon, and Bevan Koopman. Automatic boolean query refinement for systematic review literature search. *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019*, 11:1646–1656, 5 2019. doi: 10.1145/3308558.3313544. URL https://doi.org/10.1145/3308558.3313544.

[227] Harrisen Scells, Guido Zuccon, Bevan Koopman, and Justin Clark. Automatic Boolean Query Formulation for Systematic Review Literature Search. In *The Web Conference 2020 - Proceedings of the World Wide Web Conference, WWW 2020*, pages 1071–1081, New York, NY, USA, 2020. ACM. ISBN 9781450370233. doi: 10.1145/3366423.3380185. URL https://doi.org/10.1145/3366423.3380185.

[228] Harrisen Scells, Guido Zuccon, and Bevan Koopman. A comparison of automatic boolean query formulation for systematic reviews. *Information Retrieval Journal*, 24(1):3–28, 2021.

[229] Timo Schick and Hinrich Schütze. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*, 2020.

[230] Anna Mae Scott, Connor Forbes, Justin Clark, Matt Carter, Paul Glasziou, and Zachary Munn. Systematic review automation tool use by systematic reviewers, health technology assessors and clinical guideline developers: tools used, abandoned, and desired. *MedRxiv*, pages 2021–04, 2021.

[231] Hamza Sellak, Brahim Ouhbi, Bouchra Frikh, and Sidi Mohamed Ben. Using Rule-based Classifiers in Systematic Reviews: A Semantic Class Association Rules Approach. *Proceedings of the 17th International Conference on Information Integration and Web-based Applications & Services*, 2015. doi: 10.1145/2837185. URL http://dx.doi.org/10.1145/2837185.2837279.

[232] Seyedmostafa Sheikhalishahi, Riccardo Miotto, Joel T Dudley, Alberto Lavelli, Fabio Rinaldi, and Venet Osmani. Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. *JMIR Med Inform*, 7(2):e12239, Apr 2019. ISSN 2291-9694. doi: 10.2196/12239. URL https://doi.org/10.2196/12239.

[233] Ian Shemilt, Nada Khan, Sophie Park, and James Thomas. Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews. *Systematic reviews*, 5(1):1–13, 2016.

[234] Feichen Shen, Sijia Liu, Sunyang Fu, Yanshan Wang, Sam Henry, Ozlem Uzuner, and Hongfang Liu. Family History Extraction From Synthetic Clinical Narratives Using Natural Language Processing: Overview and Evaluation of a Challenge Data Set and Solutions for the 2019 National NLP Clinical Challenges (n2c2)/Open Health Natural Language Processing (OHNLP) Competition. *JMIR Med Inform*, 9(1):e24008, Jan 2021. ISSN 2291-9694. doi: 10.2196/24008. URL https://doi.org/10.2196/24008.

[235] Chaitanya Shivade, Courtney Hebert, Marcelo Lopetegui, Marie-Catherine De Marneffe, Eric Fosler-Lussier, and Albert M Lai. Textual inference for eligibility criteria resolution in clinical trials. *Journal of biomedical informatics*, 58:S211–S218, 2015.

[236] Kaveh G. Shojania, Margaret Sampson, Mohammed T. Ansari, Jun Ji, Steve Doucette, and David Moher. How quickly do systematic reviews go out of date? A survival analysis. *Annals of Internal Medicine*, 147(4):224–233, 8 2007. doi: 10.7326/0003-4819-147-4-200708210-00179.

[237] Gaurav Singh, James Thomas, and John Shawe-Taylor. Improving active learning in systematic reviews. *arXiv preprint arXiv:1801.09496*, 2018.

[238] Konstantinos C Siontis and John PA Ioannidis. Replication, duplication, and waste in a quarter million systematic reviews and meta-analyses, 2018.

[239] Konstantinos C Siontis, Tina Hernandez-Boussard, and John PA Ioannidis. Overlapping meta-analyses on the same topic: survey of published studies. *Bmj*, 347, 2013.

[240] Mary Lee Smith, Gene V Glass, and Thomas I Miller. *The benefits of psychotherapy*. Johns Hopkins University Press, 1980.

[241] Ayah Soufan, Ian Ruthven, and Leif Azzopardi. Searching the literature: an analysis of an exploratory search task. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*, pages 146–157, 2022.

[242] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

[243] Mark Stevenson and Reem Bin-Hezam. Stopping methods for technology assisted reviews based on point processes. *ACM Transactions on Information Systems*, 2023.

[244] Qianmin Su, Gaoyi Cheng, and Jihan Huang. A review of research on eligibility criteria for clinical trials. *Clinical and Experimental Medicine*, pages 1–13, 2023.

[245] Dominic Sykes, Andreas Grivas, Claire Grover, Richard Tobin, Cathie Sudlow, William Whiteley, Andrew McIntosh, Heather Whalley, and Beatrice Alex. Comparison of rule-based and neural network models for negation detection in radiology reports. *Natural Language Engineering*, 27(2):203–224, 2021.

[246] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. Arnetminer: Extraction and mining of academic social networks. In *KDD'08*, pages 990–998, 2008.

[247] Athina Tatsioni, Deborah A Zarin, Naomi Aronson, David J Samson, Carole R Flamm, Christopher Schmid, and Joseph Lau. Challenges in systematic reviews of diagnostic technologies. *Annals of internal medicine*, 142(12_Part_2):1048–1055, 2005.

[248] James Thomas, Lisa M Askie, Jesse A Berlin, Julian H Elliott, Davina Ghersi, Mark Simmonds, Yemisi Takwoingi, Jayne F Tierney, and Julian PT Higgins. Prospective approaches to accumulating evidence. *Cochrane Handbook for systematic reviews of interventions*, pages 547–568, 2019.

[249] Prem Timsina, Jun Liu, and Omar El-Gayar. Advanced analytics for the automation of medical systematic reviews. *Information Systems Frontiers*, 18(2):237–252, 4 2016. ISSN 15729419. doi: 10.1007/S10796-015-9589-7/FIGURES/1. URL https://link.springer.com/article/10.1007/s10796-015-9589-7.

[250] George Tomlin and Bernhard Borgetto. Research pyramid: A new evidence-based practice model for occupational therapy. *The American Journal of Occupational Therapy*, 65(2):189–196, 2011.

[251] Stephen Tomlinson, Douglas W Oard, Jason R Baron, and Paul Thompson. Overview of the trec 2007 legal track. In *TREC*, 2007.

[252] Andrea C. Tricco, Jamie Brehaut, Maggie H. Chen, and David Moher. Following 411 Cochrane Protocols to Completion: A Retrospective Cohort Study. *PLOS ONE*, 3(11):e3684, 11 2008. ISSN 1932-6203. doi: 10.1371/JOURNAL.PONE.0003684. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0003684.

[253] Andrew Trotman, Antti Puurula, and Blake Burgess. Improvements to BM25 and language models examined. In *Proceedings of the 2014 Australasian Document Computing Symposium*, pages 58–65, 2014.

[254] Guy Tsafnat, Paul Glasziou, Miew K. Choong, Adam Dunn, Filippo Galgani, and Enrico Coiera. Systematic review automation technologies, 4 2014. ISSN 20464053.

[255] Guy Tsafnat, Paul Glasziou, George Karystianis, and Enrico Coiera. Automated screening of research studies for systematic reviews using study characteristics. *Systematic Reviews 2018 7:1*, 7(1):1–9, 4 2018. ISSN 2046-4053. doi: 10.1186/S13643-018-0724-7. URL `https://link.springer.com/articles/10.1186/s13643-018-0724-7https://link.springer.com/article/10.1186/s13643-018-0724-7`.

[256] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. From imagenet to image classification: Contextualizing progress on benchmarks. In *International Conference on Machine Learning*, pages 9625–9635. PMLR, 2020.

[257] Amy Y Tsou, Jonathan R Treadwell, Eileen Erinoff, and Karen Schoelles. Machine learning for screening prioritization in systematic reviews: comparative performance of abstrackr and eppi-reviewer. *Systematic reviews*, 9:1–14, 2020.

[258] Peter Tugwell, Vivian Andrea Welch, Sathya Karunananthan, Lara J Maxwell, Elie A Akl, Marc T Avey, Zulfiqar A Bhutta, Melissa C Brouwers, Jocalyn P Clark, Sophie Cook, et al. When to replicate systematic reviews of interventions: consensus checklist. *Bmj*, 370, 2020.

[259] Betty Van Aken, Ivana Trajanovska, Amy Siu, Manuel Mayrdorfer, Klemens Budde, and Alexander Löser. Assertion detection in clinical notes: Medical language models to the rescue? In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 35–40, 2021.

[260] Rens van de Schoot, Jonathan de Bruin, Raoul Schram, Parisa Zahedi, Jan de Boer, Felix Weijdema, Bianca Kramer, Martijn Huijts, Maarten Hoogerwerf, Gerbrich Ferdinands, et al. An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, 3(2):125–133, 2021.

[261] Raymon van Dinter, Cagatay Catal, and Bedir Tekinerdogan. A decision support system for automating document retrieval and citation screening. *Expert Systems with Applications*, 182, 11 2021.

[262] Raymon van Dinter, Cagatay Catal, and Bedir Tekinerdogan. A Multi-Channel Convolutional Neural Network approach to automate the citation screening process. *Applied Soft Computing*, 112:107765, 11 2021. ISSN 1568-4946. doi: 10.1016/J.ASOC.2021.107765.

[263] Raymon van Dinter, Bedir Tekinerdogan, and Cagatay Catal. Automation of systematic literature reviews: A systematic literature review. *Information and Software Technology*, 136:106589, 8 2021. ISSN 09505849. doi: 10.1016/j.infsof.

2021.106589. URL https://linkinghub.elsevier.com/retrieve/pii/S0950584921000690.

[264] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.*, volume 2017-Decem, pages 5999–6009, 6 2017. URL http://arxiv.org/abs/1706.03762.

[265] Ellen M Voorhees. The TREC medical records track. In *proceedings of the international conference on bioinformatics, computational biology and biomedical informatics*, pages 239–246, 2013.

[266] Siw Waffenschmidt, Marco Knelangen, Wiebke Sieben, Stefanie Bühn, and Dawid Pieper. Single screening versus conventional double screening for study selection in systematic reviews: a methodological systematic review. *BMC medical research methodology*, 19(1):1–9, 2019.

[267] Gernot Wagner, Barbara Nussbaumer-Streit, Judith Greimel, Agustín Ciapponi, and Gerald Gartlehner. Trading certainty for speed - how much uncertainty are decisionmakers and guideline developers willing to accept when using rapid reviews: an international survey. *BMC Medical Research Methodology*, 17, 2017.

[268] Byron C. Wallace, Kevin Small, Carla E. Brodley, and Thomas A. Trikalinos. Active learning for biomedical citation screening. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 173–181, 2010. doi: 10.1145/1835804.1835829.

[269] Byron C Wallace, Thomas A Trikalinos, Joseph Lau, Carla Brodley, and Christopher H Schmid. Semi-automated screening of biomedical citations for systematic reviews. *BMC Bioinformatics 2010 11:1*, 11(1):1–11, 1 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-55. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-55.

[270] Byron C. Wallace, Kevin Small, Carla E. Brodley, Joseph Lau, and Thomas A. Trikalinos. Deploying an interactive machine learning system in an Evidence-based Practice Center: Abstrackr. *IHI'12 - Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 819–823, 2012. doi: 10.1145/2110363.2110464. URL http://github.com/bwallace/abstrackr-web.

[271] Byron C. Wallace, Sayantan Saha, Frank Soboczenski, and Iain J. Marshall. Generating (Factual?) Narrative Summaries of RCTs: Experiments with Neural Multi-Document Summarization. *arXiv*, 8 2020. URL http://arxiv.org/abs/2008.11293.

[272] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

[273] Lucy Lu Wang, Jay DeYoung, and Byron Wallace. Overview of MSLR2022: A shared task on multi-document summarization for literature reviews. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 175–180, Gyeongju, Republic of Korea, October 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.sdp-1.20.

[274] Shuai Wang, Harrisen Scells, Ahmed Mourad, and Guido Zuccon. Seed-driven Document Ranking for Systematic Reviews: A Reproducibility Study. 12 2021. URL https://arxiv.org/abs/2112.04090v1.

[275] Shuai Wang, Harrisen Scells, Justin Clark, Bevan Koopman, and Guido Zuccon. From little things big things grow: A collection with seed studies for medical systematic review literature search. *arXiv preprint arXiv:2204.03096*, 2022.

[276] Shuai Wang, Hang Li, and Guido Zuccon. Mesh suggester: A library and system for mesh term suggestion for systematic review boolean query construction. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, pages 1176–1179, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394079. doi: 10.1145/3539597.3573025. URL https://doi.org/10.1145/3539597.3573025.

[277] Shuai Wang, Harrisen Scells, Bevan Koopman, and Guido Zuccon. Can ChatGPT write a good boolean query for systematic review literature search? *arXiv preprint arXiv:2302.03495*, 2023.

[278] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.340. URL https://aclanthology.org/2022.emnlp-main.340.

[279] Zifeng Wang and Jimeng Sun. Trial2vec: Zero-shot clinical trial document similarity search using self-supervision. *arXiv preprint arXiv:2206.14719*, 2022.

[280] Leon Weber, Mario Sänger, Jannes Münchmeyer, Maryam Habibi, Ulf Leser, and Alan Akbik. Hunflair: an easy-to-use tool for state-of-the-art biomedical named entity recognition. *Bioinformatics*, 37(17):2792–2794, 2021.

[281] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

[282] Michel Wijkstra, Timo Lek, Tobias Kuhn, Kasper Welbers, and Mickey Steijaert. Living literature reviews. In *Proceedings of the 11th on Knowledge Capture Conference*, pages 241–248, 2021.

[283] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

[284] World Health Organization. Rolling updates on coronavirus disease (COVID-19), 2020.

[285] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018.

[286] Hanwei Xu, Yujun Chen, Yulun Du, Nan Shao, Yanggang Wang, Haiyu Li, and Zhilin Yang. Zeroprompt: Scaling prompt-based pretraining to 1,000 tasks improves zero-shot generalization. *arXiv preprint arXiv:2201.06910*, 2022.

[287] Leiming Yan, Yuhui Zheng, and Jie Cao. Few-shot learning for short text classification. *Multimedia Tools and Applications*, 77(22):29799–29810, 2018.

[288] Eugene Yang and David D Lewis. TARexp: A Python Framework for Technology-Assisted Review Experiments. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3256–3261, 2022.

[289] Eugene Yang, Sean MacAvaney, David D Lewis, and Ophir Frieder. Goldilocks: Just-right tuning of BERT for technology-assisted review. In *European Conference on Information Retrieval*, pages 502–517. Springer, 2022.

[290] Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. Pretrained transformers for text ranking: BERT and beyond. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 1154–1156, 2021.

[291] Hye Sun Yun, Iain J Marshall, Thomas Trikalinos, and Byron C Wallace. Appraising the potential uses and harms of LLMs for medical systematic reviews. *arXiv preprint arXiv:2305.11828*, 2023.

[292] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297, 2020.

[293] Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhifang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, Hongying Zan, Kunli Zhang, Buzhou Tang, and Qingcai Chen. CBLUE: A Chinese biomedical language understanding evaluation benchmark. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7888–7915, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.544. URL https://aclanthology.org/2022.acl-long.544.

[294] Jie Zou and Evangelos Kanoulas. Towards question-based high-recall information retrieval: Locating the last few relevant documents for technology-assisted reviews. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–35, 2020.

[295] Jie Zou, Dan Li, and Evangelos Kanoulas. Technology assisted reviews: Finding the last few relevant documents by asking yes/no questions to reviewers. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 949–952, 2018.

[296] Guido Zuccon. Understandability biased evaluation for information retrieval. In *Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016. Proceedings 38*, pages 280–292. Springer, 2016.