Natural Language Processing and Deep Learning Approaches for Systematic Review (Semi-)Automation

Dhrangadhariya, Anjani Kiritbhai

UNIVERSITÉ DE GENÈVE
Département d'informatique

FACULTÉ DES SCIENCES
Professeur Dr. Stéphane Marchand–Maillet

Service de radiologie

FACULTÉ DE MÉDECINE
Professeur Dr. Henning Müller

Natural Language Processing and Deep Learning Approaches for Systematic Review (Semi-)Automation

THÈSE

présentée à la Faculté des sciences de l'Université de Genève
pour obtenir le grade de Docteur ès sciences, mention informatique

par

Anjani Kiritbhai Dhrangadhariya

de
Jamnagar, Gujarat (India)

Thèse N° 5808

GENÈVE
2024

UNIVERSITÉ DE GENÈVE

FACULTÉ DES SCIENCES

DOCTORAT ÈS SCIENCES, MENTION INFORMATIQUE

Thèse de Madame Anjani Kiritbhai DHRANGADHARIYA

intitulée :

«Natural Language Processing and Deep Learning Approaches for Systematic Review (Semi-)Automation»

La Faculté des sciences, sur le préavis de

Monsieur S. MARCHAND-MAILLET, professeur associé et directeur de thèse
Département d'informatique

Monsieur H. MÜLLER, professeur titulaire et codirecteur de thèse
Département de radiologie et informatique médicale, Faculté de médecine

Madame D. DEMNER-FUSHMAN, docteure
National Library of Medicine, Maryland, United States

Monsieur R. HILFIKER, professeur ordinaire
Institute of Higher Education and Research in Healthcare-IUFRS,
University of Lausanne, Lausanne, Suisse

autorise l'impression de la présente thèse, sans exprimer d'opinion sur les propositions qui y sont énoncées.

Genève, le 9 avril 2024

Thèse - 5808 -

La Doyenne

N.B. - La thèse doit porter la déclaration précédente et remplir les conditions énumérées dans les "Informations relatives aux thèses de doctorat à l'Université de Genève".

Contents

Abstract

Advances in natural language processing are the talk of the town, yet these advances have not materialized into their widespread adoption into systematic review automation. Systematic reviews are resource-consuming and multifaceted processes and could cost anywhere between USD $16 - 18$ million per year for companies and research institutions. The process encompasses searching and retrieving all possible evidence to write the review, meticulously filtering studies to find relevant ones, critically appraising it for biases, extracting and collating data from these studies, performing statistical analysis and writing manuscripts. Automation is imperative to reduce workload and cut down the review cost.

Automatic citation screening methods have been suggested to reduce the initial study filtering workload, but their uptake in commercial settings has been limited due to discrepancies between existing approaches and real-world workflows. Methods in automatic information extraction could aid in chaffing multiple data types from studies. These methods, however, are limited by static hand-labelled datasets and varying data extraction needs depending on the review question. Finding a manually annotated dataset covering all necessary data types is impractical. Approaches for cheaply extending static, manually annotated datasets with new information types are necessary. Critical appraisal, particularly the bias assessment process, is the most intellectually demanding phase of review writing. The scarcity of hand-labelled datasets essential for evaluating the NLP techniques hinders their adoption into SR automation.

In this thesis we have explored methods for automation of these three stages: automatic citation screening, data extraction and bias assessment. To address the research gap in prospective methods for citation screening, we explore active citation screening methods designed and evaluated for future-facing prospective scenarios aligned with industrial processes. Additionally, we adapt and develop weak supervision methodologies to obtain labels for varied data types necessary for the data extraction stage economically. Finally, we develop a resource for evaluating state-of-the-art NLP approaches for bias assessment and provide preliminary results of language model evaluation for the resource developed. Automated methods offer the potential to make the systematic review processes cheaper, more transparent, accountable, and reproducible.

Keywords: Natural language processing, deep learning, systematic reviews, automation

Résumé

Les progrès en matière de traitement du langage naturel font couler beaucoup d'encre, mais ils ne se sont pas traduits par une adoption généralisée de l'automatisation des examens systématiques. Les examens systématiques sont des processus à multiples facettes qui consomment des ressources et peuvent coûter entre 16 et 18 millions de dollars par an aux entreprises et aux instituts de recherche. Le processus comprend la recherche et l'extraction de toutes les preuves possibles pour rédiger l'examen, le filtrage méticuleux des études pour trouver celles qui sont pertinentes, l'évaluation critique des biais, l'extraction et le rassemblement des données de ces études, la réalisation d'une analyse statistique et la rédaction de manuscrits. L'automatisation est impérative pour réduire la charge de travail et le coût de la révision.

Des méthodes de sélection automatique des citations ont été proposées pour réduire la charge de travail initiale de filtrage des études, mais leur adoption dans des contextes commerciaux a été limitée en raison des divergences entre les approches existantes et les flux de travail dans le monde réel. Les méthodes d'extraction automatique d'informations pourraient faciliter le tri de plusieurs types de données provenant d'études. Ces méthodes sont toutefois limitées par des ensembles de données statiques étiquetés à la main et par des besoins d'extraction de données qui varient en fonction de la question examinée. Il n'est pas pratique de trouver un ensemble de données annotées manuellement couvrant tous les types de données nécessaires. Des approches permettant d'étendre à peu de frais les ensembles de données statiques et annotés manuellement avec de nouveaux types d'informations sont nécessaires. L'évaluation critique, en particulier le processus d'évaluation des biais, est la phase la plus exigeante sur le plan intellectuel de la rédaction d'une revue. La rareté des ensembles de données étiquetés manuellement, essentiels à l'évaluation des techniques NLP, entrave leur adoption dans l'automatisation de la RS.

Dans cette thèse, nous avons exploré des méthodes d'automatisation de ces trois étapes : le filtrage automatique des citations, l'extraction des données et l'évaluation des biais. Pour combler les lacunes de la recherche en matière de méthodes prospectives de sélection des citations, nous explorons des méthodes actives de sélection des citations conçues et évaluées pour des scénarios prospectifs orientés vers l'avenir et alignés sur des processus industriels. En outre, nous adaptons et développons des méthodologies de supervision faible afin d'obtenir des étiquettes pour divers types de données nécessaires à l'étape d'extraction des données de manière économique. Enfin, nous développons une ressource pour évaluer les approches NLP de pointe pour l'évaluation des biais et fournissons des résultats préliminaires de l'évaluation des modèles de langage pour la ressource développée. Les méthodes automatisées offrent la possibilité de rendre les processus d'examen systématique moins coûteux, plus transparents, plus responsables et plus reproductibles.

Mots clés: Traitement du langage naturel, apprentissage profond, revues systématiques, automatisation[1]

[1]This abstract was translated by a non-native speaker (AKD) using an automated deep neural machine translator (`https://www.deepl.com/translator`). The translated abstract was checked for improvements by Mr. Adrien Bertaud, native speaker of French.

સાર

નેચરલ લેંગ્વેજ પ્રોસેસિંગમાં આવતી આધુનિકતા એ નગરચર્ચાનો વિષય છે, છતાં સિસ્ટમેટિક રિવ્યુ ઓટોમેશનમાં આ આધુનિકતાની વ્યાપક સ્વીકૃતિ સાકાર થઈ નથી. સિસ્ટમેટિક રિવ્યુ એ સંસાધન-ભોગી અને બહુપક્ષીય પ્રક્રિયાઓ છે અને કંપનીઓ અને સંશોધન સંસ્થાઓ માટે પ્રતિ-વર્ષ 16 થી 18 મિલિયન USD (અમેરિકી ડોલર) ની વચ્ચે લાગત આવી શકે છે. આ પ્રક્રિયા સમીક્ષા લેખન વિષયક તમામ સંભવિત પ્રમાણો(પ્રાથમિક અધ્યયનો) શોધવા અને પુનઃપ્રાપ્ત કરવા, સુસંગત પ્રમાણો શોધવા માટે પુનઃપ્રાપ્ત અધ્યયનનું સૂક્ષ્મતાપૂર્ણ નિસ્પંદન કરવા, પૂર્વગ્રહો જાણવા માટે તેઓનું વિવેચનાત્મક મૂલ્યાંકન કરવા, આ અધ્યયનોમાંથી માહિતી નિષ્કર્ષણ અને એકત્રિત કરવા, આંકડાકીય વિશ્લેષણ કરવા અને હસ્તપ્રતો લખવાનો સમાવેશ થાય છે. કાર્યભાર ઘટાડવા અને સમીક્ષા શુલ્ક ઘટાડવા માટે ઓટોમેશન આવશ્યક છે.

અધ્યયનનો નિસ્પંદન કાર્યભારને ઘટાડવા માટે ઓટોમેટિક સાઈટેશન સ્ક્રિનિંગ પદ્ધતિઓનું સૂચન કરવામાં આવેલ છે, પરંતુ વર્તમાન અભિગમો અને વાસ્તવિક-વિશ્વના કાર્યપ્રવાહ વચ્ચેની વિસંગતતાને કારણે વ્યવસાયિક પ્રણાલીઓમાં તેમનો ઉપયોગ સીમિત છે. ઓટોમેટિક ઇન્ફોર્મેશન એક્સ્ટ્રેક્શનની પદ્ધતિઓ અધ્યયનોમાંથી બહુવિધ માહિતી પ્રકારોની છંટણી કરવામાં સહાયક બની શકે છે. આ પદ્ધતિઓ, જોકે, સ્ટેટિક હેન્ડ-લેબલ ધરાવતા ડેટાસેટ્સ અને સમીક્ષા પ્રશ્નના આધારે વિવિધ માહિતી નિષ્કર્ષણ આવશ્યક્તાઓના કારણે સીમિત છે. તમામ આવશ્યક માહિતી પ્રકારો સમાવિષ્ટ કરતો મેન્યુઅલી એનોટેડ ડેટાસેટ શોધવો અવ્યવહારિક છે. નવા માહિતી પ્રકારો સાથેના અભિગમો સ્ટેટિક, મેન્યુઅલી એનોટેડ ડેટાસેટ્સના શુલ્ક-પ્રભાવી વિસ્તારણ માટે આવશ્યક છે. નિર્ણાયક મૂલ્યાંકન, વિશેષતઃ પૂર્વગ્રહ આંકલન પ્રક્રિયા, સમીક્ષા લેખનના સૌથી બૌદ્ધિક આવશ્યકતાની માંગ ધરાવતા ચરણ છે. NLP તકનીકોનું મૂલ્યાંકન કરવા માટે આવશ્યક હેન્ડ-લેબલવાળા ડેટાસેટ્સની ઉણપ સિસ્ટમૅટિક રિવ્યૂ ઓટોમેશનમાં તેમના અનુકૂલનને અવરોધે છે.

આ થીસીસમાં, અમે આ ત્રણ ચરણો માટેની ઓટોમેશન પદ્ધતિઓ પર અન્વેષણ કર્યું છે: ઓટોમેટિક સાઈટેશન સ્ક્રીનીંગ, ડેટા નિષ્કર્ષણ અને પૂર્વગ્રહ આંકલન. સાઈટેશન સ્ક્રિનિંગ માટેની ભવિષ્યલક્ષી પદ્ધતિઓમાં સંશોધનાત્મક અંતર ચિહ્નિત કરવા માટે, અમે ઔદ્યોગિક પ્રક્રિયાઓ સાથે સંરેખિત ભવિષ્યના સંભવિત પરિદૃશ્યો હેતુ રચિત અને આંકલિત એક્ટિવ સાઈટેશન સ્ક્રિનિંગ પદ્ધતિઓનું અન્વેષણ કરીએ છીએ. વધુમાં, અમે આર્થિક રીતે માહિતી નિષ્કર્ષણના ચરણ માટે આવશ્યક વિવિધ માહિતી પ્રકારો માટે લેબલ્સ મેળવવા માટે નિર્બળ નિરીક્ષણ પદ્ધતિ અનુકૂલિત અને વિકસિત કરી છીએ. અંતે, અમે પૂર્વગ્રહ આંકલન માટે અદ્યતન NLP અભિગમોનું મૂલ્યાંકન કરવા માટે એક સંસાધન વિકસિત કરીએ છીએ અને વિકસિત સંસાધન માટે લેંગ્વેજ મોડેલ ઇવેલ્યુએશનના પ્રારંભિક પરિણામો પ્રદાન કરીએ છીએ. ઓટોમેટેડ પદ્ધતિઓ પદ્ધતિસર સમીક્ષા પ્રક્રિયાઓને શુલ્ક-પ્રભાવી, અધિક પારદર્શક, ઉત્તરદાયી અને પુનઃઉત્પાદન સંબંધિત ક્ષમતા પ્રદાન કરે છે.

કીવર્ડ્સ: નેચરલ લેંગ્વેજ પ્રોસેસિંગ, ડીપ લર્નિંગ, સિસ્ટમૅટિક રિવ્યૂ, અટોમેશન[2]

[2]This abstract was translated by a native speaker (AKD) of Gujarati using Google Translate (`https://translate.google.com/`). The translated text was further edited by them for more natural sounding Gujarati translation.

Acknowledgments

First and foremost, I want to extend my heartfelt gratitude to Prof. Dr. Henning Müller and Prof. Dr. Stéphane Marchand-Maillet for allowing me the opportunity to undertake this thesis under their guidance. I am especially grateful to Henning, who has afforded me the freedom to expand beyond my comfort zone, explore new areas, and take ownership of my academic decisions. This has shaped me up into an independent scientist capable of not only conducting my own research but also mentoring others.

I would like to express my gratitude to Dr. Roger Hilfiker, who proposed the thesis topic and played a crucial role throughout my PhD. Thank you, Roger, for engaging in constructive discussions regarding the bias assessment project and beyond, which was instrumental in writing this thesis. It was both, Roger and Henning, whose support was crucial to the conception and funding of this Ph.D. project for which I will be immensely grateful to them.

Thanks to the MedGIFT group members, my colleagues from HES-SO, and everyone with whom I have shared the lunch and coffee hours, the aperos, hikes and many fun events. Special acknowledgement goes to Oscar, Sebas, Adrien Bertaud, Roger Schaer, Anastasiia, Gaetano, Manu, Amjad, Laura, Vatsal, Cristina, Paul, Niccolò, Ivan, Manfredo, Marek, Matteo, Adrien D., Maria, Orfeas, Nur, Cloé, Valentin, Mara, Davide, and Vincent. I must thank Anne, Claudia, Anne-Cristine, Katia and Dragana for taking care of all the administration.

I would like to thank my collaborators from the health institute, HES-SO, Valais-Wallis, Dr. Martin Sattelmayer, PhD students Katia Giacomino and Rahel Caliesch for their contribution to my understanding of bias in clinical trials and their contribution throughout the bias assessment project. I am grateful to Prof. Dr. Nona Naderi. Nona, your guidance and support significantly accelerated my decision-making process in the bias assessment project. I thank my collaborators from F. Hoffmann-La Roche AG, Dr Seye Abogunrin, Dr Andreas Witzmann, and Marie Lane, for their unwavering support during my last PhD project and for warmly welcoming me to Roche.

Shout out to my mentor, Dr. Mathew Divine, for his invaluable guidance about academia to industry transition. I also thank my former teacher, Prof. Dr. John J. Georrge.

Harsh, my dearest friend, I am deeply grateful to you for laying the foundation for this PhD journey and always being there to lend me an ear. I also thank my long-time friend, Dipali Bhatia, for her unwavering support throughout the years, even when I was a lost cause friend at times.

My time in Switzerland was wonderful thanks to my friends and Isha meditators; Dr. Nilam Jani, Dr. Sophie Cartier, and Murali Alluri. Nilam akka, thank you for always welcoming me to your home and making me feel at home. Sophie, your presence has brought healing into my life every time we've met. Thank you for that. Murali anna,

thank you for organizing Pournami poojas and graciously opening your home to fellow meditators.

I am at a loss for words to thank my family members who have been with me through all my ups and downs. I want to dedicate this thesis to my late father, Kirit Dhrangadhariya, who had so selflessly thrown his life away to see me live my dreams. His unwavering belief in me continues to inspire me every day. My dear mother, Kokila Dhrangadhariya, words fail to convey my gratitude to you. "હતો હું સુતો પારણે પુત્ર નાનો, રડું છેક તો રાખતું કોણ છાનો ? મને દુ:ખી દેખી દુ:ખી કોણ થાતું ? મહા હેતવાળી દયાળી જ મા તું." I also want to express my gratitude to my beloved siblings, sister Purvi and brother Meet. Purvi, your artistic perspective has brought a unique gentleness into my life. Meet, your smile has been a source of comfort and encouragement.

Embarking on my PhD journey coincided with the onset of the COVID-19 pandemic, bringing uncertainties as an expatriate far from my home country. Thanks to the meditation practices designed by my eternal Guru, Sadhguru, that helped guide my boat through this turbulent river. I am forever indebted to you. He rightly says, "Gratitude is not an attitude. Gratitude is something that flows out of you when you are overwhelmed by what has been given to you."

Chapter 1

Introduction

Anecdotal evidence is information or opinion or evidence based on individual stories or personal experiences [253]. In medical and health domains, such evidence stems from people's personal experiences and interactions with society and media. Human society embraces such anecdotes because storytelling is an innate ability of ours, and it is easy to comprehend anecdotes but difficult to comprehend scientific evidence. We often encounter health-related anecdotes in our daily lives, like seeking advice from a friend on ointments for bruises or burns or looking for relaxation exercises for knee pain or mobility issues. As represented in Figure 1.1, while these anecdotes may seem reliable, there is no evidence regarding whether an ointment is suitable for a skin type or whether a relaxation exercise will address the underlying cause of someone's knee pain. Relying on anecdotal evidence is risky because it is usually based on word-of-mouth or hearsay but rarely on scientific evidence and reliable statistical data. For instance, asking a relative for mouth ulcer medication might not be the best approach, as it may not consider individual health conditions or specific needs. Anecdotal evidence in medicine and health is prone to bias and subjectivity, posing several problems when it comes to public understanding and decision-making [208].

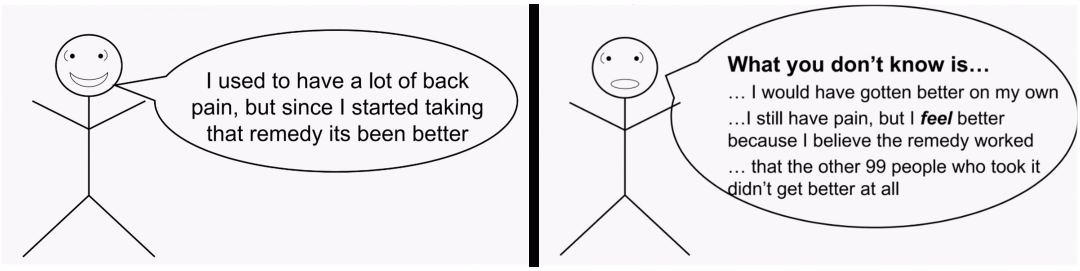


Figure 1.1: Cartoon presentation on anecdotal evidence in medicine. (Source: `https://www.slideshare.net/synchro85/cartoon-presentation-on-evidence`)

Evidence-based medicine (EBM), on the other hand, is an approach that emphasizes using the best available scientific evidence to make decisions about medical treatments, diagnosis and interventions. The term EBM was not coined in the literature until 1992, and decades before, scientists used data based on case series and anecdotal evidence [130]. In the early 1900s, physicians treated patients based on anecdotal evidence, such as using heroin as a cough suppressant for children. Until the late 19^{th} century, bloodletting was used to cure several ailments despite causing more harm than good in the vast majority

of cases [4]. With the emergence of modern scientific methods in the late 19th and early 20th centuries, there was a shift towards more systematic and controlled approaches to medical research. Randomized Controlled Trials (RCTs) began to be used to study the effects of medical interventions systematically. The milestone in this area came about in 1948 and was the use of streptomycin to treat tuberculosis, which was tested using principles of RCT methodology [57]. RCTs are clinical trials designed to compare treatment outcomes among patient groups while controlling for external factors, achieved through rigorous methods, for example, randomizing and blinding to allocating patients into the intervention groups under study [136]. As the amount of RCT literature increased (refer Figure 1.2), it became challenging for healthcare professionals to keep up with the vast amount of contrasting available evidence. In the 1970s and 1980s, researchers began to recognize the need for a method to synthesize and summarize the findings of multiple medical studies in a standardized way. In 1993, Cochrane collaboration[3] was established with 14 countries coming together to systematically review the evidence in the form of writing systematic reviews (SRs). This international network of researchers aimed to promote EBM and develop SR of RCTs across various medical interventions. SR methodology as it exists now is a collaborative effort from Cochrane collaboration in the United Kingdom, which formed a committee of 14 countries to write SRs [56]. Before that, the reviews were often narrative and lacked a structured approach to selecting, appraising, and synthesizing evidence.

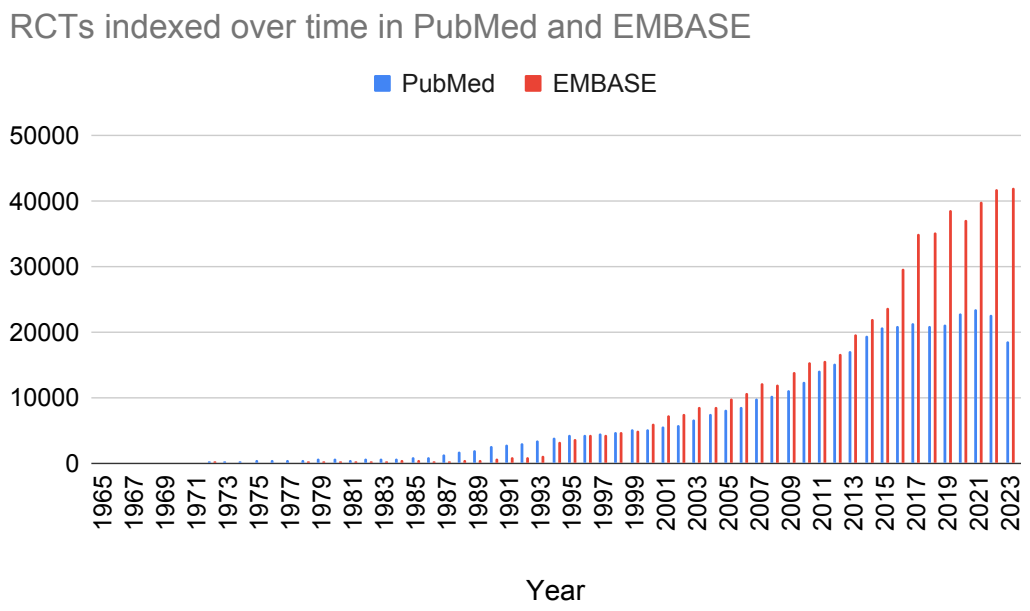


Figure 1.2: The exponential growth of RCTs indexed in PubMed and EMBASE over time. Illustration made using statistics from these literature repositories.

Systematic reviews utilize a systematic, rigorous and transparent methodology to collate all the empirical evidence in the form of primary studies that fit pre-specified eligibility criteria in order to answer a specific research question. They hold the top-most position in the evidence pyramid as shown in Figure 1.3 and are considered the most reliable form

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

of medical evidence when synthesized from primary studies like RCTs [147]. In areas where a substantial number of RCTs are lacking, Systematic Reviews (SRs) may incorporate various study types, such as non-randomized controlled trials, observational studies, quasi-experimental studies, diagnostic accuracy studies, and case series and reports [197]. As depicted in the figure 1.4, writing a systematic review begins by formulating a concrete research question the researcher wants to address using a framework called PICO (Participant, Intervention, Comparator, Outcome) [234]. Subsequently, a thorough search of relevant literature sources is performed to identify scientific studies addressing the question. A manual assessment of searched studies is conducted following the study search to ensure they meet the reviews' inclusion criteria. Data from included studies is then collected and recorded. The next step is to evaluate study quality to assess potential biases in the included studies. Finally, the results, including a detailed description of the methodology, are summarized and reported [146].

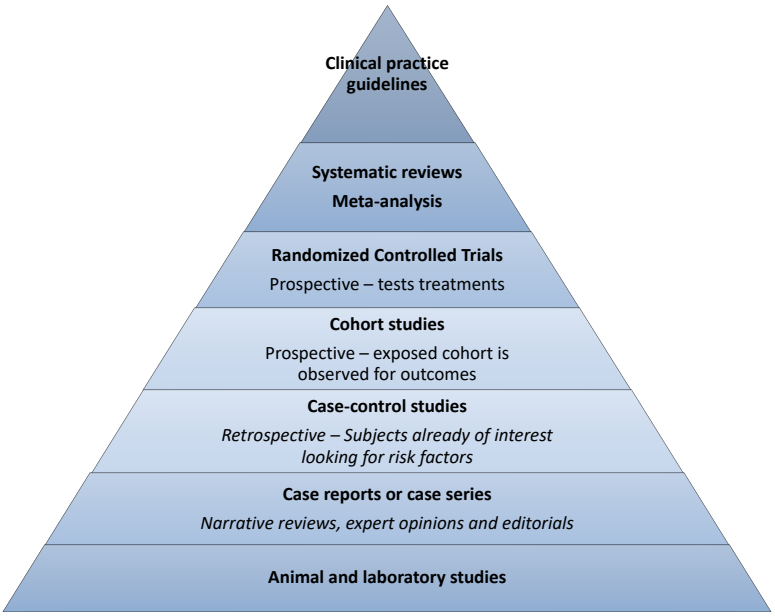


Figure 1.3: The evidence pyramid in evidence-based medicine. Adapted from [181]

Systematic reviews are the pillars of EBM and are the go-to documents for health practitioners. They help healthcare practitioners stay informed with the latest evidence, assess intervention effectiveness, and evaluate treatment impacts on patient outcomes, thereby enabling them to make better treatment and diagnosis decisions [69, 186, 205, 212]. These reviews, when combined with polemic discourse[4] are finding growing acceptance among clinicians and policymakers [109, 327]. They have also been pivotal in informing policy decisions by evaluating the effectiveness and efficiency of healthcare interventions [126, 145].

Writing a comprehensive SR takes about 6 to 24 months to complete, with a timeframe depending upon the scope of the SR question [25, 133]. However, the delay between conducting, writing and publishing an SR and the constant influx of new clinical studies necessitate regular updates to maintain their relevance. Conventional SRs are static artefacts, facing the challenge of manual updates. In contrast, living systematic reviews (LSRs),

[4]The term "polemic discourse" refers to a specific form of discourse characterized by its contentious and aggressive nature, often involving strong arguments and refutations. It is a type of communication that aims to establish a position by refuting or undermining opposing views.

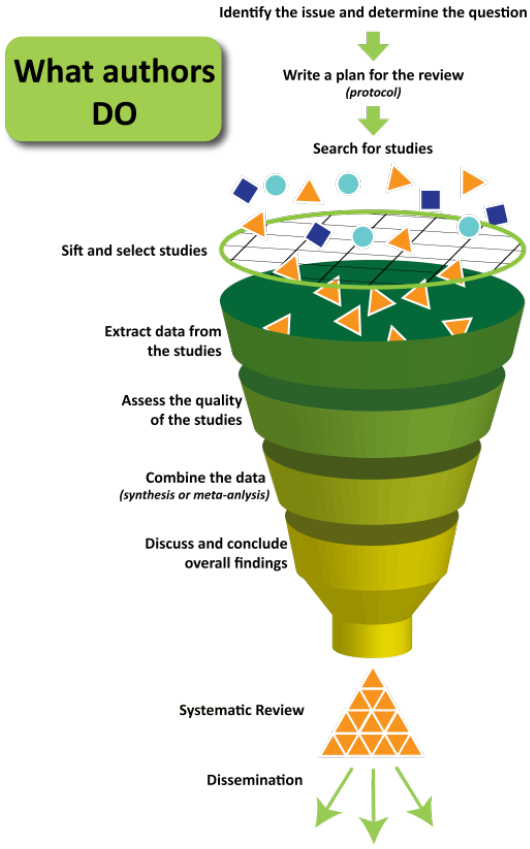


Figure 1.4: Illustration: Steps for conducting SRs (Source: `https://acesse.dev/6tlkq`)

relatively recent developments in the area, are an evolving approach to SRs in healthcare and other allied domains [100, 233]. Unlike static SRs conducted at a single point in time, LSRs are dynamic and continually updated as new studies become available. They are typically defined based on their ability to incorporate the latest research and adapt to the changes in the domain. This continuous updating ensures that the review remains a relevant and valuable resource for healthcare practitioners, policymakers, and researchers who need the most up-to-date information to inform their decisions and actions. The traditional SR and LSR remain costly, given the financial and manual resources required for conducting them. The volume of clinical studies published is growing beyond the capacity of manual tracking and reviewing, which necessitates automation. It is estimated that a single SR could cost up to $141,194.80. With academic institutions publishing at least 132 reviews a year and pharmaceutical companies publishing an average of 118.71 reviews per year, the cost could go up to $18,660,304.77 and $16,761,234.71, respectively [229]. These reasons warrant automation for conducting SRs, aiming to approach the realm of LSRs.

Writing SRs is a tedious, protracted process comprising several intricate sub-stages. In the initial stage of manually assessing the retrieved studies, also called citation screening, two reviewers must meticulously check hundreds of thousands of studies to determine their relevance to the SR, deciding whether to include them. Whenever there is a disagreement between the reviewers on making the inclusion decision for a study, a third reviewer is needed to resolve their conflict. Manual assessment for a single SR could take weeks to months depending on certain factors [332]. An important aspect of manual assessment is

checking for the inclusion criteria defined by the PICO framework, where the reviewers need to manually check for PICO information in the studies being screened [234]. It is important to note that in addition to the PICO information, the reviewers might need to control for additional important information like the type of study, study design, timeframe of the intervention, exposure, geographical location, setting (e.g., in-patient, out-patient, community care), *etc.* to ensure that the study is relevant to writing the review [32, 108, 228]. Quality assessment, also called the RoB (Risk of Bias) assessment stage, requires two reviewers[5] to go through the study full-texts carefully and comprehensively, looking for possible biases or deviations. The demand for systematic reviews is rising significantly, with Hoffmann *et al.* reporting a 20-fold increase in daily SR publications from 4 in 2000 to 80 in 2019 [153]. While the demand for SRs is increasing, the volume of published evidence in the form of primary studies is growing at break-neck pace. Taking into account the above-explained tedious process any automation tool that could help the reviewers quickly identify relevant studies, and any automation tool that could hint towards the RoB information from a study could help aid in judging the study quality, thereby reducing their burden.

In recent years, with rapid advancements in natural language processing (NLP), machine learning and the availability of labelled data resources, the research area of SR automation has also gained traction. The access and availability of the citation screening datasets opened up the avenues for SR automation. Cohen *et al.* were the ones to approach the automation of citation screening sub-task [66]. In addition, they contributed a data resource comprising 15 citation screening datasets labelled with the binary inclusion and exclusion decisions. Other than the Cohen dataset, the other citation screening datasets are scattered across the web. SYNERGY[6] initiative has compiled 26 such citation screening datasets into a single resource that could act as a benchmark for automation technologies [76]. Open-access datasets are equally, if not more, crucial to automation than the automation technique itself. However, a critical obstacle to progress in this field remains — the shortage of freely available labelled datasets to facilitate the automation of PICO and additional information extraction and risk of bias (RoB) assessment. This scarcity of labelled datasets hinders the automation for RoB assessment subtask [84, 90].

The research and development in SR automation has its roots in text mining. In its early stages, researchers devised methods to index and search large collections of scientific studies, leading to the development of large online repositories like PubMed and EMBASE. These repositories could be effortlessly searched using a combination of natural language queries and keywords to search for scientific studies of interest. However, the research in this direction eventually gained traction thanks to the breakthroughs in deep natural language processing. A decade ago, the machine learning (ML) models were trained using manually engineered text features like bag-of-words (BoW) and weighted BoW (tf-idf: term frequency-inverse document frequency) that took into account word count information and their importance but disregarded word order [21]. Later developments led to the release of dense word and paragraph embeddings taking into account word order, and the development of transformer models trained in an unsupervised fashion pushed the boundaries further [81, 231]. With the increase in hardware capacity and the advent of humongous large language models (LLMs), most notably GPT-3 (Generative Pretrained Transformers), the field of SR automation will see further breakthroughs and innovations [42].

[5]according to Cochrane, two reviewers are required

[6]`https://github.com/asreview/synergy-dataset`

1.1 Motivation

Successful design and application of NLP methodologies and frameworks for automation of SRs faces three major challenges: I) The first challenge revolves around the absence of citation screening approaches designed with real-world or "prospective" scenario in mind. To contextualize, every *de-novo* SR will come with unlabelled studies and would challenge the current "restrospective" automation techniques that consider citation screening task as fully-supervised binary classification. A fully-supervised system requires large labelled datasets which are missing for *de-novo* SR scenario. Businesses, academia, and hospitals find it impractical to label a large number of studies, necessitating a "prospective" approach. Designing such "prospective" systems typically ensures proper evaluation, unlike a "restrospective" scenario. II) The second challenge concerns automatic extraction of clinical information, especially the PICO information, from the retrieved studies. Even though a medium-sized PICO labelled dataset is available to train and evaluate NLP approaches, in the real-world scenario there is often a need to control for more information than just PICO information as described in the Section 1 [254]. Unfortunately, no datasets account for all the required new information aspects, and manual re-annotation of existing datasets is impractical due to resource constraints within the wider scientific community. In addition, the currently available benchmark for PICO information extraction (IE) is error-prone [90]. III) The third challenge relates to the lack of open-access datasets for training and evaluating NLP systems focused on extracting risk of bias information from RCTs. Currently, there are no datasets containing RCTs annotated with RoB information, making it difficult to develop and evaluate such methods.

1.2 Thesis objectives

This thesis centers around a single goal - provide solutions to the barriers impeding implementation of NLP approaches to SR automation. Although there are many ways of approaching this overarching goal, this thesis primarily contributes through three significant objectives through which it can be addressed. The first objective is to address the gaps in design and evaluation considerations for real-world automatic citation screening systems. The second objective is to design and develop inexpensive approaches concerning the information need during the citation screening stage from manually reviewing PICO and additional information. The sub-task of RoB assessment automation is still deprived of freely-available dataset and impedes development and evaluation of automation approaches. Therefore, the third objective addresses this critical gap of lack of resources, by creating open-access dataset dedicated to training and validating the automation of RoB assessment.

1.3 Summary of contributions

The main scientific contributions of this research are at the confluence of natural language processing, deep learning, corpus development and clinical sciences. I had and also actively developed the opportunities to collaborate with other researchers from the School of Health Sciences HES-SO Valais-Wallis, HEG - Genève, F. Hoffmann-La Roche AG, University of Bristol and other external researchers and be a part of seven scientific papers. The contributions are categorized as follows:

Citation Screening Automation *Chapter 3*: I explored and devised automation approaches for citation screening both in research and business domains. In [85], I evaluated the performance of two popular word embeddings by pretraining them on a large biomedical database. Subsequently, I designed an experimental framework to evaluate the efficacy of fine-tuning these biomedical embeddings for binary classification in citation screening. As outlined in [93], I introduced a prospective active learning approach tailored for commercial citation screening systems. To gain deeper insights into the system behavior and streamline decision-making, I advocated using multiple performance indicators as part of a multi-objective approach. In conjunction with inputs from the team, I was responsible for designing the experiments to ensure proper evaluation these performance indicators for decision making and shaping further research. This work also included developing a fuzzy deduplication approach that will be subsequently used to enhance the citation deduplication approach in production.

Information Extraction from Clinical Trials *Chapter 4*: I explored, proposed and extended methodologies for clinical entity extraction pertaining to SR automation. In [82], I proposed an end-to-end multitask neural network for extraction of fine-grained, semantic PICO information. In [89], I developed a novel distant supervision approach using a clinical trials knowledgebase to obtain a large pseudo-labelled dataset. The pseudo-labelled dataset was used to train an attention-based neural network model for "Intervention" information extraction. I also proposed a modified string metric for fuzzy mapping of information from the knowledgebase onto raw text. In [87], I proposed, adapted and evaluated a weak supervision approach for extraction of PICO information in absence of any labelled data. In the paper, I proposed decomposing the fuzzy entities like PICO to smaller units for effective application of weak supervision, while highlighting the need for a unified ontology for data extraction in clinical trials. In [87], I extended the approach in [90] to a new, composite "Study type and Design" entity with extended experiments and state-of-art in weakly-supervised PICOS information extraction. I introduced a simple algorithm in the paper for mapping ontology concepts to the target entity for extraction. For all the papers in this and the previous chapter, I was responsible for developing and / or adapting the methodologies, designing and executing the experiments, analysing the results and reporting the findings in form of manuscripts.

Risk of Bias Resource Development and LLM evaluation *Chapter 5*: I released visual RoB text annotation guidelines that are not only limited to corpus annotation but could also act as training material for novice bias assessors. My contribution in this project ranges from the problem identification and proposal, project planning, procuring the expert volunteers for corpus annotation, guiding the annotation instruction development and the annotation process, LLM evaluation and manuscript writing. Collaborating with Dr. Roger Hilfiker, we procured annotators for the project. In [83], I released a small corpus consisting of 10 RCT full-texts manually annotated using 22 RoB categories. Furthermore, in [84], I guided a team of five annotation experts and one linguistic expert. Together, we adapted the manual RoB assessment guidelines into RoB text annotation instructions, which were then transformed into visual placards. I was responsible for establishing the platform for annotating the corpus, supervised the annotation process, and assessed the performance of a large language model using this corpus. With this paper, I enabled development of a larger, extensively annotated RoB annotated corpus of full-text RCTs.

Complementary NLP projects: I had the privilege of collaborating with fellow researchers within my team and across different departments at the Informatics Institute

of HES-SO Valais-Wallis. This resulted into five scientific papers. In [211], I analysed German-language osteoarthritis patient notes and derived insights using text analysis. In [86], I developed a text mining approach using standardized keywords for filtering Diagnostic Light Microscopic Images (DLMI) to retrieve a set of rare cancer images. An overview of birds-eye goal of this approach was presented in [242]. In [92], I co-designed experiments for binary classification of free-text pathology reports into high gleason grade and low gleason grade class and explored interpretability methods to understand the classifier decisions. In [88], I automated a previously prototyped quality control approach for identifying incomplete radiology reports sourced from a Swiss radiology clinic.

An updated list with publications of the author and their citation impact can be accessed in the following Google Scholar URL: `https://scholar.google.com/citations?user=C4jUZ18AAAAJ`

1.4 Thesis organization

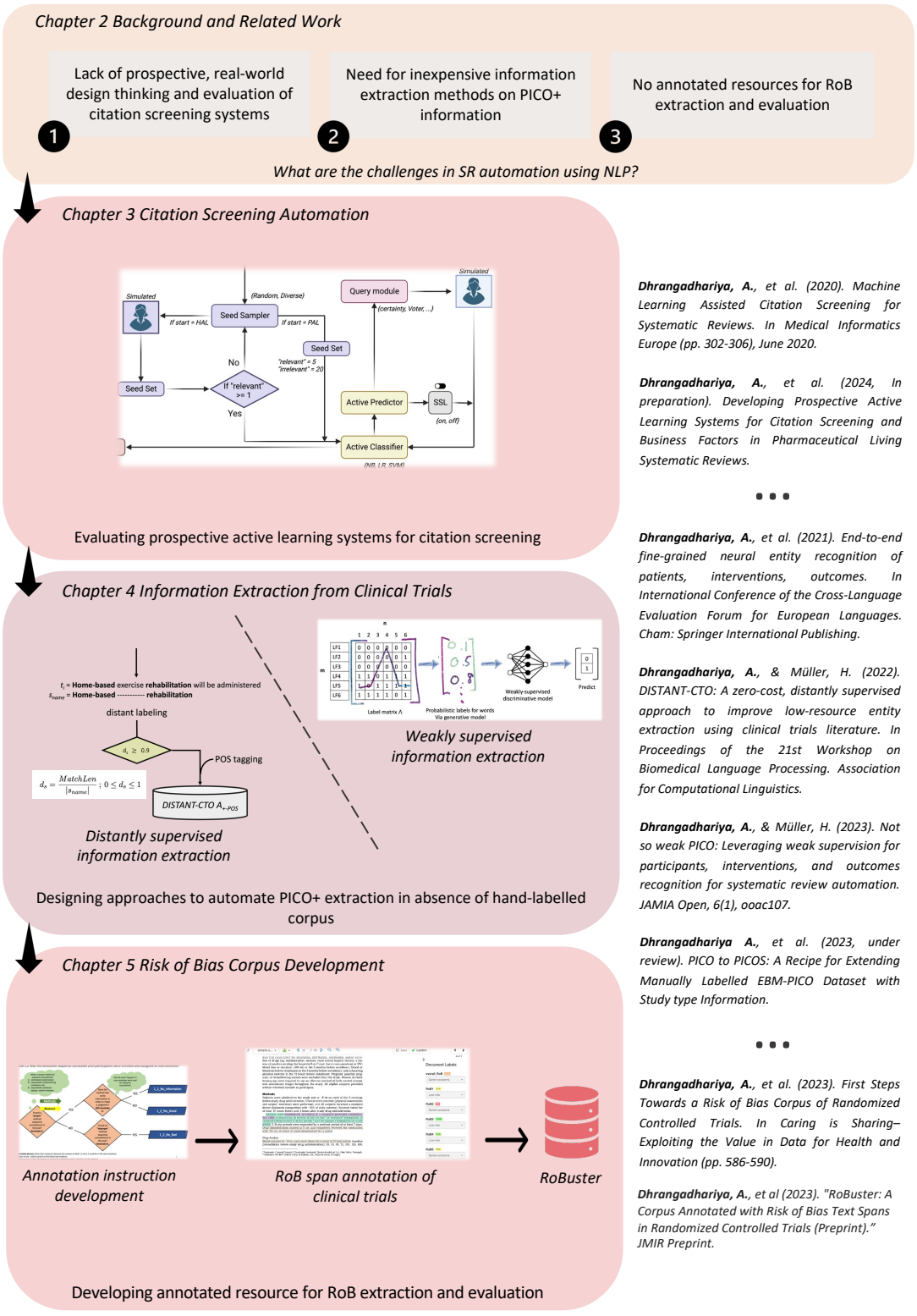
Figure 1.5 shows the overview for the organization of this thesis. In **Chapter 2**, the intricacies of SR sub-tasks and prior efforts to their automation are described in the context of up-to-date related works.

In **Chapter 3**, I begin by framing the citation screening task as a binary classification problem and explore it using classical machine learning approach. Then citing the challenges pertaining to adapting the automation approaches to real-world systems, I design a prospective active learning system and propose monitoring such systems using a comprehensive list of performance indicators. In addition, I explore a straight-forward pseudo-labelling method to reduce labelling costs for *de-novo* citation screening projects and advocate cautiousness in their adoption in real-world automation systems. My design approach to the prospective system have contributed to enhancement of specific modules for our collaborators.

The **Chapter 4** begins by first exploring the problem of fine-grained PICO information extraction through development of an end-to-end multi-task learning approach. Then citing the lack of resources, I develop a novel approach to obtain noisy, inexpensive labels for "Intervention" information extraction using clinical trials knowledgebase. The approach is extensible to the rest of the PICO entities too. Later I address the need for inexpensive information extraction for novel entities by developing a weak supervision approach for fuzzy and nested PICO entities. I demonstrate through additional experimentation that this approach could be efficiently extended to novel entities specifically the PICO + S "Study type and Design" information. These approaches successfully improve the state-of-the-art (SOTA) performance on the Participant and Intervention entities without the need for labelled datasets and establish a new benchmark on "Study type and Design" entity.

In **Chapter 5**, presents a new resource comprising 41 RCTs manually labelled by experts with risk of bias text descriptions. In this chapter, I proposed the development of corpus annotation guidelines and aided its development by leading a team of five annotation experts and a linguistic expert. I transformed these guidelines into visual placards with feedback and consistent review from the team.

Chapter 6 concludes this thesis with a discussion of the main results of the chapters in the thesis. In the end, promising research directions are proposed in light of the outstanding current developments in the field.



***Dhrangadhariya, A.**, et al. (2020). Machine Learning Assisted Citation Screening for Systematic Reviews. In Medical Informatics Europe (pp. 302-306), June 2020.*

***Dhrangadhariya, A.**, et al. (2024, In preparation). Developing Prospective Active Learning Systems for Citation Screening and Business Factors in Pharmaceutical Living Systematic Reviews.*

• • •

***Dhrangadhariya, A.**, et al. (2021). End-to-end fine-grained neural entity recognition of patients, interventions, outcomes. In International Conference of the Cross-Language Evaluation Forum for European Languages. Cham: Springer International Publishing.*

***Dhrangadhariya, A.**, & Müller, H. (2022). DISTANT-CTO: A zero-cost, distantly supervised approach to improve low-resource entity extraction using clinical trials literature. In Proceedings of the 21st Workshop on Biomedical Language Processing. Association for Computational Linguistics.*

***Dhrangadhariya, A.**, & Müller, H. (2023). Not so weak PICO: Leveraging weak supervision for participants, interventions, and outcomes recognition for systematic review automation. JAMIA Open, 6(1), ooac107.*

***Dhrangadhariya A.**, et al. (2023, under review). PICO to PICOS: A Recipe for Extending Manually Labelled EBM-PICO Dataset with Study type Information.*

• • •

***Dhrangadhariya, A.**, et al. (2023). First Steps Towards a Risk of Bias Corpus of Randomized Controlled Trials. In Caring is Sharing–Exploiting the Value in Data for Health and Innovation (pp. 586-590).*

***Dhrangadhariya, A.**, et al (2023). "RoBuster: A Corpus Annotated with Risk of Bias Text Spans in Randomized Controlled Trials (Preprint)." JMIR Preprint.*

Figure 1.5: Diagram containing the flow between the chapters in this thesis.

Chapter 2

Background and related work

2.1 Systematic Reviews

In health, biomedicine, pharmaceutical and allied domains, SRs are widely considered a gold evidence standard. They are known for their rigorous methodology, strict compliance, comprehensive literature searches, and thorough evaluation of the available evidence. These factors ensure bias minimization [144, 299]. Citing Cochrane Handbook, "A systematic review attempts to collate all empirical evidence that fits pre-specified eligibility criteria to answer a specific research question. It uses transparent, systematic methods that are selected to minimize bias, thus providing more reliable findings from which conclusions can be drawn and decisions made [324]. Compared to a generic literature review, which does not involve a comprehensive and systematic search of primary studies, an SR aims for an exhaustive, comprehensive search. A generic literature review may or may not include a quality assessment of the primary studies. In contrast, a systematic review conducts a quality assessment, which could determine the eligibility of a primary study [127]. The SR process typically involves the following detailed steps [146]:

1. **Question definition:** The SR writing process begins by formulating a precise research question that has not been previously addressed to avoid redundant efforts [149]. To identify whether the question has already been addressed in another review, a thorough investigation of published SRs and a check of the PROSPERO [7] register and Cochrane library is necessary. The question is formulated apriori to reduce the bias induced by observing the preliminary results during the review process. It specifically prevents HARKing or Hypothesizing After Results are Known, a malpractice involving modifying the review question or hypothesis based on the observed results, ultimately compromising the integrity of the review [218].

2. **Define inclusion criteria:** The systematic review question should clearly define its scope by defining eligibility or inclusion criteria. The reviews investigating interventions define their inclusion criteria using the PICO (**P**articipant, **I**ntervention, **C**omparator, **O**utcomes) framework. PICO framework (refer Table 2.1) includes defining the intervention[8] the review wants to investigate and an appropriate reference comparator, specify the patient demographics of interest, and delineating the treatment outcomes of interest that the SR in question should explore. Only

[7] `https://www.crd.york.ac.uk/prospero/`

[8] A treatment including drugs, surgery, physical exercises, etc.

those primary studies that match the target population, intervention/comparison, and outcomes specified by the review question will be eligible for answering the question. [106, 149, 225].

Element	Explanation
P	What specific participant or patient characteristics does the SR aim to investigate?
I	What treatment, intervention, or exposure does the SR intend to examine in relation to these participants?
C	What is the comparator or control group against which the effects of the investigative intervention are assessed?
O	What specific outcomes related to the intervention does the SR want to focus on?

Table 2.1: Explanation of Each Element in the PICO Framework.

3. **Determining the search databases:** The next step is to identify all the relevant databases indexing the studies relevant to the topic of the review question. To ensure a comprehensive and unbiased SR, the reviewers must search several pre-determined literature repositories [146]. These repositories include but are not limited to major literature databases, clinical trial repositories, and grey literature databases. Literature databases include PubMed, EMBASE, CINAHL (Cumulative Index to Nursing and Allied Health Literature), bioRxiv, medRxiv, DBLP (Digital Bibliography & Library Project), Google Scholar, and PsycINFO. Clinical trial repositories, including ClinicalTrials.gov, EudraCT (EU Clinical Trials Register), and WHO's ICTRP (International Clinical Trials Registration Platform), capture a range of evidence as well. Searching grey literature like `openGrey.eu` ensure the inclusion of negative results that are more likely to remain unpublished or published as grey literature [256]. Negative results are less attractive for publishing in peer-reviewed venues and are more likely to be found in grey literature. Relying solely on primary studies published in peer-reviewed venues for SRs may lead to overestimating intervention effects [295]. Therefore, it is paramount to search and include the evidence from grey literature [47, 312].

4. **Search query formulation and search:** Each literature mentioned above has its own search engine, metadata, indexing vocabulary (MeSH for PubMed vs. EMTREE for EMBASE vs. APA thesaurus for PsycINFO), and query syntax with rare interoperability. Therefore, the experts, usually the information specialists, manually formulate search queries for individual resources [328]. Search queries are formulated using the PICO inclusion criteria and adding language and study type restrictions [149]. Search queries are developed iteratively over several rounds of revisions. The aim is to formulate these queries for high sensitivity rather than specificity to ensure the inclusion of as much relevant evidence as possible [364, 367]. The final search queries used for the systematic search are published along with the SR protocol to ensure transparency [149]. After query formulation, comprehensive and systematic searches are conducted across the chosen literature repositories. The systematic search and retrieval step establishes the SR's inclusion criteria depending on the volume of studies retrieved. A broad research question may yield a substantial vol-

ume of studies, while a narrow more focused question may result in a more limited selection of relevant studies.

5. **Deduplication:** The studies are retrieved in the previous step from multiple overlapping data sources, leading to duplicates in the pool. The figure 2.1 shows the rough overlap between the different literature repositories [357]. Duplicates lead to either overestimating or underestimating treatment effects on patient outcomes [325, 342]. Additionally, reviewing the same studies multiple times wastes time and resources. Consequently, these are removed during the deduplication process.

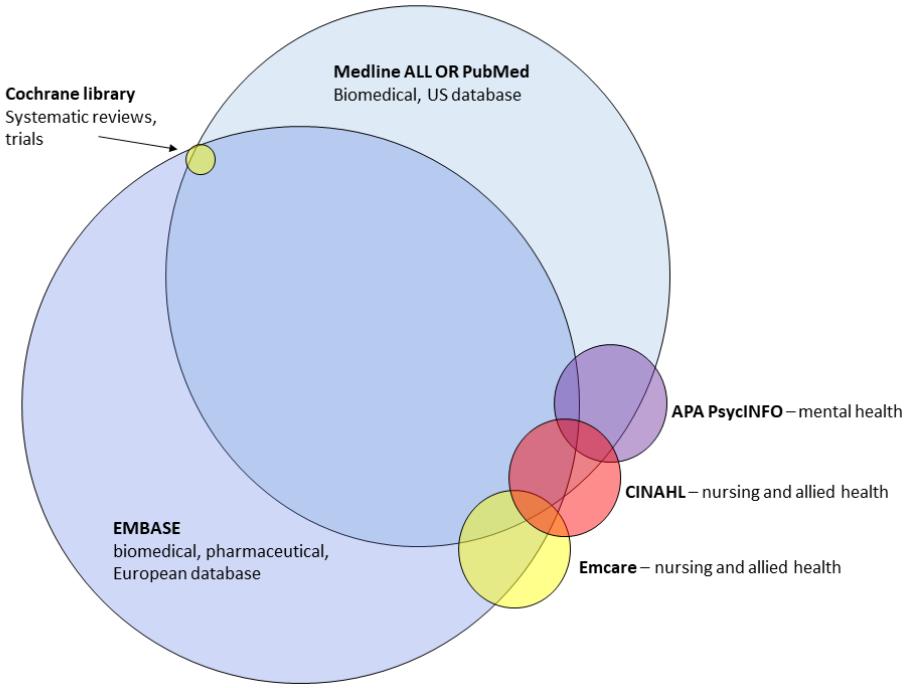


Figure 2.1: Schematic representation of overlap between a non-exhaustive list of literature databases used for literature search for SRs. Source: [357]

6. **Manual Citation screening:** Manual citation screening involves assessing retrieved studies to determine if they satisfy the pre-defined inclusion criteria. Citation screening is conducted in two steps: It begins with title and abstract screening and is followed by full-text screening. The first step involves manually looking through the title and abstract of the studies to ensure they align with the SR inclusion criteria, as in if the study addresses the relevant patient population, intervention/comparator, and treatment outcomes [234]. The inclusion criteria are defined using the PICO framework. The participants, intervention, and comparison often translate directly into inclusion criteria for an SR. If a study addresses the target population, intervention, comparator, and outcome measures specified by the SR question, it is considered relevant and included for writing the SR; otherwise, it is excluded.

7. **Data extraction:** Data extraction refers to identifying important characteristics of the included primary studies. Data of interest could include but is not limited to extracting information about study characteristics (authors, affiliations, publi-

cation date, funding sources), participant demographics such as patient condition and age, details of intervention under investigation, and data related to outcome measures like outcome type and measurement scale. Different quantitative data are extracted Depending on the analysis to perform. For example, in the case of network meta-analysis (NMA), quantitative dosage information for each treatment arm is systematically collected in distinct columns. Time series data associated with outcomes for multiple arms is then documented in separate rows for each arm [258]. Data extraction also involves extracting information regarding the study quality that could be utilized for study quality assessment [149]. The extracted data is recorded in spreadsheets, standard forms, or software like REDCap[9].

Example RoB assessment guidelines	Year
Physiotherapy Evidence Database (PEDro)	1999
Risk of Bias Assessment Tool for Nonrandomized Studies (RoBANS)	2004
Cochrane Risk of Bias assessment guidelines	2008
Risk of Bias in Non-randomized Studies of Interventions (ROBINS-I)	2016
Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2)	2017
Newcastle-Ottawa Scale (NOS)	2018
Revised Cochrane Risk of Bias for RCTs 2.0 tool (RoB 2)	2019

Table 2.2: "Evolution of Risk of Bias Assessment Guidelines Over Time"

8. **Quality assessment:** Primary studies, like clinical trials, aim to accurately measure intervention effects on participant outcomes, especially the RCTs. In theory, RCTs aim to minimize bias, but in practice, biases unintentionally affect any trial stage. When such studies with questionable biases are used to write SRs, they reduce the validity and utility of the review. Biases cannot be assessed from RCT studies, but the risk of bias can be estimated by identifying the systematic flaws in study design, planning, execution, assessment of outcomes, and reporting, among other relevant factors [313]. Quality assessment entails evaluating potential bias risk in the included studies using the risk of bias assessment tools available (refer Table 2.2). The step helps determine the reliability and validity of individual studies. Quality assessment results in each study were assessed and labelled with the identified level of bias risk, classified as low, high, or unclear. A schematic representation of the RoB assessment is shown in Figure 2.2. After quality assessment, the included studies could be grouped into different bias risk groups: low risk of bias, unclear risk of bias, and high risk of bias.

9. **Data synthesizing and Meta-analysis**: After extracting the data, the next step involves synthesizing it through statistical meta-analysis. This process includes pooling effect sizes or other pertinent data across studies, aiming to estimate the treatment or intervention effect or the association with participant outcomes. The data synthesis may involve emphasizing studies with a lower risk of bias when summarizing the evidence. Subgroup analysis could be performed to explore the impact of bias on the overall results. Common statistical techniques used in meta-analysis include fixed and random effects models [34].

[9] `https://www.project-redcap.org/`

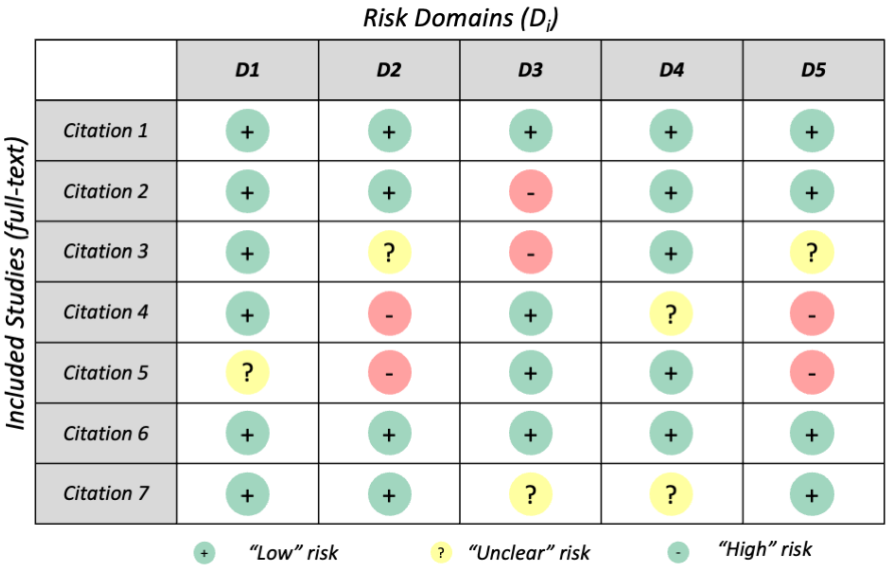


Figure 2.2: Schematic representation of the RoB assessment of n included studies (citations) over i risk domains D.

10. **Writing the SR:** Findings are presented as scientific publications that include detailed methodology to facilitate easy updates with new research findings [144, 324]. Meta-analysis findings are reported using established guidelines such as the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement along with the entire systematic procedure followed [240].

The anticipated workflow for conducting systematic reviews (SRs) follows a sequential process, as detailed above; however, convolutions and interconnections exist between these stages in practice. To put it into perspective, the deduplication process starts after search retrieval and continues until the data extraction stage. Although dedicated data extraction is shown to begin after citation screening, its preliminary stages commence when PICO information is extracted from the retrieved preliminary studies and continues through quality assessment. Different people could work on these disparate stages of writing the review. To elaborate, the review question is formulated through collaboration between reviewers and stakeholders, information specialists conduct query formulation, systematic search and retrieval by experts specializing in database systems, and very specialized reviewers focus on the risk of bias or quality assessment [252]. As a result, systematic reviews are typically conducted by a team of experts engaging in a collaborative activity that unfolds across a multi-stage and interconnected process [177].

The cost per systematic review could reach as high as 300'000 US dollars and exceed 1000 person-hours [33, 185, 266]. These reviews are methodologically rigorous, require considerable expertise and collaboration across institutes and thus could take 6 to 12 months to complete [133, 173]. By the time a systematic review is completed, it is no longer up-to-date with the current evidence given the time between its inception, planning and completion [26]. Consequently, there is a need to automate certain aspects of writing the review that are susceptible to automation.

2.2 Avenues for Automation

The stages of formulating the review question, determining the inclusion criteria, deciding upon the databases to retrieve primary studies from, and formulating the search queries for these databases are a part of the protocol writing process for SRs. Writing protocol is largely manual and requires several revisions [177]. Registering the protocol in databases like PROSPERO is imperative to avoid deduplication of efforts and is mandatory for Cochrane systematic reviews [144]. Peer review of the protocol helps identify and correct possible avoidable methodological errors before working on the actual review begins [283].

Cognitive work analysis (CWA) of medical and allied SRs revealed that search and retrieval entail a comprehensive literature search either at a single point if the review is a conventional SR or regular literature searches if the review is an LSR. An information specialist from the review team then skims the retrieved studies to check if they are broadly relevant to the review question, for example, if the study is an RCT and the disease under investigation is relevant to the question. If so, the informal data extraction stage begins, and the PICO characteristics from the study are manually extracted. The study full-text (essentially a PDF), along with the title, abstract, other meta-data, and extracted PICO information, is recorded into the research groups' local repository meant for that SR, and the metadata is recorded in a spreadsheet [177]. Full-text retrieval is a cumbersome task requiring manual efforts given license restrictions imposed by the publishers and subscription restrictions imposed by certain databases like EMBASE [38, 328]. For instance, HES-SO staff can manually query and download contents of EMBASE, including the full-texts, as the university purchases EMBASE subscriptions. The staff, however, cannot programmatically retrieve the contents of the database as EMBASE requires an additional subscription, and HES-SO does not pay for it. Automating full-text retrieval will not be the primary target of this work.

Study deduplication is the process of determining which individual study from the retrieved pool describes the same underlying clinical study [134]. It involves manually comparing the meta-data from each study with every other study in the retrieved pool. Duplicates results because the studies are retrieved into the pool from multiple overlapping repositories. Deduplication begins by exporting study data into a spreadsheet and sorting the content by sorting based on the study title. Check for citations with identical titles and assess whether other metadata helps identify them as duplicates. Examine meta-data, including abstracts, author lists, journal names, and volume information, to confirm duplicity. In case of conflicts, a third opinion is sought. Once all the duplicates are identified and marked, they are removed from the spreadsheet, leaving only one instance of each unique study, especially the one published in a later year [39, 180, 268, 275]. At this stage, the duplicates are not removed from the repository, only from the spreadsheet used to track duplicates. In a practical scenario, deduplication is carried out throughout the SR writing process to minimize bias and reduce over or under-reporting of the intervention effects. Deduplication was not the primary target of our automation efforts as automatic deduplication methods are widespread, and they perform fairly accurately [78, 134]. Widely adopted tools like EndNote, Ovid and Covidence provide fairly accurate study deduplication functionality, though advise the reviewers to skim for duplicates manually [224]. Automating study deduplication, therefore, will not be the primary target of this work.

Manual citation screening is straightforward, starting with two independent reviewers evaluating the study title and abstract free text. If it matches the predefined inclusion criteria, it is considered relevant and is included for writing the review and is otherwise

considered irrelevant and is excluded from writing the review. An inclusion criteria is typically defined by the PICO framework as explained in Table 2.1 and may include additional restrictions based on the types of studies to be incorporated, such as only RCTs or a broader range that includes non-randomized studies, case-cohort studies, and more. A two-stage citation screening involves full-text screening after title and abstract (TiAb) screening. In cases of conflict, a third opinion is sought for resolution. Wang *et al.* estimated the time required for citation screening, suggesting approximately 1 minute per abstract and 7 minutes per full-text [353]. In contrast, Polanin *et al.* reported an average of 1.2 minutes per abstract and 1.7 minutes per full-text per reviewer [266]. An experienced reviewer may take an average of 30 seconds to review a study, while an inexperienced reviewer might take longer [346]. Wallace *et al.* presented a significantly longer timeframe, estimating more than 60 hours for screening 5000 abstracts when done by a non-expert [346]. The task becomes protracted because initial literature retrieval results in hundreds of thousands of studies being retrieved due to optimised search strategies for sensitivity. A review with a narrow scope can retrieve fewer studies, while a broad question usually results in thousands of studies retrieved [329, 364, 367]. Manual citation screening is widely acknowledged as one of the most time-consuming tasks in systematic reviews. A survey of 196 reviewers revealed that about 79% of them reported using some form of automation for citation screening, underlining the recognition of the need for reliable automation approaches for the manual citation screening process [332]. An automation approach could assist in classifying the retrieved studies as "relevant" for inclusion or "irrelevant" thereby reducing the screening workload.

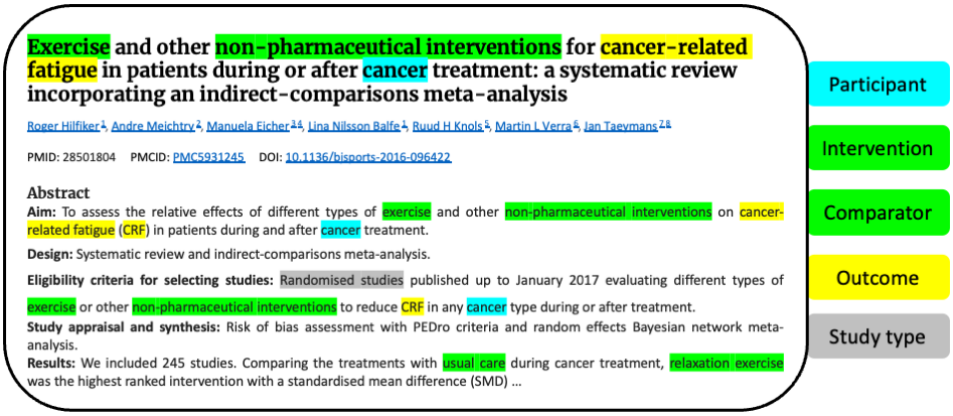


Figure 2.3: The figure shows abstract text from [150] with highlighted PICOS information.

Scanning for PICO information becomes a subtask of data extraction and citation screening. If the study's PICO information aligns with the predefined inclusion criteria, it is considered relevant and included in the review; otherwise, it is considered irrelevant to writing the review and is excluded. If the study's PICO information is absent or not pre-extracted, reviewers manually extract it and make an informed judgment [177]. According to Borah *et al.*, manually analyzing PICO information from thousands of publications to gauge its relevance often takes 2-8 months of two medical experts' time for a single SR.[33] The entire data extraction process, which includes PICO information identification and identification of other details like study type, design and geographical location, takes approximately 53 minutes per study per reviewer [352]. Automation methods can help highlight relevant PICO information in title and abstract texts, facilitating a more efficient

and effective assessment, as illustrated in Figure 2.3.

Figure 2.4: The image displays the RobotReviewer web interface [310]. It shows an RCT with text descriptions (on the right) automatically identified by RobotReviewer's underlying ML algorithm for four risk domains: Random Sequence Generation, Allocation Concealment, Participant and Personnel Blinding, and Blinding of Outcome Assessment.

An RCT might report multiple primary and secondary outcomes, but the one for which RoB assessment is conducted is usually predefined in the inclusion criteria. It's usually impractical to assess bias risk for every outcome in a trial, so usually, the focus of risk of bias assessment is on the predefined primary outcome (PIC**O**) [149]. The next most important point is to select the bias assessment tool. Numerous tools are available, and selecting the tool that best aligns with your specific research question is essential. For instance, if your SR exclusively focuses on randomized clinical trials, it's imperative to opt for a tool tailored to this type of study [148]. Different tools are designed for randomized and non-randomized studies, and a non-exhaustive list of available bias assessment tools is provided in Table 2.2. Each tool categorizes various biases into specific groups, and the reviewer must comprehensively evaluate each of these bias domains in the clinical study under investigation. RoB assessment guidelines prompt reviewers to seek text evidence in the clinical study that could indicate bias risk, aiding in decision-making about the risk level for the assessed domains. The risk judgment for each domain is categorized as "low risk", "unclear risk", or "high risk," based on the cumulative impact of the assessed domains. After assessing individual domains, reviewers assign an overall judgment of the RoB for each RCT. It's important to note that conducting RoB assessments requires reviewers to thoroughly read the entire full-text RCTs. Hartling *et al.* used Cochrane's RoB assessment tool version 1 in a systematic review of a combination of long-acting beta-agonists and inhaled corticosteroids for persistent asthma, where each reviewer took between 14 to 27 minutes per study for RoB assessment [137, 147]. In contrast, Crocker *et al.* reported a substantial 358 minutes spent per study using the revised Cochrane risk of bias assessment tool for RCTs (RoB 2) [70, 148]. The time required for bias assessment for

individual RCTs varies from a few minutes to a couple of hours, depending on the chosen bias assessment tool and the assessors' expertise. An RoB automation tool could help identify these text descriptions from clinical studies indicating bias risks, thus accelerating the assessment process as shown in the screenshot of a bias assessment tool 2.4.

The resource-intensive stages of citation screening, PICO extraction, and risk of bias assessment justify the expenditures for conducting SRs. These stages essentially revolve around manually handling a substantial volume of natural language documents which could come in various formats, including free-text (TiAb of studies), PDF format (full-text of studies), or study metadata such as bibliographic information recorded in spreadsheets that could be manipulated as CSV (comma separated values) files. It is worth noting that any automatic PICO and RoB extraction would aid parts of the data extraction step of writing reviews. Hence, These stages are positively susceptible to the NLP tools and techniques discussed in the following sections. This academic work focused on exploring automation avenues for citation screening, PICO information extraction, and RoB assessment.

2.3 Natural Language Processing (NLP)

In the 1950s, the intersection of artificial intelligence and linguistics gave rise to NLP or Natural Language Processing. NLP is a sub-field of artificial intelligence dealing with techniques that enable computers to understand, interpret and generate natural language content. In 1956, NLP heavily relied on manually crafted rules such as regular expressions. By the 1970s, heuristics like lexical analyzers and language parsers gained prominence. Lexical analyzers segmented text into tokens, and parsers were used to validate the token sequence. Recognizing the limitations of rule-based approaches and heuristics, the field evolved towards statistical NLP in the 1980s. Contemporary NLP has witnessed significant advancements attributed to the progress in modern machine learning algorithms and available annotated data [244]. Machine learning algorithms learn patterns from the input data (annotated or not) and make predictions on future data or generate more data per the patterns seen in the input data. Some of the ML algorithms used in this body of work are described in the next subsection.

2.3.1 Summary of ML Algorithms

Discriminative vs. Generative models: Discriminative models, called conditional models, learn to distinguish between different data by drawing decision boundaries. Discriminative models learn to distinguish by modelling conditional or posterior probability distribution $P(y|x)$ between the unobserved input data variables x and output (also called target) labels y. Unlike discriminative models, generative models learn joint probability $P(x, y)$ between the observed input variable x and target class y. The joint probability captures the likelihood of x and y occurring together, enabling generative models to learn underlying patterns within the data iteratively. The iterative building of a statistical model of the dataset's underlying distribution allows generative models to create new data with similar distribution (or, in layperson's terms, similar characteristics). Nonetheless, generative models like Naive Bayes can use the Bayes rule to calculate conditional probability in the case of classification tasks [250]. Figure 2.5 shows a schematic representation of generative vs. discriminative modelling.

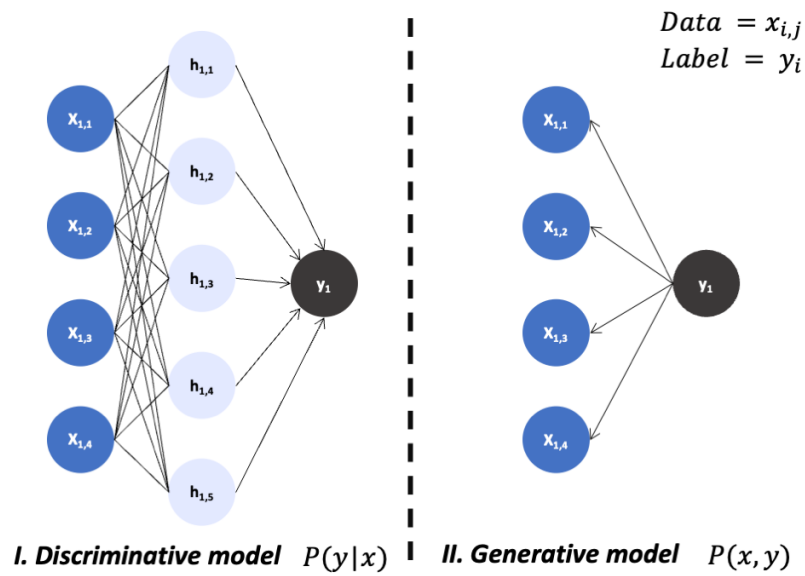


Figure 2.5: Schematic representation of I) Discrimative model, and II) Generative model.

Supervised learning: Supervised learning methods learn a statistical model between the input data variables x_i and the corresponding target labels y_i. Supervised learning models hence need to label datasets that contain a mapping between x and its associated output y, which are used to "train" the model to the labelled dataset to infer the patterns between x and y using the equation $y = f(x)$. After training, the model can be used to predict the unseen data x'. Classification is a form of supervised learning where the desired output is discrete classes like "diseased" vs. "healthy". Logistic regression, Support Vector Machines (SVMs), Decision Trees, Random forests, and K-nearest neighbours (KNN) are the most investigated classification models. Information extraction from raw text is a form of classification where parts of text are classified into representing a class vs. not representing one.

Unsupervised learning: Unsupervised learning models learn patterns and structures from unlabelled data x without explicit supervision provided by output labels y. To elaborate, it can be used to identify cluster structures from the data, but it is not enough to train the classifier on its own. Common unsupervised learning methods include clustering, dimensionality reduction, and density estimation. In unsupervised learning via clustering, the algorithm aims to group similar data points and assign them to k clusters based on their feature (x) similarity.

Semi-Supervision: Semi-supervised learning combines supervised and unsupervised learning and is used when there is an availability of a small labelled dataset and a large unlabelled dataset. Labelled training data is expensive to obtain and thus limited in quantity [237]. Semi-supervised learning can then learn the patterns representative of different categories from the labelled training data while simultaneously learning from the structure in the unlabeled data.

Distant Supervision: Distant supervision is a learning approach in which an algorithm is trained given a distantly labelled training dataset. Distant supervision uses

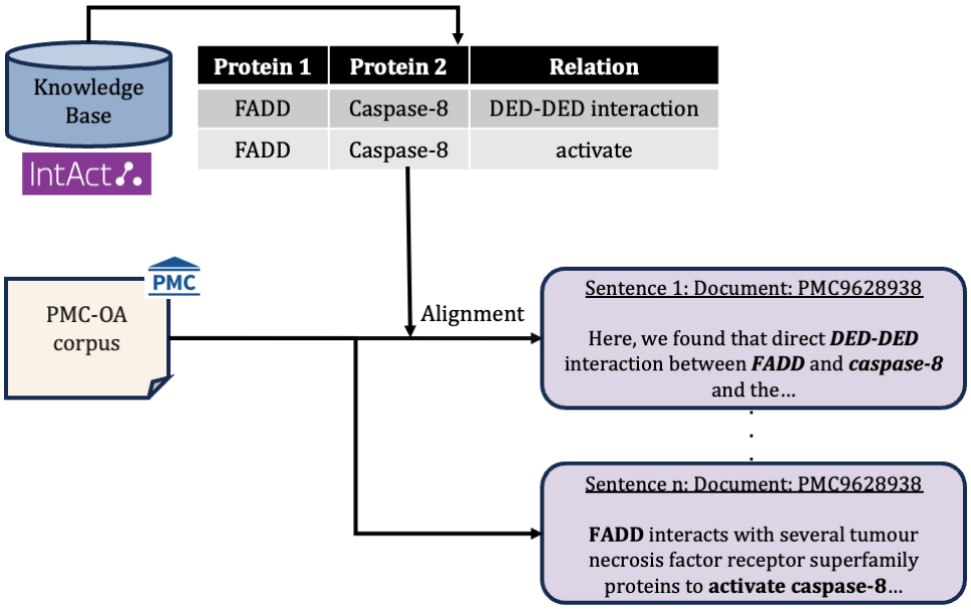


Figure 2.6: Schematic representation of distant supervision. The protein-protein interaction triples from Int Act database I) FADD - DED-DED interaction - Caspase-8 align directly to the sentence from PMC9628938 raw text and thus generate an annotated text. Another interaction between Caspase-8 and FADD, which the triple FADD represents - activates - Caspase-8 align directly to the sentence from PMC9628938 raw text and thus generates an annotated text.

existing knowledgebases to label raw text data, and the approach has been widely used for relation extraction. Mintz *et al.* used a distantly supervised approach to obtain a large relation extraction dataset using Freebase [237]. Elangovan *et al.* used IntAct PPI (Protein-Protein Interaction) knowledgebase to create a distantly supervised dataset by annotating raw text abstracts from PubMed with interacting protein pairs recorded in the IntAct [99]. A schematic representation of distant supervision is shown in Figure 2.6.

Weak Supervision: Weak supervision, similar to distant supervision, is an approach to machine learning that uses multiple noisy supervision sources to create much larger training sets much more quickly than manual labelling could otherwise produce. In contrast to distant supervision that uses only a single source of labelling, weak supervision uses multiple noisy sources of weakly labelling the raw text data. The multiple noisy labels are then consolidated using generative modelling approach modeling [276].

Self-Supervised Learning: In the self-supervised learning paradigm, the training task is designed so that the unlabelled training data can be used to train the model without any explicit human labelling. Masked language modelling (MLM) is a training task commonly used for self-supervised learning. MLM task randomly masks a certain percentage of words in a training sentence. The model is then trained to predict the masked words based on the context provided by the surrounding words, both to the left and right of the masked position. BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pretrained Transformer) are the large language models pre-trained on large datasets of unlabeled text using the MLM task [42, 81].

Active learning: Active learning is an iterative approach commonly used when unlabeled data is readily available, but manual labelling is costly. An active learning system consists of a query module that incrementally queries an oracle, usually a human, for labels. This query module encodes logic to select training data samples most informative for human labelling strategically. By identifying informative training data samples, the active learning approach could learn better on a smaller set of labelled data than required in traditional fully supervised learning approaches. Figure 2.8 provides a schematic representation of the active learning process.

Multi-task learning: Multitask learning is a machine learning paradigm where a model is trained to perform multiple tasks simultaneously. Rather than training separate models for each task, multitask learning allows the model to leverage shared information across tasks, potentially improving performance on all tasks. This approach is useful when tasks share common underlying features or when data for individual tasks is limited. Multitask learning can also help regularize the model and prevent overfitting by encouraging it to learn more generalizable representations. By jointly optimizing multiple tasks, multitask learning can lead to better generalization and efficiency in model training [49, 286]. MTL has shown to leverage performance on nested biomedical named entities, for example, for the nested entities in GENIA corpus [103, 104, 372].

2.3.2 Advances in Feature Representation

NLP uses the above-described machine learning methods to analyze text or speech data[10]. With the emergence of statistical NLP, the methods eventually transformed into core methods from machine to deep learning and, later, full-fledged generative AI (see Figure 2.7). One of the challenges faced in NLP is the inherent inability of machine learning algorithms to comprehend natural language text. Over the decades, strategies were developed to **numerically represent text** as informatively as possible, ensuring that machines can accurately interpret and analyze it. NLP as a domain has gained considerable limelight in recent years with the advent of large language models (LLMs) after evolving over the years from simple Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (tf-idf) representations utilized for information retrieval challenges.

Doc.	**the**	**cat**	**sat**	**on**	**mat**	**dog**	**barked**	**at**	**and**	**are**	**friends**
Doc1	2	1	1	1	1	0	0	0	0	0	0
Doc2	2	1	0	0	0	1	1	1	0	0	0
Doc3	2	1	0	0	0	1	0	0	1	1	1

Table 2.3: Bag-of-Words representation of documents.

Bag-of-Words as the name implies encodes pieces of text as word counts and is a sparse vector representation of text. BoW representation for these three documents is shown in the Table 2.3.

Document 1: "The cat sat on the mat."

Document 2: "The dog barked at the cat."

Document 3: "The cat and the dog are friends."

[10] `https://www.ibm.com/topics/natural-language-processing`

The BoW representation does not consider the importance of each word in the text. For instance, the article "the" in Table 2.3 scores higher than the rest because it is a grammatical necessity and not because it encodes vital information about the document subject. **Tf-idf** fills this gap and encodes weighted information about word count and effectively underweight filler words like articles (a, an, the) and conjunctions (and, but) [210].

using the formula given below.

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D)$$

Where:

$\text{TF}(t, d)$ is the Term Frequency,
$\text{IDF}(t, D)$ is the Inverse Document Frequency,
N is the total number of documents in the corpus,
n_t is the number of documents in the corpus that contain term t.

The IDF is calculated as:

$$\text{IDF}(t, D) = \log\left(\frac{N}{n_t}\right)$$

An **N-gram** is a sequence of n items - words, symbols, phonemes, or letters [244]. A single word is called a unigram, a sequence of two adjacent words is called a bigram, a sequence of three is a trigram, and so forth [170]. BoW and tf-idf are calculated using N-grams. While the lower order unigram representation captures word count, the higher order bi- and tri-grams can capture semantics, e.g. the unigram "New" *vs* the bigram "New York" in the sequence "New York is a bustling city known for its iconic skyline and diverse neighbourhoods." [210].

Both BoW and tf-idf vectors encode sparse information in word counts and importance but do not encode semantic information relating to relationships between words [210, 298]. **Word embeddings** or vectors and specifically word2vec were developed in an unsupervised fashion in the early 2010s. Word embeddings are dense vector representations of words in a continuous vector space, and they capture semantic relationships between words [230, 231]. The most notable developments in the word embedding space were Stanford NLP's GloVe and FaceBook AI's (now Meta) fastText vectors [31, 261]. Studies claimed that fastText outperformed word2vec because of its capacity to encode sub-word information. Encoding sub-word information overcomes the shortcomings of out-of-vocabulary words, which were not addressed in word2vec and GloVe [248, 272]. While word2vec and fastText embeddings were trained on open domain text, bio2vec, an extension of word2vec, and BioWordVec, an extension of fastText, have been specifically trained on PubMed to provide embeddings tailored to the biomedical domain. bio2vec and BioWordVec have outperformed word2vec on biomedical NLP benchmarks [267, 370]. Word vectors encode semantics but do not retain information about the sequential order of words in a given text.

It was the advent of memory-based contextual neural networks that revolutionized text sequence modelling in NLP. **Contextual neural networks** are chains of individual identical units that process each word in the sequence by considering the previous word, thereby incorporating the sequential memory. Take, e.g., **vanilla Recurrent Neural Networks** (RNNs) where each RNN unit takes the information about the current word i

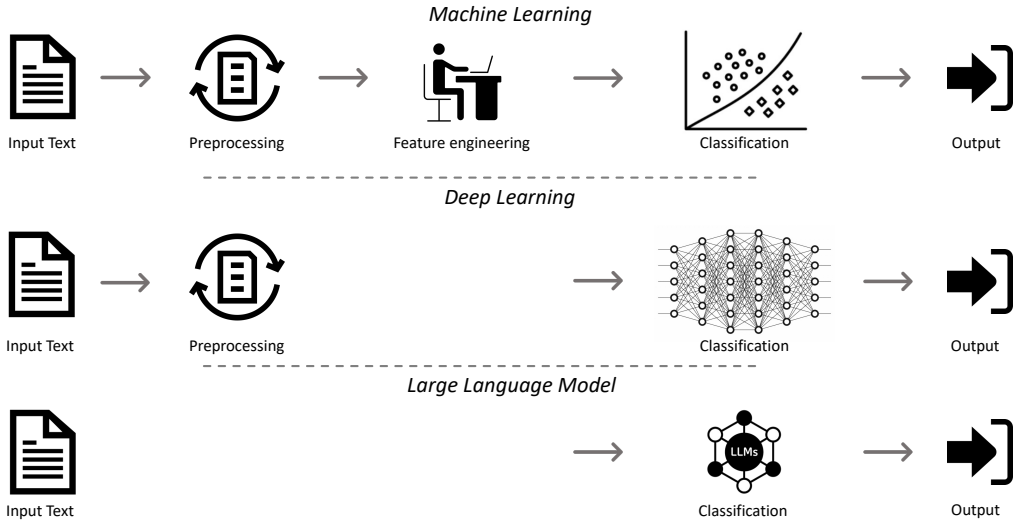


Figure 2.7: The figure illustrating progress of NLP from machine to deep learning and large language models. NOTE: The image was prepared for educational purposes using vectors available on the web and the rights to individual vector and image lies to its creator.

and the previous word $i-1$ and combines this information into a single numerical vector i_1. The numerical representation i_1 is then squished between -1 and $+1$ using a *tanh* operation. The vector is then passed to the next RNN unit, which is combined with the next word $i+1$, and continues until the end of the sequence. RNNs take into account contextual information but have a short memory window and cannot learn as much, leading to vanishing gradients during backpropagation [294, 303]. For extended sequences, Long Short-Term Memory Networks (LSTMs) step in, employing a sophisticated mechanism to combine crucial information while discarding computed data deemed irrelevant [152]. While RNNs and LSTMs were devised in the 1990s, they showed a rise in popularity in the 2010s. They were slowly adopted for biomedical and allied domains, demonstrating superior performance over traditional ML approaches on the tasks of entity recognition and classification [203, 246, 318, 361].

LSTMs can encode longer sequences and, therefore, more contextual information, but the attention mechanism's introduction further improved the sequence-to-sequence tasks. The attention mechanism allows one to attend to all the words in a sequence, up-weighting the signal from the noise words. In theory, attention can look at infinite-length sequences and attend to each word given compute power. The attention mechanism, also called self-attention, was introduced in the paper titled "Attention is all you need," which also introduced transformers, an attention-based encoder-decoder architecture. An encoder takes an input and creates a continuous vector representation out of it. The decoder then takes this representation and, step by step, generates a single output while also being fed the previous output. The transformer architecture uses positional encoding along with the input embeddings to provide the model with information about the positions of words, thereby allowing input sequences to be attended in parallel rather than sequentially [338].

BERT marked the next breakthrough by leveraging the attention mechanism to pre-train deep bidirectional text representations, significantly enhancing various NLP tasks. BERT was trained on 3300 million words from BookCorpus and English Wikipedia [81].

It is an open-domain pretrained transformer model which led to the development of similar large models for biomedical and general scientific domains like BioBERT, PubMedBERT, SciBERT, ClinicalBERT, and BlueBERT. These domain-specific models have shown competitive or even better performance on several domain-specific benchmarks, including datasets like NCBI disease, i2b2, BC4CHEMD, BioASQ, MedNLI, BC5CDR, CLEF eHealth corpora amongst many others [7, 27, 131, 189, 260]. BERT-based models only use encoders (not decoders), making them non-generative models. However, around the same time in 2018, OpenAI published a paper titled "Improving Language Understanding by Generative Pre-Training", which introduced the generative pre-trained transformer (GPT-1) system using both encoder and decoders [270].

These transformer models are trained in an unsupervised manner using masked language modelling (MLM) or other similar objectives [81]. They can thus be trained on progressively more data as resource availability and compute power permits. GPT-1 had 117 million pre-training parameters, and its subsequent GPT-2 had a staggering 1.5 billion parameters. Subsequent developments led to the emergence of larger LLMs like GPT-3 with 175 billion parameters and LLaMA (Large Language Model Meta AI) with 65 billion parameters, which excel in a wide range of NLP applications, from text generation to understanding and translation, showcasing the transformative potential of these SOTA models [42]. These models became accessible to the public with OpenAI releasing ChatGPT [11] which is a chat interface interacting with GPT-3.5, and Google's BARD chat [12]. LLMs like MEDITRON-70B built on Llama-2, continually pre-trained on curated medical domain corpora, outperformed GPT-3.5 and were within 5% of GPT-4 [59]. Dr. Hartley from EPFL, Switzerland and Harvard pioneered MEDITRON-70B, making it open-access for the scientific community. The focus is shifting to multimodal LLMs, which could concurrently handle free-text, PDFs, images, hyperlinks and speech [363]. Google BARD chat offers these capabilities free of cost, while ChatGPT has a paid version for it via accessing GPT-4, which can generate content based on visual and textual inputs. Sam Altman, the OpenAI CEO, hinted the possible release of GPT-5 in 2024 via Twitter[13].

The recent breakthroughs in LLM research have spurred research on consciousness in artificial intelligence with notable contributions from one of the leading AI experts, Dr. Yoshua Bengio [44]. These advancements have significantly broadened the scope of NLP, empowering machines to understand and produce human language with unparalleled accuracy and sophistication, all while progressing at an unprecedented pace. In the next section, the NLP approaches in SR automation are discussed in detail.

2.4 NLP approaches for SR semi-automation

2.4.1 Citation Screening

The first approach to screening reduction in citation screening is to use classification algorithms offered by text mining and NLP. Classification algorithms are trained to make binary decisions regarding the relevance of citations, categorizing them into "relevant" or to be included into writing the review *vs.* "irrelevant" or to be excluded from writing the review. The quality of a classification algorithm is typically evaluated using workload reduction metrics like work saved oversampling (WSS), which is the percentage of papers

[11] https://chat.openai.com/

[12] https://bard.google.com/chat

[13] https://twitter.com/sama/status/1738673279085457661

the reviewers do not have to read because they have been screened out by the classifier [66]. A citation screener could also serve as a second reviewer, whereby this citation screener checks whether the included citations are consistent and no citations have been missed. A citation screener as a second reviewer does not attempt to reduce the number of references that need to be screened but rather to avoid having each reference screened by multiple reviewers [255]. In fact [22–24] have advocated replacing one of the reviewers with an automatic citation screening system.

The earlier efforts to automate citation screening began with automatic database filtering using text mining techniques to filter out RCTs from the non-randomized clinical trials [146]. Cohen *et al.* and Bekhuis *et al.* developed classification models to filter RCTs and non-RCTs from the rest of the retrieved studies, respectively [23, 67]. [23] *et al.* compared kNN (k-Nearest neighbour), naive Bayes, complement naive Bayes (cNB), and evolutionary SVM (EvoSVM) using tf-idf features and MeSH/EMTREE terms. The approach was evaluated on a single citation screening dataset with 46% workload reduction and 95.5% recall using EvoSVM. Marshall *et al.* used SVMs to separate RCTs and non-RCTs, reaching an excellent AUROC (Area under Receiver Operating Curve) of 0.987, showing an improved performance in comparison to the text mining based filtering [216]. Their RCT-classifier was integrated into Covidence and EPPI-Reviewer, both widely used tools in academic and research institutions for conducting SRs spanning disciplines [43]. These efforts were limited to the binary classification of randomized versus non-randomized studies without considering the other inclusion criteria.

In 2006, Cohen *et al.* released DERP (Drug Evaluation Review Project), a set of 15 citation screening datasets with their inclusion decisions [66]. They used DERP to train a voting perceptron to classify studies for inclusion, and the approach was evaluated on the WSS metric measuring the amount of workload reduction. They reported work savings between 0% - 67.95%. Later, in 2010, Cohen *et al.* extended DERP to 18 and 24 labelled citation screening datasets [64, 65].

The availability of the Cohen dataset played a crucial role in accelerating the research efforts exploring classical machine learning models like naive Bayes and SVMs trained for binary classification [62, 63, 168, 219–221]. In [63], Cohen *et al.* evaluated the use of SVMs with n-gram features, MeSH and UMLS terms. The authors from Matwin *et al.* used the DERP dataset to train and evaluate FCNB (Factorized Complement Naive Bayes) for binary classification, reaching WSS of 8.5% to 62.2%. In [168], the authors simulated active learning with random indexing. The authors simulated the system on the DERP dataset and reported work savings between 6%–30%. Bannach *et al.* [18], used SVM with and without clustering for workload reduction in two systematic reviews of preclinical animal studies reporting 70.5% and 69.3% WSS for over them.

Previous automation studies utilised classical machine learning approaches involving massive text preprocessing and feature engineering for a very long. Hand-crafted features must be heavily fine-tuned to achieve good performance, which is a tedious and time-consuming task and must be performed by an expert. Progress in deep learning saw the application of shallow and deep learning methods again, considering citation screening automation as binary classification. In their approach, Lerner *et al.* used four medical citation screening datasets to evaluate binary LR models trained on bio2vec word embeddings [190, 267]. The approach achieved an excellent 100% recall on two datasets and more than 93% In a different approach, van Dinter *et al.* trained multiple binary multi-channel CNN models using GloVe word representations of citations, presenting work savings across 21 citation screening datasets introduced by Cohen *et al.* [336]. Despite their comprehen-

sive assessment of workload reduction over 21 datasets, they did not report the most vital recall metric. Qin *et al.* trained a gradient boosting ensemble (LightBGM) integrating four BERT-based representations, achieving the expected recall of 96% on one internal dataset [269], but did not release the dataset. Citation screening automation has also been explored with PICO information extraction [41].

By training a machine learning model, binary classification involves effectively categorizing studies as either "relevant" or "irrelevant". This training is typically conducted on 50-90% of manually labelled citations from the screening dataset, with evaluation on the remaining 10-50% [337]. In practice, addressing each new SR question requires manual labelling of approximately 50-90% of the dataset, depending on the chosen training dataset size. This becomes impractical for businesses due to the associated labelling costs.

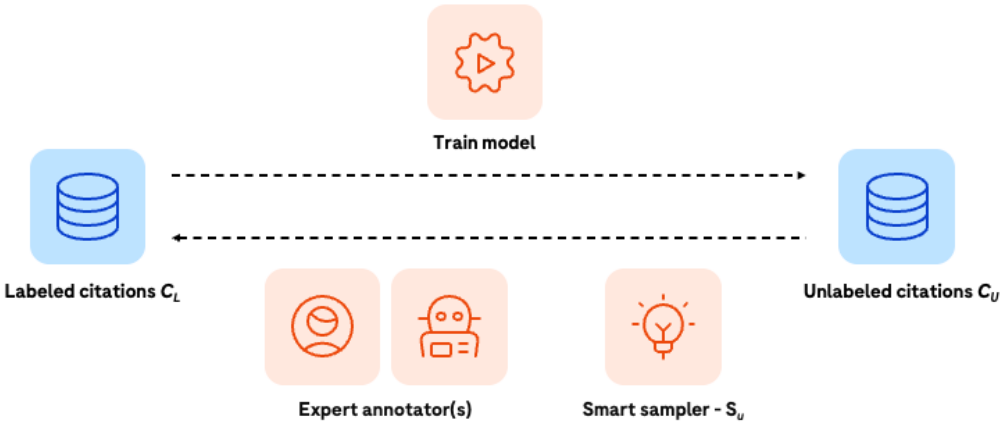


Figure 2.8: Schematic representation of active learning approach for citation screening or active citation screening.

Active learning (AL) is an approach that aims to reduce the labelling cost per *de-novo* SR by selecting the few most informative citations for classifier training. As schematically shown in the Figure 2.8, an active citation screening system has a query module also called a smart sampler module S_U that encodes the logic to select these most informative citations from the unlabelled citation set C_U. The smart sampler S_U interacts with an oracle (e.g., a human expert) to obtain labels for these selected citations, and once a certain number of labelled citations are collected, it triggers the active classifier model training. The goal is to select the samples that are the most informative to the model in order to improve its performance with a minimal number of labelled citations. AL is quite suitable for *de novo* SRs, which involve a fresh set of unlabeled citations. In such cases, AL could significantly reduce the labelling workload by assisting in selecting the most informative citations for human labelling. This approach could lead to more efficient resource utilization, improved work savings and thus cost-effectiveness.

AL approaches have been extensively explored in retrospective citation screening scenarios using publicly available datasets across domains like biomedical, public health and software engineering. These approaches use either classical ML, deep learning or a combination of both to test multiple active learning system settings like starting and stopping criteria, query strategies, methods of seed set sampling, and sampling methods amongst others [50, 51, 192, 201, 238, 334, 345, 346, 365]. None of these studies explores active learning in a prospective scenario, a crucial business requirement. Optimizing prospective active

citation screening systems involves considering multiple objectives, including core metrics like work savings (WSS), recall of "relevant" citations, and decision-enhancing metrics such as AUROC and the labelling cost to initiate AL system training. These decision-enhancing metrics assist in selecting better AL system settings. None of the above AL studies evaluate their approaches using all these crucial metrics required to assess AL systems. Additionally, the current studies do not statistically test the impact of different active learning system settings on performance. We identify a lack of citation screening studies employing and comprehensively evaluating active learning systems in a prospective scenario, which poses the challenge in deploying such systems.

2.4.2 PICO+ Information Identification

Automatic identification of PICO information involves employing information extraction techniques from NLP, such as named entity recognition (NER), entity recognition, and paragraph retrieval. Demner-Fushman *et al.* developed knowledge-based extractors supported by concepts from UMLS (Unified Medical Language System) to extract PICO terms from scientific abstracts. The extractors outperformed location-based baselines on an internal gold standard corpus [80]. Location or position baselines assume that PICO information is present in specific positions within the abstract, such as participant information being in the first three sentences. Boudin *et al.* employed structural constructs from PubMed abstracts to create a training set for classifying sentences describing PICO information from the rest. They used the test dataset obtained from Demner-Fushman *et al.*, and Chung *et al.* comprised 14279 abstracts [60, 80]. They used random forest, SVM, naive Bayes, and a shallow neural network - multilayer perceptron (MLP), with SVM outperforming in extraction precision while MLP is consistently demonstrating strong performance in F1-score citeboudin2010combining. These structural constructs included, for instance, extracting the first sentence from semi-structured abstracts following the Population/Participant, Intervention, and Outcome headings. A similar strategy was employed by [157], emphasizing that relying solely on the first sentences following the P/I/O heading for training a classification model might not always be optimal compared to using all sentences following these headings. Wallace *et al.* used supervised distant supervision to learn to identify PICO sentences from full-text studies [343]. Chabou *et al.* used conditional random fields (CRF) for PICO element detection and noticed an acceptable precision but a low recall rate. Combined with a rule-based approach, they noticed an increase in recall compared to a machine learning-based CRF approach alone [54]. With the introduction of deep, contextual neural networks, Jin *et al.* trained a single LSTM model on thousands of abstracts to classify sentences to one of the PICO classes. Theirs was the first large sentence-level PICO dataset made publicly available [166]. However, this decade of automatic PICO information identification was constrained to identifying information at the sentence level and, if not, then relying on hand-engineered features and extensive text preprocessing. This limitation was due to the unavailability of a larger corpus annotated with PICO entities [158, 159]. There were no publicly-available While sentence-level recognition and summarization provide some degree of automation, the potential for full automation lies in semantic-level recognition of PICO-describing phrases, potentially granting machines the ability to reason.

The availability of a comparatively large, and probably the only, PICO benchmark corpus (EBM-PICO corpus hereafter) from [254] with multi-grained PICO annotations opened up possibilities to explore the neural models. EBM-PICO corpus consists of 4,993

titles and abstracts comprehensively annotated with two levels of PICO descriptions. The first level is the coarse-grained descriptions, and the next level is the fine-grained PICO descriptions. Each PICO class further decomposes into more refined semantic units in the fine-grained descriptions. For example, the 'P' in PICO, which represents Participant class, can be decomposed into specifics such as participant age, gender, ethnicity, disease status, comorbidity, and so forth. Nye *et al.* used EBM-PICO to train baseline models using hand-engineered features and shallow neural models for separately detecting fine- and coarse-grained entities. Several peer-reviewed studies consequently used EBM-PICO as a benchmark for PICO entity or span extraction going beyond the previously dominant sentence-level extraction [28, 41, 131, 202, 369].

Labelling PICO entities is tricky because of the high disagreement between human annotators on the exact text spans constituting PICO, leading to human errors in hand-labelled corpora[41]. Abaho*et al.* and Lee *et al.* have explored errors in the EBM-PICO dataset, attempting to manually correct them for their applications [1, 188]. Hand-labelled datasets are static and prohibit quick manual re-labelling in case of human errors. More importantly, PICO analysis frequently extends to analysing other information like PICOS (**S = Study type and design**), PICOT (**T = Timeframe**), PICOC (**C = Context**), EDR (**E = Exposure**, **D = Duration**, **R = Results**), PIBOSO (**B = Background**, **S = Study type and design**, **O = Other**), *etc.* depending upon the SR question [8, 228, 280, 333]. Study-type information is vital, for example, in conducting systematic reviews that aggregate evidence from selected clinical study types. Trial duration information is essential for establishing the long-term efficacy of the treatment [227]. Authors in [74] had to extract 835 PECODR text manually from 20 EBM journal synopsis. After manual analysis, they proposed rules and linguistic patterns to extract the PECODR items automatically. Similarly, Kim *et al.* hand-labelled a corpus (NICTA-PIBOSO) of 1,000 medical abstracts annotated with PIBOSO entities which were used to train a PIBOSO sentence classifier using CRF model [175]. NICTA-PIBOSO was subsequently used by Hassanzadeh *et al.*, Verbeke *et al.* and Sarkar *et al.* to improve the PICO sentence extraction using hand-crafted features and using CRF, kernel-based learning (kLog), SVM classifiers, respectively [138, 292, 339]. The challenge arises again in the expensive and resource-intensive process of manually re-labelling large datasets that do not provide annotations for these additional entities.

The challenge of manual annotation, particularly in cases involving human errors and the need for additional entities in annotated corpora, has led to a shift in focus towards distant supervision and weakly supervised learning approaches. These approaches leverage programmatic labelling sources, offering more cost-effective options for obtaining training data. One of the first applications of distant supervision to open domain relation extraction was proposed by Mintz *et al*, who used Freebase to extract 10,000 relation instances from Wikipedia articles with a precision of 67.6% They assumed that if Freebase has an instance of two entities participating in a relation, then any sentence containing these entities will likely express that relation too [237]. Zheng *et al.* used DS augmented training dataset using UniProt [14] for extraction of protein subcellular localizations with an 82% accuracy [373]. Weak supervision too has demonstrated strengths for clinical document classification, biomedical information and relation extraction, but clinical entity extraction tasks like PICO have heavily relied on fully supervised (FS) approaches [98, 99, 209, 226, 351, 354]. The weakly or distantly supervised entity recognition approaches to PICO could

[14] `https://www.uniprot.org/`

be more challenging than the entity recognition approach to protein or chemical names, which are more or less standardized. PICO terms are not standard, and even the experts disagree on the exact tokens constituting them [41]. Thus, the area of procuring affordable PICO+ entity labels in the absence of manually annotated data presents a research gap.

2.4.3 RoB Assessment

Risk of bias assessment in clinical studies is traditionally a manual and challenging process led by experts. Marshall *et al.* used distantly supervision using SVMs to classify RCTs into the risk of bias assessment classes, "low" risk of bias and "high" or "unclear" risk of bias. Their distantly supervised models were trained using labelled data from the Cochrane Database of Systematic Reviews (CDSR). CDSR [15], though a valuable, high-quality resource, is a subscription-based service, and access to the full content requires a subscription or institutional access through universities, libraries, or other organizations [213, 214]. Around the same time in 2016, Millard *et al.* published a paper on automating RoB assessment using supervised machine learning. They trained one supervised logistic regression model to predict whether a sentence contained information pertinent to a risk of bias and used another supervised logistic regression model to predict one of the three bias classes. The data used for training the models was obtained from Cochrane Collaboration, specifically CDSR and was proprietary [232].

The research utilizing this pay-walled data was used to develop RobotReviewer that has been evaluated by several studies for its human-competent performance [11, 12, 151, 164, 215, 310, 317, 340]. RobotReviewer was developed using proprietary data from CDSR and the older risk of bias guidelines (Cochrane Collaboration's tool for assessing the risk of bias in randomized trials - RoB 1) [146]. Even though RoB 1 is the most frequently used to assess RCT quality, a recently revised Cochrane RoB 2 offers significant differences in comparison [313]. Compared to the original RoB version released in 2008, the RoB 2 version provides a more reliable and concrete structure to the RoB evaluation by developing comprehensive guidelines that aim to increase consistency [206]. A study analyzing Cochrane systematic reviews and protocols found that the use of RoB 2 increased from 0% in 2019 to 24.1% in 2022, indicating the importance of using an updated tool [217].

Wang *et al.* recently released three RoB annotated datasets but for preclinical studies with RoB assessments about animals [348]. A manually annotated corpus of RoB spans for human clinical trials is necessary but is unavailable. This gap in annotated resources is a significant bottleneck for training and evaluating machine learning models in RoB assessment. The recent advancements in LLMs suggest that fully supervised training of machine learning models may no longer be necessary. However, a benchmark corpus is necessary to evaluate language models for the expert-led task of bias assessment. Recognizing this need, Rose *et al.* very recently published a protocol for RoB annotation of an *in-house* dataset and proposed using this annotated dataset for evaluating LLMs [284]. This indicates the pressing need for the risk of bias annotated resources to assess powerful LLMs.

[15] `https://www.cochranelibrary.com/cdsr/about-cdsr`

2.5 Summary

- Systematic review answers clinical and research questions by analyzing and collating all published evidence systematically and transparently.

- Being foundational to evidence-based medicine, systematic reviews are vital in informing clinical guidelines and shaping health policies.

- However, the increasing cost of conducting these reviews, fueled by the increase in the published literature, poses a significant financial challenge.

- The steps of citation screening, data extraction and risk of bias assessment are three of the most time-consuming steps in conducting the review. These steps involve chaffing out information from volumes of clinical studies.

- Natural language processing and text mining approaches offer potential solutions to streamline these labour-intensive stages.

- The decade from 2014 to 2024 witnessed significant advancements in natural language processing and machine learning approaches, evolving from simple numerical representations like bag-of-words to sophisticated techniques such as semantic word embeddings and large language models capable of capturing extensive knowledge.

- Several approaches to citation screening automation have been developed since 2006. However, these methods have not been developed considering the prospective scenario and are thus under-evaluated for a real-world automation scenario.

- There is a research gap in procuring affordable labels for PICO and more entities in the absence of manually annotated data, underscoring a need for innovative solutions powered by weak and distant supervision methodologies.

- The absence of a publicly available annotated dataset for Risk of Bias assessment in human clinical studies poses a significant obstacle to automating the process and advancing machine learning models in this domain.

Chapter 3

Citation Screening Automation

This chapter details the machine learning approaches used to investigate citation screening automation, considering both the retrospective and prospective scenarios. Section 3.2 details the classical machine learning approach to explore manual citation screening as a binary classification task. Section 3.3 details the active learning approach developed to address the challenges associated with prospective citation screening systems in the pharmaceutical industry. This chapter also explored the considerations associated with implementing a semi-supervision approach in real-world active citation screening systems. Parts of this chapter have been published as a conference paper, and another segment is currently being prepared for submission as a journal paper [85, 93]. In [85] and [93], my contribution was to procuring and cleaning the datasets, devising approaches and system evaluations aligned with research and business requirements, designing and executing experiments, analyzing results, and presenting findings in the form of conference and journal papers, respectively. The results of [93] were also presented at the BioTechX Europe 2023.

The following resources are made available via this research:

1. The citation screening dataset used in the section 3.2 is available on Zenodo.
    - `https://zenodo.org/records/10423427`
2. The citation screening approach explored in the section 3.2 is available on GitHub.
    - `https://github.com/anjani-dhrangadhariya/citation-screening-ml`
3. The active citation screening approaches explored in the section 3.3 is available on GitHub.
    - `https://github.com/anjani-dhrangadhariya/active-ssl-citation-screening`

3.1 Introduction

SRs involve synthesizing and summarising relevant data from hundreds of thousands of primary studies to answer specific clinical questions. Among all stages of a systematic review, citation screening is known to be one of the most time-consuming and labor-intensive steps. Citation screening involves manually evaluating a bulk of studies to determine their relevance for answering the SR question. Manual citation screening begins after the literature search and retrieval phase, which involves gathering as many studies as possible to answer a systematic review question. The process involves two independent reviewers reading

through hundreds of thousands of studies to comprehend whether the study is relevant to writing the SR. The decision about the relevance of the study is made based on comparing the study title and abstract with the pre-defined inclusion criteria. The studies that fulfil the inclusion criteria are included in writing the review, and the rest are excluded, thus narrowing down the evidence pool. In case where one of the two reviewers disagrees with the study's eligibility, a third opinion is sought as a tiebreak [18, 146, 173, 174].

The process is not only redundant but also resource-consuming. One of the more recent studies shows that writing a single SR takes about 67 weeks, but the studies in the past reported this time to be anywhere between 2.4 to 3 years, mainly depending on how narrow or broad an SR question [33, 133, 326, 331]. About 23% of static SR need updating within two years of completion as new primary studies become available [304]. As previously described, LSRs or living systematic reviews are dynamic and are continually updated as new studies become available, keeping up to date with the newly published evidence [100, 233]. Manual citation screening becomes impractical considering situations like a pandemic that require rapidly assessing evidence to formulate policies and make treatment decisions.

With the exponentially increasing primary studies, the urgency to rapidly carry out SRs to answer pressing questions, and the need to update a review as soon as new evidence is available, the reviewers often can not keep up with the manual process of screening the studies and constantly updating outdated SRs [173]. In the area of physiotherapy and rehabilitation, such an exponential increase in the number of publications is also observed[16]. For an ongoing update to the review on exercise and non-exercise interventions in reducing cancer-related fatigue, the database search retrieved over 30,000 references, about 2,000 of which were published in 2017 alone. The two independent reviewers took more than 200 hours each to manually assess the titles and abstracts for relevance to the research question before the studies were taken for further meta-analysis [150]. Automation is, therefore, imperative.

3.2 Machine Learning Assisted Citation Screening

Supervised machine learning based classification approaches are successfully applied for automation of citation screening but either only for broad and shallow SRs [17] or SRs that retrieved fewer than 6,000 studies [190]. However, there are SRs that address very specific research questions leading to narrow, predefined criteria for selection of relevant studies to be included for meta-analysis. For such narrow SRs, inclusion prevalence becomes as low as 10%, which means that out of all the studies retrieved during the search phase, 90% are excluded as non-relevant. A narrow research question combined with a low inclusion prevalence leads to class imbalance and class overlap problems for classification tasks that generally reduce classifier performance [117]. Class overlap cannot be artificially controlled but class imbalance can be tackled using oversampling or undersampling [289]. Oversampling or undersampling aim to bring the number of instances in the minority class equal to the number of instances in the majority class. In this work, we aim to explore machine learning and natural language processing to assist citation screening in SRs with a narrow research question and low inclusion prevalence using word embeddings and random oversampling.

[16] https://www.ncbi.nlm.nih.gov/pubmed?term=(physiotherapy)%20OR%20rehabilitation

3.2.1 Methodology

This section describes the machine learning approach used to explore citation screening as a binary classification task and also describes the dataset used to generate word embeddings and the dataset used to test the machine learning approach.

3.2.1.1 Datasets

Datasets from the open access literature were used in the work described here. These datasets were used primarily for two tasks: I) PubMed Central Open-Access subset (PMC-OA) subset was used to generate task-specific word vectors or embeddings. II) A citation screening dataset from Hilfiker *et al.* was used to test the machine learning approach. In order to generate word embeddings, PMC-OA subset was used. Articles in the PMC-OA are made available under Creative Commons or similar licenses that lets one share and reuse the information more openly for research than a regular copyrighted work [17]. As of 2019, PMC-OA contained titles and abstracts (TiAb) of 2.09 million studies that were used to generate semantic word embedding using the two most common architectures: word2vec and fastText [31, 231].

A physiotherapy citation screening dataset was used to test automation approaches and the dataset includes the studies identified for citation screening in an update to the systematic review by Hilfiker *et al.* [150]. The dataset included TiAb from 31,279 studies identified during the search phase of this SR. These studies were already manually assessed for relevance and labelled by two reviewers into two mutually exclusive labels. 2259 studies assessed as relevant were labelled "include" and 23,279 studies assessed as non-relevant were labelled "exclude". The inclusion prevalence for this case is only about 8.84% leading to class imbalance. Inclusion prevalence refers to the percentage of studies were relevant for writing the SR and meet the inclusion criteria, relative to the total number of studies in the citation screening dataset.

3.2.1.2 Screening Automation Approach

We framed the citation screening automation as a fully supervised binary classification task whereby we trained six classifiers to learn the difference between "relevant" (include) and "irrelevant" (exclude) citations using the Hilfiker dataset represented using corpus-specific static word embedding. This exercise let us gauge the effect of class imbalance too. The approach follows steps enumerated below. 1) Generation of word embedding, 2) Dataset and text preprocessing, 3) Random oversampling, 4) Feature extraction, and 5) Classifier training and evaluation.

1. Word embedding generation To generate word embeddings, the TiAb from PMC-OA subset were lower-cased and all punctuation except the hyphens were removed. Phrase generation was then performed using the word2phrase tool[18] to identify frequently occurring bi-grams. The output of phrase generation along with the unigrams was fed to gensim's word2vec[19] and to fastText[20] using the hyperparamters in Table 3.1 to obtain two dense, semantic, real-valued word embeddings [31, 125, 231].

[17] `https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/`
[18] https://github.com/travisbrady/word2phrase
[19] https://radimrehurek.com/gensim/models/word2vec.html
[20] https://radimrehurek.com/gensim/models/fasttext.html

Parameter	Value	Parameter	value	Parameter	value
Size	300	Alpha	0.05	bucket*	2000000
Window	5	min_alpha	0.0001	minn*	3
min_count	5	sample	0.0001	maxn*	6
sg	1	iter/epoch	5	seed	1
hs	1	negative	5		

Table 3.1: Hyperparameter values used to generate word embeddings using gensim's word2vec and the fastText functionality. (*) means that these parameters were available only for the fastText embeddings.

2. Text pre-processing Manual deduplication was carried out on the Hilfiker dataset to remove identical citations. After deduplication and removal of non-English language studies, 25,540 studies remained. For these remaining studies, the text was lower-cased and tokenized into words using NLTK[21] (Natural Langauge ToolKit). Irrelevant tokens were removed using a predefined set of stop-words provided by NLTK and PubMed[22]. Stop-words were removed using a predefined set of words provided by NLTK, PubMed, and corpus-specific stop-words identified during the experiments. Additional corpus-specific stop-words identified during the experiments with the training set were removed accordingly. The text normalization process converted British English terms into American English. After token lemmatization, a corpus vocabulary was constructed from all the unique unigram and bigram tokens. To scale this vocabulary down, we removed tokens with fewer than five characters and a vocabulary count below five, as these were deemed uninformative and not representative of the classes. We scaled down the vocabulary to reduce the vector dimensionality and enhance efficiency in subsequent analyses.

3. Random oversampling Class imbalance often deteriorates the classifier performance, so in the present dataset it was addressed using naive random oversampling [97]. This method randomly duplicates data points from the minority class and brings the total number of instances in the minority class equal to the majority class in binary classification settings [117]. A class which has a disproportionately low number of instances is termed the minority class.

4. Feature extraction Feature extraction was performed to generate real-valued, dense feature vectors from the tokenized text using both the pretrained word embedding types. These resulting features served as input for training non-neural supervised machine learning classifiers. Then, all vectors corresponding to each token within an individual study underwent an averaging process over the entire study, normalized by the study's length. For the Convolutional Neural Network (CNN), feature extraction was part of the model. A static, non-trainable weight matrix, derived from the word embedding, was incorporated with the embedding layer. This matrix facilitated the extraction of token word vectors during the training phase of the CNN model. Typically for image processing regular CNNs are used, but we used 1-dimensional CNN model here, aligning with the 1D dimensionality inherent in text.

[21] https://www.nltk.org/api/nltk.tokenize.html
[22] https://www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T.stopwords/

Figure 3.1: A t-SNE projection for 25,540 studies labelled "include" vs. "exclude" before and after oversampling.

5. Classifier training and evaluation Six classifiers including Logistic Regression (LR), Support Vector Machines (SVM), k-nearest neighbour (KNN), Decision Trees-CART (DT), Random Forest (RF), and CNNs were trained and evaluated for binary classification ("include" vs. "exclude") over the generated embeddings. Estimated probabilities for whether a study belonged to class "include" or not are output by all the classifiers. Hyperparameter tuning was performed using GridSearch. The CNN was trained using hyperparamters loosely based on suggestions from Zhang *et al.* [371]. Classifier performance before and after oversampling was evaluated on an unseen evaluation set and metrics pertinent to imbalanced classification like precision, recall, F1 and precision-recall AUC (PR-AUC) score were tracked [289].

3.2.2 Results and Discussion

The Hilfiker dataset used for training and evaluating the classifiers was highly imbalanced and comprised 2,259 relevant studies labelled "include" and 23,281 labelled "exclude". To measure how similar both the classes were, we calculated cosine similarity between the class centroid vectors for the "include" and "exclude" classes. Cosine similarity between class centroid vectors for the classes before and after oversampling was 0.985814 and 0.985824 respectively. A high cosine similarity is indicative of high class overlap even after tackling the class imbalance using oversampling. These high imbalance and high overlap of our dataset are demonstrated in the Figure 3.1 using t-SNE (t-distributed Stochastic Neighbor Embedding) representation.

The evaluation metrics obtained by applying classifiers on the dataset before and after random oversampling are summarized in Table 3.2 and 3.3. Before oversampling, the classifiers focused on improving the performance for the majority class but in reality they are simply predicting the majority class as noticeable from the relatively high F1 score for the exclude class. Upon oversampling, the overall classifier performance drastically improves for the minority class especially the precision (see Table 3.3), while the precision for the majority class is reduced with a small improvement in recall.

The most naive LR classifier performs the best in terms of the binary recall for the "include" class before oversampling. If the task is considered a classification task, a high class overlap still leads to unacceptable precision and recall values for citation screening implying the inclusion of false-positive or "irrelevant" studies and exclusion of false-negative or "relevant" studies. Note that in the Table, we detail only the best performing word embeddings of the two and fastText performs consistently better than word2vec. This

could be caused by the ability of fastText to represent out-of-vocabulary (OVV) words, which word2vec cannot. FastText can provide better embeddings for morphologically rich languages compared to word2vec as it uses the hierarchical classifier to train the model. Also, the dataset vocabulary coverage for both the word embeddings was about 60.629% which meant the rest of the words were OOV.

		Class "include"				Class "exclude"		
Model	embed	P	R	F1	PR-AUC	P	R	F1
LR	fastText	0.4044	**0.8990**	0.5576	0.6008	0.9891	0.8746	0.9283
SVM	fastText	0.6538	0.4640	0.5428	0.6317	0.9463	0.9746	0.9602
KNN	word2vec	0.6536	0.6066	0.6288	0.6512	0.9619	0.9685	0.9652
DT	fastText	0.2961	0.8627	0.4394	0.4287	0.9837	0.8000	0.8821
RF	fastText	0.4995	0.7892	0.6108	0.5921	0.9780	0.9209	0.9485
CNN	fastText	0.6545	0.5032	0.5690	0.6388	0.9511	0.9732	0.9620

Table 3.2: Classifier performance before random oversampling for the "include" and "exclude" classes. P = Precision, R = Recall

		Class "include"				Class "exclude"		
Model	embed	P	R	F1	PR-AUC	P	R	F1
LR	word2vec	0.8981	0.8850	0.9116	0.9342	0.8951	0.9090	0.8816
SVM	fastText	0.8914	0.8818	0.9012	0.9378	0.8990	0.8792	0.8890
KNN	word2vec	0.8860	0.8303	0.9500	0.9321	0.9418	0.8055	0.8682
DT	fastText	0.8348	0.8201	0.8510	0.8734	0.8285	0.8463	0.8126
RF	word2vec	0.8695	0.8918	0.8488	0.9279	0.8966	0.8757	0.8560
CNN	word2vec	0.9034	0.7480	0.8183	0.9318	0.7850	0.9200	0.8471

Table 3.3: Classifier performance after random oversampling for the "include" and "exclude" classes.

3.2.3 Conclusion and Future Work

To the best of our knowledge, this was the first attempt to explore citation screening automation for a narrow, physiotherapy SR topic using domain-specific word embedding on a range of ML classifiers. We also shed a light on the impact of class imbalance and class overlap on the classifier performance before and after oversampling as also discussed by Garcìa *et al.* and Prati *et al.* [117, 289]. Knowledge of these challenges could be immensely useful for further development of citation screening automation approaches. However, a fully-supervised machine learning approach as we explored in this section does not lend itself to accelerating systematic reviews in real-world scenario. A full supervision approach as previously explained requires a large labelled training data which is impractical for academia as well as businesses. Especially for citation screening automation whereby with every *de-novo* SR question, comes a fresh set of citations and new inclusion criteria requiring training a new model after labelling almost the complete dataset. It defeats the purpose of automation. Active learning could help overcome this bottleneck and the next section details how we used active learning for developing a citation screening system for businesses.

3.3 Active Citation Screening: Business Scenario

As explained in the previous section, a supervised machine learning approach tackles citation screening as a binary classification task whereby the ML model in question learns to distinguish between relevant and irrelevant studies or citations. Supervised ML could help automate the process but requires large, labelled datasets to ensure good performance. In a real-world practical setting, whether in business or academia, for every *de-novo* SR, there's a need to label a fresh batch of citations to train the classifier. Hence, supervised ML does not lend itself to faster SRs in practice. However, the majority of the existing research uses supervised ML as a testing ground to retrospectively simulate the citation screening process [18, 43, 85, 269, 329, 337, 346].

Active learning (AL) is an approach that aims to reduce the labelling cost by selecting the few most informative citations for classifier training. An AL system has a query module encoding logic to select these most informative citations. The query module interacts with an oracle (e.g., a human expert) to obtain labels for these selected citations. The goal is to select the samples that are the most informative to the model in order to improve its performance with a minimal number of labeled citations. Active learning is quite suitable for *de novo* SRs, which involve a fresh set of unlabeled citations. In such cases, AL can significantly reduce the labelling workload by assisting in selecting the most informative citations for human labelling. This approach could lead to more efficient resource utilization, improved work savings and thus cost-effectiveness. AL approaches have been extensively tested in retrospective citation screening scenarios using publicly available datasets across domains like biomedical, public health and software engineering. These approaches use either classical ML, deep learning or a combination of both to test multiple active learning parameters like starting and stopping criteria, query strategies, methods of initial training set sampling, and sampling methods, amongst others [50, 51, 192, 201, 238, 334, 345, 346, 365]. However, none of these approaches explore active learning in a prospective scenario, a business requirement.

Optimizing a prospective active learning system for citation screening requires considering multiple objectives. The first objective of such a system is to minimize the inclusion of irrelevant citations by excluding as many of them as possible. The second objective is to ensure a high recall for relevant citations, aiming to identify and retain at least 95% of them correctly. For medical and allied domains, the most used threshold is 95% to ensure satisfactory performance, and the threshold could be lower for other domains like software engineering. Performance in the first objective is commonly assessed using WSS (work-saved oversampling), which measures the reduction in effort achieved by automatically excluding irrelevant citations [66]. The recall of the minority class measures the second objective and assesses the ability to retain relevant citations correctly. In binary classification settings, a class with a disproportionately low number of instances is termed the minority class. This is often the case in citation screening datasets, and the count of relevant citations is significantly lower than irrelevant ones. Therefore, a successful active citation screening algorithm should retain as many relevant citations as possible and save time for the reviewers by removing irrelevant citations. In addition, the receiver operating characteristic area under the curve (ROC-AUC) score is an important evaluation measure that quantifies a classifier's effectiveness in distinguishing relevant and irrelevant citations. There could be instances where both WSS and minority recall are identical for the active classifiers. In such cases, ROC-AUC could be used as a tiebreaker. None of the above-mentioned research evaluates their approaches using all three crucial metrics for assessing

active classifiers.

Active learning tools like Abstrakr, asreview and EPPI reviewer have shown efficiency improvements, resulting in workload reductions ranging from 9% to 57% [122, 332, 334, 335]. van de Schoot *et al.* simulates an AL system that starts training only after the reviewer provides at least one relevant and irrelevant citation; however, they do not provide the initial cost incurred for selecting the relevant citations during the citation screening [334]. The researchers test default settings in their experiments and encourage practitioners to simulate the impact of different AL system settings like start criteria, query strategies, class balancing options, and seed sampling strategies. The current studies do not statistically test the impact of different AL system settings on performance. An active query sampler chooses citations and sends them to a human oracle for labelling, which incurs some cost. Semi-Supervision based pseudo-labeling could be effective in such scenarios where obtaining labeled data is costly, but unlabeled data is readily available. Nonetheless, none of those mentioned above studies listed except of [178, 201] explore the advantages and disadvantages of semi-supervision on active learning. These two studies do employ semi-supervision and show improvements in comparison to the results when using AL alone, but they do not address that a semi-supervision module uses pseudo-labels as ground truth to train a classifier and thereby could introduce possible errors into a prospective system, reducing its trust for deployment. Though comprehensive, none of the approaches assesses a prospective scenario for citation screening where active learning might not yield any results for very low prevalence studies with narrow research questions. We conduct this work to explore some research questions businesses have.

1. Does active learning save work and have expected recall of the relevant studies?
2. Is a smart query sampling methodology better than random sampling?
3. Is semi-Supervision effective?
4. Do some AL system settings perform better over the others?
5. In a prospective scenario, what could it cost to start training an AL system?

In a nutshell, this work contributes the following: I) The work systematically explores different AL system settings on 25 citation screening datasets to measure the effectiveness of each setting on performance metrics. II) The work explores semi-supervision to support active learning for citation screening. Semi-supervision strategies have the potential to reduce manual labelling costs. III) As we develop and retrospectively simulate the approach, we detail our experience in adapting such a methodology to prospective *de-novo* SRs for businesses. These details could help businesses focus on the relevant aspects of implementation. IV) The work proposes tracking multiple performance metrics and indicators to evaluate an active citation screening system comprehensively. V) We also provide open access to the code and our *in-house* citation screening datasets from the fields of physiotherapy and pharmaceuticals which can be used as test sets in future methodologies.

3.3.1 Methodology

In this section, we will explain the datasets used and delve into the working of the modules of the AL system. An overview of our methodology is shown in the Figure 3.2.

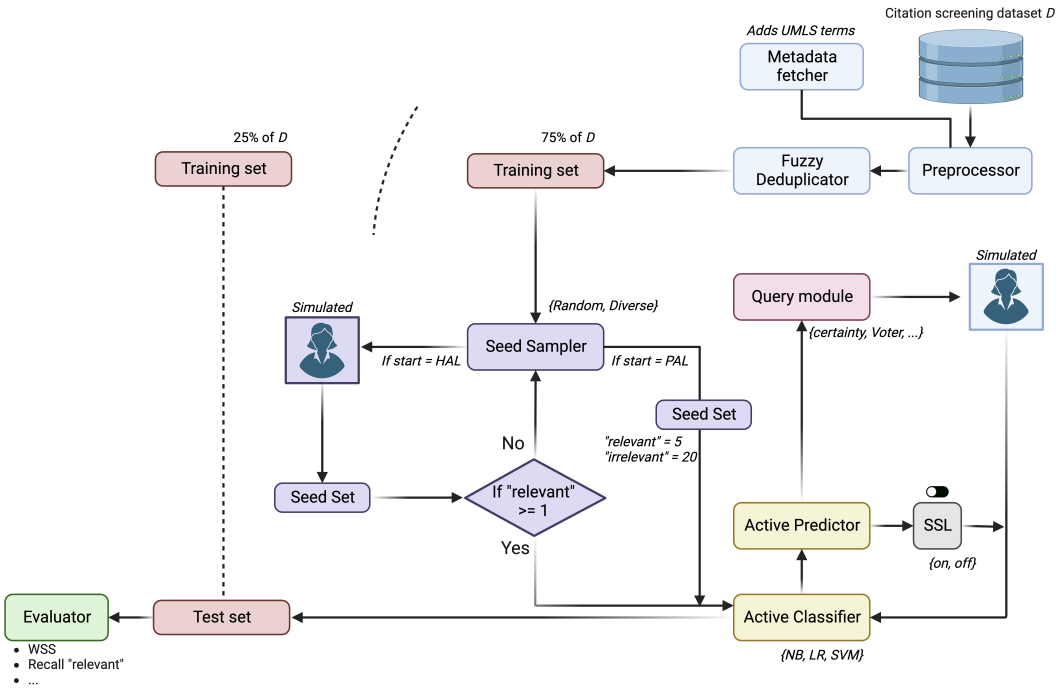


Figure 3.2: Semi-supervision supported active citation screening approach.

3.3.1.1 Datasets

To evaluate the effectiveness of different settings of the AL system, three distinct sets of citation screening datasets were employed. The first set was a subset of and obtained from the publicly accessible SYNERGY project, which is an openly available citation screening dataset with manually entered inclusion and exclusion decisions [76]. For this particular set, only those citation screening datasets where all citations contained a PubMed Identifier (PMID) were selected. This criterion facilitated the retrieval of corresponding titles and abstracts for the citations. The second set comprised four citation screening datasets that were *in-house* to F. Hoffmann-La Roche. The third set was sourced from the School of Health Sciences at HES-SO and centred around physiotherapy, rehabilitation and health education, thereby adding to the diversity of the evaluation process. This set encompassed two citation-screening datasets from published systematic reviews. One focused on the effectiveness of using Peyton's scale for health education, while the other compared the effectiveness of power training versus conventional resistance training on elderly patients [119, 330]. Each citation in the dataset had metadata information like author list, journal name, publication year, volume number, publisher, etc. The general characteristics of the datasets are summarized in Table 3.4.

3.3.1.2 System Modules

Preprocessor In the preprocessing phase, steps were implemented to transform the citations [23]. The first step involved tokenization and lemmatization, which were performed

[23] An individual citation comprises a title and an abstract

Sr.	Dataset (**identifier/topic**)	Prev	I	E
	Roche datasets			
1	**PoC1 NSCLC**	10	905	7911
2	**PoC2 CRPC**	22	878	3012
3	**SLR04 TNBC**	2.5	23	928
4	**SLR09 NSCLC**	6.5	1356	18783
	HES-SO datasets			
5	Giacomino **Peyton Physiotherapy**	3.17	14	442
6	**Tschopp Resistance training**	1.89	13	771
	Public datasets			
7	**Nelson** 2002 **Hormone therapy**	27.78	80	288
8	**Cohen** 2006 **Triptans**	3.71	24	647
9	**Walker** 2018 **Transgenerational inheritance**	1.2	765	47873
10	**Cohen** 2006 **Proton Pump Inhibitors**	3.98	51	1282
11	**Sep** 2021 **Rodent, Psychology**	17.17	40	233
12	**Cohen** 2006 **NSAIDS**	11.65	41	352
13	**Cohen** 2006 **Calcium Channel Blockers**	8.94	100	1118
14	**Wassenaar** 2017 **Bisphenol A**	1.46	111	7589
15	**Cohen** 2006 **Beta Blockers**	2.07	42	2030
16	**Cohen** 2006 **Statins**	2.51	85	3380
17	**Cohen** 2006 **ACE Inhibitors**	1.64	41	2503
18	**Chou** 2004 **Skeletal muscle relaxants**	0.55	9	1634
19	**Rooney** 2015 **Immunotoxicity**	2.93	54	1846
20	**Cohen** 2006 **Oral Hypoglycemics**	37.06	136	367
21	**Cohen** 2006 **ADHD**	2.41	20	831
22	**Cohen** 2006 **Antihistamines**	5.44	16	294
23	**Cohen** 2006 **Urinary Incontinence**	13.94	40	287
24	**Cohen** 2006 **Atypical Antipsychotics**	14.99	146	974
25	**Chou** 2003 **Oral opioids**	0.79	15	1900

Table 3.4: Gold standard citation screening datasets. Inclusion prevalence (Prev.) is the ratio of “includes” to “excludes”. I = “Includes” and E = “Excludes”

using the powerful spaCy [24] package [155]. Lemmatization normalizes individual words to their base form, ensuring consistency and coherence in the subsequent analysis. Preprocessing is a time-consuming step, and therefore, the datasets were preprocessed beforehand and stored locally for use in the experiments. Considering a prospective scenario, preprocessing can be performed on the fly, allowing each preprocessed citation to be indexed in a local database, e.g., Lucene [25], using a unique identifier, ensuring effective use in future citation screening projects. These citations would have a unique identifier and additional curated metadata such as author list, journal name, publication year, volume number, publisher, etc.

UMLS (Unified Medical Language System) keyword extraction was conducted for both the study titles and abstract texts using a third-party MetaMap [26] Python API [13, 311, 350]. By incorporating UMLS, the preprocessing method sought to capture the normalized medical topics for each study. A list of default stopwords from `en_core_web_sm` was utilized to eliminate insignificant words that might hinder the analysis. `en_core_web_sm` is a pre-trained spaCy model for the English language. As our experiments were a simulation, UMLS retrieval involved programmatic bulk requests to the MetaMap API. However, it is worth noting that this approach could become impractical when dealing with a large volume of texts due to the limitations UMLS imposes on the number of requests. However, a real-world scenario anticipates requesting UMLS API for smaller batches of studies at irregular intervals, closely mimicking human behaviour. This makes it more feasible to integrate UMLS retrieval into an automatic citation screening system for *de-novo* systematic reviews. The system could index UMLS terms as meta-data to a citation screening record.

Deduplicator module Duplicates refer to multiple records or entries in the citation screening dataset representing the same research study. These can occur due to multiple database searches (EMBASE, PubMed, etc.), indexed conference vs journal papers, updates or the same paper, errata, *etc.*. In certain studies, erratums are not considered duplicates. Duplicates can inflate the results of an SR, and consequently, deduplication is the process of removing such duplicates using the citation meta-data. State-of-the-art deduplication automation approaches require clean datasets with missing author names imputed with unknown, normalized author names (replace initials with full names, additional middle name initials), normalized journal names, add missing information about journal volume, page number, issue, etc [134]. Not everyone has ready access to clean datasets, and cleaning such data is time-intensive, making it a secondary priority. In business operations, duplicates can be found throughout the SR process and the aforementioned citation metadata is not cleaned until the data extraction process. Therefore, the reviewers deal with messy and incomplete data during the deduplication process. Noisy data anticipates journal name writing variations (abbreviation vs full-form), author name writing variations (initials of middle or last name or first name) and partial study title. Using DOI (Document Object Identifier) alone for deduplication could lead to false negatives, given a conference proceeding with many abstracts assigned the same DOI. Using PMIDs as unique deduplication identifiers is not feasible in all cases because not all studies are indexed in PubMed. Keeping this in mind, we used the following fuzzy deduplication module.

[24] `https://spacy.io/`

[25] `https://lucene.apache.org/`

[26] `https://github.com/lhncbc/skr_web_python_api`

Algorithm 1 Pseudo-code for fuzzy citation deduplication

Require:

Input: Dataset D of citations, where each citation C is represented as a tuple (T, A, J, Au, PD), where:

T: Title of the citation (string)
A: Abstract of the citation (string)
J: Journal name (string)
Au: List of authors (list of strings)
PD: Publication date (date)

Ensure:

Output: *Unique_citations*, a filtered dataset containing non-duplicate citations.

function CitationScreening(D)
  Initialize an empty list *Unique_citations*.
  for each citation C in Dataset D **do**
    if A is not empty **then**
      for each citation C' in Dataset D **do**
        Calculate Levenshtein similarity between abstracts A and A'.
        Calculate string similarity between titles T and T'.
        Calculate Jaro-Winkler similarity between journal names J and J'.
        Calculate Jaro-Winkler similarity between author lists Au and Au'.
        if (T is identical to T') **and** (A similarity $> 80\%$) **and** (J similarity $> 60\%$) **and** (Au similarity $> 80\%$)) **then**
          Identify the citation with the latest publication date among C and C'.
          Add the identified citation to *Unique_citations*.
        end if
      end for
    end if
  end for
  return *Unique_citations* as the filtered dataset containing non-duplicate citations.
end function

The deduplication algorithm takes a dataset of citations with its metadata. Each citation within the dataset is defined as a tuple, including the following items: the title (T) represented as a string, the abstract (A) in string format, the journal name (J) as a string, a list of authors (Au) in the form of strings, and the publication date (PD) as a date. The algorithm takes these tuples as input and aims to output a filtered dataset containing unique citations, removing duplicates. First, an empty list called "Unique citations" is initialised to store the unique references. Next, check whether the abstract (A) is empty for each citation C in dataset D. If A is not empty, it calculates the similarities between the titles, abstract, journal and author names for every citation pair. If the titles of these citations are identical, it checks whether the abstracts are more than 80% similar, the journal names are more than 60% similar and author names are at least 80% similar. If these conditions are met between the two candidate citations, the algorithm retains the latest published citation and adds it to the unique citations list. After the process is repeated for the citation pairs in the dataset, return the unique citations list as the filtered

dataset containing non-duplicate citations. For the similarity calculations, Levenshtein distance and Jaro-Winkler distance were used [191, 273, 347, 358]. It has to be noted that we remove citations with empty abstracts only from the active training process. This decision was based on the understanding that a citation with only a title lacks sufficient information for training a model or making an informed decision during human labelling. Details regarding the choice of similarity threshold are in the Supplementary material.

Seed Sampler A **seed sampler** module is employed for seed selection, which is the process of selecting a seed set. It takes citation data as input and produces a seed set comprising selected citations. The choice of seed sampling method can impact the performance of the Active Learning (AL) system. Two seed sampling methods were employed, namely random sampling and diversity sampling. Random sampling involves selecting a seed set through a random process, and diversity sampling selects the most representative citations from the pool. Diversity sampling achieved this via clustering the numerical representation of citations and selecting the citations near the centroid. We employed Affinity Propagation Clustering (APC), a clustering algorithm that measures the Euclidean distance between numerical representations of citation pairs as a similarity measure. We used Sklearn's Affinity Propagation module [27] without altering any default parameter settings. AP clustering, distinct from methods like K-means, doesn't require specifying the number of clusters as a hyperparameter. This made it suitable for our task, where the number of clusters was unknown and determining it through hyperparameter tuning would have been resource-consuming and impractical. The AP algorithm iteratively identified cluster centroids, also called exemplars, which could be used as the seed set [110]. APC outputs a variable number of cluster centroids, potentially resulting in more citations selected than required for the seed set. In such instances, the seed sampler randomly clipped the total number of selected citations to match the desired size of the seed set.

Start Criteria Functionality We explored the two most common active learning strategies based on the **start criteria**, hasty active learning (hasty start) and patient active learning (patient start). In patient active learning (PAL), the training begins only when at least five "relevant" citations were selected in the seed set. In hasty active learning (HAL), the training process begins with including at least one "relevant" citation. In real-world SRs, particularly when the research question is narrow, it is not uncommon to find only a small number of "relevant" studies among the total retrieved citations, ranging from one to ten. E.g., the dataset *Chou 2004 Skeletal muscle relaxants* has only 9 "relevant" citations out of a total of 1643 (refer Table 3.4). Applying PAL could pose a challenge because it requires at least five "relevant" citations to initiate training, leaving only four "relevant" citations for evaluating the experimental setup. Citation screening datasets with less than five citations leave no room to experiment with PAL.

Our experiments were divided into retrospective and prospective scenarios based on the **start criterion** they employed. A prospective scenario simulates a situation assuming an unlabeled citation screening dataset, while the retrospective scenario assumes a labelled dataset. We applied PAL exclusively in the retrospective scenario, assuming that our citations were labelled and selecting at least five "relevant" citations during the seed selection. In PAL, the seed set consisted of 25 fixed citations, comprising five "relevant"

[27] `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AffinityPropagation.html`

and 20 “irrelevant” ones. For HAL, we simulated a prospective scenario assuming an unlabeled dataset. During HAL seed selection, we sampled 25 studies in multiple iterations, and in each iteration, we kept track of whether at least one “relevant” citation was included. HAL-based seed sampler ensured at least one “relevant” citation was sampled in the seed set. In some datasets, the prevalence of “relevant” citations was very low, making it unlikely to find even a single “relevant” citation in a single iteration of 25 samples. Therefore, we limited the iterations to a maximum of 12, sampling 25 studies at each iteration, resulting in a maximum of 300 studies being sampled before initiating the active classifier training. It must be noted that inclusion prevalence is unknown in a prospective scenario; hence, setting a constant of max 300 studies to be sampled does not ensure that at least one “relevant” citation will be retrieved. We selected 300 using a simulation experiment, as explained in the Appendix. The settings for the start criterion were merged with the seed sampler.

Active classifier module Considering citation screening as a binary classification task of separating the input citations into “relevant” and “irrelevant”, we experimented with three prevalent binary classifiers: naive Bayes (NB), logistic regression (LR) and support vector machines (SVM). These classifiers took as input the numerical representation of a citation and output prediction probabilities for the respective classes. We used tf-idf (term frequency-inverse document frequency) features to represent the citations numerically. To optimize the performance of both the feature vectorizer and the classifier, we utilized scikit learn grid search cross-validation module [28] allowing us to tune the hyperparameters of the vectorizer and classifier simultaneously (details in Appendix).

In our study, we introduce the concept of a **data view**, representing the modalities on which an active classifier was trained. We experimented with three different data views, each obtained by concatenating individual vectors for the four modalities: title, abstract, title UMLS, and abstract UMLS. The first view comprised the title and abstract of a citation; the second view was represented by the UMLS terms from the title and abstract, and the third view comprised the title, abstract and their corresponding UMLS terms for a citation. Our active classifier module was a homogeneous ensemble of three classifiers to simulate the QBC (Query-by-Committee) approach. One classifier is trained on the first data view, another classifier on the second data view, and the third classifier incorporates all available data views, including titles, title UMLS, abstracts, and abstract UMLS (third data view). The ensemble is called homogeneous because all three classifiers used the same vector representation (tf-idf) and model architecture but were trained on different data views.

The predictor module of the trained classifiers made predictions on the remaining training set citations, with each predictor in the ensemble producing prediction probabilities for “relevant” and ”irrelevant” classes. These probabilities were aggregated by averaging across the three classifiers, and the resulting aggregated probability is used as the prediction probability for the citation. The final predicted label is marked as “relevant” if at least one of the three classifiers votes the citation as ”1”, as the citation screening process is recall-oriented for “relevant” citations.

[28] `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html`

Query sampler module The active classifier module produces prediction probabilities on the training citations, which are then used by the **query sampler module** to select the most informative citations depending upon the query strategy. The citations sampled by the query module are assumed to be more information and are used to train the active classifier. In this work, we explored five different query strategies: random sampling, uncertainty sampling (least confident, margin and entropy), certainty sampling, query by committee/simple voter, and diversity sampling. As the name goes, random sampling randomly samples the citations. Random sampling also acted as a reference comparison against the other smarter query samplers.

The least confident, margin, entropy, and certainty query samplers used predicted posterior probabilities of the training citations to select the samples. *Least confident query sampler* calculated the least confident score as $1 - P_\theta(y^x|x)$, where $P_\theta(y^x|x)$ is the maximum softmax probability. This query strategy selected the citations with the highest Φ_{LC} scores as they are deemed lowest of confidence, i.e., where the maximum probability was close to 0.5. For example, if we consider binary classification and the model predicts a 0.90 probability value for class-0 for a certain data sample, and thus for class-1, the probability value will be 0.1, we can say that the model was confident with the class membership or class assignment for this data sample. If the model predicted both the classes with almost similar probablities of 0.5, we can say that the model was uncertain on its predictions. The least confidence strategy considers only the most probable class for evaluation. *Margin sampling* strategy considers the two most probable classes, i.e., the classes having the highest and second-highest probabilities:

$$\Phi_M(x) = 1 - P_\theta(y_1^*|x) - P_\theta(y_2^*|x).$$

The *Entropy Sampler* prioritizes data points for which the model's predictions have the highest entropy. Higher entropy indicates a greater level of uncertainty in the model's predictions:

$$\Phi_{ENT}(x) = -\sum_y P_\theta(y|x) \log_2 P_\theta(y|x).$$

Certainty sampler selected the citations with the highest cumulative predicted probability Φ_C [241, 301]:

$$\Phi_C(x) = \sum_{i=1}^{n} P_\theta(y|x_i)$$

For the *diversity sampler*, we used the Affinity Propagation sampler as previously explained (See 3.3.1.2). The key idea behind the *voter query sampler* was to select data points that caused the most disagreement among the voting committee of the classifier ensemble. Given we had three members on the committee, even if one disagreed with the rest, we used it as a criterion to select the citation. When the training budget did not yield sufficient citations, and the voters disagreed, we calculated the entropy in the agreement using prediction probabilities. The entropy in the agreement between the voters was an average of the differences between the prediction probabilities of each voter over each predicted class.

Semi-Supervision Strategy The semi-supervision (SSL) module leveraged prediction probabilities from the active classifier on training citations to choose and label citations with the most confident predictions the module selected citations where the active classifier

Module	Experiment settings
Seed sampling	Random, Diversity
Start criteria	HAL, PAL
Active classifier	ND, LR, SVM
Query sampler	Random, Least Confident, Entropy, Margin, Certainty, Diversity, Voter
SSL benefit	on, off

Table 3.5: The experiment settings explored in the AL system.

had a prediction probability greater than or equal to 0.99 for either the "relevant" or "irrelevant" classes. This threshold helped ensure high confidence in the assigned pseudo-labels. The semi-supervision module can be toggled on or off, and when activated, it selects and assigns pseudo-labels to half of the training citations in the AL system. An SSL module could further reduce the manual labelling effort by automatically assigning high-confidence pseudo-labels to training instances.

3.3.1.3 Approach

For each dataset d listed in Table 3.4, UMLS (Unified Medical Language System) keyword extraction was performed on all citations using a third-party MetaMap Python API, followed by additional preprocessing steps [13, 311, 350]. The preprocessing involved deduplication using both citations and citation metadata. The AL system received the preprocessed and deduplicated citations, simulating the system in both retrospective and prospective scenarios. Before sampling the seed set and following deduplication, the citation dataset d was divided into 75% for training (t) and internal validation, and 25% for evaluation (e) (refer Figure 3.2). Depending on the sampling procedure, the seed sampler module selected a seed set. If PAL was the start criterion, the seed sampler simulated a retrospective scenario, assuming labels were available for querying and selecting 5 "relevant" and 20 "irrelevant" citations from t. For HAL seed selection, the seed sampler assumed the unavailability of labels, sampling 25 studies in multiple iterations over t". In each iteration, it kept track of whether at least one "relevant" citation was included. It's important to note that in our work, the citation screening datasets came with inclusion decisions and were pre-labelled. However, the seed set is sent to a human oracle for labelling. Once the start criterion was fulfilled and a predefined number of "relevant" citations were sampled, the active classifier module initiated the classifier training. The classifier predictor was then used to predict training citations t. The query sampler module then selected 10% of the training citations, then retraining the active classifier and regenerating probabilities on the remaining training citations. This process continued until 30% of training citations were sampled, after which the system halted training. The evaluator tracked performance indicators on the evaluation set at each 10% of the training steps. To test the effectiveness of each module, we conducted 168 experiments for each dataset corresponding to the two start criteria, three active classifier types, and seven query methods, with and without semi-supervision (ref Table 3.5). Each experiment was done over five random seeds, and the averaged results were reported. These experiments were designed to ensure we could probe the effect of individual modules, thereby addressing all the questions we aimed to answer.

3.3.1.4 Evaluation

Our system was multi-objective, and we tracked eight performance indicators throughout the experimentation. While most of these metrics were obtained on the 25% of the test set, some metrics, like Coverage, were monitored on the training set. These specific metrics are tracked at 10% intervals of the training dataset, up to a 30% training data threshold, but the results are reported based on the 30% of the training data. These metrics are collected across five-fold cross-validation for the five most common Python random seeds (0, 1, 42, 123, 1234).

Performance Indicators We tracked work saved over sampling @95% recall (or WSS@95% r) to evaluate the system on the amount of manual work saved by removing "irrelevant" citations [66]. The choice of 0.95 recall r was based on the following reasons: a) Achieving a perfect recall of 1.00 is not pragmatic by any text mining method since it would require reviewing all candidate studies manually. b) In the context of evidence-based medicine (EBM), a recall of 0.95 is generally deemed acceptable. This level of recall ensures that a significant proportion of "relevant" evidence is captured while still accounting for the limitations and practical considerations of the screening process.

$$\text{WSS@95\%} = \frac{\text{True Negative} + \text{True Positive}}{N} - (1 - r); r = 0.95 \tag{3.1}$$

Additionally, we kept a close eye on all the traditional metrics outlined in Sklearn's classification report, but placed an emphasis on the recall of "relevant" citations (micro-recall on class "1"). Together, WSS@95%, and micro-recall of the "relevant" class were vital evaluators of our system. ROC-AUC curve score was tracked and reported to measure a classifier's ability to separate "relevant" and "irrelevant" citations. ROC-AUC score can also act as a tiebreaker for identifying best-performing modules in cases where both WSS and recall are identical. Error of the semi-supervision module (hereafter SSL module error) helped us assess whether an SSL module should be implemented in a real-world citation screening system. SSL module error was measured as a $(1 - \textit{Accuracy})$. We also measured Coverage, as Miwa *et al* explained. Coverage indicates the ratio of "relevant" instances in the data pool annotated during active learning [238]. It must be noticed that Coverage was calculated on 75% of the dataset partitioned as the training set. The original formula has a FN^L in the denominator, but assume all labelled instances to be true. thereby setting FN^L to zero.

$$\text{Coverage} = \frac{\text{TP}}{\text{TP}^L + \text{FN}^L + \text{FP} + \text{TN}}; \text{FN}^L = 0 \tag{3.2}$$

We tracked the seed cost for HAL experiments, which is the number of citations used to initialize the active classifier training. It essentially tracks how many citations the reviewers need to annotate until they identify at least one "relevant" citation. Finally, we tracked the time to train a system, given that this could impact the production systems and human interaction for providing the next set of labels. We initially planned to include the tracking of Burden and Yield metrics as used in certain reference papers but decided to omit them to maintain simplicity [238, 345].

Significance Tests We employed t-test and Kruskal-Wallis tests to evaluate the significance of different modules. The Kruskal-Wallis test was employed to determine whether

significant differences existed among the performance metrics for the various query methods and classifiers we used. If the p-value was ≤ 0.05, we concluded that there were significant differences between the performance metrics of the module groups being investigated. To understand the differences between each module pair, we conducted Dunn's post-hoc analysis [95, 123]. Dunn's test involves comparing each group to every other group. The test calculated a p-value that assessed whether there were significant differences between those specific groups. We used a trimmed t-test to compare the means (for performance metrics) of two groups, which returned a p-statistic that we inferred the same as we inferred it for the Kruskal-Wallis test [366]. A trimmed t-test was used to compare seed sampling strategies, start criteria, and the effect of the semi-supervision module. The primary reason for selecting both these tests was the unequal sample sizes for the groups being compared.

Effect Sizes Additionally, we measured the effect size of different modules on the performance metrics using Cohen's-d. Cohen's d is a statistical measure used to quantify the effect size of a particular intervention on patient outcomes in clinical trials. For instance, consider [278], who utilized Cohen's d to assess the impact of occupational stress management intervention programs (intervention), reported effect sizes indicating a significant medium to large effect on clinical outcomes. Other clinical trials used it as well [142, 285]. In AI or machine learning, we could consider a module as an intervention or a variable and assess its effect on system's performance metrics and indicators. It provides valuable information on performance differences between the two groups. There were cases where one module did not significantly differ from the other according to these tests. However, in practical situations, we might need to select one module. We can effectively choose a module over the other if its effect size over a performance metric is higher than the other.

3.3.2 Results

Table 3.6 displays the citation counts in both the public institute and private company citation screening datasets before and after the deduplication and preprocessing steps. There are no empty abstracts in the public institute datasets, with the Physiotherapy dataset being the cleanest without duplicates.

Dataset	Deduplication		Empty Abstracts
	Before	After	
Public Institute Datasets			
Physiotherapy	456	456	0
Resistance training	724	701	0
Private Company Datasets			
PoC1 NSCLC	8817	8747	519
PoC2 CRPC	3890	3371	43
SLR04 TNBC	951	900	50
SLR09 NSCLC	20139	18476	1509

Table 3.6: Dataset statistics before and after deduplication step including removal of empty abstracts.

3.3.2.1 Question 1: Does active learning save work and have expected recall on the minority class?

	Dataset	Prev.	WSS	Recall 1 (std.)	Recall 0
	Private Company Datasets				
1	PoC1 NSCLC	11.43	0.78	0.71 ± 0.23	0.9
2	PoC2 CRPC	34.14	0.66	0.76 ± 0.15	0.86
3	SLR04 TNBC	2.62	0.92	0.15 ± 0.14	0.97
4	SLR09 NSCLC	7.16	0.81	0.56 ± 0.32	0.89
	Public Institute Datasets				
5	Physiotherapy	3.17	0.92	0.2 ± 0.22	0.97
6	Resistance training	1.89	0.93	0.23 ± 0.18	0.99
	Public datasets				
7	NSAIDS	11.65	0.84	0.53 ± 0.21	0.94
8	Proton pump inhibitor	3.98	0.9	0.28 ± 0.24	0.97
9	Calcium channel blockers	8.94	0.88	0.24 ± 0.2	0.94
10	ACE inhibitors	1.64	0.92	0.31 ± 0.25	0.97
11	Beta blockers	2.07	0.91	0.21 ± 0.21	0.97
12	Statins	2.51	0.9	0.3 ± 0.25	0.95
13	ADHD	2.41	0.93	0.1 ± 0.14	0.98
14	Antihistamines	5.44	0.93	0.03 ± 0.04	0.98
15	Atypical antipsychotics	14.99	0.84	0.31 ± 0.2	0.92
16	Triptans	3.71	0.84	0.16 ± 0.13	0.97
17	Oral hypoglycemics	37.06	0.76	0.32 ± 0.15	0.85
18	Urinary incontinence	13.94	0.83	0.46 ± 0.2	0.93
19	Oral opioids	0.79	0.93	0.13 ± 0.18	0.98
20	Skeletal muscle relaxants	0.55	0.94	0.01 ± 0.03	0.99
21	Transgenerational inheritance	1.2	0.91	0.22 ± 0.24	0.96
22	Immunotoxicity	2.93	0.91	0.36 ± 0.22	0.97
23	Rodent psychology	17.17	0.89	0.18 ± 0.11	0.96
24	Bisphenol A	1.46	0.9	0.44 ± 0.24	0.95
25	Hormone therapy	27.78	0.79	0.34 ± 0.14	0.89

Table 3.7: The table showing average WSS and binary recalls for "relevant" and "irrelevant" classes over all experimental configurations. Note: Prev. = Inclusion prevalence, Recall 1 = Recall of "relevant" class, std. = standard deviation, Recall 0 = Recall of "irrelevant" class.

All the results were generated from $168X5$ experiments per dataset for different AL setting combinations across five random seeds. The results presented in Table 3.7 address our first question with insights into the average work saved (WSS@95%r), recall on the "relevant" class (Recall 1), and recall on the "irrelevant" class (Recall 0) across the evaluation set e. These scores are averaged across all experiment combinations to allow us to examine the overall trends. The observed work saved ranges from 66% to 94%, while the average binary recall, for most datasets, remains consistently low ceiling at 76%. For

none of the datasets, the average recall reaches the expected 95%; in fact, none of the individual experimental configurations achieves a 95% recall either. Despite a satisfactory WSS, active learning falls short of the expected 95% recall. Additionally, the standard deviation for the recall ranges from 3% to 34%. This variability underscores the influence of different modules on recall and justifies the need for an in-depth analysis to identify the most impactful module settings. Notice the high recall values for the "irrelevant" class, revealing why there were high work savings. To reiterate, WSS measures how effectively the classifiers eliminate "irrelevant" citations, reducing the workload for human reviewers.

3.3.2.2 Question 2: Is a smart query sampling methodology better than random sampling?

While the overall work saved was generally satisfactory, the recall for the "relevant" class fell short of expectations. Therefore, to address the second question regarding whether smart query sampling outperformed naive random sampling, our focus shifts to evaluating the retention of the "relevant" class. Table 3.8 compares the performance of various query samplers *wrt.* (with respect to) the recall of the "relevant" class. When considering absolute recall scores, certainty and diversity sampling stand out, performing the best on 10 (40% of all datasets) and 11 (44% of all datasets) out of 25 datasets, respectively, of which on 3 of the 25 datasets both the query methods had tied. One of the other sampling methods tends to perform the best on the remaining datasets. However, when evaluating significance through the Kruskal-Wallis test, diversity (via clustering) sampling only demonstrates a significant improvement over certainty sampling in 7 out of 11 datasets (63.63% of datasets). Conversely, certainty sampling significantly outperforms cluster sampling for 4 out of 10 (40%) datasets. Surprisingly, none of the query methods outperformed the rest on physiotherapy and pharmaceutical datasets. Notably, for the Beta Blockers, ACE inhibitors dataset and oral opioids, both certainty and diversity sampling perform significantly better than the other methods but are on par with each other. For the Physiotherapy dataset, certainty sampling doesn't have a significant advantage over cluster sampling, while both vastly outperform the rest of the query methods. For the Rooney dataset, even though Least Confident Sampling performs better in absolute numbers, it does not significantly outperform margin, entropy, and voter sampling methods. Although our expected recall of 95% is not achieved for either of the datasets using any of the query strategies, decomposing and inspecting the recalls for individual query methods reveals that specific query methods do impact recall. Looking at Tables 3.7 and 3.8 show that each of the best-performing query methods exhibited a consistent increase in overall recall, leading up to a 4.8 per cent points increase. For example, the overall recall of 0.49 for NSAID increases by 14 percent points to 0.63 using diversity sampling. These results confirm that Diversity and Certainty sampling methods outperform naive random sampling across these biomedical, pharmaceutical, and physiotherapy datasets.

3.3.2.3 Question 3: Is semi-supervision effective?

Table 3.9 displays the average recall and WSS values with and without semi-supervision benefits to the AL system. The table also illustrates the SSL module's contribution to the error in the system. Regarding average recall, aiding the AL system with SSL benefit outperformed on 20 out of 25 datasets; overall recall with and without SSL benefit was only 0.266 percentage points. The standard deviation over absolute recall with and without the SSL module is between $3.2\% - 34.3\%$, pointing towards the variable capability

	Dataset	Cer	Div	Ent	LC	Mar	Rand	Vote
	Private Company Datasets							
1	PoC1 NSCLC	0.63	0.74	0.73	0.73	0.73	0.73	0.69
2	PoC2 CRPC	0.65	0.77	0.78	0.77	0.78	0.78	0.76
3	SLR04 TNBC	0.2	0.13	0.13	0.15	0.14	0.15	0.15
4	SLR09 NSCLC	0.54	0.58	0.58	0.57	0.58	0.57	0.54
	Public Institute Datasets							
5	Physiotherapy	0.23	0.22	0.05	0.04	0.04	0.2	0.12
6	Resistance training	0.29	0.22	0.24	0.22	0.21	0.24	0.21
	Public datasets							
7	ACE inhibitors	0.37*	0.37*	0.27	0.3	0.28	0.31	0.3
8	ADHD	0.20*	0.12	0.06	0.06	0.05	0.11	0.07
9	Antihistamines	0.06	0.02	0.02	0.02	0.02	0.02	0.02
10	Atypical antipsychotics	0.33	0.37*	0.3	0.29	0.29	0.31	0.3
11	Beta blockers	0.27*	0.26*	0.17	0.18	0.19	0.19	0.19
12	Calcium channel blockers	0.28	0.27	0.23	0.22	0.22	0.26	0.22
13	NSAIDS	0.51	0.63*	0.53	0.51	0.53	0.51	0.52
14	Oral hypoglycemics	0.32	0.39*	0.29	0.32	0.31	0.3	0.3
15	Proton pump inhibitor	0.26	0.23	0.32	0.3	0.31	0.27	0.29
16	Statins	0.31	0.34	0.29	0.3	0.29	0.29	0.29
17	Triptans	0.18	0.14	0.14	0.14	0.16	0.16	0.18
18	Urinary incontinence	0.45	0.5	0.46	0.45	0.47	0.45	0.47
19	Oral opioids	0.21*	0.21*	0.09	0.08	0.08	0.12	0.14
20	Skeletal muscle relaxants	0.01	0.01	0.01	0.01	0.02	0	0
21	Immunotoxicity	0.36	0.36	0.37	0.37	0.38	0.32	0.37
22	Rodent psychology	0.18	0.21	0.2	0.16	0.19	0.16	0.19
23	Hormone therapy	0.34	0.42*	0.33	0.31	0.33	0.36	0.31
24	Transgenerational inheritance	0.23	0.2	0.25	0.23	0.23	0.22	0.23
25	Bisphenol A	0.31	0.45	0.46	0.48	0.47	0.41	0.48

Table 3.8: Table reporting the binary recall on "Relevant" class for different query sampling methods. Note: Cer. = Certainty, Div. = Diversity, Ent. = Entropy, LC. = Least Confident, Mar. = Margin, Rand. = Random, Vote. = Voter. Underline denotes the query method performs the best for a dataset in terms of absolute recall. An asterisk (*) denotes the query sampling method performs significantly better than the rest.

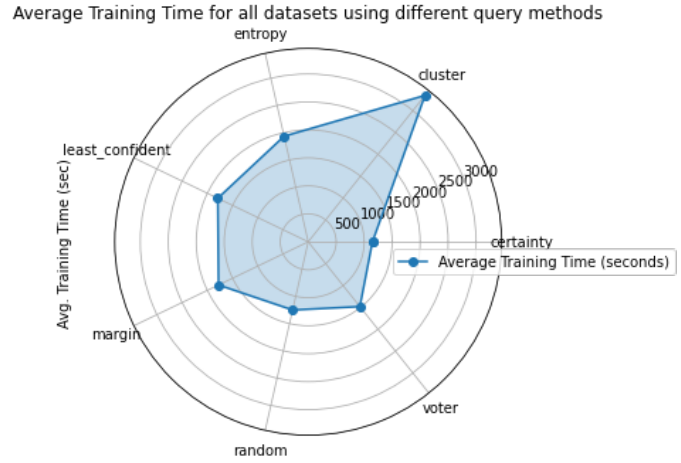


Figure 3.3: Radar chart plotting the average time taken by each query sampling method. Diversity sampling method is represented in the figure as cluster sampling.

of active classifier algorithms employed in this work. Similarly, overall WSS with and without the SSL benefit was only 0.156 percentage points. Aiding active learning with the semi-supervision module did not significantly [29] increase either the work saved or recall, thus bringing no clear benefit to the system. In contrast, in the SSL-aided experiments, the SSL module error ranged from 1.3% to 29.4%, and the error values had a standard deviation value between 2% to 8.3% again pointing towards the variable capability of active classifiers. Across different domains, the SSL module did not show significantly superior performance in physiotherapy, biomedicine, and pharmaceutical datasets. Our experiments and results show that the SSL module, while saving half the labelling cost, also has accumulated errors in the system. This raises questions about the deployment worthiness of adding an SSL module to the system, addressing our next question regarding the practicality and effectiveness of such an addition.

3.3.2.4 Question 4: Do some AL system settings perform better over the others?

In addressing the question of whether certain settings in an AL system yield better performance than others, we conducted tests to evaluate the impact of different active classifiers on performance, comparing the outcomes of the hasty (HAL) and patient (PAL) start approaches and finally the impact of seed sampler choice. Table 3.10 presents the average recall and WSS scores for both the hasty (HAL) and patient (PAL) start approaches. Patient start, which triggers classifier training with five "relevant" and 20 "irrelevant" citations, surpassed the hasty start. PAL performed significantly better on 15 of the 25 (60%) datasets and had a higher effect size on recall over 18 out of 25 (72%) datasets. On the other hand, HAL significantly outperformed PAL in terms of work saved over 12 out of the 25 datasets. On 11 of 25 datasets where HAL does not significantly outperform PAL, HAL and PAL have identical absolute work saved. Significance was calculated using a trimmed t-test. These results confirm that PAL outperforms HAL regarding recall across these biomedical, pharmaceutical, and physiotherapy datasets.

[29] Significance was calculated using trimmed t-test.

	Dataset	Recall "relevant" SSL	No SSL	Error
	Private Company Datasets			
1	PoC1 NSCLC	**0.722** (±0.238)	0.706 (±0.216)	0.128 (±0.055)
2	PoC2 CRPC	**0.763** (±0.155)	0.748 (±0.147)	0.17 (±0.037)
3	SLR04 TNBC	**0.154** (±0.15)	0.146 (±0.138)	0.048 (±0.035)
4	SLR09 NSCLC	**0.573** (±0.336)	0.557 (±0.314)	0.143 (±0.076)
	Public Institute Datasets			
5	Physiotherapy	**0.206** (±0.223)	0.195 (±0.214)	0.051 (±0.022)
6	Resistance training	**0.263** (±0.188)	0.203 (±0.174)	0.03 (±0.016)
	Public datasets			
7	Hormone therapy	**0.348** (±0.148)	0.337 (±0.135)	0.234 (±0.026)
8	Triptans	**0.159** (±0.127)	0.156 (±0.134)	0.046 (±0.045)
9	Transgenerational inheritance	**0.222** (±0.242)	0.221 (±0.233)	0.046 (±0.055)
10	Proton pump inhibitors	0.28 (±0.241)	0.283 (±0.237)	0.068 (±0.039)
11	Rodent psychology	0.18 (±0.114)	0.185 (±0.115)	0.156 (±0.028)
12	NSAIDS	**0.539** (±0.211)	0.529 (±0.206)	0.106 (±0.027)
13	Calcium channel blockers	**0.245** (±0.214)	0.238 (±0.196)	0.121 (±0.066)
14	Bisphenol A	**0.446** (±0.253)	0.428 (±0.233)	0.058 (±0.083)
15	Beta blockers	**0.223** (±0.223)	0.192 (±0.188)	0.049 (±0.054)
16	Statins	**0.309** (±0.255)	0.294 (±0.242)	0.066 (±0.073)
17	ACE inhibitors	**0.331** (±0.256)	0.297 (±0.248)	0.037 (±0.039)
18	Skeletal muscle relaxants	0.009 (±0.032)	0.009 (±0.032)	0.013 (±0.016)
19	Immunotoxicity	**0.373** (±0.224)	0.35 (±0.225)	0.049 (±0.04)
20	Oral hypoglycemics	**0.329** (±0.149)	0.311 (±0.143)	0.294 (±0.036)
21	ADHD	0.095 (±0.146)	0.097 (±0.141)	0.043 (±0.034)
22	Antihistamines	**0.027** (±0.046)	0.025 (±0.042)	0.055 (±0.02)
23	Urinary incontinence	**0.474** (±0.195)	0.456 (±0.204)	0.137 (±0.028)
24	Atypical antipsychotics	**0.312** (±0.209)	0.311 (±0.193)	0.161 (±0.056)
25	Oral opioids	0.115 (±0.173)	0.153 (±0.19)	0.022 (±0.041)

Table 3.9: The table demonstrating the recall values for "relevant" class and WSS with and without semi-supervision benefit. **Bold** values in the recall column denote the best absolute recall with and without semi-supervision benefit.

	Dataset	Prevalence	Recall "Relevant" HAL	Recall "Relevant" PAL	WSS HAL	WSS PAL
	Private Company Datasets					
1	PoC1 NSCLC	11.43	0.72	0.71	0.78	0.79
2	PoC2 CRPC	34.14	0.76	0.76	0.66	0.66
3	SLR04 TNBC	2.62	0.12	**0.18***	**0.93***	0.92
4	SLR09 NSCLC	7.16	**0.57**	0.55	0.81	0.81
	Public Institute Datasets					
5	Physiotherapy	3.17	0.17	**0.21***	0.92*	0.92
6	Resistance training	1.89	0.21	**0.25***	**0.94***	0.93
	Public datasets					
7	ACE inhibitors	1.64	0.28	**0.34***	**0.93***	0.92
8	ADHD	2.41	0.07	**0.11***	0.93*	0.93
9	Antihistamines	5.44	0.03	0.02	**0.94***	0.93
10	Atypical antipsychotics	14.99	0.31	0.31	0.84	0.84
11	Beta blockers	2.07	0.18	**0.23***	**0.92***	0.91
12	Calcium channel blockers	8.94	0.23	**0.25**	0.88	0.88
13	NSAIDS	11.65	0.51	**0.55***	**0.84***	0.83
14	Oral hypoglycemics	37.06	0.32	0.31	0.76	0.76
15	Proton pump inhibitor	3.98	0.26	**0.31***	**0.91***	0.9
16	Statins	2.51	0.29	**0.32***	0.9	0.9
17	Triptans	3.71	0.12	**0.19***	0.83	**0.85**
18	Urinary incontinence	13.94	0.47	0.46	0.83	0.83
19	Oral opioids	0.79	0.09	**0.16***	**0.94***	0.93
20	Skeletal muscle relaxants	0.55	0.01	0.01	0.94	0.94
21	Immunotoxicity	2.93	0.34	**0.39***	**0.92***	0.91
22	Rodent psychology	17.17	0.17	**0.19***	**0.89***	0.88
23	Hormone therapy	27.78	0.34	0.33	0.79	0.79
24	Transgenerational inheritance	1.2	0.22	**0.23***	0.91*	0.91
25	Bisphenol A	1.46	0.42	**0.46***	0.9	0.9

Table 3.10: Average Recall for the "relevant" class and WSS for hasty and patient start. Bold means the absolute performance was the best (HAL vs. PAL) and asterisk (*) means the module performed significantly better. **Bold** denotes the start criterion functionary performs the best for a dataset in terms of absolute recall and WSS.

As shown in the scatter plot 3.4, diversity seed sampling significantly outperformed in absolute recall on 20 of 25 datasets. In comparison, the random seed sampling had significantly better work saved over 20 of 25 datasets. The good performance of the diversity seed sampler was quite unsurprising because the diversity query sampler too outperformed the random sampler by a large margin, evident from Table 3.8.

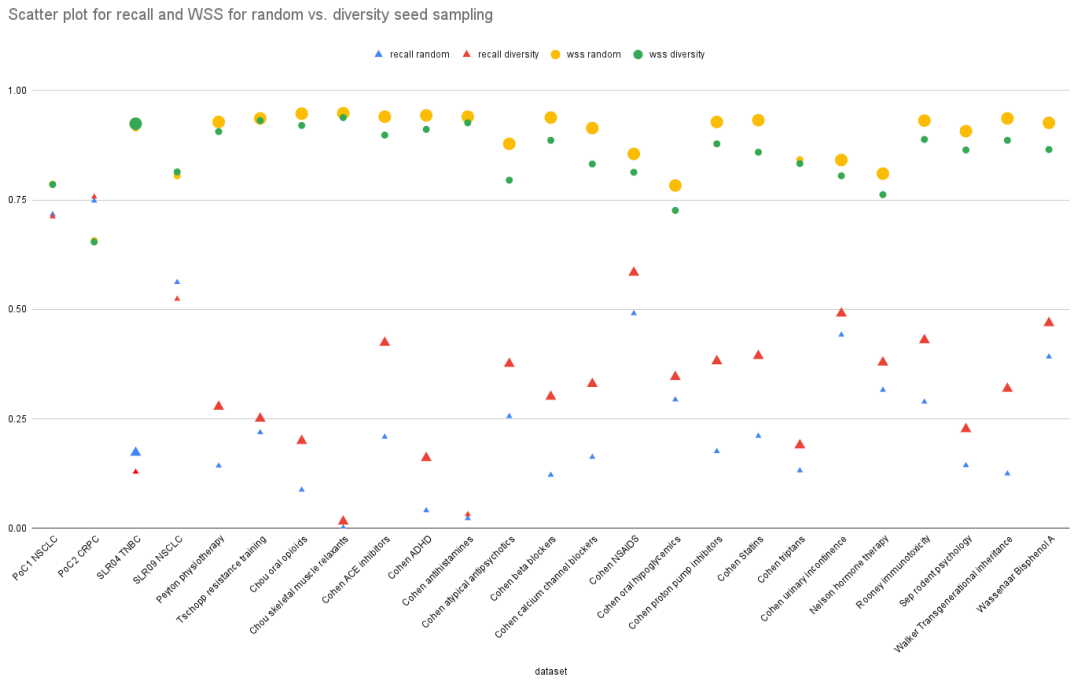


Figure 3.4: Scatter plot for the average absolute work saved and recall for both the seed sampling methods. Note: A larger circle shows the method performed significantly better over work saved over a dataset and a larger triangle shows the method performed significantly better on the recall of the "relevant" citations over a dataset.

Table 3.11 shows the average recall for the "relevant" class and semi-supervision module error values over individual classifiers: NB, LR and SVM. We present SSL error for classifiers because the quality of pseudo-labels depends on the active classifier's ability to distinguish between "relevant" and "irrelevant" instances. These results confirm that Naive Bayes outperforms LR and SVM across these biomedical, pharmaceutical, and physiotherapy datasets. Furthermore, we do not observe domain-wise differences in the performance of NB over recall.

3.3.2.5 Question 5: In a prospective scenario, what could it cost to start training an active learning system?

In the context of prospective business scenarios, we assume that inclusion decisions are unavailable for the citation screening dataset, whether a patient start or a hasty start. Consequently, sampling a seed set incurs a seed cost, determined by the number of citations a reviewer must label before identifying one "relevant" citation for a hasty start and five "relevant" citations for a patient start. We present the average seed cost plotted over ten

	Dataset	Recall "Relevant" class NB	LR	SVM	SSL Error NB	LR	SVM
	Private Company Datasets						
1	PoC1 NSCLC	**0.938***	0.833	0.401	0.187	0.11	0.074
2	PoC2 CRPC	**0.894***	0.803	0.591	0.212	0.149	0.151
3	SLR04 TNBC	**0.297***	0.147	0.017	0.081	0.035	0.027
4	SLR09 NSCLC	**0.849***	0.754	0.108	0.225	0.128	0.063
	Public Institute Datasets						
5	Physiotherapy	**0.398***	0.217	0.004	0.074	0.045	0.034
6	Resistance training	**0.401**	0.283	0.105	0.044	0.024	0.024
	Public datasets						
7	ACE inhibitors	**0.547***	0.361	0.084	0.068	0.026	0.016
8	ADHD	**0.168***	0.094	0.021	0.07	0.033	0.027
9	Antihistamines	**0.046***	0.035	0.001	0.07	0.053	0.042
10	Atypical antipsychotics	**0.472***	0.389	0.076	0.207	0.155	0.122
11	Beta blockers	**0.398***	0.234	0.037	0.093	0.035	0.018
12	Calcium channel blockers	**0.413***	0.289	0.033	0.177	0.101	0.085
13	NSAIDS	**0.693***	0.597	0.327	0.124	0.094	0.102
14	Oral hypoglycemics	**0.453***	0.384	0.148	0.329	0.29	0.262
15	Proton pump inhibitors	**0.459***	0.352	0.029	0.098	0.058	0.047
16	Statins	**0.569***	0.323	0.036	0.127	0.044	0.027
17	Triptans	**0.25***	0.175	0.054	0.078	0.036	0.024
18	Urinary incontinence	**0.638***	0.551	0.232	0.156	0.128	0.126
19	Hormone therapy	**0.485***	0.364	0.196	0.247	0.238	0.217
20	Immunotoxicity	**0.571***	0.371	0.177	0.085	0.032	0.03
21	Rodent psychology	**0.233***	0.246	0.062	0.176	0.146	0.146
22	Oral opioids	**0.222***	0.108	0.015	0.048	0.01	0.008
23	Skeletal muscle relaxants	**0.023***	0.003	0	0.023	0.009	0.007
24	Transgenerational inheritance	**0.487***	0.187	0.001	0.097	0.028	0.012
25	Bisphenol A	**0.709***	0.41	0.22	0.14	0.023	0.012

Table 3.11: The table demonstrating classifier-wise recall for "relevant" class and error values for semi-supervision module. An asterisk (*) denotes a classifier significantly outperformed the rest. **Bold** values in the SSL error columns denote lowest average error.

random seeds against the inclusion prevalence values of all datasets in Figure 3.5 (left side figure). The top-left graph illustrates the average seed cost for PAL against the inclusion prevalence, while the bottom-left graph depicts the corresponding values for HAL. As anticipated, the average seed cost decreased with increased inclusion prevalence for both HAL and PAL. Specifically, for HAL, manual screening was required for an average of 159.72 citations for the lowest prevalence dataset (prevalence = 0.551). In contrast, PAL necessitated labelling over 275 citations on average for the lowest prevalence dataset. For the highest prevalence dataset (prevalence = 37.057), PAL demanded an average iteration through 26.38 citations, while HAL required sampling at least 25 citations to encounter one "relevant" citation. On the right side of Figure 3.5 are the line graphs plotting the average number of "relevant" citations sampled in the seed set against inclusion prevalence. The top-right graph displays these details for PAL, while the bottom-right graph shows the corresponding details for HAL. As inclusion prevalence increased, the number of "relevant" citations sampled in the seed set also increased. To find at least one "relevant" citation, a minimum inclusion prevalence of 0.789 was required. To find five or more "relevant" citations, the minimum inclusion prevalence was 3.709, as indicated by the red dotted line.

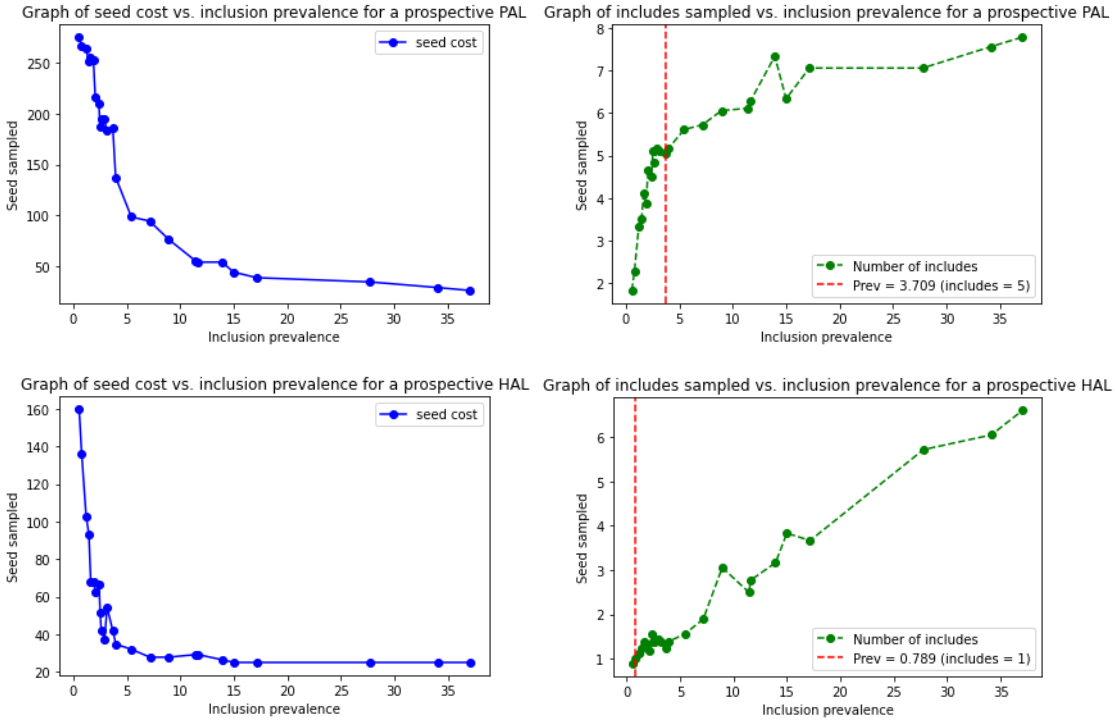


Figure 3.5: Average seed cost for hasty and patient active sampling vs. inclusion prevalence (Left). Line graph showing number of "relevant" samples sampled in the seed set vs. the inclusion prevalence (Right).

3.3.3 Discussion

We first discuss some aspects of our methodology, especially the UMLS retrieval and the preprocessing steps. As mentioned, UMLS was retrieved in bulk for all the citation screening datasets. However, this bulk retrieval approach could become impractical when dealing with a large volume of text due to the limitations imposed by UMLS on the number of

requests. In a real-world scenario, the expectation would be to request UMLS API for smaller batches of studies at irregular intervals, which closely mimics human behaviour. This makes it more feasible to integrate UMLS retrieval into an automatic citation screening system for *de-novo* systematic reviews. Such a system could index UMLS terms as metadata to a citation screening record stored as a unique record in a local database, e.g., Lucene [30]. In our simulation, preliminary preprocessing too was performed on all the citation screening datasets in bulk. Considering a prospective scenario, preprocessing can be performed on the fly, allowing each preprocessed citation to be indexed in the aforementioned local database, using a unique identifier, ensuring effective use in future citation screening projects. This makes it easier for the diversity seed sampler to sample the seed set citations efficiently. In addition, these citations would have additional curated metadata such as author list, journal name, publication year, volume number, publisher, etc.

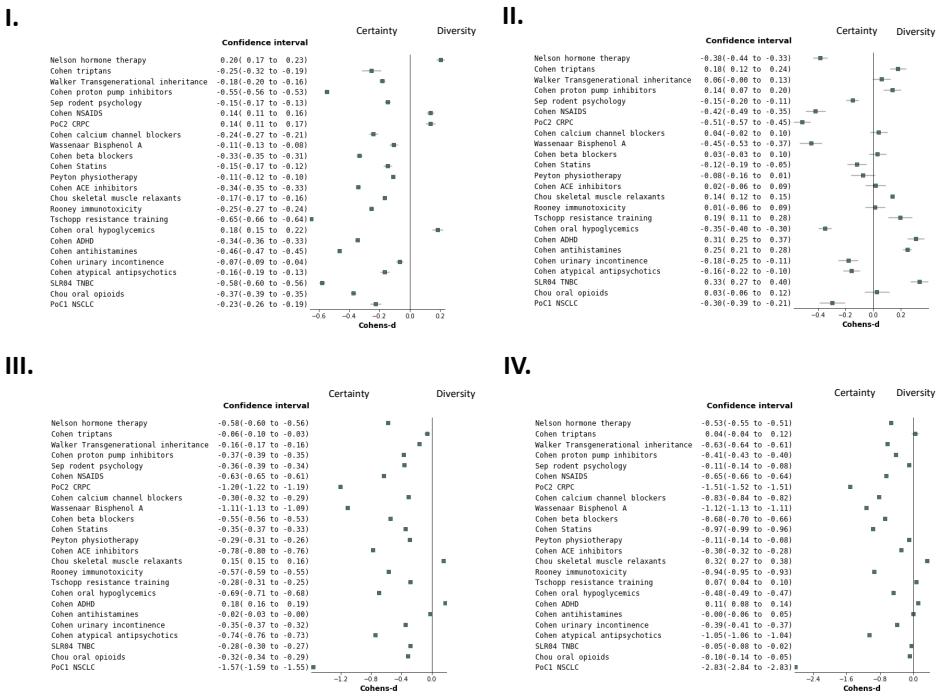


Figure 3.6: Forest plot comparing the effect size of certainty sampling and cluster sampling on I) WSS@95% r, II) Recall of "relevant" class, III) Macro-F1 score, and IV) ROC-AUC score.

The Forest plot in figure 3.6 shows a pragmatic comparison between certainty and diversity sampling methods. Considering the effect size of both sampling methods on the multiple performance indicators using Cohen's-d, it became evident that certainty sampling outperformed diversity sampling over 21 of 25 datasets in terms of effect size on WSS, over 23 of 25 datasets in terms of effect size on macro-F1 and 22 of 25 datasets on the ROC-AUC score. Certainty sampling thus showcased a higher overall algorithmic efficacy. Diversity sampling has a higher effect size on recall for the "relevant" class over 13 of the 25 (52%) datasets, while certainty sampling is on the other half. Lightly.ai showed that using diversity sampling to select instances for few-shot training ML models

[30] `https://lucene.apache.org/`

improved the model performance by up to 4.6x per additional labelled batch compared to random selection [316]. Apart from the performance comparison, it must be noted that the diversity sampling tool takes almost four times as much time as the other query methods. Diversity sampling takes an average of 55.78 minutes per experiment, requiring nearly three times as much time as certainty sampling, which takes an average of 19.13 minutes. This observation suggests significant scaling challenges in practical applications.

We investigated the strength of the correlation between the Coverage of "relevant" instances and their recall. Figure 3.7 displays the Pearson correlation coefficient scores between Coverage and recall for the "relevant" class over all the datasets. Among the 25 datasets, nine demonstrated a negative correlation between Coverage and recall. Meanwhile, two out of 25 datasets displayed a high correlation, six had a moderate correlation, and the remaining datasets exhibited varying degrees of low correlation. It is safe to assume that Coverage does not consistently impact the recall of "relevant" citations. More than the Coverage, it could be the proportion of "relevant" and "irrelevant" training citations causing a class imbalance that could impact the recall.

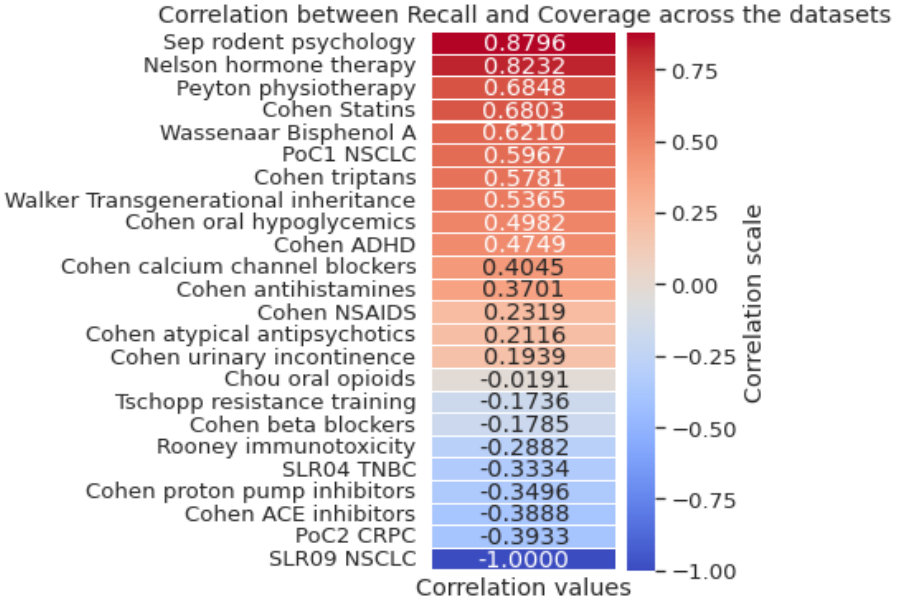


Figure 3.7: The figure illustrates Pearson correlation coefficient scores between individual coverage and recall for the "relevant" class over all experiments.

Class imbalance has been a challenge in binary classification tasks and is a persistent challenge in citation screening too [92, 162, 183, 368]. In the context of our study, it might be a contributing factor to HAL's lower recall compared to PAL in capturing relevant instances. PAL assumes a retrospective scenario and samples five fixed "relevant" and 20 "irrelevant" citations in the seed set. In contrast, HAL assumes a prospective scenario and requires a sample of a seed set only with at least one "relevant" citation and n "irrelevant" citations. This difference in approach may result in HAL having a smaller ratio of relevant to irrelevant citations in its seed set, leading to a comparatively larger class imbalance. The right part of the figure 3.8 shows inclusion prevalence for seed set across all the prospective hasty start experiments in this work compared to the average seed set inclusion prevalence of retrospective patient start (red dotted bar). The figure shows that

about 90% (green dotted bar) of the HAL experiments have a higher class imbalance in the seed set compared to the PAL experiments. This initial class imbalance could contribute to HAL's lower retention of the minority "relevant" class and, hence, the lower recall. On the left side of Figure 3.8, the seed set inclusion prevalence is depicted across both HAL and PAL in a prospective scenario, demonstrating a similar class imbalance across experiments. The observation also raises whether PAL outperformed HAL because it was tested in a retrospective setting, maintaining a manageable class imbalance. We leave this as an open question, suggesting the need for further exploration and testing of start criteria in prospective settings in future research.

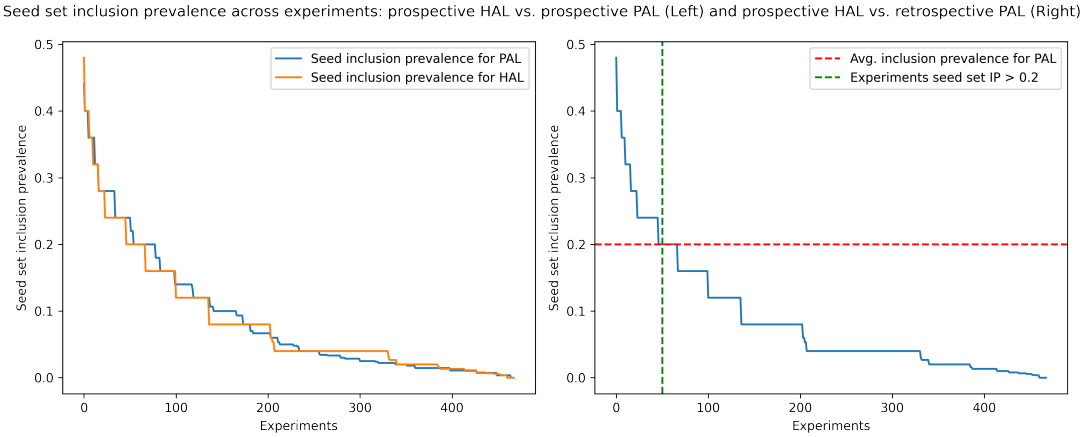


Figure 3.8: Line graph plotting seed cost inclusion prevalence across hasty active learning experiments.

Naive Bayes had the highest recall values for the "relevant" class but also high on error values. Regarding semi-supervision module error, SVM outperformed the other classifiers and consistently had the lowest overall error, while NB had consistently higher error, as shown in Figure 3.9. It is vital to track the error of the SSL module in biomedical business systems as this error could propagate and impact health policies. However, we have yet to encounter an active learning study that uses semi-supervision for citation screening and reports the error propagated by semi-supervision modules. The superior recall values achieved by Naive Bayes, a generative algorithm, raise the possibility of testing active learning for few-shot training language models for citation screening. This remains an open question, suggesting a potential avenue for future exploration.

Limitation: One limitation of our study is the utilization of English-language citation screening datasets. Therefore, the NLP components of the system, like the stopword removal component, relied on the English language transformer model. There are multiple candidate solutions to overcome this limitation. The easiest solution is mapping the language of the citation screening dataset to the corresponding spaCy language model. E.g., for a Spanish citation screening dataset, use `es_core_web_sm` [31] instead of `en_core_web_sm`. Another solution is translating the citations from the source language to English using neural machine translation approaches, albeit understanding that the quality of translation could depend upon language pairs for translation and also upon the length of the input text [55, 167, 323].

[31] `https://spacy.io/models/es`

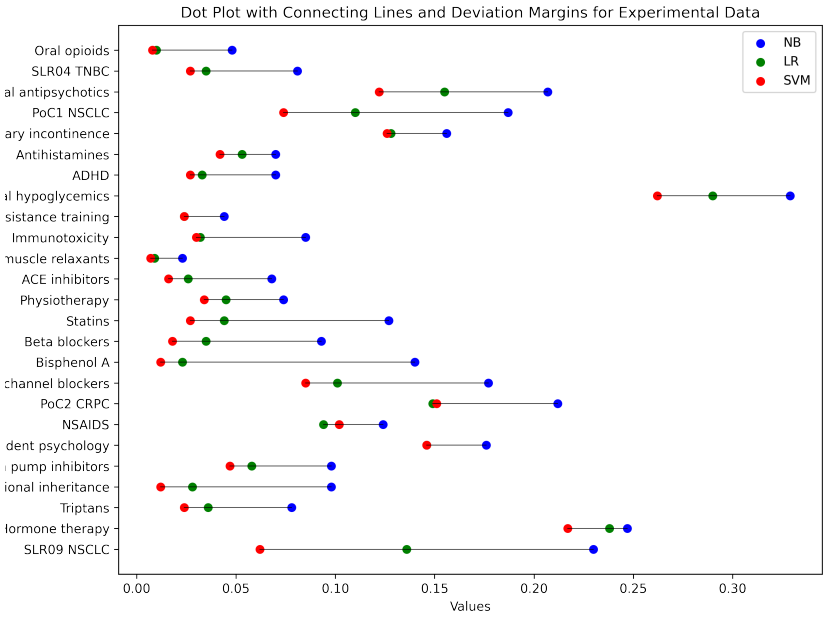


Figure 3.9: Dot plot plotting average semi-supervision module error for NB, LR and SVM for the datasets.

Our fuzzy deduplication module logic was based on empirical observations, as explained in the Supplementary material. As deduplication was not the focal point of this research, we aim in future to benchmark our deduplicator module on a larger dataset and optimize it with and without clean meta-data.

3.3.4 Conclusion and Future Work

We draw careful conclusions from our results and propose a multi-objective AL system with the following components. We also introduced a metric called seed cost, which measures the average number of citations needing manual labelling before at least one "relevant" citation is discovered for starting to train the prospective active learning system. We suggest a diversity sampler compared to naive random sampling for the seed sampling process. At this point, NB, a generative learning algorithm, performed best; therefore, we suggest NB as a classifier. The performance of NB has prompted our future work. We plan to test whether large language models (LLMs) like GPT-4 (Generative Pretrained Transformer) could be prompted to chaff out "relevant" from "irrelevant" citations using few-shot learning. Given the error it could propagate in the results, we do not suggest using an SSL benefit for the system, which is unaffordable for medical SRs.

The aim of this work investigating the active learning systems was to reduce the number of citations needing hand labelling for every *de-novo* SR question. AL still requires labelling a subset of the retrieved citations. Instead, using the PICO framework, LLMs could be employed to generate synthetic citation screening datasets representative of "relevant" and "irrelevant" classes. This eliminates the need for any manual annotation of citations. The synthetic dataset could be used to train active learning or machine learning

models for binary classification. Abogunrin *et al.* recently carried out the first steps in this direction, testing the feasibility of using ChatGPT to synthetically generate abstracts that mimic peer-reviewed journal-looking abstracts [2]. Finally, we will evaluate our fuzzy deduplication module for a larger data set in the absence of clean and complete metadata.

3.4 Chapter Conclusions

This chapter delved into methodologies for citation screening, specifically focusing on semi-supervised supported prospective active learning. This approach provided an early insight into the system through various performance indicators. In Section 3.2, we explored word embeddings and highlighted the potential impact of class imbalance and overlap on performance. In Section 3.3, we introduced a multi-objective active learning system simulated in a prospective scenario. The system results indicated that employing smart sampling methods improved recall when using Naive Bayes, a generative model. Given their generative nature, these sampling methods could be effectively applied to few-shot prompt Language Model (LLM) approaches for citation screening.

Chapter 4

PICO Information Extraction from Clinical Trials

4.1 Introduction

The previous chapter explored a prospective active learning-based binary classification approach to automate citation screening. Another perspective for automating citation screening involves examining it through the lens of PICO. The criteria for including a citation into an SR is decomposed into whether all or most predetermined (inclusion criteria) PICO information is present in the citation [287]. ML algorithms can help automate the recognition of PICO information from clinical trial studies by directly pointing the human reviewers to the correct PICO descriptions in a document. To effectively design and employ ML approaches to PICO extraction, one must first understand the nature of PICO. PICO information comprises broad categories (refer Table 2.1), illustrated in Figure 4.10, which further include subcategories. For instance, "P" represents clinical trial participant information, which is then decomposed into details such as participants' disease condition, gender, age, sexual orientation, ethnicity, social status, overall sample size, and sample sizes within different clinical trial groups. The intervention information category is subdivided into the type of intervention, its role in the clinical trial (active or placebo), dose and duration of administration, mode of administration, and the intervention administrator. Similarly, the outcome information category is broken down into whether the outcome is subjective, objective, or a hard outcome like mortality (natural or unnatural causes). It also includes details about the outcome measurement device used and the unit of measurement [90]. In summary, PICO information is inherently characterized by its highly fuzzy and compositional nature.

In this chapter, we explore strategies and methodologies for PICO+[32] information extraction from clinical studies, mainly RCTs. In the chapter section 4.2, the possibility of using multi-task learning (MTL) for fine-grained PICO information extraction was explored. Section 4.3 introduces a distant supervision methodology for extracting the "Intervention" term using freely available resources. Section 4.4 outlines developing and evaluating a weakly supervised information extraction workflow focused on PICO entities. In the next and the final section 4.5, the PICO weak supervision workflow was extended to an additional "study type and design" entity.

Segments of this chapter have been published as conference papers and a journal

[32] "+" denotes additional entities

paper [82, 87, 89, 90]. [87] is submitted to a conference. In [82], my contribution was formulating the tasks as multi-task learning, designing the experimental setup with transformer and LSTM models, executing the experiments, analyzing the results, and reporting the findings in a conference paper. In [89], my contribution was conceptualizing the idea and framework. This involved crawling and locally dumping the data from `https://clinicaltrials.gov/` repository into the local Lucene database. I also designed and executed distant supervision framework and experiments, devised and adapted the string matching approach for "Intervention" information labelling, analyzed results, and presented the findings in a conference paper. In [90] and [87], I adapted PICOS information extraction as a fuzzy, weakly supervised task. The process involved defining the weak supervision framework, identifying and repurposing the sources of weak supervision, manually and semi-automatically mapping them to PICO targets, designing experiments using generative label model and transformer models, analyzing results, and reporting findings in a journal paper and a conference paper.

1. The code and Hilfiker physiotherapy dataset manually annotated with coarse and fine-grained PICO information and introduced in [82] could be found on GitHub and Zenodo, respectively.
    - `https://zenodo.org/records/6961986`
    - `https://github.com/anjani-dhrangadhariya/multitask-pico-detection`
2. The distantly-labelled dataset and the code for the DISTANT-CTO approach introduced in [89] could be found on Zenodo and GitHub, respectively.
    - `https://zenodo.org/records/6961986`
    - `https://github.com/anjani-dhrangadhariya/distant-cto`
3. The silver standard pseudo-labelled dataset and the weak supervision approach introduced in [90] could be found on DRYAD and GitHub, respectively.
    - `https://datadryad.org/stash/dataset/doi:10.5061/dryad.ncjsxkszr`
    - `https://github.com/anjani-dhrangadhariya/distant-PICO`
4. The code to reproduce the results in [87] could be found at GitHub.
    - `https://github.com/anjani-dhrangadhariya/distant-studytype`

4.2 Multitask learning for PICO Information Extraction

This section details experiments conducted to explore the potential of using MTL to extract PICO and its subcategories.

An ML model can help automate PICO information extraction from clinical trial studies by directly pointing the human reviewers to the correct PICO descriptions in a document. However, as explained in the preface of Section 4.1, the detected coarse-grained PICO descriptions are further delineated into fine-grained semantic units (see Figure 4.1). This means that even after a machine points a human reviewer to the correct coarse-grained PICO description, the reviewer must manually read and understand its finer aspects to screen the study for relevance. This leads to the semi-automation of the process. Fully

I.	II.
... A semistructured interview was used to obtain qualitative information on the effect of the intervention. The convenience sample included {15 adult Oncology outpatients, 13 female and 2 male, ranging in age from 20 to 87} [PARTICIPANT]...	... A semistructured interview was used to obtain qualitative information on the effect of the intervention. The convenience sample included {15} [P:SAMPLE SIZE] {adult} [P:AGE] {Oncology} [P:CONDITION] outpatients, 13 {female} [P:SEX] and 2 {male} [P:SEX], ranging in age from {20 to 87} [P:AGE] ...

Figure 4.1: Example of I. coarse-grained annotated participant span and II. further delineated fine-grained "Participant" entities (P = Participant).

automating the citation screening process requires identifying, delineating, and normalizing the fine-grained PICO mentions, allowing for machine reasoning over the extracted semantic units. Unlike in many biomedical journals, fine-grained PICO mentions in the broader health literature are neither clearly identified nor standardized as semantic units (e.g. naming conventions for interventions and outcome measurement), making it an even more tedious process for the reviewers [141]. This hampers machine reasoning over the semantic units, leading to barriers for full automation.

This work explores and proposes end-to-end neural attention models that require no hand-engineered features, unlike the previous approaches and are trained to improve the recognition of fine-grained PICO entities. This approach achieves state-of-the-art (SOTA) performance for fine-grained "Participant" and "Outcome" entity recognition. In this work, the approach to fine-grained PICO recognition was considered a sequence labelling task for which two different setups were tested: single-task learning (STL) setup and multi-task learning (MTL) setup. It was investigated whether these model setups trained on the PICO benchmark corpus extend to reaching similar performance for an *in-house* PICO-annotated corpus from the physical therapy domain (hereafter: physiotherapy corpus). The key takeaway from the error analysis and corpus exploration was that the PICO benchmark corpus over-represents pharmaceutical entity labels, leading to poor performance on any low-frequency entities, especially the non-pharmaceutical entities coming from domains of physiotherapy, complementary therapies and in the more general health domain. Automating PICO recognition is far more challenging than open-domain named-entity recognition (NER) because there are disagreements between human experts on the exact words that make up PICO elements. Additionally, PICO recognition cannot be purely labelled as a NER task because "Participant" entities span entire sentences.

4.2.1 Related Work

As our work focuses on recognizing fine-grained semantic PICO mentions, this section reviews previous PICO recognition approaches.

PICO elements were first proposed for building structured clinical questions [159, 279]. Since then, studies have explored their use for IR with the potential to automate relevant screening for SRs. Research towards automatic PICO detection peaked with the exploration of several methods including rule-based lexical approaches [74, 80], language models (LM) [36], support vector machines (SVMs) [37, 75], graphical models like CRF [54, 175], shallow neural (Multilayer Perceptron MLP) approaches [35, 135], a combination of machine learning and rules [53, 54] and deep neural approach like LSTMs [166]. These studies, however, used small annotated corpora, heavy text pre-processing and hand-engineered features.

The availability of a comparatively large, and probably the only, PICO benchmark corpus from [254] with multi-grained (fine and coarse-grained) PICO annotations opened up possibilities to explore the neural models. They used this corpus to train baseline models using hand-engineered features to detect fine- and coarse-grained entities separately. Their baselines achieved a competitive performance on the coarse-grained PICO but a poor performance on the more difficult, semantic, fine-grained entities [33]. SciBERT, through domain-adaptation improved [34] the overall coarse-grained PICO recognition for the EBM-NLP corpus [28]. [171] used the EBM-NLP pretrained parameters to improve coarse-grained PICO recognition over a small *in-house* dataset compared to a randomly initialized LSTM-CRF model. [41] combined sentence-level PICO recognition with relevance screening, leveraging the use of this corpus.

A few studies dived into recognizing finer aspects of PICO but did not focus on all of them together. For instance, the DNER (Disease NER) [369] neural model focused on disease-mention extraction, [359] focused on extraction of patient demographics (sex, sample size, disease), [61] explored extraction of different intervention arms from RCTs (randomized controlled trials). Except [254], prior work either focused on coarse-grained or sentence-level PICO detection. Fine-grained PICO detection has not yet garnered as much attention as it should, given its potential for fully automating the SR screening phase through machine reasoning.

The focus of our work is to improve recognition of fine-grained PICO entities, test feasibility and competency of MTL models utilizing joint information from the fine- and coarse-entity annotation, and improve generalization by introducing inductive bias [20, 49]. Additionally, the MTL and STL models trained on the EBM-NLP benchmark corpus were used to evaluate fine-grained performance using an *in-house* corpus from physiotherapy and rehabilitation.

4.2.2 Method

This section describes our motivation for using the MTL approach, the datasets used to train and evaluate it, the end-to-end MTL system (refer Figure 4.2), and its components.

4.2.2.1 Multitask learning

We assume that both fine- and coarse-grained PICO entity recognition are closely related tasks because fine- entities are essentially nested under the coarse spans (see Figure 4.1). These closely related tasks could serve as mutual sources of inductive bias for each other opening up the possibility to jointly train them using a multitask learning approach [20, 49]. MTL has also shown to leverage performance on nested biomedical named-entities (NEs) for example for the GENIA corpus [103, 104, 372].

In the STL setup, multiple models separately learn to detect the coarse- and fine-grained entities each for the population, intervention/control and outcome using individual label structure. In contrast, an end-to-end MTL system jointly learns to recognize fine- and coarse-grained entities by exploiting the similarities and differences between the task characteristics. MTL is particularly suitable for this case because the fine-grained entities are nested under the coarse-grained spans both drawn from the same chunk of text. This opens up the possibility to improve recognition of poorly performing [35] fine-

[33] https://ebm-nlp.herokuapp.com/

[34] https://paperswithcode.com/sota/participant-intervention-comparison-outcome

[35] https://ebm-nlp.herokuapp.com/#Leaderboard

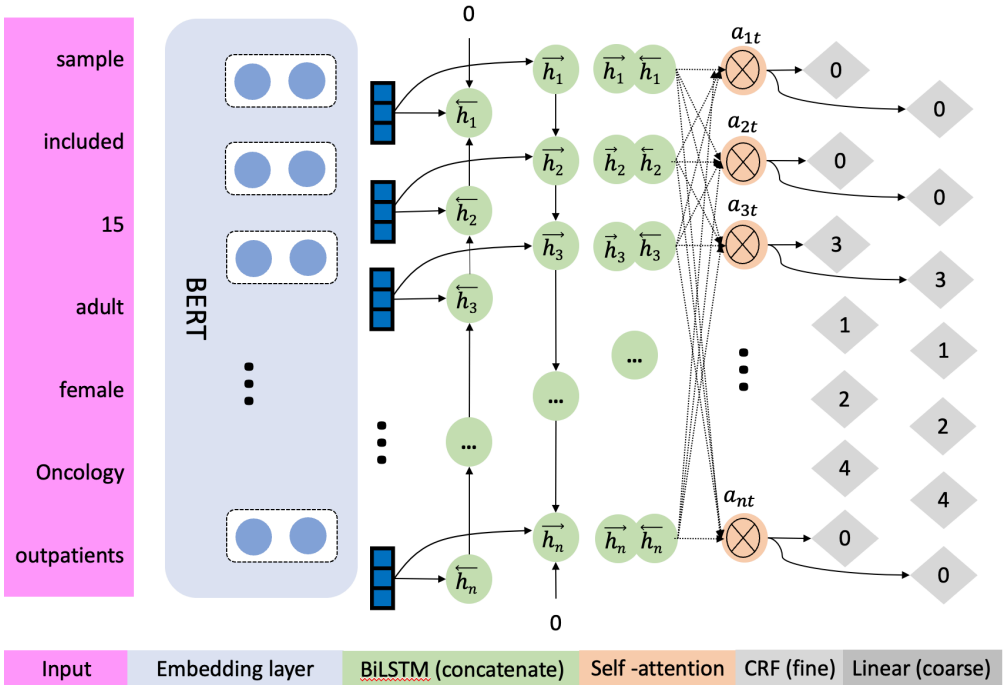


Figure 4.2: The proposed end-to-end MTL approach with fine-grained recognition as the main-task and coarse-grained as the auxiliary task. Removing either of the CRF decoder heads gives the respective STL setups.

grained recognition by sharing the hidden representation with the far better performing coarse-grained task. For comprehensive details on the MTL algorithms in NLP read [286].

In our MTL setup, fine-grained PICO recognition was considered as the main-task and involved assigning each token in the input text with the fine-grained PICO class labels (see Table 4.1). Coarse-grained recognition was considered as an auxiliary or helping task and involved assigning each token in the input text with either 1 ("Participant" or "Intervention" or "Outcome") or 0 ("No Label"). For both tasks, 0 ("No Label") was considered as the out-of-the-span label. Both tasks shared the encoder system components using the concepts from hard-parameter sharing but used separate decoder heads and loss calculation. To understand the effect of each of the functional components for the MTL setup, we began training simple models and sequentially added more layers to understand the improvement effect. To compare performance of each MTL setup, a corresponding STL setup was used. To probe the effect of the self-attention component individually on both the tasks in the MTL setup two ablation experiments were performed.

4.2.2.2 Dataset

EBM-PICO test set: We used the EBM-PICO corpus comprising ~5000 coarse- and fine-grained PICO-annotated documents [36] to train and test the end-to-end system (see Figure 4.1 and Table 4.1). A part of the dataset was annotated by crowd-sourcing and a small part by medical experts. It comes pre-divided into a training set comprising 4,993 documents and a test set comprising 191 that was used for evaluation. More details about the dataset can be found in [254].

[36] A single document consists of a title and an abstract.

	Participant	count	Intervention/Comparator	count	Outcome	count
0	No label	124372	No label	120453	No label	115578
1	Age	708	Surgical	659	Physical	7215
2	Sex	157	Physical	1988	Pain	180
3	Sample size	661	Drug	4424	Mortality	261
4	Condition	3893	Educational	1328	Side effect	540
5			Psychological	62	Mental	1657
6			Other	323	Other	2064
7			Control	542		

Table 4.1: Coarse-grained P (Participant), I (Intervention) and O (Outcome) labels are delineated into respective fine-grained labels. Annotation counts are shown in the table.

Physiotherapy and Rehabilitation test set: An additional test set comprising 153 documents in an *in-house* SR titled "Exercise and other non-pharmaceutical interventions for cancer-related fatigue in patients during or after cancer treatment: an SR incorporating an indirect-comparisons meta-analysis" was manually annotated by the first author using the annotation instructions[37] available from [150, 254]. The primary purpose of this additional test dataset was not to establish any inter-annotator agreement (IAA) but 1) to understand the complexity and noise encompassed in the multi-grained PICO annotation process and 2) to test the feasibility of the proposed setups trained on the general medical (EBM-PICO) dataset to predict PICO classes for a corpus from physiotherapy and rehabilitation domains. The vitality of this annotation exercise will be apparent in the discussion section (see Section 4.2.5). IO (Inside, Outside) or raw labelling was used for both sequence labelling tasks.

4.2.2.3 System Components

1. Embeddings: Instead of random weight initialization or using word embeddings, initialization with pretrained contextual language models (LM) have proven to benefit multiple NLP tasks [169]. Contextual representations like BERT [81], ULMFit [156], and GPT [271] encode rich syntactic and semantic information from the English language text into vectors eliminating the need for heavy feature engineering. The proposed model setups used BERT to extract dense, contextual embeddings. Healthcare corpora contain several domain-specific vocabularies often absent in word embeddings generated from general-purpose corpora. In this case, the corpus words not present in the word embeddings are either dropped during vector computation or a specialized OOV (out-of-the-vocabulary) or UNK (unknown) token vector is used instead [169]. BERT mitigates the problem of OOV words by using WordPiece tokenizer [115] and byte pair encoding (BPE) [300], respectively.

$$e_t = BERT(x_t) \tag{4.1}$$

In Equation 4.1, e_t are the vectors extracted from the encoded text tokens x_t using BERT and is used as the input for the downstream layers. t corresponds to the individual time steps (tokens) and ranges from 1 to n.

[37] `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6174533/bin/NIHMS988059-supplement-Appendix.pdf`

2. Feature transformer: Each document in the training set was composed of ∼500 tokens. To encode long-term dependencies and learn a task-specific text structure, the model stacked a single bidirectional LSTM (BiLSTM) layer on top of the embedding layer [152]. A forward LSTM ran from left-to-right (LTR), encoding the text into a ($\overrightarrow{h}$) vector using the current token embedding input e_t and the previous hidden state h_{t-1}. A backward LSTM does the same from right to left (RTL). Both outputs were shallow concatenated ($[\overrightarrow{h}; \overleftarrow{h}]$) into h_t and used as the input for the next layer.

$$(\overrightarrow{h}) = LSTM(e_t, h_{t-1}) \tag{4.2}$$

$$(\overleftarrow{h}) = LSTM(e_t, h_{t+1}) \tag{4.3}$$

3. Self-attention: Next, the model stacked a softmax-based multihead self-attention layer that calculated for each token in the sequence a weighted average of the feature representation of all other tokens in the sequence [338]. A self-attention layer was added with the aim to improve the signal-to-noise ratio by out-weighting important tokens in comparison to the noise tokens in the text. Self-attention weights for each token were calculated by multiplying hidden representation h_t with randomly initialized Query q and Key k weights, which were further multiplied with each other to obtain attention weights. Finally, the obtained scaled attention weights were multiplied with the Value (V) matrix which was obtained by multiplication between a randomly initialized weight matrix v and h_t.

$$Q = h_t * q \qquad K = h_t * k \qquad V = h_t * v \tag{4.4}$$

$$a_t = Attention(Q * K^T) * V \tag{4.5}$$

4. Decoder: After weighting BiLSTM output using self-attention, the representation is either fed to a fully-connected layer to predict tag emission sequence followed by calculation of weighted cross-entropy loss or to a CRF layer along with the ground-truth tag sequence y_t. CRF is a graph-based model suitable for learning tag sequence dependencies from the training set and has shown to perform better than a softmax classifier [5, 160, 207]. During training, the CRF layer learns to predict the correct tag sequence and computes loss. During inference, the trained CRF layer is used to predict the tag sequence $\hat{y}_t$. In the MTL setup, only the final CRF and/or fully-connected layers were different for both the tasks while all the other layers were shared.

$$CRF_{loss} = CRF(linear(a_t), y_t) \tag{4.6}$$

4.2.2.4 Experiments

To compare our proposed methodology on fine-grained PICO recognition, two strong baselines from Nye *et al.* were used. The baselines use a combination of n-grams, part-of-speech tags, and character embeddings as features and used them to separately train a logistic regression model and a neural LSTM-CRF. To demonstrate the feasibility of the MTL approach for improving fine-grained recognition using the auxiliary coarse-grained task and to compare the performance of each MTL setup, exactly identical STL setups were used. The setups are:

I. BERT Linear setup includes a linear transformation layer stacked on top of the $\text{BERT}_{\text{BASE}}$ model followed by weight-balanced cross-entropy loss calculation.

II. BERT LSTM CRF setup uses $\text{BERT}_{\text{BASE}}$ for feature extraction followed by an LSTM and a linear layer to generate emission probabilities that feed into the CRF decoder head that learns tag sequence dependencies and calculates loss.

III. BERT BiLSTM CRF setup is identical to setup II, but BiLSTM replaces the LSTM layer.

IV. BERT LSTM atten CRF setup incorporates a single self-attention head. Attention weights calculated by the attention head are applied to the output of the LSTM layer followed by a linear transformation to generate emission probabilities. These probabilities feed into the CRF decoder.

V. BERT BiLSTM atten CRF setup is identical to the setup IV, but BiLSTM replaces the LSTM layer.

VI. BERT BiLSTM Multihead atten CRF setup differs from setup V in how attention-weights are applied. For MTL, this setup uses a single-head attention-weighted BiLSTM representation to decode coarse-grained entities while a two-head attention-weighted BiLSTM representation is used to decode the fine-grained entities. This was to over-weigh the fine-grained signals.

VII. BERT BiLSTM Multihead atten: setup has specific settings for the MTL and STL. In the MTL setup, CRF is used as a decoder for the fine-grained task. The coarse-grained task includes a linear layer followed by a weighted cross-entropy loss calculation. As STL cannot have a coarse-grained task, the encoder setup was used with a linear layer as the decoder for the fine-grained task. Similar to the previous setup, to decode the coarse-grained sequence, a single-head attention-weighted BiLSTM representation was used, while it was a two-head attention-weighted BiLSTM representation to decode the fine-grained entities.

In the MTL setup, all except the final decoding layer shared the parameters for the main and auxiliary tasks. For decoding, the final shared hidden representations were fed to two separate decoding heads that calculated the losses separately for both tasks. The back-propagated loss was a linear combination of both task losses ($\mathcal{L}oss = \mathcal{L}oss_{coarse} + \mathcal{L}oss_{fine}$). For the STL setups without any shared representation between the tasks, the models were optimized using these individual task losses.

Experimental Details: Each model was trained for 15 epochs with a mini-batch size of 6 and maximum sequence length of 512. Last four layers of BERT embeddings were summed before passing to the next layer. BERT was fine-tuned without freezing the weights. Hidden size for LSTM/BiLSTM was set to 512/1024. Model training was optimized using AdamW using a learning rate of 5e-5. The gradients were clipped to 1.0 to mitigate exploding gradients problem. For the models that used weighted cross

entropy loss calculation, balanced class weights were calculated using sklearns' compute class weights function [38].

Ablation experiments: To probe the effect of attention weights individually on the fine- and coarse-grained tasks in the MTL setup, two ablation experiments each were performed. For the experiments, the linear transformation was directly applied to the BiLSTM layer without attention-weighting and this unweighted BiLSTM output was first used for the main task and in the second experiment for the auxiliary task.

4.2.3 Evaluation

Similar to the other PICO recognition studies, the F1 score was evaluated and reported per token for comparison. Each F1 score is an average of individual fine-grained categories for PICO. The F1 score serves to compare: 1) the performance of our methodology with the baseline, 2) the performance of STL *vs.* MTL for the fine-grained PICO recognition, and 3) the performance improvement brought by the additional functional layers for the MTL and STL setups. A t-test was applied as a significance test with a Bonferroni corrected p-value ($\alpha_{altered}$) threshold set to 0.007 to the normally distributed F1 scores for each MTL model and its corresponding STL counterpart for the fine-grained task [94, 114].

4.2.4 Results

F1 scores for the EBM-PICO and physiotherapy corpus are reported in Table 4.2. In most setups, STL significantly outperforms MTL. For the EBM-PICO corpus, in terms of the cumulative PICO F1, the MTL setup VII outperforms the STL counterpart, but only by gaining a 4% boost in F1 for the "Intervention" recognition while deprecating the performance on the "Participant" entity. Compared to the MTL setup V, setup VI gains 3% F1 on the "Participant" and "Outcome" recognition by exploiting the two-head attention-weighted BiLSTM outputs exclusively for decoding the fine-grained output *vs.* only a single head for decoding the coarse-grained output. Setup VII further improves the performance for the "Intervention" by switching to a linear decoding layer that uses the weighted cross-entropy loss. In comparison to the baseline, both setups outperform for "Participant" and "Outcome".

For evaluation on the physiotherapy corpus, MTL again seems to exploit the two-head self-attention exclusively on the fine-grained task (*vs.* only a single head on the coarse-grained task) and linear decoding followed by weighted cross-entropy loss calculation for the coarse-grained task to achieve a similar performance as STL. The MTL setup VII obtains 2% better F1 scores for the "Participant" and "Intervention" classes. MTL outperforms STL only by carefully exploiting task weights, weighted loss, task-specific decoder heads. Ablation experiments (see Table 4.3) show that the performance boost for the MTL setup is brought by cumulative attention weighting for both decoding tasks. Removing attention weights from either of the decoding heads reduces the F1 score. This effect of weights on the tasks was also observed in the experiments of [49] where the MTL benefited from the weighted hidden layers on the input, the rationale being that weighted input when backpropagated carried more information.

[38] `https://scikit-learn.org/stable/modules/generated/sklearn.utils.class\_weight.compute\_class\_weight.html`

	Setup	MTL F1			STL F1		
	Fine-grained	P	I/C	O	P	I/C	O
	EBM-PICO evaluation corpus						
b1	logistic regression	-	-	-	0.45	0.25	0.38
b2	LSTM-CRF	-	-	-	0.4	**0.5**	0.48
I	BERT Linear	0.21	0.07	0.09	0.20	0.08	0.12
II	BERT LSTM CRF	0.33	0.24	0.37	0.45	0.27	0.45
III	BERT BiLSTM CRF	0.39	0.28	0.40	0.52	0.27	0.53
IV	BERT LSTM attn CRF	0.34	0.28	0.47	0.53	0.25	0.49
V	BERT BiLSTM attn CRF	0.51	0.30	0.53	**0.54**	**0.30**	**0.57**
VI	BERT BiLSTM multihead attn CRF	**0.54**	0.30	**0.56**	**0.54**	0.29	0.55
VII	BERT BiLSTM multihead attn linear	0.52	**0.34**	**0.56**	**0.54**	**0.30**	**0.56**
	Physiotherapy corpus						
I	BERT Linear	0.23	0.07	0.05	0.22	0.07	0.06
II	BERT LSTM CRF	0.36	0.15	0.20	0.52	0.15	0.27
III	BERT BiLSTM CRF	0.40	0.17	0.24	0.57	0.19	0.27
IV	BERT LSTM attn CRF	0.37	0.14	0.28	0.56	0.17	0.27
V	BERT BiLSTM attn CRF	0.57	0.17	**0.30**	**0.60**	0.19	**0.30**
VI	BERT BiLSTM multihead attn CRF	**0.62**	0.18	**0.30**	0.56	0.18	0.29
VII	BERT BiLSTM multihead attn linear	**0.62**	**0.23**	**0.30**	**0.60**	**0.21**	**0.30**

Table 4.2: F1-score comparison for the fine-grained (main task) PICO labels for multitask learning vs. single task learning for the EBM-PICO evaluation corpus and the physiotherapy corpus. The EBM-PICO baseline F1 scores for the fine-grained PICO recognition are annotated as b1 and b2. The best F1 score for an entity in its series of experiments is shown in bold. Underlined scores show that the setup performed significantly better than its counterpart.

Setup	F1 (Physiotherapy)			F1 (EBM-PICO)		
Fine-grained	P	I/C	O	P	I/C	O
BERT BiLSTM attn CRF	**0.57**	**0.17**	**0.30**	**0.51**	**0.30**	**0.53**
BERT BiLSTM attn (on coarse) CRF	0.44	0.11	0.19	0.39	0.21	0.37
BERT BiLSTM attn (on fine) CRF	0.43	0.15	0.23	0.31	0.29	0.42

Table 4.3: F1 score for the ablation experiments in the MTL setup (BERT BiLSTM attention CRF) for both test corpora

In general, it was observed that 1) using BERT alone gave very poor performance (See Table 4.2 Experiment I), 2) the addition of a single head self-attention layer brought a significant performance boost for both setups (See Table 4.2 Experiment V), 3) the approaches have poor generalization on the physiotherapy corpus for the "Intervention" entity, and 4) though most MTL setups did not outperform the STL setups, it cannot be concluded that MTL is ineffective. These results warrant further investigation into task-weighting, appropriate task decoders, loss weighting strategies, especially for the label-imbalanced tasks.

4.2.5 Discussion

As apparent from Table 4.2, the "Intervention" entity showed the most dissatisfying overall F1-score and was the only entity unable to pass the baseline. For the EBM-PICO corpus, performance on the "Intervention" entity had saturated at 0.30 F1 and was even worse for the physiotherapy corpus. Upon the confusion matrix inspection for "Intervention" for both setups and evaluation corpora it was identified that all the sequence taggers failed to correctly identify any of the "Other" and "Psychological" fine-grained classes (see red box in Figure 4.3).

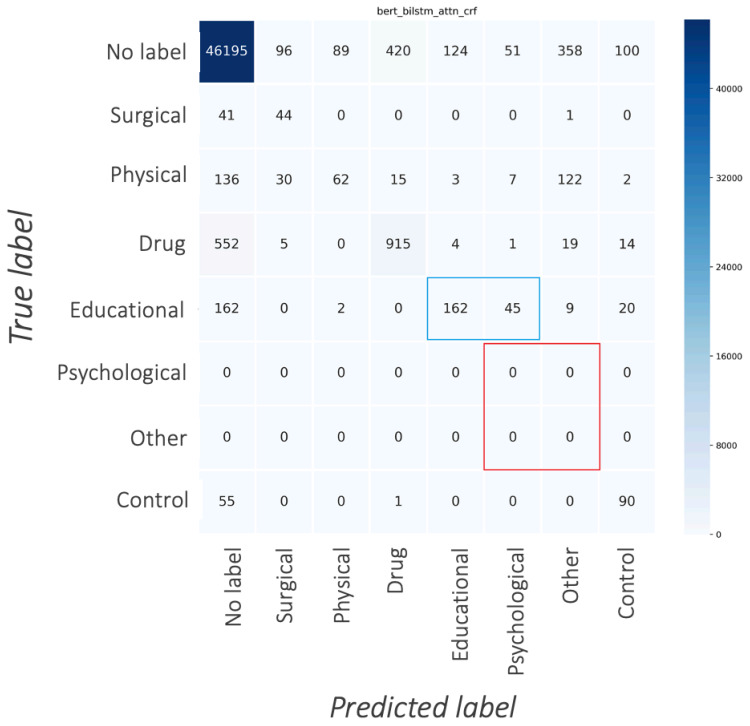


Figure 4.3: "Intervention" entity example error matrix for the MTL experimental setup V (BERT BiLSTM attention CRF)

The most obvious reason for this is the comparatively lower number of label annotations for these classes. It was apparent during the manual annotation of the physiotherapy corpus that the "Other" entity encompassed any intervention mention that did not fall into the rest of "Intervention" classes making this class highly heterogeneous with a mixture of diverse entities that followed several patterns (see Table4.1). Heterogeneous entities are a challenge for IR [165].

All the taggers were consistently confused between the physiological and educational intervention classes (see the blue box in Figure 4.3), which are important for our field of interest. This challenge is related to the "Intervention" class definition. During manual annotation, it was rather difficult, even as a human annotator, whether to classify certain interventions as educational or psychological (for example, the psycho-educational intervention if administered by a psychologist is considered as psychological intervention and if administered by a nurse it is classified as an educational intervention). The performance of automatic labelling was just a direct reflection of the difficulty emanating from class

definitions. General analysis of all the PICO confusion matrices shows several out-of-the-span entities were mislabelled as PICO and vice versa. If it was merely PICO being miss-tagged as OOS, it could have pointed to the class-imbalance problem given that OOS forms the majority class. However, consistently even the OOS entities were mislabelled as PICO which points to the class-overlap problem. Error inspection showed that the overall limited performance of these classifiers might result from the class-overlap between the PICO and OOS classes and ambiguities in how each coarse-grained PICO was divided further into fine-grained PICO classes, especially for the health entities.

Confusion matrix for the fine-grained "Intervention" entity for the best performing STL (refer Figure 4.4) setup V. is given below. STL model makes errors similar to the MTL by failing to recognize "Intervention:Other" and "Intervention:Psychological" entities altogether (see red box) and confusing between "Intervention:Psychological" and "Intervention:Educational" (see blue box).

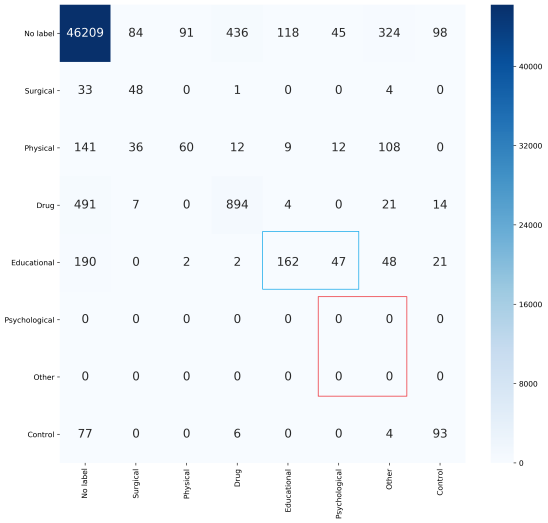


Figure 4.4: "Intervention" entity example error matrix for the STL experimental setup V (BERT BiLSTM attn CRF)

4.2.6 Conclusion

This work proposes two end-to-end neural model setups for fine-grained PICO recognition that outperform the previous SOTA for the fine-grained "Participant" and "Outcome" entities without any need for hand-engineered features. We show that MTL is not only feasible but also a good alternative to the STL setup. However, combining even the seemingly related tasks in MTL might not directly boost the performance. To perform similar to or outperform its STL counterpart, MTL could require rather careful individual weighting of the involved tasks and task losses. As part of our contribution, we provide a manually annotated dataset with multi-level PICO annotations, complementing existing resources [39]. Our error analysis warrants rethinking of semantically solid class definitions for fine-grained PICO entities along with ontology development for the health domain. The code is available on Github [40].

[39] `https://zenodo.org/records/6961986`
[40] `https://github.com/anjani-dhrangadhariya/multitask-pico-detection`

4.3 DISTANT-CTO: Distantly Supervised Intervention Extraction

Automating PICO entity detection has garnered lower interest than other biomedical NER tasks because of the lack of publicly available entity annotated corpora. The largest publicly-available PICO entity/span dataset (EBM-PICO) contains only 5000 annotated abstracts, some of which were annotated through crowd-sourcing and others by hired medical experts [254]. Crowd-sourcing, involving non-expert workers, necessitates intensive training, making it a less commonly affordable option. On the other hand, hiring medical experts for annotation is often prohibitively expensive. Extracting PICO entities or spans is somewhat tricky because of high disagreement between human annotators on the exact spans constituting the mentions. This variability results in human errors within hand-labeled corpora. Hand-labeled datasets are static and prohibit quick manual re-labelling in case of human errors or when a downstream task requires new entities. For example, PICO entities extend to PICOS, where S denotes the "study type" of included evidence. In such instances, additional efforts are often required. For instance, if an application necessitates the inclusion of this additional study type entity, one can choose to annotate EBM-PICO with this entity or utilize another corpus, training and evaluating separate models for the new entity. Both options require some form of resource investment.

Distant supervision (DS) is a data-centric approach that allows generating massive weakly annotated datasets without human annotators and has previously been used to create large relation extraction corpora for the general and biomedical domains. To address the challenges above and democratize PICO entity recognition, we propose DISTANT-CTO, a distantly supervised and scalable approach to obtaining clinical trials annotations. We take an integrative approach combining methods of semi-supervised learning (SSL) and gestalt pattern matching (GPM) to develop a continuously extensible dataset. We successfully demonstrate this approach for the "Intervention" and "Comparator" entity annotations as proof of concept (POC). We specifically chose "Intervention" entity because in prior studies utilizing EBM-PICO as a benchmark, the extraction of the "Intervention" entity consistently exhibited the poorest performance. This challenge stems from the class heterogeneity within the "Intervention" class, which comprises several intervention subclasses, presenting a significant obstacle to accurate extraction [48]. Our assumption is that if the DISTANT-CTO approach demonstrates promising results on the challenging "Intervention" entity, it can be extended to tackle other, more homogeneous entities such as "Participant" and "Study type".

The contributions of this work are as follows:

- A zero-cost, data-centric approach using DS to obtain "Intervention" and "Comparator" entity annotations was developed.

- The work develops and makes publicly available a large weakly-labeled dataset from more than 300,000 clinical trials. The dataset offers about a million sentences with more than 977,682 annotations across 11 semantic types.

- The work improves the state-of-the-art by 2% macro-F1 on the previously most poor-performing "Intervention" entity extraction on the EBM-PICO benchmark corpus without using costly manually labeled data and by 5% when combined with manually labeled data.

4.3.1 Related Work

A decade of automatic PICO information extraction was limited to sentence-level due to the unavailability of entity-annotated corpora [35, 157, 158, 166, 343]. The release of the EBM-PICO corpus paved the way for the community to improve upon the PICO entity/span extraction task. [254]. The corpus is biased towards pharma intervention classes overshadowing non-pharma ones leading to a substandard performance on it in the previous SOTA fully-supervised PICO entity/span recognition models [28, 41, 369] and weakly supervised model [202]. Small-scale annotation projects cannot capture the range and variation of the PICO descriptions spanning the entirety of clinical trials literature. At some point, applications of such static corpora will confront the problem of insufficient and irrelevant annotations. Manual annotation projects are neither affordable nor scalable for every lab, limiting innovation.

A plethora of DS methods have been previously explored for large-scale relation extraction but not for (named) entity extraction [3, 101, 309]. Entity extraction in high-impact clinical and biomedical domains largely relies on small expert annotated datasets. Commonly, obtaining weak annotations using DS rely on aligning terms (a word or phrase) from ontologies onto the unstructured text [120, 143, 259, 362]. Ontologies are structured, standardized data sources that do not capture various writing variations from clinical literature. Weak annotations obtained using custom-built rules like regular expressions are restricted by either task or worse even by entity type [112, 276, 288]. Bootstrapping approaches like label propagation (LP) still require an expert annotated dataset to obtain pseudo annotations for previously unlabeled data samples [29]. It is hence not zero-cost.

This work focuses on overcoming the discussed bottlenecks using a data-centric DS approach to generate a large clinical entity annotated corpus and train a downstream NER model to assess if it yields adequate results. Unlike the reviewed DS approaches, this approach does not use ontologies or rules or LP but rather uses GPM for flexibly aligning structured text in a clinical trials database to the free-text fields in the same database using an adaptable internal scoring scheme.

4.3.2 Methods

The approach is schematically illustrated in Figure 4.6 and is described below.

4.3.2.1 Data

`https://clinicaltrials.gov/` (CTO hereafter) documents more than 350,000 human clinical studies conducted around the globe. The trial's principal investigator enters and updates information about each study stored in CTO. It includes the title and description of the clinical trial, participant's eligibility criteria, participant disease and demographics, interventions evaluated, outcomes, *etc.* CTO allows programmatic access to this vast amount of information in the JSON (JavaScript Object Notation) format. The information is stored as a combination of structured tabular and unstructured free-text (see Figure 4.5). The 'OfficialTitle' and 'BriefTitle' tags in the JSON respectively store the official and shorter version of the study title in an unstructured free-text format. The 'BriefSummary' and 'DetailedDescription' tags store study summaries. Interventions used in the study are stored under the 'InterventionName' tag and their synonyms under 'InterventionOtherName' tag each of which could be linked to their broad semantic type (drug, device, behavioral, procedural, biological, dietary supplement, diagnostic test, radi-

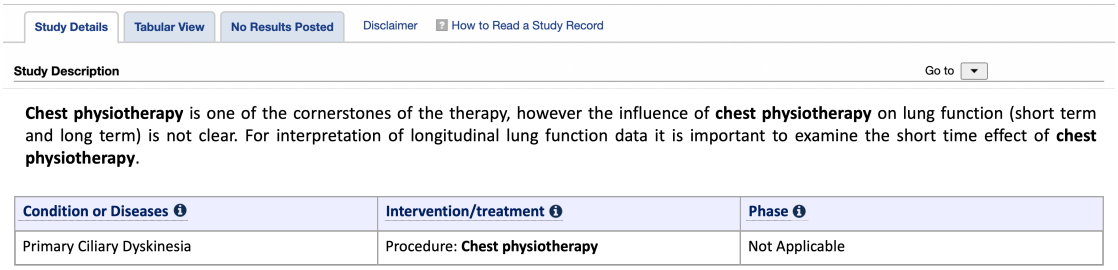


Figure 4.5: An example CTO record (ID - NCT01929356) to demonstrate the information storage format which is a combination of structured table and unstructured text.

ation, genetic, combination product, other) mentioned under the 'InterventionType' tag. As each intervention name is linked to its semantic type, this becomes a structured information store. The 'InterventionDescription' tag describes intervention administration procedures often in a detailed passage.

4.3.2.2 Distant Supervision

Distantly supervised (DS) information extraction (IE) is an efficient SSL method [101, 356]. It is used when the task at hand has 1) some strongly-labeled data, 2) abundant unlabeled data, and 3) a weak-labelling function that could sample from this unlabeled data and label them using a heuristic function. This labelling function is a heuristic algorithm that uses a heuristic to label the unlabeled data [128, 263]. It results in a weakly-labeled dataset with potential label noise. DS-IE models can then collectively use this strongly-labeled and weakly-labeled training data to give the final output.

4.3.2.3 Gestalt Pattern Matching

In entity extraction, the most common form of DS is to heuristically align terms from a structured information source onto the unstructured text [356]. When flexible, this heuristic boils down to a substring matching problem. The weak-labelling function matches the longest common substring (LCS) between the structured term and unstructured text. Gestalt Pattern Matching (GPM), also known as Ratcliff/Obershelp similarity algorithm, is a string-matching algorithm for determining the similarity of two strings. The similarity between two strings S_1 and S_2 is measured by the formula, calculating twice the number of matching characters K_m divided by the total length $|S_1| + |S_2|$ of both strings. Matching characters are identified by the LCS algorithm followed by recursively finding matching characters in the non-matching regions on either side from both strings [274]. *Similarity* ranges between 0, which means no match, and 1, which means a complete match of the two strings.

$$Similarity(S) = \frac{2K_m}{|S_1| + |S_2|} \; ; \; 0 \leq S \leq 1 \tag{4.7}$$

Difflib: It is a python module providing a `sequencematcher` function that extends the GPM algorithm for comparing pairs of strings. `sequencematcher` finds the longest contiguous subsequence between the sequence pair without the "junk" elements such as blank lines or white spaces. The same idea is then applied recursively to the flanks of the

sequences to the left and the right of the matching subsequence. This yields matching sequences that appear normal to the human eye.

4.3.2.4 Candidate Generation

We define candidate generation as the process of automatically generating entity-annotated sentences.

Assumption and Problem formulation: As "Intervention" and "Comparator" entities represent interventions in two different roles in clinical trials and semantically the same classes, they are clubbed into a single "Intervention" entity class. Let each CTO record JSON file be $r_i \in \boldsymbol{R}, i = \{1, 2, ..., I\}$. Let the intervention terms in 'InterventionName' tags and 'InterventionOtherName' tags be the intervention source $S = \{s_1, s_2, ..., s_m\}$ used in the study r_i. Each intervention term $s_i \in S$ is linked to intervention class from 'InterventionType' tag converting it into a tuple of $\langle s_{class}, s_{name} \rangle$, s_{name} = intervention term and s_{class} = intervention category. s_{name} is a sequence of words $\{y_1, y_2, ..., y_n\}, n = \{1, 2, ..., N\}$. Let each sentence $t_i = \{x_1, x_2, ..., x_m\}, m = \{1, 2, ..., M\}$ in the 'BriefSummary', 'DetailedDescription', 'BriefTitle', 'OfficialTitle' and 'InterventionDescription' be a part of the intervention target set T. We assume that for each s_{name} in r_i there could exist a mapping to t_i meaning s_{name} is possibly either completely or partially mentioned in the t_i (see Figure 4.5). Our goal is to build a scalable and adaptable candidate generation pipeline that maps each s_{name} from the structured intervention source S to the target sentences $t_i \in T$ (if a loose mapping exists). In this prototypical work, we focus on *almost* direct matches between the s_{name} and t_i and keep the order-free matches for future work.

Approach For each individual CTO record r_i, we extract all $s_{name} \in S$ and $t_i \in T$ from the locally stored CTO dump. Both S and T are preprocessed by lower-casing, replacing hyphens and multiple trailing spaces with a single space and removal of Unicode characters. Given a s_{name} and t_i, our aim is to identify and score (if identified) the mapping between both sequences. To map and score alignment from the s_{name} to t_i, we use a distant supervision labelling function LF_{ds} which is a combination of the `sequencematcher` function and an internal scoring function to fetch almost direct annotations. The `sequencematcher` function takes as input s_{name} and t_i and outputs several matching blocks $d_{block} \in D_{blocks}$ between both strings. These matching blocks between the two strings are calculated using a modified gestalt pattern matching algorithm as elaborated in 4.3.2.3. Each $d_{block} = \langle MatchPos_t, MatchPos_s, MatchLen \rangle$. $MatchPos_t$ is the start of the match in t_i, $MatchPos_s$ is the start of the match in s_{name} and $MatchLen$ is number of characters matching between the both. `sequencematcher` provides an internal scoring function called as `ratio` that returns a similarity score between the two sequences being matched. We do not use `ratio` because it returns an overall matching score between the two full sequences s_{name} and t_i rather than a match score for s_{name} and d_{block}. Instead, to identify the matching blocks that correspond to an exact match between an entire s_{name} and a part of t_i, we calculate a match score d_s for each matching block output by `sequencematcher` using equation 4.8 which is dividing the number of matching characters in the match block d_{block} by number of characters in s_{name}.

$$d_s = \frac{MatchLen}{|s_{name}|} \; ; \; 0 \leq d_s \leq 1 \tag{4.8}$$

Any d_{block} with the d_s score of 1.0 is considered as complete match and then the s_{name} corresponding to the d_{block} is mapped onto sentence t_i to generate a positive annotation sentence $a_+ \in A_+$. Using the d_{block} with only the match score 1.0 leads to missing out on several entities leading to an incomplete noisy weakly annotated dataset. Taking this into consideration, we retrieve the d_{block} matching with d_s score of 0.9 as fairly-accurate partial matches. We used a validation set to relax the choice of similarity match score d_s to 0.9. We relax the labelling function LF_{ds} to match bigrams in source terms to the targets. In the real-world data, not all sentences in clinical trial literature mention the intervention name and therefore in addition to the positive annotation sentences we require negative annotation sentences. We take t_i and s_{name} where no parts of d_{block} scored d_s more than 0.2 to generate the negative annotation sentences $a_- \in A_-$. We call all these sequences comprised of the positive and the negative entity annotated sentences A_{+-} our weakly annotated dataset. Next, for all A_{+-} instances we fetch part-of-the-speech (POS) tags using `POS-tagger` from NLTK (Natural Language Toolkit) resulting into A_{+-POS}. We call the resulting dataset DISTANT-CTO set. POS tags are added as additional features as they have shown to help model generalization [15]. `difflib` in combination with the internal scoring function are previously unexplored for automatic entity annotation generation. It has to be noted that the method depends on availability of short source texts with the possibility that they will be mentioned in longer target texts.

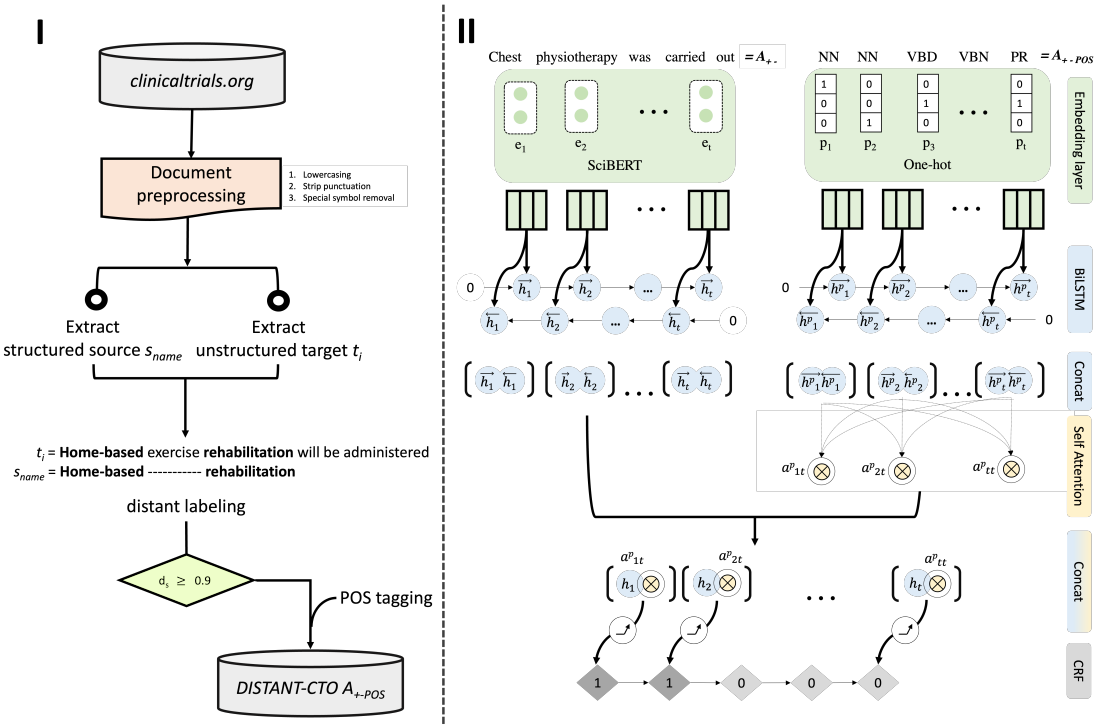


Figure 4.6: DISTANT-CTO approach - I) Distantly-supervised candidate generation approach, and II) Distantly-supervised NER model architecture.

4.3.2.5 Model Training

We train an end-to-end distant NER model on A_{+-POS} using the architecture explained below.

1. Feature Extraction: Open-domain pre-trained language models (LM) like BERT, ULMFit and GPT rule out the need for heavy feature engineering and also tackle the the challenge of out-of-vocabulary (OOV) words using the WordPiece tokenizer and byte pair encoding (BPE) [81, 169]. Unfortunately, they encode limited semantic and syntactic information for domain-specific tasks. To capture the domain-specific information, we used SciBERT, which was continually pretrained and domain adapted on the scientific literature from semantic scholar [132]. The models used SciBERT to tokenize the text input A_{+-} into encoded tokens x_t and extract dense, contextual vectors e_t from x_t at each time-step t [28]. POS-inputs A_{+-POS} were one-hot encoded into p_t vectors.

2. Feature transformation: To further fine-tune to the training corpus, the model stacked a bidirectional LSTM (BiLSTM) on top of the SciBERT [152]. A BiLSTM layer encodes the text into a ($\overrightarrow{h}$) and ($\overleftarrow{h}$) vector using the current token embedding input e_t and the previous hidden state h_{t-1} in both the directions. $\overrightarrow{h}$ and $\overleftarrow{h}$ were shallow concatenated ($[\overrightarrow{h}; \overleftarrow{h}]$) into h_t and used as the input for the next layer. Similarly, the one-hot encoded POS-vectors p_t underwent feature transformation and were concatenated ($[\overrightarrow{h}_{POS}; \overleftarrow{h}_{POS}]$) into POS-features h_t^p.

3. Self-attention: Next, the model stacked a single-head self-attention layer that calculated for each POS-tag feature at time t in the sequence a weighted average of the feature representation of all other POS-tag features in the sequence a_t^p [338]. This improves the signal-to-noise ratio by out-weighting important POS features. Self-attention weights for each POS-tag were calculated by multiplying hidden representation h_t^p with randomly initialized Query q and Key k weights, which were further multiplied with each other to obtain attention weighted vectors. Finally, the obtained attention weights were multiplied with the Value (V) matrix which was obtained by multiplication between a randomly initialized weight matrix v and h_t finally obtaining scaled attention-weighted vectors a_t^p. Attention-weighted POS features and h_t were shallow concatenated into ($[a_t^p; h_t]$) vector.

4. Decoder: The attention-weighted representation ($[a_t^p; h_t]$) was fed to a linear layer to predict the tag emission sequence $\hat{y}_t$ followed by a CRF layer that takes as input the $\hat{y}_t$ sequence along with the true tag y_t sequence [160]. CRF is a graph-based model suitable for learning tag sequence dependencies from the training set and has shown to outperform softmax classifiers.

4.3.2.6 Experiments

The experiments were designed to evaluate the performance of the distant NER models trained with the DISTANT-CTO set alone *vs.* DISTANT-CTO set in combination with the EBM-PICO training set. The EBM-PICO training set is naturally composed of both positive and negative annotation sentences, but for the DISTANT-CTO, we artificially generated the negative sentences A_-. To evaluate the impact of these negative annotation sentences, we perform ablation experiments, training the models only with positive annotation sentences A_+. Finally, we also evaluate the performance when training using the entity annotations with match score $d_s = 1.0$ alone *vs.* entity annotations with $d_s \geq 0.9$. A simple SciBERT-CRF model trained using positive annotation sentences A_+ was used as the baseline. Transformer-based models incorporate sequence order and self-attention

components, so our baseline served to check the impact of removing costly BiLSTM and self-attention modules.

Benchmark datasets We evaluate our weakly annotated dataset and the NER model on the following PICO benchmarks.

1. **EBM-PICO gold.** The EBM-PICO dataset developed by Nye *et al.* consists of 5000 PICO entity/span annotated documents [41]. It comes pre-divided into a training set (n=4,933) annotated through crowd-sourcing and an expert annotated test set (n=191) for evaluation purposes. We use the training set for combined training experiments and the test set for evaluation. More information regarding the dataset could be found in [254].

2. **Physio set.** A test set comprising 153 PICO entity/span annotated documents from Physiotherapy and Rehabilitation RCTs (Randomized Controlled Trials) was used as an additional benchmark to evaluate the generalization power of our approach for this sub-domain [82].

Experimental Setup We define the following experimental setups based on the motivations described in section 4.3.2.6:

- **Exp 1.0 distant A_{+-} c[1,0.9] wPOS** The setup is composed of SciBERT BiLSTM CRF trained on the surface form (text) and attention-weighted POS inputs using DISTANT-CTO set comprising entity-annotated sentences A_{+-} with $d_s \geq 0.9$.

- **Exp 1.1 distant A_{+-} c[1] wPOS** The setup is composed of SciBERT BiLSTM CRF trained on the surface form and attention-weighted POS inputs using the DISTANT-CTO set comprising only the entity-annotated sentences A_{+-} with $d_s = 1.0$.

- **Exp 1.2 distant A_{+} c[1] wPOS** The setup is composed of SciBERT BiLSTM CRF trained on the surface forms and attention-weighted POS inputs using DISTANT-CTO set comprising only the $d_s = 1.0$ annotations. The negative annotation sentences were removed in this case and the system was trained with positive annotated candidates A_{+} only.

- **Exp 1.3 distant A_{+} c[1] POS ¬ BiLSTM attention** The setup is composed of SciBERT CRF trained on the surface form inputs using DISTANT-CTO set comprising only the $d_s = 1.0$ annotations with only positive annotated candidates A_{+}. Attention weights were removed from the POS inputs. This setup was used as the baseline.

- **Exp 2.0 - Exp 2.3** These experiments are identical to their series 1.x counterparts except that the models are trained on a combination of the DISTANT-CTO with the EBM-PICO training set. Exp 2.3 using SciBERT-CRF architecture was used as another baseline.

[41] A single document consists of a title and an abstract.

4.3.2.7 Evaluation

To evaluate the quality of automatic annotation using the DISTANT-CTO approach, we performed manual annotation of the "Intervention" class over 200 randomly selected samples from the dataset and compared it to the automatic annotations.

Model evaluation was carried out by predicting the "Intervention" tokens for both benchmarks. Each experiment was conducted thrice with three random seeds (0, 1, and 42), and the average metrics (Precision, Recall, and F1) over three repetitions were reported. We evaluated the statistical significance of our best model using the paired student's t-test as described in [94]. Further experimental details are in the Appendix.

4.3.2.8 Experiments

This section reports empirical results for the candidate generation process, evaluation for the annotation quality of DISTANT-CTO approach using the validation sets (see Table 4.8), and the average of the performance metrics and standard deviation σ over three random seeds on both benchmark datasets for the described NER experiments (see Table 4.10). We compare the performance of our weakly-supervised NER models with the previous SOTA fully supervised (FS) methods that train on the EBM-PICO training set and evaluate on EBM-PICO gold and also a weakly supervised approach (see Table 4.9). These models were separately trained for each of the PICO entities/spans and also clubbed the "Intervention" and "Comparator" together.

4.3.3 Results

This section reports empirical results for the candidate generation process, evaluation for the annotation quality of DISTANT-CTO approach using the validation sets (see Table 4.8), and the average of the performance metrics and standard deviation σ over three random seeds on both benchmark datasets for the described NER experiments (see Table 4.10). We compare the performance of our weakly-supervised NER models with the previous SOTA fully supervised (FS) methods that train on the EBM-PICO training set and evaluate on EBM-PICO gold and also a weakly supervised approach (see Table 4.9). These models were separately trained for each of the PICO entities/spans and also clubbed the "Intervention" and "Comparator" together. Therefore, the comparison is valid. Sentence-level PICO recognition methods are not comparable to that of entity-level.

4.3.3.1 Candidate Generation

A total of 360,395 CTO records were downloaded as of March 2021. From all the downloaded CTO records, we extract 200,545 unique (391,286 redundant) intervention names from the aforementioned intervention sources. Out of the 391,286 intervention terms retrieved, 104,433 terms were successfully mapped to one of the target sentences with the $d_s = 1.0$, and 3084 more were mapped with a score of 0.9. Adding $d_s \geq 0.9$ mappings did not increase the total number of annotated sentences, but it did increase the number of annotations obtained in each sentence. Table 4.4 shows the total number of intervention annotations obtained from mapping the source terms to target sentences.

The total number of entity-level "Intervention" mentions in DISTANT-CTO are almost 30 times more than in the EBM-PICO dataset as shown in Table 4.5. For the EBM-PICO training set, 57.48% of mentions fell under the "drug" class and the rest under the six remaining classes.

Annotation level	$d_s = 1.0$	$1.0 < d_s \geq 0.9$
mention-level	943,284	17,199
token-level	1,515,868	43,096

Table 4.4: Token-level and mention-level intervention annotations obtained in the weakly annotated DISTANT-CTO dataset grouped by their d_s scores.

Total	**DISTANT-CTO**	**EBM-PICO**
mention-level	977,682	32,890
token-level	1,558,964	125,920

Table 4.5: Comparing the number of "Intervention" annotations in DISTANT-CTO *vs.* EBM-PICO.

Out of all the mention-level annotations in the DISTANT-CTO dataset, 59.90% corresponded to "drug" class and 40% to the rest of 10 classes. The pie chart (upper pie in Figure 4.7) shows the class distribution of the semantic classes for the retrieved "Intervention" mentions s_{name} about half of which fall under the "drug" (or Pharma) class and the rest under the remaining 10 non-pharma classes. Out of the total retrieved mentions, almost two-thirds that get mapped to a target t sentences also fall under the "drug" class (lower pie in Figure 4.7).

Table 4.6 and 4.7 shows the number of retrieved intervention mentions by their semantic class *vs.* the percentage of these intervention mentions that get mapped to some target sentence with the match score d_s of 1.0 and score 0.9 respectively. Notice that collectively the intervention mentions that fall under the non-pharma classes outnumber the pharma ("drug") mentions.

Domain	**retrieved - (mapped)**
drug	184835 (35.50%)
device	43134 (20.09%)
other	51703 (16.19%)
procedure	31630 (21.38%)
behavioral	33590 (16.03%)
biological	21225 (22.86%)
dietary supplement	11699 (25.46%)
radiation	4134 (20.44%)
diagnostic test	6742 (10.13%)
combination product	1070 (14.39%)
genetic	1524 (07.94%)
all non-pharma	206,451 (18.80%)

Table 4.6: Number of intervention mentions retrieved *vs.* percentage mapped with $d_s = 1.0$

Metrics for the manual evaluation of DISTANT-CTO using the validation set show that adding annotations with $d_s \geq 0.9$ increases the recall by 3%, but lead to an expected drop in the precision (see Table 4.8).

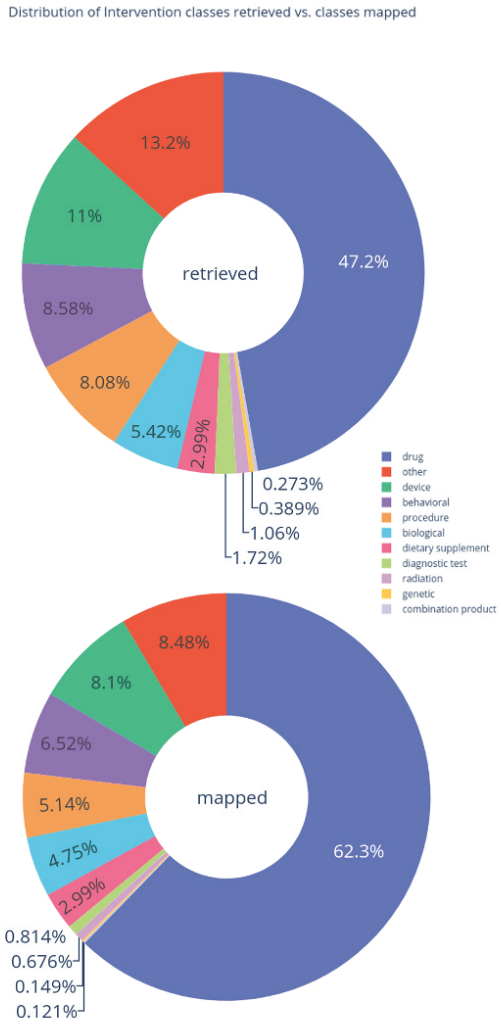


Figure 4.7: upper) Class distribution for the retrieved "Intervention" mentions, and lower) Class distribution for the mapped "Intervention" mention.

.

4.3.3.2 Model Training

Using the DISTANT-CTO set alone with the NER approach (Exp 1.1 Table 4.9 and 4.10) crosses the previous SOTA F1 on the EBM-PICO benchmark by 2%. The best overall F1 for both benchmarks is reached upon training the NER models with combined weakly-labeled DISTANT-CTO with the strongly-labeled EBM-PICO dataset (Exp 2.0 Table 4.10) crossing the previous SOTA F1 by 5% on the EBM-PICO benchmark. The improvement in F1 for the combined experiments (see Exp 2.1 and 2.0 Table 4.10)) is significant when compared to the their best DISTANT-CTO counterparts (see Exp 1.1 Table 4.10)). Using DISTANT-CTO alone has good precision across the experiment series 1.x, but combining it with the EBM-PICO further improves the recall and balances out the F1 in the experiment series 2.x. Adding the artificially generated A_{-} sentences increases the previous F1 by 5.71% and 3.77% (compare Exp 2.2 with Exp 2.1) for both the benchmarks. Note that adding these negative sentences results in an important im-

Domain	retrieved - mapped
drug	184835 (36.22%)
device	43134 (21.13%)
other	51703 (16.84%)
procedure	31630 (22.16%)
behavioral	33590 (16.44%)
biological	21225 (24.07%)
dietary supplement	11699 (27.44%)
radiation	4134 (21.17%)
diagnostic test	6742 (10.78%)
combination product	1070 (14.95%)
genetic	1524 (08.53%)
all non-pharma	206,451 (19.64%)

Table 4.7: Number of intervention mentions retrieved *vs.* percentage mapped with a d_s of 0.9

Match score	P	R	F1
$d_s = 1.0$	0.86	0.80	0.83
$d_s \geq 0.9$	0.84	0.83	0.84

Table 4.8: Macro-averaged evaluation metrics for the $d_s = 1.0$ and ≥ 0.9 entity annotations for the validation set detailed in the section 4.3.2.7

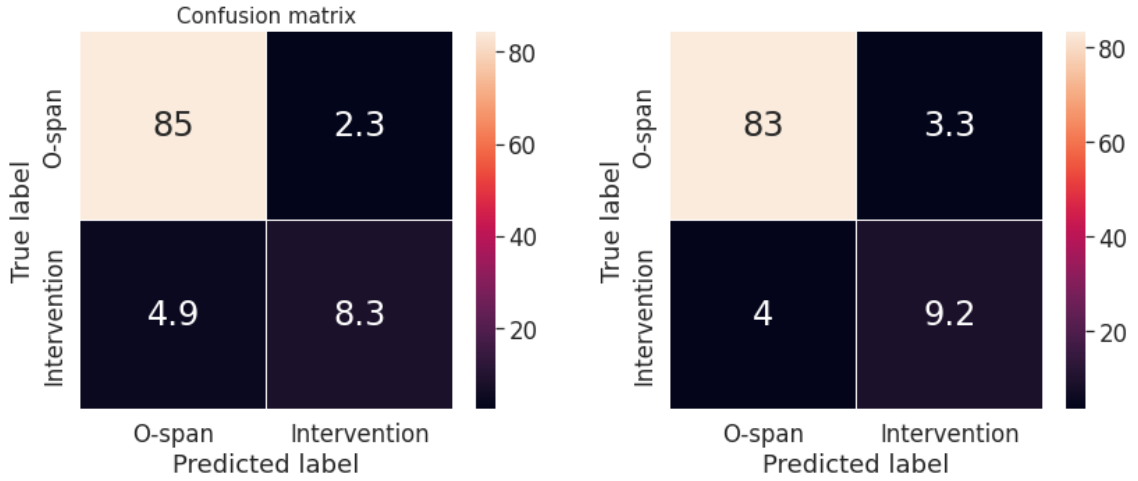
provement of about 9% in the F1 for the Physio dataset that is specific for the domain of physiotherapy and rehabilitation. For the combined experiment, the addition of the $d_s \geq 0.9$ annotations improves the F1 as well by a small margin for the EBM-PICO benchmark (Exp 2.0 I.) but has a marginal performance loss for the Physio benchmark (Exp 2.0 II.). While using the DISTANT-CTO alone with the $d_s \geq 0.9$ annotations boosts the precision but downgrades recall thereby reducing the F1 for both benchmarks.

Type	Method	P	R	F1
FS	Nye [254]	**84.00**	61.00	70.00
FS	Beltagy [28]	61.00	70.00	65.00
FS	Brockmeier [41]	69.00	47.00	56.00
FS	Stylianou [315]	69.04	79.24	73.29
WS	Liu [202]	22.00	54.00	31.00
WS	distant-cto (our)	83.36	70.38	75.02
HS	combined (our)	76.93	**80.17**	**78.44**

Table 4.9: Comparison of DISTANT-CTO NER models against the previous SOTA NER methods for "Intervention" recognition in terms of macro-averaged precision (P), recall (R), and F1 scores. Boldface represents the best score. Note: FS = Fully Supervised, WS = Weakly Supervised, HS = Hybrid Supervision.

Dataset	Experimental setup	P	R	F1 $\pm\sigma$
			I. EBM-PICO gold	
distant	A_{+-} c[1,0.9] wPOS	88.85	65.39	71.27 $\pm$0.007
distant	A_{+-} c[1] wPOS	83.36	70.38	**75.02** $\pm$0.013
distant	A_{+} c[1] wPOS	74.85	68.74	71.25 $\pm$0.005
distant	A_{+} c[1] POS $\neg$ BiLSTM att	85.82	64.84	70.31 $\pm$0.002
combined	A_{+-} c[1,0.9] wPOS	76.93	80.17	**78.44*** $\pm$0.006
combined	A_{+-} c[1] wPOS	77.10	78.83	77.89 $\pm$0.007
combined	A_{+} c[1] wPOS	67.65	85.02	72.18 $\pm$0.009
combined	A_{+} c[1] POS $\neg$ BiLSTM att	70.91	77.38	73.60 $\pm$0.025
			II. Physio set	
distant	A_{+-} c[1,0.9] wPOS	86.13	63.70	69.14 $\pm$0.003
distant	A_{+-} c[1] wPOS	79.45	66.28	**70.63** $\pm$0.008
distant	A_{+} c[1] wPOS	70.52	66.37	68.14 $\pm$0.002
distant	A_{+} c[1] POS $\neg$ BiLSTM att	79.97	60.79	65.14 $\pm$0.005
combined	A_{+-} c[1,0.9] wPOS	75.55	79.42	77.32 $\pm$0.010
combined	A_{+-} c[1] wPOS	76.29	80.18	**78.07*** $\pm$0.009
combined	A_{+} c[1] wPOS	64.80	83.69	68.75 $\pm$0.011
combined	A_{+} c[1] POS $\neg$ BiLSTM att	71.50	78.40	74.38 $\pm$0.020

Table 4.10: Macro-averaged performance metrics for the NER models trained on weakly annotated DISTANT-CTO alone *vs.* in combination to the strongly annotated EBM-PICO on the two described benchmarks (EBM-PICO gold and the Physio corpus). "att" = attention. Bold is the best experiment score. Asterisk (*) denotes a significant F1-score of the experiment to its counterpart in the series 1.x. Significance tested using the paired student's t-test.



(a) $d_s = 1.0$ (b) $d_s \geq 0.9$

Figure 4.8: Confusion matrices for the evaluation of DISTANT-CTO validation set annotations with a) $d_s = 1.0$ and b) $d_s \geq 0.9$.

4.3.4 Error Analysis

4.3.4.1 Candidate Generation

Confusion matrices (see Figures 4.8a and 4.8b) for manual evaluation of DISTANT-CTO validation set show that relaxing d_s from 1.0 to 0.9 does improve the true positives (TP)

and reduce false negatives (FN) by 0.9% for the "Intervention" class but also reduce the precision by increasing false positives by 1%. Improved recall for the "Intervention" class is undoubtedly preferred, and hence it is vital to inspect the cause of false negatives. A considerable chunk of false negatives was either i) missed intervention abbreviations and the synonyms not mentioned under the sources, or ii) when only the partial intervention name was mentioned in the source, or iii) if specific intervention terms from the source were mentioned in the target but with different word order (see Table 4.11). This detailed post-hoc error analysis also revealed that 67% false negatives fell under non-drug type composite intervention mentions (phrase mentions of more than two words). For instance, although the term *'Home-based Rehabilitation using Interactive devices'* is expressed in the sentence *'This study investigates clinical outcomes after the rehabilitation by interactive home-based devices.'*, it will remain unmapped to it because the term does not map to the target text using our alignment heuristic. The problem lies in the lack of naming conventions for non-pharma treatment mentions that are neither clearly identified nor standardized as semantic units[82]. There are two possible programmatic solutions to this. The first is using additional external ontologies as sources of distant supervision which improves coverage of our labelling function to detect further writing variations within the text. Another solution to matching such source and target text is using order-free string matching algorithms [10]. Using external ontologies solves the issues of missed synonyms, and adding an external dictionary of treatment abbreviations could solve the problem of missed abbreviations [112]. We noticed that the "Comparator" terms (e.g., placebo, sham, saline, etc.) were often not mentioned as structured sources. The development of a general comparator term dictionary could improve this. Improving the coverage and reducing the false negatives (thereby improving recall) using these methodologies suggests an area where future work would be valuable. Most false positives were a result of bigram matching. We will modify fuzzy bigram matching to relevant bigram matching, thereby reducing the occurrences of spurious false-positive bigrams as matches. Only frequently occurring bigrams from the source will be matched to the targets. We plan to explore the quality of DISTANT-CTO for $d_s \leq 0.9$.

Category	FN count
Missed synonym	168
Missed abbreviation	77
Partial match (incl. boundary errors)	361
Missed comparator term	43
Reorder	39
Total	688

Table 4.11: Distribution of the false negatives in the DISTANT-CTO evaluation corpus.

4.3.4.2 Model Training

Manual error analysis was carried out for both the PICO benchmarks, and the error counts for EBM-PICO gold are reported in Table 4.12. Each token level error was divided into either of the four classes: 1) false negative (FN) - if the entire entity that the token as part of was missed out by the NER model prediction, 2) false positive (FP) - if the entire entity that the token was part of was falsely recognized as "Intervention", 3) boundary error (BE) - if the boundary tokens were missed out but otherwise the entity was identified by the

NER model prediction, and 4) overlapping error (OE) - if the NER model made an error in the non-peripheral tokens of an otherwise identified entity mention. Non-peripheral tokens are all the tokens except the first and the last token of the multi-token entity/span.

Models trained on DISTANT-CTO alone had a fewer boundary and overlapping errors, meaning they missed out on many "Intervention" entity signals leading to high precision but compromised recall. On the contrary, NER models trained on combined datasets made twice the more BE and six times more OE. While most BE and OE in the 1.x series were false negatives, they were false positives in the 2.x series leading to a higher recall. This could be because the EBM-PICO training set annotated the longest possible intervention span resulting in spans rather than pure entities in the DISTANT-CTO approach. Combined training set models also picked out names of treatments, surgeries, and enzymes not used as treatments in the RCT as intervention mentions. A huge chunk of overall FN (including the FN tokens in BE and OE) was for entities with composite intervention terms containing two or more tokens. We noticed that the NER system also missed several short intervention names and abbreviations. Overlapping errors occurred when multiple intervention names were mentioned together, separated by either comma or punctuation, or other conjunctions. The error analysis revealed some issues within EBM-PICO ground truth, which had inconsistencies with the intervention boundaries for whether intervention frequency, dose, and the way of administration should be marked as "Intervention". Several times, the ground truth marked articles preceding the entity and prepositions and punctuation succeeding the entity.

Exp	FP	FN	BE	OE
	EBM-PICO gold			
Exp 1.0	819	1688	559	66
Exp 2.0	759	1112	1278	515
Exp 1.1	790	1152	650	55
Exp 2.1	793	1039	1327	517

Table 4.12: Distribution of the token-level errors made by the corresponding NER models on EBM-PICO gold.

Manual error analysis results for Physio corpus are reported in the Table 4.13. FP error count was always lower than the FN error count in the EBM-PICO gold but for the Physio set, the combined NER experiments (series 2.x) lead to a higher FP compared the FN. The ratio of BE in Exp series 2.x is on an average 1.2 times that of series 1.x. However, a large chunk of BE in series 1.x are false negatives in contrast to the BE in series 2.x which are false positives. Upon closer inspection of false-negative BE in series 1.x, we found that they were either missed intervention synonyms inside brackets, missed information accompanying intervention terms like dose, type, medium of intervention, administrator of intervention, or location of administration. This is due to the fact that distantly supervised annotation does not take into account labelling the additional intervention information except the name. The addition of the manually annotated EBM-PICO in the combined training experiments reduces the number of false-negative BE. This is due to the fact that EBM-PICO guidelines required the annotators to mark the longest possible phrase describing intervention including the additional information like dose, mode, medium, and location of administration.

For both the evaluation corpora, the combined NER experiments lead to more TP for

the "Intervention" class which is vital to PICO entity/span recognition. This could be the case because the combination of weakly and strongly annotations reduce the percentage of unseen surface forms (words) from both test sets. 27.70% of the intervention entity surface forms in the EBM-PICO gold benchmark remain unseen in the EBM-PICO training set while for the DISTANT-CTO training set it drops to 21.38%. 27.29% of the intervention entity surface forms in the Physio benchmark remain unseen in the EBM-PICO training set while for the DISTANT-CTO training set it drops to 22.97%. Combining both training sets leads to a reduction in unseen surface forms to 16.29% and 15.13% for the EBM-PICO gold and Physio benchmarks respectively. [15] has shown that recall on unseen surface forms is significantly lower than on seen surface forms for NER tasks.

Exp	**FP**	**FN**	**BE**	**OE**
	Physio set			
Exp 1.0	963	1586	654	20
Exp 2.0	1168	897	867	347
Exp 1.1	990	1420	723	19
Exp 2.1	1116	904	1025	228

Table 4.13: Distribution of the token-level errors made by the corresponding NER models on Physio set.

4.3.5 Conclusion

This work exploits the freely-available `https://clinicaltrials.gov/` (CTO) and distant supervision for developing the largest available weakly annotated database of Intervention-Comparator entities across 11 sub-types. Using these weak annotations combined with the manual annotations, an "Intervention" NER model was trained that surpasses current approaches by more than 5% in terms of F1 on the EBM-PICO gold benchmark and demonstrates strong generalizability on a domain-specific physiotherapy benchmark. When the same NER model was trained with the weakly annotated dataset alone, it surpassed other approaches by 2%.

DISTANT-CTO successfully demonstrated the feasibility of using distant supervision for "Intervention" extraction, but they did not extend it to other PICO+ entities. In the next Section 4.4, a weak supervision approach is developed and explored for PICO+ extraction, thus bridging the gap of the previous work.

4.4 Weakly Supervised PICO+ Information Extraction

4.4.1 Background and Significance

Supervised ML requires hand-labelled data, but hand-labelling PICO information requires experts with combined medical and informatics skills, which is expensive and time-consuming in terms of intensive annotator training and the actual annotation. Labelling PICO is tricky because of the high disagreement between human annotators on the exact spans constituting PICO, leading to human errors in hand-labelled corpora [41]. Some studies examine the errors in the publicly-available EBM-PICO benchmark [1, 188, 254]. More importantly, depending upon the SR question, PICO criteria extend to PICOS

(S-Study design), PICOC (C-Context), PICOT (T-timeframe), *etc* [228, 280, 333]. Hand-labelled datasets are static and prohibit quick manual re-labelling in case of human errors or when a downstream task requires new entities. This annotation bottleneck has pivoted attention towards weakly supervised (WS) learning that relies on programmatic labelling sources to obtain training data. Programmatic labelling is quick and allows efficient modifications to the training data labels per the downstream application changes.

Weakly-supervised learning has demonstrated strengths for clinical document classification and relation extraction, but clinical entity extraction tasks have heavily relied on fully supervised (FS) approaches [98, 209, 226, 237, 351, 354]. Despite the availability of UMLS (Unified Medical Language System), a large compendium of medical ontologies, which can be re-purposed for weak entity labelling, it has not been extensively applied to clinical entity labelling [161]. Several legacy clinical applications are also supported by rule-based *if-else* systems relying on keyword cues that aid weak labelling [111, 176, 360]. With so many weak labelling sources available, the challenge for weak supervision is efficiently aggregating these sources of varying accuracy. Compare this to crowd sourcing, where an important task is to model the worker's accuracy without the ground truth [163]. Though crowd sourcing requires annotator training and quality control, programmatic labelling does not [107].

Data programming is a domain-agnostic generative modelling approach combining multiple weak labelling sources and estimating their accuracies. The effectiveness of data programming for biomedical entity recognition has been explored by Fries *et al.* in their Trove system [112]. However, Trove only explores well-defined entities like chemical, disease, disorder and drug. PICO categories are highly compositional spans by definition, fuzzier in comparison and much more intricate in that they can be divided into subclasses. A shortcoming of span extraction is that even after a machine points a human reviewer to the correct PICO span, the reviewer requires to manually read and understand its finer aspects to screen the study for relevance. Span extraction hence leads to semi-automation but hinders full-automation. The entity recognition approach to PICO is more challenging than the entity recognition approach to disease or chemical names which are more or less standardized. PICO terms are not standard, and even the experts disagree on the exact tokens constituting them [41]. Weakly-supervised PICO entity recognition has not garnered as much attention as supervised span recognition. As far as our knowledge goes, only two studies exist for weakly-supervised PICO recognition. One of these approaches only explores distant supervision for intervention extraction using a single labelling source [89]. The other approach studies weak supervision for PICO span extraction but still utilizes some supervised annotation signals about whether a sentence includes PICO information [202].

The challenges to developing weak supervision approaches to PICO entity recognition are first defining the subclasses within PICO spans and then mapping several available ontologies and terminologies to these. The next challenge is developing weakly-supervised classifiers by optimally combining ontologies and evaluating their performance compared to full supervision. Another challenge is developing higher-cost expert-generated rules corresponding to these subclasses to aid ontology classifiers and evaluate their combined performance. The study also identified limitations in the currently available EBM-PICO training set and corrected them in the EBM-PICO test set for reliable evaluation of the WS approaches. This work demonstrates the feasibility of using weak supervision for PICO entity extraction using the EBM-PICO benchmark and shows how weak supervision overtakes full supervision in certain instances. This work also shows how using only

ontology-dependent classifiers vs combining them with expert-generated rules compares to fully-supervised extraction and, in some instances overtaking it.

4.4.2 Methods

The birds-eye view of our approach is shown in Figure 4.9.

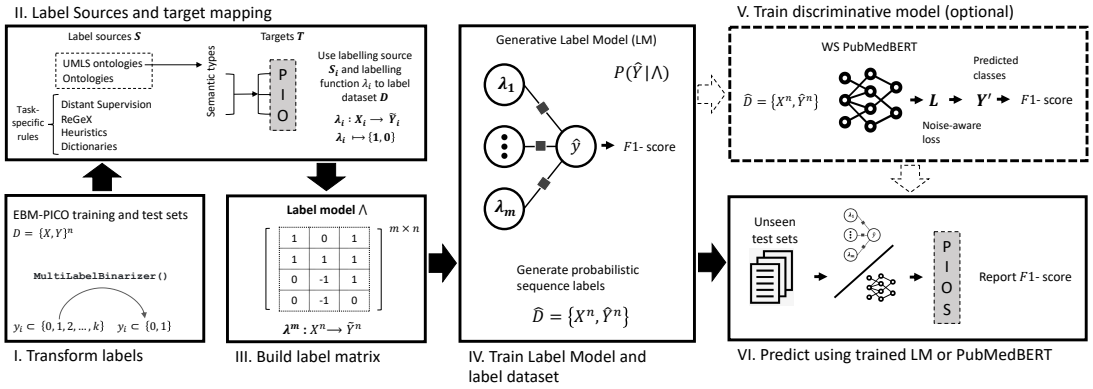


Figure 4.9: Weak PICO entity extraction approach: I) Multi-class labels in the EBM-PICO benchmark are binarized. II) Low-cost UMLS vocabularies are re-purposed as labelling sources, and experts design rules as high-cost labelling sources. III) Labelling functions map the training sequences to class labels using labelling sources resulting in an $m \times n$ label matrix. IV-V) The label matrix is used to train a generative model that outputs probabilistic labels that a downstream transformer model can use for entity recognition.

4.4.2.1 Datasets

EBM-PICO is a widely used dataset with multi-level PICO annotations: span-level or coarse-grained and entity-level or fine-grained (see Table 4.14). Span-level annotations encompass the maximum information about each class. Entity-level annotations cover the more fine-grained information at the entity level, with PICO classes further divided into more semantic subclasses. The dataset comes pre-divided into a training set (n=4,933) annotated through crowdsourcing and an expert annotated gold test set (n=191) for evaluation [254]

The EBM-PICO original paper and annotation guidelines caution about variable annotation quality[42]. Abaho *et al.* developed a framework to post-hoc correct EBM-PICO outcomes annotation inconsistencies [1] Lee *et al.* studied annotation disagreements suggesting variability across the annotators [188] Low annotation quality in the training dataset is excusable, but the errors in the test set can lead to faulty evaluation of the downstream ML methods. About~1% of the EBM-PICO training set tokens were evaluated to gauge the possible reasons for the fine-grained labelling errors and use this exercise to conduct an error-focused PICO re-annotation for the EBM-PICO gold set. The paper's first author, who has a master's in life science informatics and relevant experience in manual curation projects, carried out the re-annotation.

[42] `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6174533/bin/NIHMS988059-supplement-Appendix.pdf`

The dataset is pre-tokenized and did not require additional preprocessing except the addition of POS tags and token lemma using spaCy[43]. Multi-class fine-grained PICO annotations were binarized, i.e. a token label was reset to 1 if the token represented a fine-grained entity.

	P	I/C	O
0	No label	No label	No label
1	Age	Surgical	Physical
2	Sex	Physical	Pain
3	Sample size	Drug	Mortality
4	Condition	Educational	Side effect
5		Psychological	Mental
6		Other	Other
7		Control	

Table 4.14: P (Participant), I (Intervention) and O (Outcome) represent the coarse-grained labels that are further divided into respective fine-grained labels. The table is taken from Nye *et al.*[254]

4.4.2.2 Binary Token Labelling

Automatic PICO entity labelling was considered a classical binary token labelling problem whereby a labeller maps an input sequence of n text tokens, $\boldsymbol{X} = (x_1, x_2, \dots, x_n)$ to output sequence $\boldsymbol{Y} = (y_1, y_2, \dots, y_n)$, where $y_i \subset y; y = \{1, 0\}$ is the label for token x_i. In weak supervision, $\boldsymbol{Y}$ is latent and should be estimated by aggregating several weak labellers of variable accuracy. The estimates $\hat{\boldsymbol{Y}}$ of $\boldsymbol{Y}$ are assigned as probabilistic token labels of $\boldsymbol{X}$ leading to a weakly labelled dataset that can be used to train discriminative models.

4.4.2.3 Labelling Functions

In a binary token labelling task, a labeller, also called a labelling function (LF) is a weak classifier λ that uses domain-specific labelling sources $\boldsymbol{S}$ and some logic to emit binary token labels $\widetilde{\boldsymbol{Y_i}}$ where $\widetilde{y} \in \{-1, 0, +1\}$ for a subset of input $\boldsymbol{X_i}$ tokens. An LF designed for a particular target class $t \in \boldsymbol{T}$ (here; $\boldsymbol{T} \subset \{Participant, Intervention, Outcome\}$) should output 1 for the positive token class label, 0 for the negative token class label, and abstain (-1) on the tokens where it is uncertain $\lambda \mapsto \{-1, 0, +1\}$. Three LF types depending on the types of labelling sources were defined and designed. 1) The ontology LFs for a target class take a dictionary of terminologies with each terminology mapped to one of $y \subset \{0, +1\}$ token target labels. Any labelling function using terminologies used string matching as the labelling heuristic. Relevant bigram word co-occurrences were used to account for fuzzy span matching from the terminologies. A bigram was considered relevant for a vocabulary if it occurred ≥ 25 times in that vocabulary. 2) A ReGeX (regular expression) labelling function for a target class uses regular expression sources for both negative and positive token class $\{0, +1\}$ labels and abstains from the rest. 3) A heuristic labelling function is personalized for each target class and takes a generic regex pattern and specific POS (part-of-speech) tag signals. Abbreviations in clinical studies

[43] `https://spacy.io/`

are considered using a heuristics abbreviation identifier, and the identified abbreviations were mapped to their respective target classes. Stopwords from Natural Language Tookit (NLTK)[44], spaCy, Gensim[45], and scikit learn[46] were used to initialize negative token label templates.

4.4.2.4 Labelling Sources

This section describes the labelling sources $\boldsymbol{S}$ used and their mapping to the PICO targets $\boldsymbol{T}$. The 2021AB-full release of the UMLS Metathesaurus English subset with 223 vocabularies was used. After removing non-English and zoonotic vocabulary and the vocabularies containing fewer than 500 terms, 127 vocabularies remained [161]. Terms in the selected vocabularies were preprocessed by removing stopwords, numbers, and punctuation. The following non-UMLS vocabularies were used: Disease Ontology (DO), Human Phenotype Ontology (HPO), Ontology of Adverse Events (OAE), Chemical entities of biological interest (ChEBI), Comparative Toxicogenomics Database (CTD) - Chemical and Disease subclasses, Gender, Sex, and Sexual Orientation Ontology (GSSO), Chemotherapy Toxicities Ontology (ONTOTOX), Cancer Care: Treatment Outcomes Ontology (CCTOO), Symptoms Ontology (SYMP), Non-pharmacological Interventions Ontology (NPI), Nursing Care Coordination Ontology (NCCO) [77, 118, 140, 179, 198, 239, 251, 281, 282, 297]. The Table4.15 details links to the following non-UMLS labelling sources used. ReGeX and heuristics like POS tag cues were used to capture recurring class-specific PICO patterns otherwise not captured by standardized terminologies. Vocabularies are structured, standardized data sources that do not capture writing variations from clinical literature and custom-built ReGeX are restricted by either task or entity type [276, 288] Distant supervision dictionaries were created from the structured fields of `https://clinicaltrials.gov/` (CTO) as described by Dhrangadhariya *et al.*[89] Principal investigators of the clinical study manually enter data in CTO, thereby incorporating large-scale writing variations [46].

4.4.2.5 Sources to Targets

Along with the source $\boldsymbol{S}$ and the logic to map S_i to token labels, an LF needs information about which target T_i label and binary token class to map the source. This section reports on how the LF sources were mapped to PICO targets. UMLS 2021AB-full release contains 16,543,671 concept names, making direct concept to PICO target mapping impractical. These concepts are organized under semantic type categories (e.g. disease, signs and symptoms, age group, etc.)[47], which allows mapping these semantic categories to PICO targets invariably mapping the concepts from the vocabularies to target classes [222] It is a challenging expert-led activity, though decomposing PICO into subclasses greatly helps map sources to a target. A semantic category was marked 1 to represent a positive token label for that target class or 0 to represent a negative token label for that target class. Non-UMLS vocabularies were obtained from NCBO bioportal[48] and were chosen to be PICO target specific and assigned to a single label. Structured fields from CTO were used to create target-specific distant supervision dictionaries. The structured

[44] `https://www.nltk.org`
[45] `https://radimrehurek.com/gensim/`
[46] `https://scikit-learn.org`
[47] `https://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html`
[48] `https://bioportal.bioontology.org/`

Target	Ontology	Source
P	DO	bioportal.bioontology.org/ontologies/DOID
P	HPO	bioportal.bioontology.org/ontologies/HP
O	OAE	bioportal.bioontology.org/ontologies/OAE
I	ChEBI - Chemical	bioportal.bioontology.org/ontologies/CHEBI
P	ChEBI - Disease	bioportal.bioontology.org/ontologies/CHEBI
O	CTD	ctdbase.org/
P	GSSO	bioportal.bioontology.org/ontologies/GSSO
O	CCTOO	bioportal.bioontology.org/ontologies/CCTOO
O	ONTOTOX	bioportal.bioontology.org/ontologies/ONTOTOX
P	SYMP	bioportal.bioontology.org/ontologies/SYMP
I	NPI	bioportal.bioontology.org/ontologies/NPI
I	NCCO	bioportal.bioontology.org/ontologies/NCCO

Table 4.15: The table lists the links for the non-UMLS ontologies used in work along with the PICO (P = Participant, I = Intervention and O = Outcome) target class the ontology was mapped.

CTO field "Condition or Disease" was mapped to the participant target, and the "Intervention/Treatment" field was mapped to the intervention target. The semi-structured "Primary Outcome Measures" and "Secondary Outcome Measures" fields were mapped to the outcome target. The hand-crafted dictionaries for outcomes were designed using the official websites listing patient-reported outcome (PROM) questionnaires[49] and PROMs[50]. Other hand-crafted dictionaries were separately designed for participant gender and sexuality and intervention comparator terms.

4.4.2.6 LF Aggregation

Depending upon the number of sources $\boldsymbol{S}$ for each $\boldsymbol{T}$, we obtained several LFs. Each LF $\lambda_i \in \boldsymbol{\Lambda^m}; \boldsymbol{\Lambda} = \{\lambda_1, \lambda_2, \ldots, \lambda_m\}$ maps a subset of inputs $\boldsymbol{X^n}$ to output sequence $\widetilde{\boldsymbol{Y^n}}$ with labels $\widetilde{y} \in \{-1, 0, +1\}$ yielding a label matrix $\lambda \subset \{-1, 0, +1\}^{m \times n}$. The weakly-generated labels might have conflicts and overlaps and are generally noisy. These LFs can be ensembled using the majority vote (MV) rule, where a token label is elected only when a majority of λ_i votes for it. Ties and abstains lead to the selection of the majority label.

$$\hat{\boldsymbol{Y}}_{\boldsymbol{MV}} = \max_{y \subset \{0,1\}} \sum_{i=1}^{m} \mathbb{1}(\lambda_i = y_i) \tag{4.9}$$

However, MV considers each labelling function as conditionally independent and does not consider the variable accuracies of different labelling sources weighing them equally. Snorkel implements data programming paradigm into the label model (LM) that re-weights and aggregates labelling functions into probabilistic labels $\hat{y_i}$. To do this, the label model trains a generative model $\boldsymbol{P}(\Lambda, Y)$ to estimate LF accuracies θ_j using stochastic gradient descent to minimize log loss in the absence of labelled data [96, 276] Even though the ground truth is not observable to estimate accuracies, they can be estimated using observed

[49] https://www.thoracic.org/members/assemblies/assemblies/bshsr/patient-outcome/

[50] https://www.safetyandquality.gov.au/our-work/indicators-measurement-and-reporting/patient-reported-outcomes/proms-lists

agreement and disagreement rates between labelling function pairs λ_i, λ_j in Λ. Generative modeling ultimately results into token label probablities $\hat{\boldsymbol{Y}}$ for label classes $\{0, 1\}$. GridSearch was used to fine-tune the label model parameters using the hand-labelled validation set from the EBM-PICO. The parameters are listed in the Experimental details section of the supplementary material. Once the pseudo-labels are generated by majority voting or the label model, these could be used to train a discriminative model.

$$\hat{\boldsymbol{\theta}} = \arg\min_{\theta} \big(- \log \sum_{Y} p_\theta(\Lambda, Y) \big) \tag{4.10}$$

4.4.2.7 Experiments

The labelling functions λ_m were used to label the EBM-PICO training set and obtain Λ. Methods like MV and LM were tested to aggregate LFs. LM output probabilistic labels for the training set were used as weak supervision signals to train downstream PubMedBERT to minimize noise-aware cross-entropy loss. PubMedBERT was trained on PubMed literature and was chosen because of its domain similarity to our training data (PubMed abstracts) and task [131]. It was tuned on fixed parameters listed in the experimental details section in the supplementary material.

$$\hat{\boldsymbol{\omega}} = \arg\min_{\omega} \frac{1}{N} \sum_{i=1}^{n} \mathbb{E}_{\hat{y} \sim \hat{Y}} \big[l\big(f(x, w), \hat{y}\big) \big] \tag{4.11}$$

UMLS ontologies are readily-available sources of weak supervision, while searching the non-UMLS ontologies requires an additional effort and understanding of the target class and domain. On the contrary, designing the rules requires understanding the idiosyncratic clinical patterns for the target classes. Therefore, we experiment and report results on three "expense" tiers to gauge the performance changes: 1) UMLS labelling sources, 2) UMLS and non-UMLS labelling sources, and 3) UMLS, non-UMLS and expert-generated rules. Label aggregation via MV and LM along with WS PubMedBERT for the above tiers was tested. The weakly supervised experiments were compared against a competitive, fully-supervised PubMedBERT trained using the hand-labelled EBM-PICO training set. For all the experiments, 80% of the EBM-PICO dataset was used for training and 20% for validation.

UMLS ontologies were ranked and sorted based on the number of n-gram overlaps between the respective terminology and the EBM-PICO validation set. These were then partitioned into 127 partitions, where the first partition combined the entire UMLS into a single LF and was used as the baseline. The last partition kept all the terminologies as separate LFs. Partition-wise performance over the validation set was tracked.

4.4.2.8 Evaluation

The classical macro-averaged F1 and recall for MV, LM, weakly-supervised (WS) PubMedBERT model and the fully-supervised (FS) PubMedBERT model was reported. Token-level macro-F1 was chosen to make it comparable to the PICO extraction literature. Mean macro-averaged scores are reported over three runs of each model, with the top three random seeds (0, 1, and 42) used in Python. The models were separately trained for each target class recognition task using the raw (IO) tagging scheme. Students t-test with an alpha α threshold of 0.05 was used to measure the statistical significance.

4.4.3 Results

We extended the EBM-PICO subclasses (see Table 4.14) to better query the labelling sources and design LFs (see Figure 4.10). For a more comprehensive subgrouping, we propose developing a PICO ontology [291] It is more straightforward to search for ontologies representing adverse events or diseases rather than fending for an ontology describing the entire participant or outcome span. It is easier to grasp cues separately for outcome terms and instruments of outcome measurement to develop heuristics. The intervention span can include the intervention name, role (primary intervention or comparator), dosage, frequency, mode of administration and administrator. The outcome span can include the outcome names, the scales, techniques or instruments used to measure them and the absolute outcome measurement values. The EBM-NLP guidelines restrict annotating the outcome name, how it was measured, and the intervention's name and role (control, placebo), leaving out the other subclasses.

Figure 4.10: Hierarchical representation of PICO subclasses. The categories marked in bold italic are the same as the fine-grained categories in the EBM-PICO corpus.

4.4.3.1 Error Rectification

The errors in the EBM-PICO validation subset were rectified and categorized them for each PI(C)O class, as shown in Table 4.16. Of 12,960 (~1% of 1,303,169) validation tokens evaluated to gauge the errors, 18.30% of the intervention class tokens, 23.39% of the participant class tokens and 20.21% of the outcome class tokens were errors. The error analysis was used to correct fine-grained annotation errors in the EBM-PICO test

set, and both the EBM-PICO and its updated version were used for evaluation. We were constrained with obtaining multiple annotators for the re-annotation to calculate inter-annotator agreement (IAA). Therefore, the Cohen's κ_{new} was calculated between the original EBM-PICO gold set and our re-annotation over 200 documents and compared it to Cohen's κ (see Figure 4.11) provided by the authors of the original corpus [254]

Table 4.17 reports macro-averaged F1 for the experiments detailed in the experiments section compared to the fully-supervised approach. Error rectification leads to an overall average F1 improvement of 4.88% across the experiments using a weakly labelled training set with the highest average improvement of 8.25% (7.15%-9.52%) for participants and 2.68% (-0.11%-4.28%) for outcomes. For the participant class, both the LM and the WS F1 scores increase the full supervision score by 0.90% - 1.71%. It has to be noticed that weak supervision outperforms full supervision on the rectified benchmark only for the participant entity.

Error category	Participant	Intervention	Outcome
Repeat mention unmarked	213	227	207
Remain un-annotated	47	59	71
Inconsistency	46	18	85
Punctuation/article	15	23	48
Conjunction connector	30	36	57
Junk	53	79	30
Extra information	80	146	58
Generic mention	70	120	85
Total errors	554	708	641

Table 4.16: Error distribution and error categories in the analysed tokens (~1%) of EBM-PICO corpus.

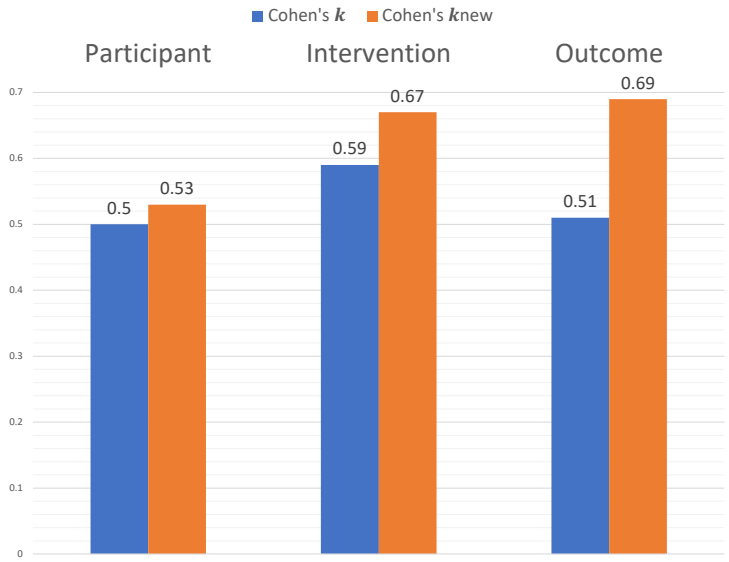


Figure 4.11: Cohen's κ_{new} between the expert annotated EBM-PICO gold test set and EBM-PICO compared to the Cohen's κ for EBM-PICO gold test set annotations.

4.4.3.2 Model Training Results

			MV		LM		WS		FS	
Target	LF source	#LF	Fine	Corr	Fine	Corr	Fine	Corr	Fine	Corr
P	UMLS	3-4	62.13	69.28	64.28	72.22	65.32	73.49	72.99	74.41
	+Ontology	4	61.72	69.32	64.23	72.18	64.76	72.31		
	+Rules	19-119	63.08	72.06	65.79	75.31	66.73	**76.12**		
I/C	UMLS	8-95	59.7	63.94	60.11	64.28	59.17	61.72	**83.37**	81.06
	+Ontology	5-101	62.14	66.92	62.83	67.09	67.06	69.76		
	+Rules	4-35	58.51	63.45	64.34	68.17	70.27	72.39		
O	UMLS	5-6	55.79	59.85	58.76	62.36	60.83	63.55	**81.2**	80.53
	+Ontology	4-5	56.006	59.64	59.27	62.34	59.55	60.46		
	+Rules	3-5	55.08	59.36	60.9	62.87	60.5	60.39		

Table 4.17: Macro-averaged F1 scores for UMLS, UMLS+other and rule-based weak supervision. Underlined values show the best score without manually labelled training data. Bold values show the best overall F1 score in any category. Note: Fine = EBM-PICO fine-grained annotations, Corr = EBM-PICO fine-grained annotations (EBM-PICO updated)

MV vs. LM vs. WS The label model improved the average performance by 2.74% (0.17%-5.83%) compared to majority voting. However, PubMedBERT did not guarantee improved performance across the targets leading to performance drops between 0.4 - 2.56%. Though the weakly-supervised PubMedBERT models did not always improve the performance compared to their label model counterparts, they had the best F1 score for each target class. The majority voting had higher recall across experiments compared to precision, while LM focused on precision (see Figure 4.13), making it a possible choice for recall-oriented PICO extraction tasks.

LF tiers Adding non-UMLS LFs to the UMLS tier increases performance for the intervention target by an average of 4.48%, but leads to performance drops for the participants and outcomes targets by 0.36% and 0.64%, respectively. Adding task-specific LFs increased the overall F1 by a negligible 0.98%. Heuristics improved performance for the interventions LM by 11.1%.

UMLS partitions To investigate the optimal number of UMLS labelling functions required, we used the same methodology as in Trove, holding all non-UMLS and heuristics LFs fixed across all ablation tiers and computed performance across $s = (1, 2, \ldots, 127)$ partitions of the UMLS terminology. We noticed an increased performance for the first few partitions. However, We did not see the performance drop with a further increase in the participant and intervention target partition number. Partitions with more than 100 LFs performed better. This situation contrasts with Trove, where an increase in partitions leads to a drop in performance across targets (see Figure 4.12). For the outcomes target, an increase in the number of partitions leads to an increased performance initially but a drop with a further increase in the partition numbers. LM outperforms MV on training performance across the two targets and experiments except for the intervention target,

where the MV model combining UMLS and additional ontologies outperforms LM. The simple baseline collapsing UMLS into a single LF usually did not perform better than the others in UMLS partitions for any of the three experiment tiers (refer to the #LF columns in Table 4.17).

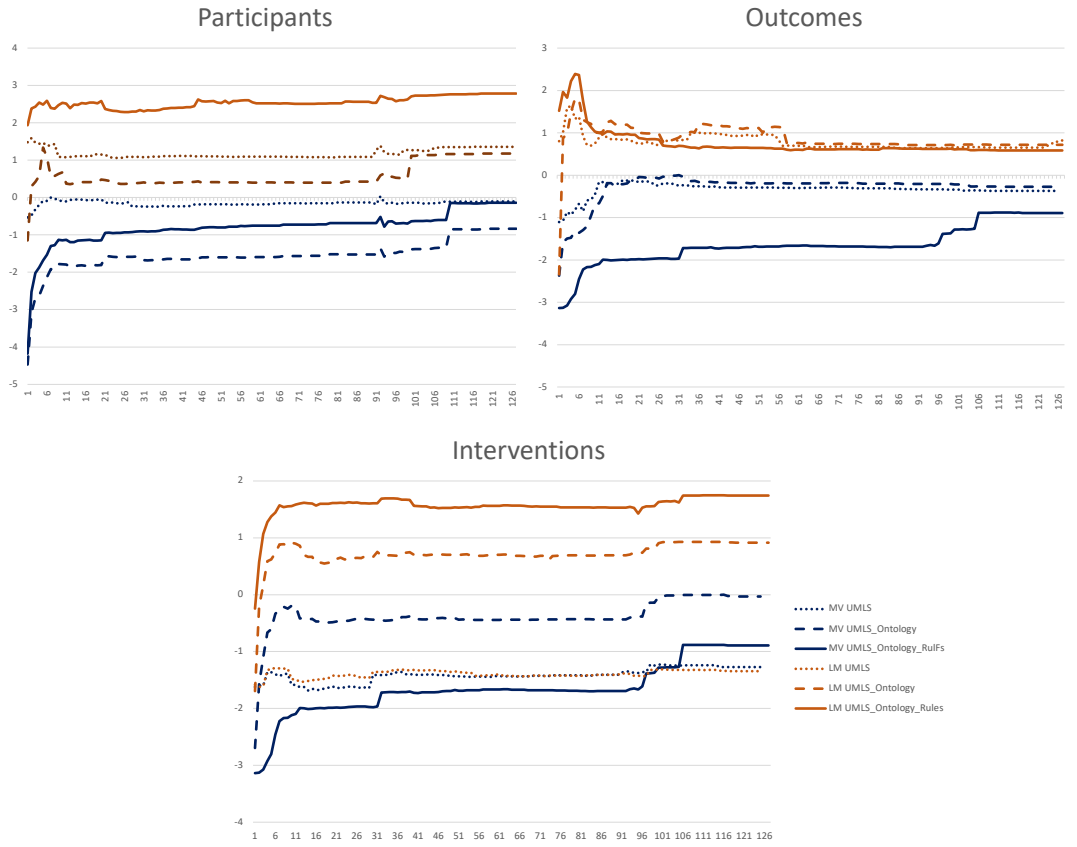


Figure 4.12: The relationship between the number of UMLS partitions and the macro-averaged F1 score for i) participants target, ii) interventions target and iii) outcomes target.

4.4.4 Discussion

Our study results show the promising performance of weak supervision compared to full supervision, surpassing it for participant extraction. It has to be noticed that weak supervision requires careful LF design consideration to surpass full supervision, primarily due to the compositional nature of PICO classes. In another study, we use this weak supervision approach to successfully extend PICO to PICOS extraction (S-Study type) without needing additional annotated "Study type" data to quickly power applications [91]

Although it is easy to re-purpose the vocabulary for labelling, it is challenging to map them to the correct PICO targets. A decreased or stagnant F1 after adding non-UMLS LFs to the UMLS tier indicates this. The performance boost using rule-based LFs was only observed in the participants and interventions and was detrimental to the outcomes.

Even though LM improves performance compared to MV, MV has a higher recall across experiments indicating a good corpus coverage of the LFs (refer to Figure 4.13).

While some studies press on PICO extraction being a recall-oriented task, this is debated in practice. In practice, high recall might lead to a high false positive (FP) rate, leading to the reviewers spending more time weeding out FP noise than reading and annotating the abstract with the entities [202]

LM only considers the information encoded in the weak sources to label phrases from the training text but does not consider the contextual information. Transformers consider the contextual information and should generalize beyond the label models in theory. It is empirically confirmed by the performance boost that PubMedBERT brings this on top of the label model for some instances, but the weakly-supervised outcomes extraction results refute it.

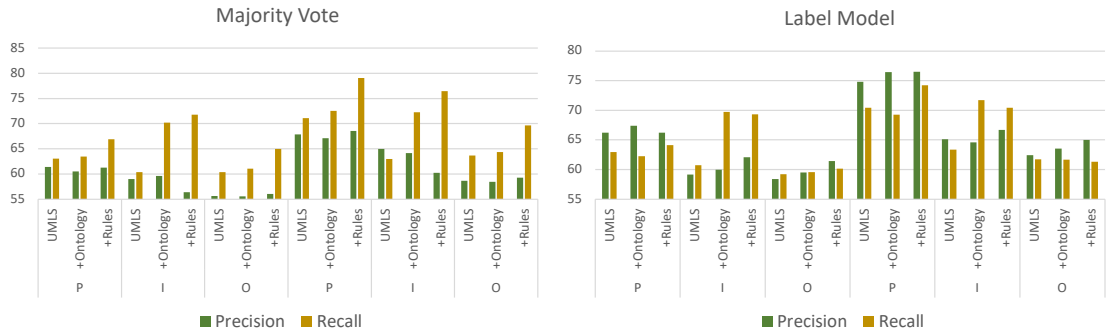


Figure 4.13: Precision and recall across the experiments for the I. Majority vote models (left) and II. Label models (right)

4.4.4.1 Error Analysis

We conducted an error analysis on 18 (n = 5,291 tokens) out of 200 EBM-PICO gold test set documents to contextualize the weak supervision models. Table4.18 shows token level errors divided into either of the four classes: 1) false negative (FN) - if the entire entity that the token was part of was missed out by the LM, 2) false positive (FP) - if the entire entity that the token was part of was falsely recognized as an entity, 3) boundary error (BE) - if the boundary tokens were missed out, but otherwise the entity was identified, and 4) overlapping error (OE) - if LM made an error in the non-peripheral tokens of an otherwise identified entity mention. Non-peripheral tokens are all tokens except the first and last of the multi-token entity.

In future, we aim to reduce FNs and dig deeper into this category. Besides participant disease, tokens representing participant sample size, age group, gender, and symptoms subclasses went unrecognized. The LM labelled these FNs with low confidence, meaning the LFs did encode this information, but the signals were not strong enough for correct classification. Such FNs could be mitigated by weighting LFs for these subclasses. Considerable standard and unusual abbreviation terms were missed out, especially the ones encompassed by brackets, e.g., MetS (Metabolic Syndrome), TC+TSE (testicular cancer + testicular self-examination), LVH (left ventricular hypertrophy), *etc.*. The model did not pick some of the standard abbreviations, e.g. LVH and MetS, due to a faulty mapping of these abbreviations to the incorrect PICO target. A similar pattern was observed for the intervention (e.g., IBA (Inference-based Approach), RT (Radiotherapy)) and outcomes class abbreviations too. The mismapping is now amended. LM did not capture the

abbreviations enclosed in a bracket (e.g., "(COPD)") as the LFs were not designed to tag these brackets.

Intervention LMs did not recognize common drug names, e.g. Fenofibrate and CP-529,414. In addition, many non-standardized treatment names went unrecognized, e.g. substance abuse prevention program, inference-based approach, high-concentration contrast agents, epigastric impedance, *etc.* Such terms are absent from UMLS and non-UMLS vocabularies leading to FNs, so the LFs do not encode them. Similarly, intervention BEs were the non-idiosyncratic tokens partially misrecognized because the vocabulary did not encode this partial information. E.g., the term "internal stenting" is partially recognized because "stenting" is a UMLS concept but not "internal". Similarly, in the term "endopyelotomy stent placement," only the UMLS concept token "stent" was identified. Participant BE FNs were usually the extra information that described more about the participant's disease, e.g., the information about disease staging went unrecognized in the participant's disease entity (in "advanced carcinoma", the word "advanced" was a BE FN). Such entities not encoded by the LFs contribute to the FNs and could be mitigated by adding relevant vocabulary and rules [15] It is straightforward to add vocabulary but challenging to map a semantic group or a vocabulary to PICO categories, especially for the outcomes class. The current source-to-target mapping approach is manual and based on subjective expert judgment which an objective algorithm can improve. This mapping could have led to several unexplainable errors, especially for the outcomes and, to a less extent, for the intervention class. Additionally, it took time to identify semantic categories and UMLS vocabularies corresponding to the study outcomes pointing towards the gap in developing one.

	FP	FN	BE	OE
Participants	160	76	80	10
Interventions	308	119	60	0
Outcomes	233	306	139	7

Table 4.18: Distribution of the token-level errors made by the best label models on EBM-PICO gold.

Some of the error categories identified in this work were also identified by Abaho *et al.*, but this work adds more classes on top of them. While they limit their error exploration to the outcomes class, we extend it to the rest. An error falls under *repeated mention* if one instance of an entity is marked, but another identical instance of the same entity in the same context is not marked within the abstract. The reason can be the EBM-PICO guidelines flaw where the annotators of fine-grained entity annotation were confined to only annotate within the longer span-level annotation. Hence, any annotation error missed by the coarse-grained annotators was continued by the fine-grained annotators.

An error falls under *remains unannotated* if a token should have been annotated as an entity but was not. In the intervention class, a large portion of this category was constituted by the generic mentions of controls (placebo, saline), which were not annotated. In the participant class, patient ethnicity and other information like smoking status and pregnancy (marked in the coarse-grained span) were not marked in the fine-grained entity. The reason could be that there was no fine-grained class to categorize this information. The annotators missed multiple important outcomes and repeated mentions in the outcome class.

Conjunction connector errors are the conjunctions occurring between two semantically separate entities but are falsely marked as entities. For example, "Nausea and vomiting" are two different outcomes marked as one by annotating the conjunction between them. When falsely marked as an entity, punctuation succeeding the entity, an article, or a preposition preceding the entity fall under the *punctuation/article* errors. Extraneous tokens marked along with the entity tokens fall under *extra information* category. For example, in the phrase "This trial demonstrated short-term efficacy of smokeless tobacco in combination with", the annotators had marked "short-term efficacy of smokeless tobacco" as an outcome entity, but only "short-term efficacy" is an outcome entity. In contrast "smokeless tobacco" is an intervention entity. In the intervention class, the annotation guidelines mentioned not annotating any part of the text that did not mention the intervention name. The annotators often marked extraneous information like intervention dosage, frequency, route of administration and information about the intervention administrator.

A generic reference is a co-reference of an entity mentioned using different or similar (but not identical) words. A generic reference of an entity (and its repeated mention) in the same abstract was several times left unmarked by the annotators constituting a *generic reference* error. For example, if the outcome endpoint "smoking cessation" was referred to in the same abstract elsewhere as "quitting smoking", it was not marked even though it is a reference to the outcome phrase "smoking cessation". If "aerobic exercise" mentioned as "exercise intervention" was not marked, this also constitutes a generic reference error. For instance, in an RCT, if the "breast cancer risk counselling" intervention was referred to as "risk counselling", the former was marked, and the latter was missed. This error was pronounced specially for the non-pharmaceutical interventions and outcomes.

An *inconsistency* error arises when an entity is fully marked in some abstracts vs when in the other abstracts the same entity is partially marked. For example, if an exercise intervention involved aerobic exercise involving stretching and running, this information ("stretching", "running") was marked in some studies while not in all the other studies. In the case of the participant class the sample size sub-grouping information needed to be more consistently marked. In another example, participant sample size information was either partially or entirely marked ("59" vs "59 subjects", "200" vs "200 controls"). We consider a phrase as a sample size only when the absolute value quantifying sample size follows an appropriate unit (subject, controls, patients, participants, women, men). Only in some cases, when the unit was unavailable, did we consider a standalone number as a sample size participant span. In the outcome class, the annotation guidelines marked "what was measured and how it was measured". Often, the method used for measuring outcomes was inconsistently marked.

A *junk* error is a token entirely irrelevant to an entity but is marked as one. For example, in the outcomes evaluation, the phrase "evaluate and compare" from the larger phrase "This study aimed to evaluate and compare" was marked as an outcome entity even though it is not a valid outcome.

These errors and inconsistencies in the EBM-PICO gold test (and training) set can cause faulty evaluation of the machine learning approaches defying the purpose of the corpus. A possible reason behind these inconsistencies in the corpus could be that the annotators had clinical background but lacked an informatics background. This situation could undermine the importance of semantic consistency required for annotating such corpora. Hiring annotators with a combined knowledge of clinical and informatics domains might improve the manual annotation quality. Aggregation of crowd annotations for spans

with fuzzy span boundaries might lead to many boundary errors, for example, when two disparate entities are linked by a *conjunction connector*.

4.4.5 Conclusion

This work successfully adapted weak supervision for PICO spans and developed models for predicting PICO entities without a hand-labelled corpus. Another contribution of this work was the errors identified from the current PICO benchmark; rectified them and used both datasets to evaluate the recognition models. Compared to full supervision, the approach achieves promising performance and warrants further research into weak supervision for challenging PICO extraction. In the future, we will work on extending the data programming approach to inspect strategies for objectively mapping ontologies to PICO subclasses and experiment using external models like MetaMap as LFs. The approach can be extended to more clinical SR entities without a manually labelled corpus, thereby being a starting point to overcome the annotation bottleneck. The next section explains how the weak supervision approach was extended to extract an additional "study type" entity without hand-labelled data.

4.5 PICO to PICOS: Using Weak Supervision to Extend EBM-PICO Dataset with Study type (S) Information

4.5.1 Background and Significance

During the citation screening phase, PICO analysis frequently extends to analysing other information like Study type and design, study context, timeframe, trial duration and background, *etc.* depending upon the SR question and inclusion criteria [8, 228, 280, 333]. Study-type information is vital, for example, in conducting systematic reviews that aggregate evidence from selected clinical study types, for e.g., Randomized Controlled Trials. Trial duration information is essential for establishing the long-term efficacy of the treatment [227]. Ethnicity of participants in important for the pharmaceutical SRs concerning intervention effects on particular patient populations [9, 245]. The challenge lies in the expensive and labor-intensive manual re-annotation of extensive datasets like EBM-PICO, which currently lack annotations for these additional entities.

Weakly supervised (WS) information extraction is a powerful technique that allows for the programmatic labelling of datasets using noisy and imprecise data labelling sources with varying accuracies. Common freely-available sources include ontology compendiums such as UMLS (Unified Medical Language System) and other terminologies available from NCBO BioPortal, which can be repurposed for programmatic labelling [111, 161]. Weak supervision sources utilize multiple imprecise labelling sources to label datasets followed by aggregating these annotations into consensus labels. The task of aggregating these labels obtained from multiple noisy sources was earlier addressed using methods such as majority voting and joint conditional probability models [154, 302, 322]. Soft majority voting techniques takes into account the confidence or probability scores assigned by each classifier or voter. Instead of considering only the most common prediction, soft voting assigns weights to each classifier's prediction based on its confidence level. For example, max-margin majority voting takes into account the distance from the decision boundary to the closest data point of each class for each classifier in the ensemble. Larger margins indicate higher confidence in the prediction and higher the weight a voter gets. Soft majority

voting does not explicitly model the relationships or dependencies among labels or features, potentially leading to sub-optimal label aggregation. Joint conditional probability can capture the intricate relationships and dependencies among labels or features Label aggregation has now transitioned to training generative models that estimate the accuracy of each labelling source to obtain a weighted programmatic label [195, 200, 277, 288]. Generative modelling allows estimating the accuracy of each labelling source to obtain a weighted programmatic label.

Previous studies have explored weakly-supervised biomedical and clinical document classification and relation extraction. However, clinical information extraction (IE) still heavily relies on hand-labelled data [98, 209, 226, 237, 351]. For instance, Wang *et al.* employed DS for biological entity extraction by combining multiple dictionaries into a single labelling operator [349]. Zhou *et al.* too used a single source to weakly label chemical and disease information using PubTator pipeline followed by correcting the labels using knowledge-bases [374]. Dhrangadhariya *et al.* explored distantly supervised extraction of clinical trial intervention extraction but relied only on the clinical trials database[51] for the noisy labelling [89]. Fries *et al.* used biomedical ontologies and regular expressions to train their weakly-supervised system called Trove for extracting entities like diseases, chemicals, and drugs [112]. WS requires a set of programmatic labelling sources that have at least some representational value for the entity being weakly labelled. Such sources are commonly available for semantically well-defined biomedical entities but are difficult to obtain and utilize for the fuzzy clinical entities like PICO. Fries *et al.* [112] succeeded in extracting the strictly standardized biomedical entities but extracting entities in the clinical domain is more challenging due to lack of standardization, language diversity, and fuzzy class definitions. Our previous work (see Section 4.4) addressed the challenge of repurposing the weak supervision sources for the fuzzy, compositional PICO information but did not extend it to "Study type and design" and other before-mentioned entities.

Drawing inspiration from the prior work [90], this work demonstrates the feasibility of weak supervision for enriching large hand-labelled corpora, such as EBM-PICO, with new clinical entities. Specifically, we successfully applied weak supervision techniques to generate programmatic labels for the "Study type and design" entity for all 4,081 EBM-PICO documents. This approach was validated using an additional 912 manually labeled documents, serving as validation and test sets for future methodologies. In addition, [90, 112] use expert knowledge to repurpose the UMLS labelling source to the target entity classes. In the absence of expert knowledge, no approach exists to map the labelling sources to these new classes of interest. We have also provided a straightforward algorithm for mapping the weak labelling sources to the new entity class, offering a pragmatic solution in the absence of domain experts. Furthermore, many studies in this domain focus on RCT vs. non-RCT classification [105, 216, 296, 344]. However, these studies do not delve deeper to provide information about the study design, such as whether an RCT was a cluster or parallel trial, which could be crucial for writing a systematic review and so the current research landscape has a gap in addressing this broader aspect of "Study type and design".

4.5.2 Methods

Figure 4.14 schematically represents our approach.

[51] `https://clinicaltrials.gov/`

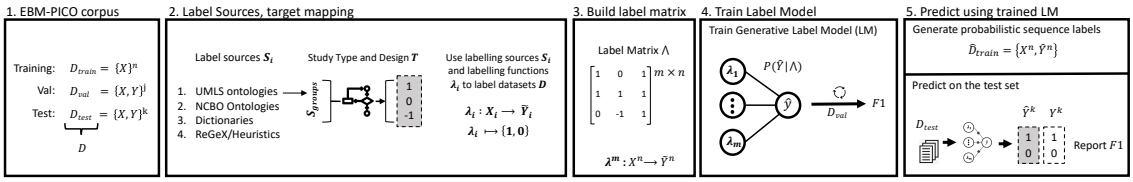


Figure 4.14: WS entity extraction approach: 1. Define the training, validation, and test datasets. 2. Define labelling sources S_i. UMLS vocabularies are reused as labelling sources and mapped to the"Study type and design" class labels (see Algorithm 2). 3. Labelling functions λ_i map the datasets to class labels using S_i resulting in an $m \times n$ (training) label matrix. 4-5) The training label matrix is used to train a generative model that could be used to label unlabelled training sets with probabilistic labels or can be used to predict class labels on unseen test set to evaluate.

4.5.2.1 Dataset

We used the EBM-PICO dataset comprising manually annotated RCTs (randomized controlled trials) to demonstrate the effectiveness of programmatic labelling and the weak supervision approach. The dataset comes pre-divided into a training set (n=4,933) annotated through crowd-sourcing and an expert annotated gold test set (n=191) for evaluation [254]. We further segmented EBM-PICO into a validation set comprising 721 documents ($\sim$ 15% of $4,933$) and reserved the remaining documents to train the weak supervision model. EBM-PICO comes pre-tokenized and no additional preprocessing was done except enrichment with the part-of-the-speech (POS) tags using spaCy[52] [155].

4.5.2.2 Dataset Annotation

We need gold standard annotations to evaluate the WS approach to "Study type and design" information extraction. Specifically, we need a hand-labelled validation set to tune hyperparameters and an annotated test set for final evaluation. However, the EBM-PICO test set is not annotated with the "Study type and design" entity.

We first defined the "Study type and design" class to design the annotation guidelines efficiently. According to the National Institute of Health (NIH), "Clinical studies are the medical research involving people to evaluate the safety and effectiveness of medical interventions, including drugs, devices, procedures, and behavioural interventions, as well as studies that aim to understand the mechanisms of disease, develop new diagnostic tools, or identify risk factors for health conditions"[53]. As such, the "Study type and design" class should comprise text descriptions that provide information on the clinical study's type and design features. For example, the "Study type and design" text description from PMID:36116481 "randomised, double-blind, placebo-controlled, parallel-group, phase 3 trial" elaborates on this clinical trial's type and design.

Next, we defined clear annotation guidelines for the "Study type and design" entity. Following this, the test set (n = 191) was doubly-annotated by two health informatics experts[54] to calculate pairwise F1 measure as inter-annotator agreement (IAA). F1 measure disregards out-of-the-span tokens (unannotated tokens) during agreement calculation and is an ideal measure of annotation reliability for the token-level annotation tasks. It

[52] `https://spacy.io/`

[53] https://www.nih.gov/health-information/nih-clinical-research-trials-you/basics

[54] a Ph.D. student and a postdoc

measures the F1 measure, as shown below, for each pair of annotators, treating one annotator's labels as the "true" labels and the other annotator's labels as the "predicted" labels. It assesses the agreement or overlap between annotations and thus the annotators at the token level [79].

$$\text{F1 measure} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \tag{4.12}$$

where:

$$\begin{aligned} &\text{TP} : \text{True Positives} \\ &\text{FP} : \text{False Positives} \\ &\text{FN} : \text{False Negatives} \end{aligned}$$

The IAA for the 191 test documents between the expert annotators was 78.33%. Deeming the IAA sufficient as per the F1 agreement interpretation in Dhrangadhariya *et al*, the validation set (n=721) was singly annotated [84].

4.5.2.3 Task definition

We define the entity tagging task as a binary sequence labelling task where given an input sequence of tokenized words $X = (x_1, x_2, \ldots, x_n)$ and an output sequence of binary labels $Y = (y_1, y_2, \ldots, y_n)$ where $y \subset \{0, 1\}$ (here 1 = "Study type and design", 0 = OOS token), the task is to train a supervised ML model using the token-label pair.

In weak supervision, the task is to design m weak labelers or labelling functions (LF) λ_m, each of which is a function that takes input sequence X and produces an integer label sequence $\widetilde{Y} = (\widetilde{y_1}, \widetilde{y_2}, \ldots, \widetilde{y_n})$; $\widetilde{y_i} \subset \{1, 0, -1\}$ for "Study type and design" class labels. Output label 1 represents "Study type and design" or positive class label, 0 represents not a "Study type and design" or OOS or negative class label, and -1 are abstains.

As the groundtruth Y is latent in the absence of labelled data, it should be estimated by aggregating several weak labellers of variable accuracy. It has to be noticed that aggregating model can apply differential weights to the abstain labels compared to the 1 and 0 labels. These weights reflect the confidence level associated with each label, allowing the model to appropriately weigh the contributions of abstain labels in the final labelling decision. The estimates $\hat{Y}$ of Y are assigned as probabilistic token labels of X, leading to a weakly labelled dataset that can be used to train discriminative models.

4.5.2.4 Programmatic labelling

In this section, we describe the process of designing labelling functions using labelling sources used to programmatically label EBM-PICO dataset.

Labelling Sources A labelling source s could be a set of terms (including ontology, terminology, vocabulary, dictionary), expert-designed ReGeX, heuristics, or a combination of these sources that encode some domain-specific knowledge. Our labelling sources included available biomedical vocabularies, expert-led rules like ReGeX and heuristics. We used the 2021AB-full release of the UMLS Metathesaurus English subset with 224 vocabularies[55]. After removing the zoonotic and non-English vocabularies, we were left with

[55] `https://www.nlm.nih.gov/research/umls/sourcereleasedocs/index.html`

112 vocabularies. The task to label "Study type and design" class entails using terms (or concepts as they are called in UMLS) within these UMLS vocabularies as representative of this class. Terms are organized under vocabulary like MeSH (medical subject headings) or SNOMED-CT (SNOMED Clinical Terms), *etc.* which are not very helpful for labelling "Study type and design" class labels $\{1, 0, -1\}$. These terms are also organized under 127 semantic groups $S_{groups} = (s_{group_1}, s_{group_2}, ..., s_{group_n}); n = 127$ like "disease", "age group", "geographical location" denoting whether a term represents a disease name or a geographic location. As such semantic groupings impart meaning to these terms. Thus our task was to map these semantic groups to "Study type and design" class as per their representational value, ultimately mapping the terms to "Study type and design" class labels.

Non-UMLS ontologies refer to ontological frameworks not part of the UMLS Metathesaurus. Non-UMLS ontologies like Clinical Trial Ontology (CTO), Randomized Controlled Trials Ontology (RCTONT), Ontology of Clinical Research (OCRe), and Clinical Trials Ontology (CTONT) [199, 307] were used to represent positive class (+1) labels[56]. We provide the links to the following non-UMLS labelling sources used in our work in the Table4.19.

Ontology	Source
RCTONT	`https://bioportal.bioontology.org/ontologies/RCTONT`
OCRe	`https://bioportal.bioontology.org/ontologies/OCRE`
CTO	`https://bioportal.bioontology.org/ontologies/CTO`
RCTONT - Chemical	`https://bioportal.bioontology.org/ontologies/RCTONT`

Table 4.19: The table lists the links for the non-UMLS ontologies used in work.

Handcrafted dictionaries were designed using key-phrases from MeSH containing the generic term "trial"[57]. The terms in this dictionary were used to label positive (+1) "Study type and design" labels. Some example terms include "random allocation", "randomization", "controlled clinical trial", "quasi-experimental study", and "crossover trial". We included hand-crafted dictionaries provided by [90] for the PICO classes as labelling sources for negative/(0) "Study type and design" labels.

The hand labelled validation set was used to develop ReGeX. We examined the most common keyword patterns in "Study type and design" class. These class-specific keyword patterns were used as ReGeX hooks along with the observed POS patterns to emit the positive class label. For example, the trial design information precedes the hook pattern "randomized controlled trial", for example, "multi-arm, double-blind, non-inferiority, **randomized controlled trial**". To identify such domain-specific patterns, a ReGeX was developed to identify the hook pattern "randomized controlled trial" and was combined with position and POS tags to identify preceding trial design information.

Sources to targets mapping We do not use expert knowledge to map UMLS S_{groups} to "Study type and design" targets, instead we conducted a separate experiment using the manually-labelled validation set using the following steps to obtain the mapping.

1. Label the hand-labelled validation set using all the S_{groups}.

[56] `https://bioportal.bioontology.org/`

[57] `https://meshb.nlm.nih.gov/search?searchMethod=FullWord&searchInField=termDescriptor&sort=&size=20&searchType=allWords&from=0&q=trial`

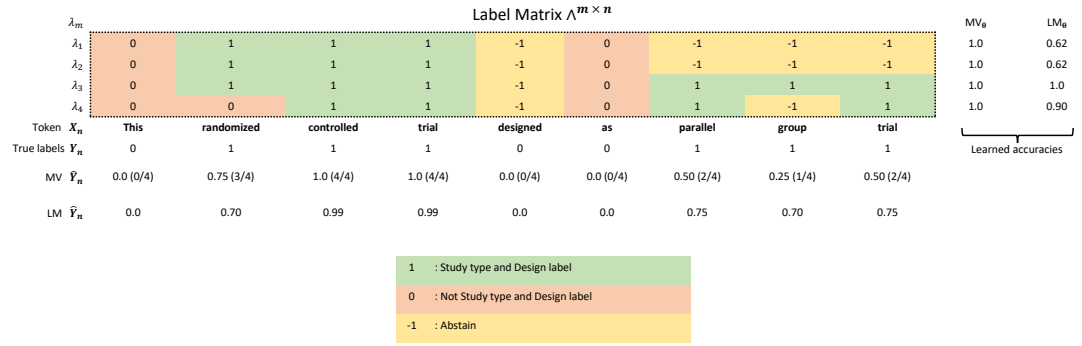


Figure 4.15: The figure shows four LFs used to label token sequence $\boldsymbol{X_n}$ with entities "randomized controlled trial" and "parallel group trial". MV assigns labels based on the equally weighting the importance of each LF. LM uses the $\Lambda_{m \times n}$ to estimate latent accuracies θ_j of LFs using agreement and disagreement rates between LFs. These accuracies are then used to re-weight the labels generating more accurate probablities $\hat{\boldsymbol{Y}}$.

2. Calculate recall for the positive class[58]
3. Rank and sort S_{groups} based on their recall.
4. Next, label the validation set using the S_{group} that ranked 1 and calculate the initial recall r and f1-score $f1$.
5. Then loop through the ranked S_{group} starting at rank 2 and sequentially add labels to the validation set (already labelled with S_{group} rank 1) and calculate the new recall r_i and f1-score $f1_i$ with the combined labels.
6. After looping through all the S_{groups}, Algorithm 2 is used to classify S_{group} into representing the positive (+1) class, negative/OOS (-1) class or abstain (0) class.

We consider a S_{group} representative of the "Study type and design" class if the change in the recall Δr is greater than equal to 1 without impacting the f1-score. Such S_{group} are marked as +1 and the rest as abstain or negative per algorithm 2. The concepts within non-UMLS ontologies and the hand-crafted dictionary entries were mapped to target class +1.

Labelling functions We categorize our LFs into three types depending on the labelling sources.

1. Any ontology LF takes a set of terms (vocabularies, ontologies, *etc.*) each mapped to one of $y \subset \{0, +1, -1\}$ class labels. Dictionary LFs used string matching as the labelling heuristic.
2. A ReGeX LF used only regular expressions representative of the positive token label $\{+1\}$ and abstained from the rest.
3. A heuristic LF often took a generic ReGeX pattern, specific POS (part-of-speech) tag signals, and token positions to label tokens with the positive token class $\{+1\}$ labels and abstained from the rest.

[58]The recall and the F1 score are binary metrics calculated for the "Study type and design" (positive) class.

Algorithm 2 An algorithm to map UMLS S_{groups} to "Study type and design" labels

Require: $D =< S_{groups}, r_{i+j}, f1_{i+j} >$
Ensure: $< S_{groups}, S_{class} >$
Initialize $r_0 \leftarrow r_{S_1}$, $f_0 \leftarrow f_{S_1}$, $S_{class} \leftarrow []$
for S_i, r_i, f_i in $D[1:]$ **do**
    Calculate $\Delta r = r_i - r_0$, $\Delta f = f_i - f_0$
    if $\Delta r \geq 1$ **and** $\Delta f > 0$ **then**
        S_{class}.insert(1)
    else if $0 \geq \Delta r < 1$ **and** $\Delta f > 0$ **then**
        S_{class}.insert(-1)
    else if $\Delta r = 0$ **and** $\Delta f < 0$ **then**
        S_{class}.insert(0)
    else if $\Delta r < 0$ **then**
        S_{class}.insert(0)
    end if
end for

Labelling Consider $S = (s_1, s_2, \ldots, s_x)$ set of labelling sources used by m LFs to programmatically label the X_n EBM-PICO training tokens to the aforementioned integer labels $(-1, 0, 1)$. The LFs map X_n input tokens to the integer label sequence $\widetilde{Y}_n$ leading to a label matrix $\Lambda^{m \times n}$. It has to be noted that these labels are noisy in nature, and their accuracy depends upon the labelling source and function. The next task is to aggregate the labels from these LFs to obtain a consensus label sequence $\hat{Y}$.

Label aggregation Majority voting (MV) and generative label model (LM) are tested to aggregate labels in the label matrix. Snorkel implements data programming paradigm into the label model (LM). LM learns varying accuracies of LFs, weights them accordingly, and aggregates them into probabilistic labels $\hat{y_i}$. To do this, LM trains a generative model $\boldsymbol{P}(\Lambda, Y)$ to estimate LF accuracies θ_j using stochastic gradient descent to minimize log loss in absence of labelled data. [96, 276] The true labels are not observable to estimate accuracies, but they can be estimated using observed agreement and disagreement rates between LF pairs λ_i, λ_j in Λ. Generative modeling ultimately results into token label probablities $\hat{Y}$ for label classes $\{0, 1\}$. An example of combining LFs using MV and LM into consensus labels is shown in Figure 4.15. Figure 4.15 shows an example of LFs using MV and LM into consensus labels.

4.5.3 Experiments

The experiment aimed to evaluate the impact of sequentially adding labelling sources on the label aggregation methods of MV and Snorkel's LM. We ranked various labelling sources based on their costs, with UMLS labelling sources being the least expensive as they require no extra effort of web searching to obtain. Non-UMLS sources follow them in terms of cost, and then come manually-crafted dictionaries and ReGeX, which are the most expensive and require expert knowledge. The experiments were carried out in seven tiers, with tiers 1-4 testing the addition of non-UMLS, dictionaries, and rule-based LFs to UMLS LFs in sequence (least to most costly labelling sources). Tier 5 examined whether up-weighting rules could improve performance, while tiers 6 and 7 measured the effect of

removing non-UMLS and dictionaries from tier 4 on performance. We performed addi-

Tier	Labelling sources s
1	UMLS
2	UMLS + Non-UMLS
3	UMLS + Non-UMLS + dictionaries
4	UMLS + Non-UMLS + dictionaries + ReGeX
5	UMLS + Non-UMLS + dictionaries + weighted ReGeX (ReGeX $\times$ 2)
6	UMLS + Non-UMLS + ReGeX - dictionaries
7	UMLS + dictionaries + ReGeX - Non-UMLS

Table 4.20: The table enumerates seven experiment tiers and describes what labelling sources the programmatic labelling module used.

tional experiments to compare the Snorkel LM experiments with another weak supervision methodology, FlyingSquid [113]. Previously, we had planned to compare Snorkel with skweak, another generative modelling methodology for combining LFs [200]. However, we switched to FlyingSquid for two reasons: firstly, the label matrix created for Snorkel interoperable with FlyingSquid, and secondly, both Snorkel and FlyingSquid label matrices use the IO (raw) labelling scheme, whereas skweak's label matrix includes additional span boundary information for entities, which makes the comparison less meaningful.

UMLS vocabularies were ranked based on the number of n-gram overlaps between the respective vocabulary and the EBM-PICO validation set. The ranked vocabularies were divided into $p = (1, 2, \ldots, 112)$ partitions, with partition one aggregating all the ontologies into a single LF and partition 112 with all the ontologies as separate LFs. The above-mentioned experiments were carried out with all the partitions to evaluate the performance of the number of UMLS LFs. We evaluate performance using token-level macro F1, precision and recall over three runs of experiment tiers with three random seeds.

Baseline: We used a CRF model trained on the hand-labelled validation set (n=721) as the fully supervised baseline. The CRF model was trained on token-level entity recognition task with with IO (Inside - Outside) labelling scheme.

4.5.4 Results

In both the labelled validation and test sets, the "Study type" class constitute the minority class with token strength between 2.24% - 2.34% of all the tokens (see Table 4.21).

Of 178 unique tokens representing the minority class in the validation set and 103 in the test set, 135 were unique. The top 14 out of 15 tokens are common between both sets (see Figures 4.16 and 4.17).

Token class	Validation set	Test set
1	4,650	1,257
0	207,220	53,667

Table 4.21: Simple statistics for the EBM-PICO validation and test set annotations for "Study type" (class = 1) entity and out of the span (class = 0) entity.

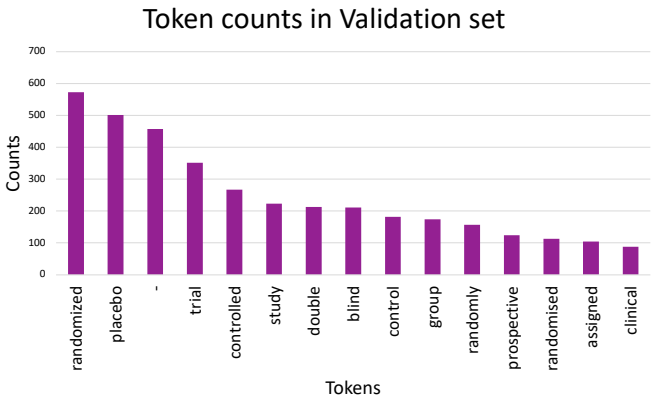


Figure 4.16: The top 15 most common tokens from the "Study type" class in the EBM-PICO test set.

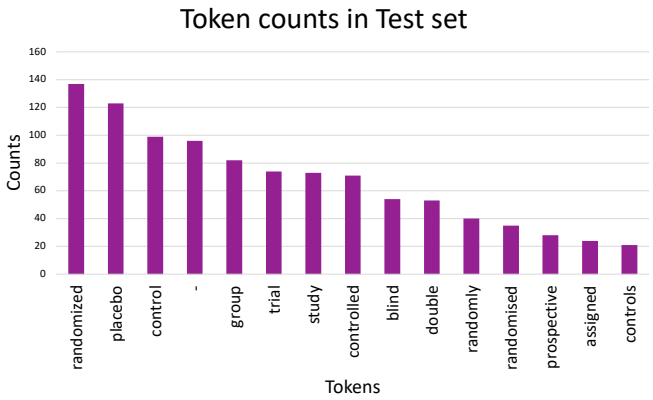


Figure 4.17: The top 15 most common tokens from the "Study type" class in the EBM-PICO validation set.

Using the described labelling sources and functions, we developed a total of 144 LFs (refer Table 4.22).

The results of the experiments are listed in Table 4.17. LF aggregation via MV fails to detect any meaningful signals and performs at a level close to or even worse than random. For tier 7, however, removing non-UMLS LFs boosts the recall and therefore the F1 for MV. The performance of UMLS alone for the LM tier 1 is poor. Incorporating non-UMLS sources into the model results in a significant drop in F1 score by as much as 8.4% again pointing towards the low representational value of this labelling source. Adding dictionaries increases the previous F1 by 13.52%, but it hits the performance glass ceiling at 63.16%. As expected, adding generic rules in tier 4 boosts the recall by 18.25% from tier 1. Up-weighting rule-based LFs in tier 5 leads to a nominal F1 increase by 0.55%. In the ablation tier 6, removing handcrafted dictionaries decreases the previous best recall by 6.87%, demonstrating performance contribution. In the ablation tier 7, removing the non-UMLS labelling sources improves the F1 by 3.06% improving both recall and precision. The best performing weakly supervised LM in Tier 7 outperforms the fully supervised CRF model by 16% on F1-score as shown in Table 4.24.

LF source	Number of LFs
UMLS	112
non-UMLS	10
dictionary	2
ReGeX	20

Table 4.22: Table enumerates the number of labelling functions for each of the labelling sources.

	Experiments		MV			LM		
Sr.	LF tier	Part.	P	R	F1	P	R	F1 (stdev)
Tier 1	UMLS	75-94	48.64	50.00	49.31	61.03	56.42	58.02 (4.4×10^{-5})
Tier 2	+ non UMLS	1-2	51.58	50.01	49.37	50.21	50.02	49.62 (2.2×10^{-4})
Tier 3	+ Dictionaries	2	48.64	49.99	49.31	64.87	62.23	63.16 (3.6×10^{-2})
Tier 4	+ Rules	2-4	48.64	50.00	49.31	86.03	78.50	81.41 (4.2×10^{-3})
Tier 5	+ Weighted rules	2-4	98.64	50.17	49.66	85.09	79.42	81.96 (7.4×10^{-3})
Tier 6	- Dictionaries	1	98.64	50.13	49.59	81.40	72.55	75.37 (1.5×10^{-3})
Tier 7	- non UMLS	1	96.22	53.31	55.56	89.96	81.41	**85.02** (1.7×10^{-2})

Table 4.23: Macro-averaged recall, precision and F1 % for "Study type and design" extraction models. The best F1 score is shown in bold. Standard deviation (stdev) is reported for average over three runs. Part. = Partition.

A lower number of partitions lead to good performance (also refer to Table 4.19), except for the UMLS tier 1, which required between 75-94 partitions to perform its best as shown in Figure 4.19. Snorkel's LM consistently and substantially outperformed FlyingSquid as shown in Figure 4.18 for tiers 4-7.

4.5.5 Discussion

Our results showed that even large ontology databases such as UMLS may be inadequate for representing an entity and may require expert-led rules to perform optimally. Both studies by Fries *et al.* and Dhrangadhariya *et al.* observed an improvement in F1 scores between entity classes when non-UMLS labelling sources were included, indicating that these sources have value in representing these classes. However, we found the opposite effect, as the F1 score for the "Study type and design" entity improved upon removing non-UMLS labelling functions, suggesting that these functions were not typical for "Study type and design" entity. It must be noted that while the ReGeX and heuristics developed for the "Study type and design" class cannot be directly applied to other entity classes, the method of developing such ReGeX using the hook patterns and the small labelled

Model	Type	P	R	F1	Δ F1
CRF	FS	82.44	63.66	69.02	
LM	WS	89.96	81.41	**85.02**	+16

Table 4.24: The table shows results of the fully supervised CRF model in comparison to the Tier 7 weakly supervised LM (the best performing model)

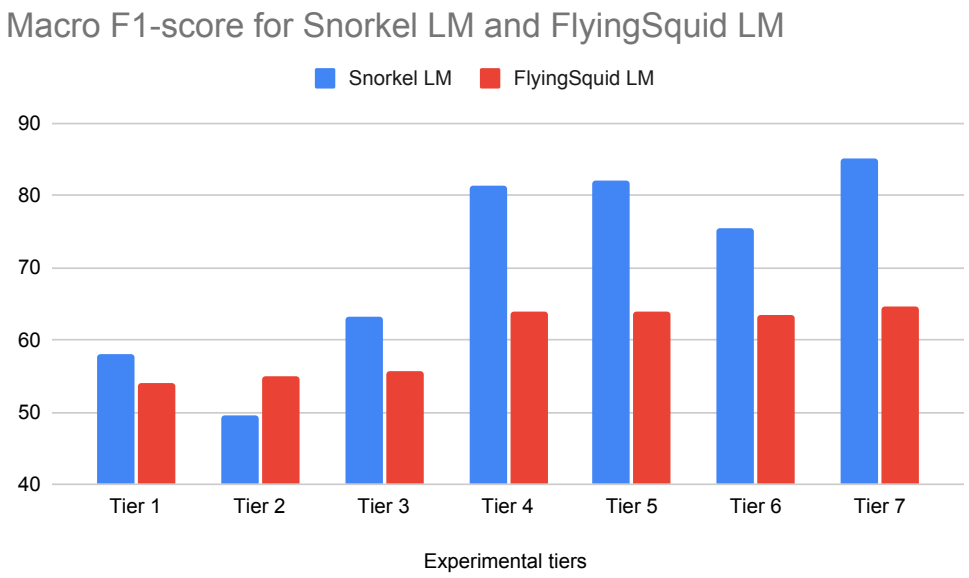


Figure 4.18: Graph comparing macro F1 scores for aggregating the designed LFs using Snorkel's LM vs FlyingSquid's LM for all experiment tiers.

validation set can be very well extended to other entity classes.

We conducted experiments across multiple tiers and determined the optimal number of UMLS partitions necessary to achieve the best macro F1 score. As depicted in Figure 4.19, for all experiment tiers except tier 1, the first and second partitions achieved high scores, after which the F1 score began to decline with consecutive increases in the number of partitions. However, for tier 1, which exclusively used UMLS labelling functions, increasing the number of partitions actually led to an increase in the F1 score. In particular, the higher-end partitions exhibited the best performance. We did not have any experiment tier without UMLS LFs, as doing so either resulted in disproportionate abstentions during prediction or worse than random performance.

Previous research has shown that full supervision typically performs better than weak supervision, but it is much slower and requires significantly more manual labor. In fact, researchers at Stanford found that by using Snorkel's weak supervision techniques, models could be built 2.8 times faster and with an average of 45.5% better predictive performance than with hand-labeled data alone [276]. Labelling clinical information is a challenging task due to the high level of disagreement between human annotators regarding the exact spans that constitute these entities. As a result, hand-labelled corpora may contain errors. For instance, recent studies have identified errors in the EBM-PICO dataset, correcting which requires costly reannotation [1, 41, 90, 188]. A data-centric approach like this can be used to analyze errors and programmatically re-annotate the dataset quickly.

Conflicts in manual annotation: Despite an acceptable IAA, there were specific conflicting scenarios where the agreement between the annotators fails, and these are highlighted here. Annotator 2 always missed marking the phrase "clinical trial" where its corresponding protocol in external repositories is mentioned. E.g., in the sentence "Clinical Trial registration: `https://clinicaltrials.gov/` Identifier: NCT01467843." the

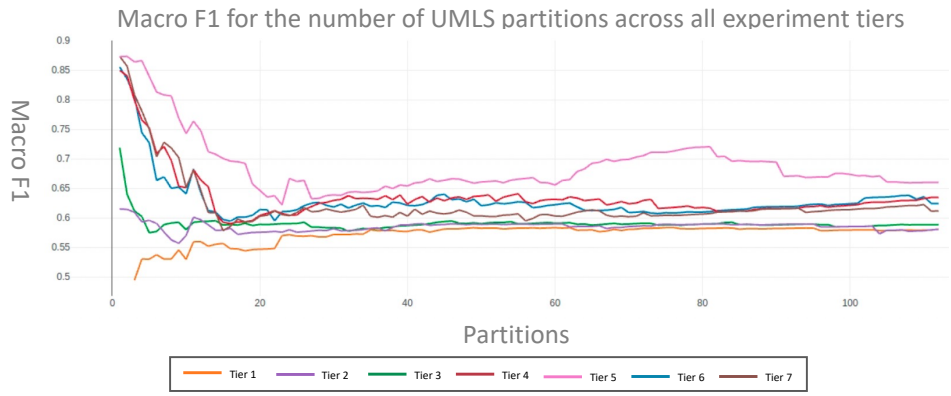


Figure 4.19: Macro-averaged F1 scores for UMLS partitions across the experiment tiers.

phrase "Clinical Trial" explicitly mentions the study registration as a clinical trial. So this should have been marked but was not marked by the annotator 2. This is attributed to our annotation guidelines' lack of explicit instruction. Our annotation guidelines require marking information about the number of arms used in a clinical study. Many conflicting tokens resulted when Annotator 1 completely missed marking information about the number of arms in the clinical trials. In contrast, annotator 2 marked details additional to the number of arms used in the clinical trial. For example, in the phrase "a four-arm randomized controlled trial", " four-arm" should be marked as an entity along with "randomized controlled trial". However, annotator 2 marked additional information not corresponding to the number of arms in a trial, e.g., marking the phrase "40K arm" from the sentence, "eighty-three patients were assigned to the 40K arm", but 40k here refers to the treatment "epoetin alfa 40,000 U". Annotator 2 did not mark whether a study was a "superiority" or a "non-inferiority" study which is study type feature, but the annotator 1 marked it. These was not clarified in the annotation guidelines, but will be improved.

Error Analysis We conducted an error analysis to investigate the impact of high-cost expert-generated rules on tiers 3 and 4, and the results are presented in Table 4.25. Token-level errors were categorized into four types: 1) false negative (FN), which occurred when the model failed to predict the entire entity that the token belonged to; 2) false positive (FP), which occurred when the model mistakenly recognized the token as part of an entity; 3) boundary error (BE), which occurred when the model missed the boundary tokens but otherwise correctly identified the entity; and 4) overlapping error (OE), which occurred when the model falsely captured out-of-span tokens between adjacent entities. Tier 3

Errors	Tier 4	Tier 3
FP	30 (5.24%)	920 (57.57%)
FN	232 (40.55%)	502 (31.41%)
BE	209 (36.53%)	170 (10.63%)
OE	59 (10.31%)	6 (0.73%)
Total	530	1598

Table 4.25: Distribution of the token-level errors for the Tier 3 and Tier 4 LMs on EBM-PICO test set.

had plenty of FP and FN errors completely missing an entity, while tier 4 had abundant BEs, indicating its ability to capture more signals. For instance, in the entity "randomly assigned", the generic term "randomly" was captured, but "assigned" was missed by tier 4, while tier 3 completely overlooked it. It indicates the ability of the rule-based LFs to capture the characteristic patterns not detected by the vocabularies. Rule-based LFs reduces FPs by diminishing the probability of irrelevant vocabulary concepts being signals.

One of the common mistakes across both tiers was several FPs for the token "control", which did not occur in a context denoting that a particular study is a controlled trial. For e.g., the term "control" was identified from "BP control" and "electrical control activity". Tier 3, based on vocabularies alone, captures lots of irrelevant FP tokens like "man", "document", "baseline", "treatment", *etc.*

Interestingly, vocabulary-based LFs capture more information irrelevant to the "Study type and design" entity. For instance, some terms related to "allocation", "block randomization", "protocol", and "blinding" suggest their usage to the other clinical entities beyond study type.

OE are usually the commas connecting multiple components of study design falsely identified as entities. For instance, commas are detected as entities in the term "Multicentre, double-blind, double-dummy, placebo-controlled trial.'

Another limitation of the label model is that it relies on dictionaries, and the ReGeX only learns about label agreements and disagreements, ignoring contextual information in the sentence. Consequently, entities may be marked incorrectly if they appear in an inappropriate context within the document. For instance, the phrase "randomized controlled trials" from the sentence "Larger randomized controlled trials are needed to evaluate the results of the case-control study further" could be labeled as a "Study type and design" entity in the future tense, even though the context of the present study does not mention study types.

Interestingly, tier 3 performed better than tier 4 in identifying generic terms such as "placebo," "controls," and "control," despite being less powerful overall. Tier 3 also correctly identified signals from clinical trial phases like "Phase III," "randomized phase I clinical trial," and "clinical trial phase I," but was unable to identify the entire entity. However, the poor performance of tier 3 also suggests that the vocabularies do not adequately represent the "Study type and design" entity, missing common patterns such as "randomization" and "randomly." Additionally, both tiers failed to identify entities such as "controlled trials," "prospective," "open-labelled," "multicentre," "prospectively," "superiority," and "retrospectively." Further analysis is needed to compare all the tiers, along with measuring the seen and unseen entities in each tier, to understand the contribution of each approach.

4.5.6 Conclusion and Future Work

We adopt a weak supervision approach to enhance existing datasets, such as EBM-PICO, by incorporating additional categories, like "Study type and design" without relying on manual annotation. This is achieved through the application of weak supervision techniques using Snorkel. Our approach achieved exceptional performance, with an F1 score of 85.05% on the hand-labelled EBM-PICO test set, highlighting the potential of this method for rapidly generating large amounts of annotated data compared to traditional supervised approaches. We also provide a straightforward algorithm for mapping UMLS terms to entity classes, even in cases where expert guidance is unavailable, and semantic groupings

are unclear. Our code, weakly-labelled EBM-PICO training set, doubly-annotated EBM-PICO test set, and hand-crafted dictionaries are openly accessible.

In the future, we aim to extend the methodology to cover more entities and improve the mapping methodology of UMLS semantic groups to clinical classes. Currently, we employ hard-coded rules decided based on observed performance changes over recall and F1 scores, and such rules require experimentation on other entities. Our study is limited to exploring the "Study type and design" entity class, which has less heterogeneity and is more standardized than the less standardized and fuzzy classes like PICO. However, by limiting the scope to a single, well-defined clinical entity class, our study can provide a thorough and detailed analysis of the relevant steps and variables critical to weak supervision's success in more complicated clinical entity classes. The corpus we labeled, EBM-PICO, is expected to contain only RCTs leaving out many other trial types; hence our weakly-labelled corpus and labelling functions might not represent all the trial types. We suggest extending the approach to other wide variety of clinical studies.

4.6 Chapter Conclusions

The contributions of this chapter are the methodologies in information extraction, specifically weak-supervision and distant-supervision that enable clinical information extraction in absence of hand-labelled corpora. These methods primarily focus on enhancing the extraction of PICOS-related entities. In the section 4.2, a multi-task learning (MTL) approach was investigated, where coarse-grained PICO extraction served as the primary task, while fine-grained PICO extraction functioned as an auxiliary task. The method improved performance on fine-grained "Participant" and "Outcome" extraction. Moving on to Section 4.3, the DISTANT-CTO methodology was developed. This methodology leveraged distant supervision principles for successful "Intervention" information extraction, surpassing state-of-the-art (SOTA) results. Then, in the section 4.4, a weak-supervision framework as developed for PICO information extraction using freely-available resources. The methodology improved extraction of the "Participant" entity. Finally, in the section 4.5, the previously introduced weak supervision framework was extended to extract "Study type and design" information showing extensibility of the methodology.

Chapter 5

Risk of Bias Corpus Development

This chapter addresses the research gap of the lack of a publicly available corpus to train and evaluate the RoB assessment automation. The section 5.3 describes a pilot annotation project that aimed to test whether existing RoB assessment guidelines could be used as RoB corpus annotation guidelines. The pilot project introduces a small RoB span annotated corpus of 10 RCTs. The section 5.4 describes in detail the annotation guidelines developed for RoB corpus annotation and their adaptation as visual placards. The section also describes the RoBuster corpus annotated using these guidelines. Parts from this chapter are published as a conference paper and as a journal paper (preprint) [83, 84]. In [83] and [84], my contribution was to conceive the project, guide a team of five annotation experts to adapt RoB corpus annotation guidelines, conceptualize and aid the development of visual placards, set up the annotation platforms, analyze the results and report the findings in the form of the papers. The following resources are made available via this research:

1. The dataset developed in section 5.3 is available on Zenodo.
    - `https://zenodo.org/records/7924466`
2. The dataset developed in the section 5.4 is available on GitHub.
    - `https://github.com/anjani-dhrangadhariya/RoBuster/tree/main/RoBuster`
3. The dataset parsers and the python notebooks exploring the corpus characteristics for corpora in the section 5.3 and 5.4 are found on GitHub.
    - `https://github.com/anjani-dhrangadhariya/rob-preliminary-annotation`
    - `https://github.com/anjani-dhrangadhariya/RoBuster/`
4. The visual annotation placards developed in section 5.4 are available on GitHub.
    - `https://github.com/anjani-dhrangadhariya/rob-annotation-placards`

5.1 Introduction

SRs synthesized from RCTs are the highest quality of evidence in the hierarchy of evidence (see figure 1.3). SRs are used for health-care policymaking, effective drug formulation, and primary care physicians and health professionals could use them to make treatment decisions [6, 187, 196]. An RCT is a scientific experiment in which a group of patients is

randomly divided into two or more groups and allocated to either an intervention under investigation or a control intervention group or other another comparator intervention to compare the effect of the interventions being studied [305]. In theory, an RCT accurately measures the intervention effect on patient outcomes. However, it can be biased in practice due to expected and unexpected flaws in the study design, execution, analysis or outcome reporting. Biases in clinical trials lead to systematically overestimating or underestimating the intervention effect. Therefore, when the RCTs with questionable biases are used to write SRs, it can diminish their validity and reliability. Such SRs can lead to faulty clinical care guidelines, ultimately harming the patients [136]. Therefore, researchers conducting SRs must rigorously look for possible biases in the RCTs before using them for writing SRs.

The biases in RCTs cannot be measured, but an RCT can be assessed for biases to minimize the overall risk and judge the RCT quality. The revised Cochrane risk of bias tool for randomized trials, also called RoB 2 [59] provides RoB assessment guidelines and has been extensively used for bias evaluation in RCTs [19, 30, 184, 262, 313]. Published RCTs are exponentially increasing [60] over time, so manual RoB assessment for every study becomes a protracted process. RoB assessment is a part of writing systematic reviews, which is highly resource-heavy, taking in most cases about six months to several years to complete [173, 331].

ML approaches can help accelerate the RoB assessment process by directly pointing the reviewers to the parts of the RCT text relevant to identifying bias, leading to quickly judging the trial quality. Supervised ML models require RoB span annotated data, but unfortunately, to date, there's a lack of publicly available manually labelled RoB corpora or any established guidelines aiding in corpus annotation. RoB assessment is a knowledge-heavy task in which even highly trained experts are prone to subjective judgments. The primary requirement to develop such a corpus entails creating a well-thought-out annotation scheme and clear annotation guidelines. As neither the corpus annotation guidelines nor the annotation scheme exists for risk of bias, this work is focused on the following primary concerns. The main goals for the works in this chapter were:

1. To test whether RoB 2 assessment guidelines could be used as RoB corpus annotation guidelines. If so, these guidelines could be used to develop a corpus that could be utilized for training supervised ML models.

2. To develop and test a RoB annotation scheme that closely mimics the RoB 2 guidelines [313].

3. To develop concrete RoB corpus annotation guidelines using RoB 2 and adapt them into visual annotation placards.

4. To develop a corpus of RCT full-texts manually annotated with RoB text spans.

5. To ensure these resources are publicly available for the community to build upon and improve.

[59] https://sites.google.com/site/riskofbiastool/welcome/rob-2-0-tool

[60] https://pubmed.ncbi.nlm.nih.gov/?term=randomized\%20controlled\%20trial&filter=pubt.randomizedcontrolledtrial

5.2 Related Work

Marshall *et al.* [214] attempted automation of RoB assessment using a distant supervision approach supported by proprietary data from the Cochrane Database of Systematic Reviews (CDSR). They classified the trial quality assessment as binary into *low-risk* and *unclear-risk/high-risk* quality attributes for each risk domain. The study was supported by the manually-entered data from CDSR, which is behind a paywall and automates based on Cochrane's RoB 1.0 guidelines and not the latest RoB 2 [147]. Even though Cochrane's RoB tool (version 1) is the most frequently used to assess RCT quality, a recently revised Cochrane RoB 2 offers significant benefits in comparison [206]. Compared to the original RoB version released in 2008, the RoB 2 version provides a more reliable and concrete structure to the RoB evaluation by developing comprehensive guidelines that enforce consistency [147, 313]. A study analyzing Cochrane systematic reviews and protocols found that the use of RoB 2 increased from 0% in 2019 to 24.1% in 2022 [217]. This indicates the importance of using an updated and standardized tool to assess bias in RCTs.

Millard *et al.* attempted automating RoB assessment using supervised machine learning trained on proprietary data as well [232]. The research utilizing this pay-walled data was used to develop RobotReviewer that several studies have evaluated for its human-competent performance [151, 164, 215, 310, 340]. The question, however, remains of the unavailability of a publicly available RoB annotated corpus that hinders community efforts for automation and evaluation. Wang *et al.* recently released three RoB annotated datasets but for preclinical studies with RoB assessments about animals [348]. A manually annotated corpus of RoB spans for human clinical trials is still necessary but is unavailable. The next section, 5.3, describes the initial steps taken to work towards developing a RoB span annotated corpus.

5.3 First Steps Towards Developing a Risk of Bias Corpus

Manual RoB assessment is a complex, expert-led task with subjective judgements. Systematically translating this process for developing a RoB annotated corpus requires a carefully designed annotation scheme and detailed annotation guidelines. To our knowledge, there exist no detailed corpus annotation guidelines. To address this gap and build upon existing resources, a pilot annotation study was conducted to assess the applicability of the currently available RoB 2 guidelines in the manual annotation of a RoB corpus. The RoB 2 guidelines were directly applied to annotate a small set (n = 10) of RCTs by multiple experts to assess their utility as corpus annotation guidelines. The results were analyzed to confirm the suitability of these guidelines as corpus annotation instructions.

5.3.1 Methods

The methodology section provides an overview of our annotation scheme, explaining its design rationale, the annotation software used, and the common annotation guidelines adhered to in addition to the RoB 2 assessment guidelines in our annotation process. The section also describes our expert annotation team for this work.

5.3.1.1 Formulating Annotation Scheme

The RoB annotation scheme is formulated as a function of the revised Cochrane RoB tool for randomized trials (RoB 2). Understanding the tool structure is essential to under-

standing the proposed RoB annotation scheme [313]. Version 2.0 divides biases into five risk domains, each corresponding to different parts of the trial design. Each risk domain decomposes into several signalling questions, each aiming to prompt a relevant response to bias assessment (refer to Table 5.1). The response options are restricted to "Yes", "Probably yes", "No", "Probably no", or "No information". A visual representation of our annotation scheme is illustrated in Figure 5.1.

Bias class	Bias domain	Signalling questions
RoB 1	biases arising from the randomization process	3
RoB 2	biases due to deviations from intended interventions	7
RoB 3	bias due to missing outcome data	4
RoB 4	bias in the measurement of the outcome	5
RoB 5	bias in the selection of the reported result	3

Table 5.1: The table lists down the bias domains as structured in the revised Cochrane RoB assessment tool (RoB 2) and the number of signalling questions in each domain.

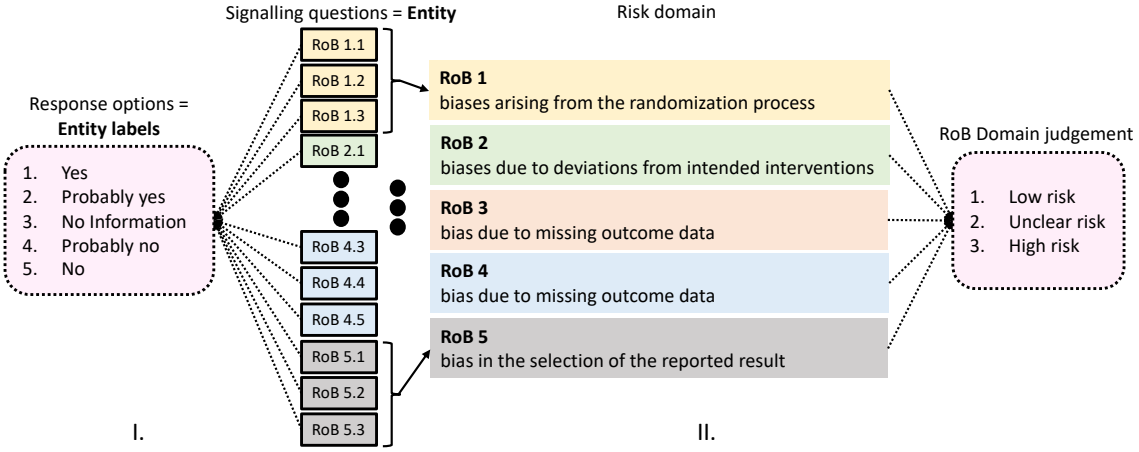


Figure 5.1: Annotation scheme. I. SQ level: each SQ (RoB 1.1, 1.2, ...) is an entity that could take either of five response options (entity labels). SQ response judgements for individual risk domains (RoB 1-5) could be combined to arrive at risk domain judgement. Note: Risk domain judgments are not addressed in this work.

To respond to the signalling questions, the reviewers must go through individual RCTs and inspect the evidence required to respond with one of the five previously mentioned response options. For instance, to respond to the signalling question "Was the allocation sequence random?", the reviewers read through the clinical trial study to identify the methodology used to randomize the allocation of participants into the intervention groups. If a clinical study described a proper allocation sequence randomization, the reviewer responds to this question as "Yes" and otherwise "No". Similarly, each signalling question prompts the reviewer to look for a piece(s) of factual evidence in the clinical study to respond with one of the five response options. An annotation scheme where each signalling question is an entity was formulated. The factual evidence in the RCT helps decide the response to that question. Each entity could have one of the five response options incorporating the reviewer's judgment of the answer. The reviewer needs to mark the identified text evidence (a phrase, sentence (s), or paragraph) with the RoB entity along

with one of the five response options. In this regard, this makes it a hierarchical annotation scheme comprising 22 entities corresponding to the 22 signalling questions, each with five response options.

The cumulative risk judgment from each domain could be estimated using decision flowcharts combining the responses from all the signalling questions. The flowchart allows this risk judgment to be classified as either "Low-risk", "High-risk" or "Some-concerns" (refer the Figure 5.2). To accommodate this, an additional document-level annotation scheme for each risk domain (as outlined in Table 5.1) allows reviewers to select one of the three risk judgments.

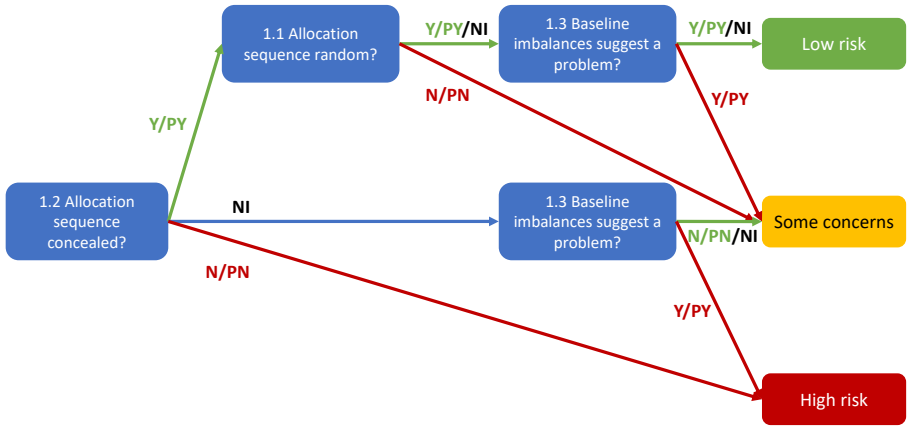


Figure 5.2: Algorithm for suggested judgement of risk of bias arising from the randomization process. The figure is recreated from the revised Cochrane's risk of bias tool (RoB 2) [313].

5.3.1.2 Preliminary Annotation Guidelines

The four reviewers used RoB 2 guidelines to annotate RCT full-texts. [61] The reviewers include two doctoral researchers and two postdoctoral researchers who have previously performed RoB ratings in several systematic reviews. Some generic annotation guidelines were developed by a natural language processing (NLP) expert in conjunction with four experienced physiotherapists writing systematic reviews.

These are the generic guidelines; If an entire sentence is relevant to answering a signalling question, then annotate this whole sentence, including the full stop at the end. If a sentence phrase is relevant to answering this particular RoB signalling question, then annotate only that phrase. When there is no information relevant to answering a signalling question with any of the response options, do not annotate. Be sure to annotate all parts of the information you used to respond to the questions, even if the information is found in disparate parts of the RCT full text. Every signalling question must be answered invariant to the flowchart structure in the RoB 2 assessment manual. If the caption of a particular table or figure leads to answering a question, annotate the caption. If the caption does not provide text relevant to the answer, then annotate the table's contents. If the reference to the table or figure leads to answering the signalling question, annotate it. If all three are relevant to answer the question, annotate all.

[61] `https://drive.google.com/file/d/19R9savfPdCHC8XLz2iiMvL_71lPJERWK/view`

5.3.1.3 Pilot Annotation

Our aim with the pilot annotation was to assess if RoB 2 assessment guidelines could be used as annotation guidelines to obtain an RoB annotated dataset. We also tested the suitability of the previously detailed annotation scheme 5.3.1.1. One of the authors, the most experienced in RoB assessment, developed an R script to build the corpus. This entailed an Entrez search [62] using the search query "(randomized[title] or randomized[title]) and (rehabilitation or (physical therapy))" that searched for the term "randomized" in the study title. The search query was restricted to retrieving first 1000 hits and for one-year time spans. Ten such searches were made for 10 time spans each (2000 - 2001, 2002 - 2003, 2004 - 2005, 2006 - 2007, 2008 - 2009, 2010 - 2011, 2012 - 2013, 2014 - 2015, 2016 - 2017, 2018 - 2019). The code then used a function to randomly choose ten studies from the retrieved 1000 for that particular year. Out of the ten sampled studies, the author in question took the first possible study with a freely available PDF.

Four experienced annotators carried out pilot annotation following the RoB 2 tool, generic annotation guidelines and the developed annotation scheme. Tagtog [63], a commercial text annotation web application, allows for annotating PDF (Portable Document Format) documents, was used for pilot annotation [52]. We chose to annotate PDFs rather than plain text because RCT PDFs have a visual format (maintains the structure of sections and subsections, tables, and figures) that makes the annotation task quicker for the annotators and increases annotation quality. Tagtog allows customized annotation schemes at entity and document levels and has functionality for parsing PDF documents to plain HTML for annotation extraction, allowing for easy quality control for the annotations. Tagtog has an internal IAA (inter-annotator agreement) scoring scheme and a visual display to report the agreement. This setup streamlines the iterative annotation projects. Each annotator was given access to the Tagtog project with ten corpus RCTs after a brief training session with Tagtog. The task for the annotator was to annotate text relevant to answering each signalling question entity and choose a signalling question response option, each risk domain judgment for the five risk domains.

5.3.1.4 Annotation Evaluation

Cohen's kappa is the standard annotation reliability measure for many classification annotation tasks, but it is not a relevant measure for token-level annotation tasks. We report the pairwise F1 measure that disregards out-of-the-span tokens (unannotated tokens), which is the ideal measure of annotation reliability for the token-level annotation tasks [40, 79].

Our first aim was to determine if RoB 2 assessment guidelines could be reliably used as RoB corpus annotation guidelines. To this end, we measure how consistently the annotators identified chunks of text in the RCTs to answer each signalling question and report inter-annotator agreement IAA_{sq}. IAA_{sq} measures the pairwise agreement between annotators for identifying the same chunk of text to answer each signalling question. We calculate $IAA_{response}$ to determine how reliable the RoB 2 guidelines were for the annotators to make a "response" judgment for each signalling question after identifying relevant chunks of text. Document-level agreements (Cohen's Kappa) for risk domain

[62] The Entrez Global Query Cross-Database Search System is a federated search engine, or web portal that allows users to search PubMed database

[63] `https://www.tagtog.net/`

judgment IAA_{rd}. We interpret IAA values (F1 scores as well as Cohen's Kappa) as shown in the Table 5.2 [182].

Agreement interpretation	IAA range
Poor	0-0.99
Slight	1 - 20.99
Fair	21 - 40.99
Good	41 - 60.99
Substantial	61 - 80.99
Almost perfect	81 - 99.99
Perfect	100

Table 5.2: The table details interpretation of pairwise F1-measure and Cohen's Kappa.

5.3.2 Results

Our pilot annotation corpus comprised ten RCTs between the years 2000-2019. The document PDF lengths varied between 4-20 pages, with the smallest RCT having 2836 tokens and the largest having 20,290 tokens. Annotators also annotated figures and tables because answering specific signalling questions requires looking into these modalities. Four annotators annotated these ten RCTs, resulting in 902 RoB entity labels corresponding to the signalling question categories and 220 risk domain class labels at the document level. It took each annotator minimum of 20 minutes to annotate a single RCT with all the bias classes. An example of tagtog annotation is shown in Figure 5.3. All the annotations were stored as JSON files and parsed accordingly for analysis. The distribution of labels for each signalling question entity label is shown in Figure 5.4.

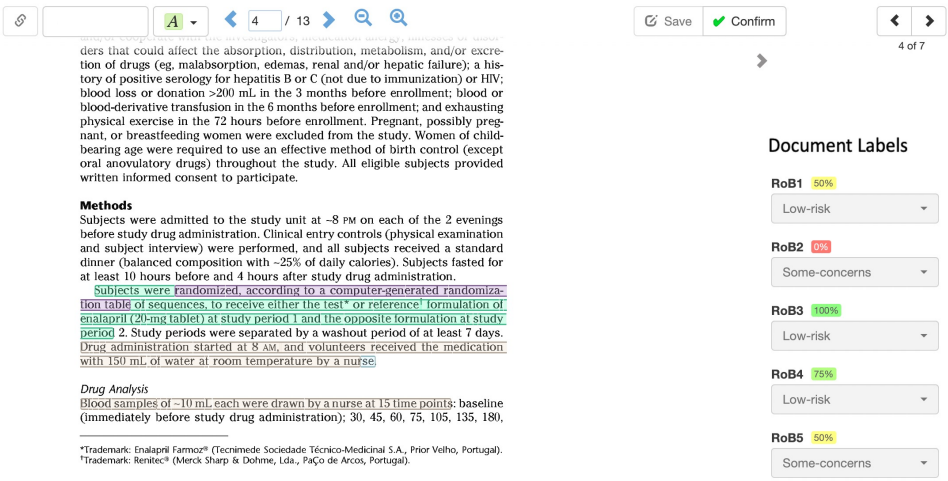


Figure 5.3: A screenshot of tagtog interface with text evidence annotations for the RoB signalling questions (in the left) and risk of bias judgment labels (in the right) in display.

Table 5.3 reports pairwise IAA_{sq} between the six annotator pairs (Left). Individual pairwise agreements range between 0% (poor) and 75% (substantial), with most agreement

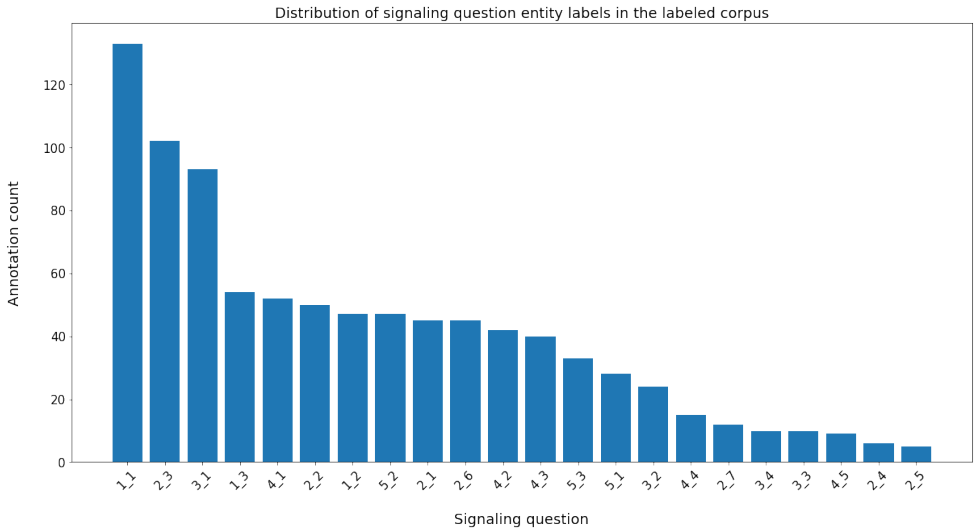


Figure 5.4: Distribution of signalling question entity labels in the labeled RCT corpus.

values falling under the poor category and very few under the substantial agreement (refer Figure 5.5). Signalling questions RoB 1.1, 1.2, 1.3, 2.6, and 3.1 fared well in terms of the average pairwise agreement between all pairs, but none of these categories had a substantial agreement. Signalling questions 2.3, 2.4, 2.5, 2.7, 3.4, 4.4, 4.5, and entire domain 5 fared extremely poorly or with no agreement or annotation. Table 5.3 (Right) reports the $IAA_{response}$ averaged over all the annotator pairs at the signalling question response option level. The $IAA_{response}$ scores are considerably lower (to zero) than agreement at the signalling question level (IAA_{sq}), hinting that annotators assign different response options for the text relevant to answering a signalling question.

Table 5.4 reports document-level IAA_{rd} for each risk domain. The agreement scores range between poor (0%) and substantial ($\geq$ 80%). Risk domains 2 and 5 are the most challenging at the signalling question, response option and risk domain judgment level.

Table 5.5 details the $IAA_{response}$ agreement specifically at the signaling question response level for risk domain 1, which focuses on bias arising from the randomization process. The agreement values for signalling question 1.2 are the lowest and mostly zero. The highest agreement values in domain 1 signalling questions are between good and perfect. Annotator pair P4 has a surprising 100% (perfect) agreement on the judgment "No information" for RoB 1.3.

Tables 5.6 and 5.7 report the $IAA_{response}$ agreement at the signalling question response level for the risk domains 2 (biases due to deviations from intended interventions). Except for the signalling question 2.6, agreement at other signalling questions fare poorly. Answering the signalling question 2.6 ("Was an appropriate analysis used to estimate the effect of assignment to intervention?") is straightforward and prompts the reviewers to look for the appropriate analysis method(s) in the RCT literature. In contrast, answering RoB 2.1 and 2.2 require the annotators to read between the lines about whether the participants, carers and intervention administrators were blinded to the administered intervention. Except for vague terms like "single-blind", "double-blind", and "open-label," there is no single method or process that could help determine answers to these signalling questions. For questions 2.4, 2.5 and 2.7, either the agreement is zero, or there are no annotations for the response.

SQ	P1	P2	P3	P4	P5	P6	Avg.	Y	PY	NI	PN	N
1.1	23.1	24.5	52.2	57.0	48.0	21.5	37.7	21.8	7.1	0.0	-	-
1.2	**66.1**	50.3	72.8	50.7	46.0	50.5	56.1	4.9	11.5	10.2	0.0	-
1.3	**69.5**	20.5	16.1	31.6	59.9	53.5	41.8	-	-	41.8	11.4	9.9
2.1	1.0	1.4	0.0	9.1	19.1	0.0	5.1	8.2	0.0	-	3.0	0.0
2.2	18.3	7.3	11.1	0.0	23.0	7.4	11.2	3.6	0.0	0.0	0.0	0.0
2.3	20.6	5.5	13.4	0.0	0.0	0.0	6.6	-	0.0	-	1.0	0.0
2.4	0	-	-	0	0	-	0	-	0	-	0	-
2.5	0	0	0	0	0	-	0	0	0	-	0	-
2.6	**75.3**	**68.9**	19.3	**63.9**	12.9	19.6	43.3	39.4	0.0	0.0	0.0	3.6
2.7	0.0	6.6	0.0	0.0	0.0	0.0	1.1	0.0	0.0	-	0.0	0.0
3.1	45.8	23.6	32.2	43.4	22.9	14.8	30.4	47.6	0.6	-	1.3	3.3
3.2	1.4	0.0	0.0	3.3	7.4	0.9	2.2	0.0	0.0	-	0.0	0.0
3.3	0.0	0.0	0.0	16.4	0.0	0.0	2.8	-	0.0	31.4	0.0	0.0
3.4	-	0	-	0	0	0	0	0	0	0	0	0
4.1	4.0	6.6	14.2	25.6	22.3	6.3	13.2	-	-	-	0.8	12.0
4.2	1.8	0.0	0.4	0.0	40.1	0.0	7.1	-	-	-	0.3	0.0
4.3	7.6	13.9	5.0	10.5	39.5	8.4	14.2	0.0	0.0	0.0	13.1	20.5
4.4	0	0	0	0	0	0	0	0	0	-	0	0
4.5	0	0	0	0	0	0	0	0	0	-	0	-
5.1	0.0	0.0	0.0	0.0	0.0	4.2	0.7	0.0	0.0	0.0	0.0	0.0
5.2	23.9	0.0	0.0	0.0	0.0	2.4	4.4	-	0.0	0.0	0.0	0.0
5.3	0.2	0.0	0.0	0.4	8.1	42.0	8.4	-	0.0	0.6	0.0	0.0

Table 5.3: Left: Table lists down IAA_{sq} between the six annotator pairs (P1-P6) for the RoB SQs. Substantial ($\geq$ 61) agreements are in bold. Right: Table lists down IAA_{sq} averaged over the six annotator pairs for the SQs at the entity label level. Note: Y = Yes, PY = Probably Yes, NI = No Information, N = No and PN = Probably No, Avg. = Average. "-" shows that one of the annotators did not annotate any text for a particular SQ.

Table 5.8, 5.9, and 5.10 report the $IAA_{response}$ agreement at the signalling question response level for the risk domains 3 (bias due to missing outcome data), domain 4 (bias in the measurement of the outcome), and domain 5 (bias in the selection of the reported result) respectively. For question responses for these domains, the agreement either remains zero, or there are no annotations. Signalling question 3.1 has fair to almost perfect agreements between the pairs. The reason for this is "almost clear" assessment guidelines for this signalling question. The question asks whether the outcomes data is available for all, or nearly all, participants randomized. If the outcomes data is available for nearly all randomized patients, the annotator chooses the response option "Yes" and otherwise "No". Annotator pair P4 again has a surprising 94.11% (almost perfect) agreement on the judgment "No information" for RoB 3.3, and it is the only non-zero agreement for this signalling question. Generally, the agreement values at the response level vary across the annotator pairs. Prospectively, the reason for no annotation was the lack of clear corpus annotation guidelines.

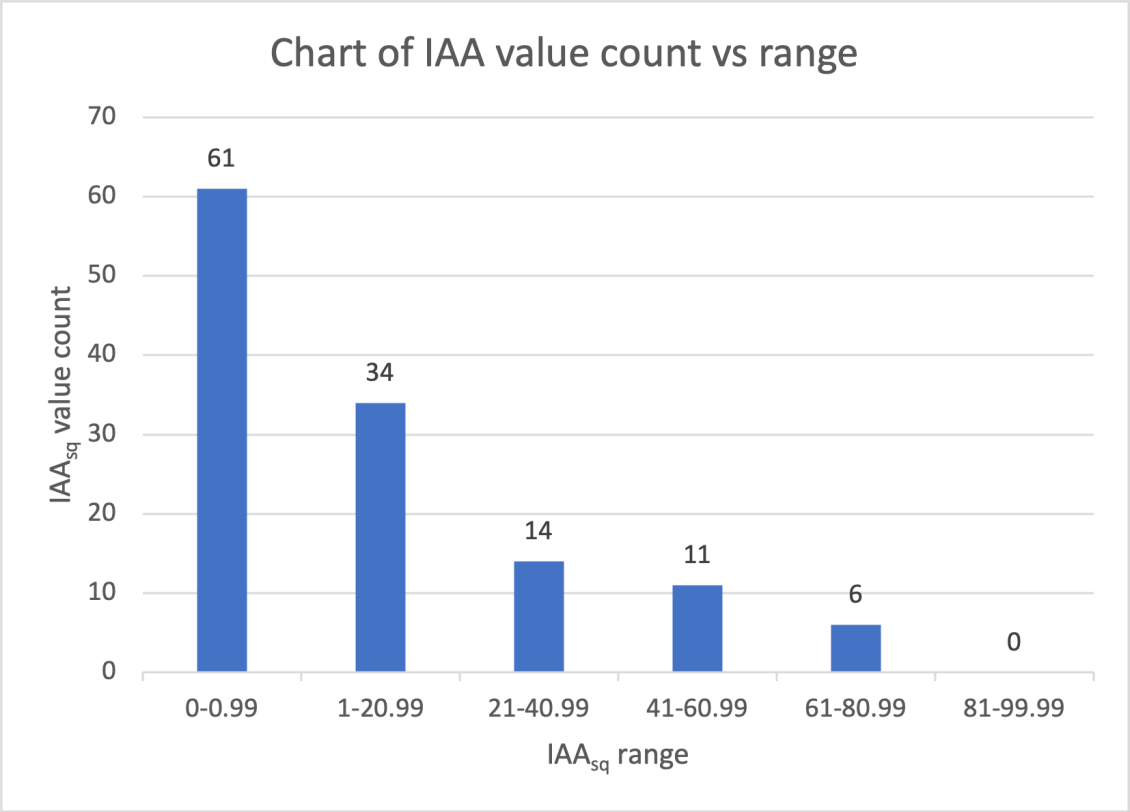


Figure 5.5: Distribution of IAA_{sq} agreement values in the labeled corpus.

5.3.3 Discussion

In this section discusses the four types of annotation disagreements identified.

Polarity disagreement: A polarity disagreement arises when two annotators choose the same chunk of text to answer a SQ but choose polar opposite entity labels ("Yes" or "Probably yes" vs "No" or "Probably no" vs "No information"). In one of the documents, all four annotators chose the same text evidence ("71 allocated routine services, 67 allocated intervention service, ...") to answer the SQ 3.1. However, three of the four annotators responded to this question with "Yes", but one chose "Probably no". This SQ asks whether the outcomes data were available for all, or nearly all, participants randomized but does not clarify the exact cut-off for how many participant dropouts increase the risk? Therefore, the annotators make subjective response judgments depending upon what exact percentage of participant dropout is considered valid in their experience. In another example example, for corpus document number 5, annotator pair P1 chose the text evidence, "Statistical analyses were performed using STATA version 10.0 (Statcorp, College Station,TX). The Shapiro-Wilk W test for normal data was performed on continuous outcome measures. The distribution of categorical variables in each group was compared..." to respond to the SQ 2.6. Even though both annotators in the pair selected the same evidence for answering 2.6, one chose the "Probably no" response option, and another chose "No information". For the SQ 2.6 ("Was an appropriate analysis used to estimate the effect of assignment to intervention?"), RoB 2 guidelines do not instruct when

Risk domain (class)	P1	P2	P3	P4	P5	P6	Total
RoB 1	57.14	57.14	57.14	57.14	85.71	71.42	64.28
RoB 2	14.28	0	0	57.14	14.28	42.85	21.42
RoB 3	50	33.33	42.85	83.33	57.14	71.42	56.34
RoB 4	66.66	14.28	42.85	28.57	71.42	28.57	42.058
RoB 5	50	0	28.57	0	42.85	0	20.23

Table 5.4: The table lists down IAA_{rd} between the pair of annotators at the risk domain level.

RoB 1.1					
-	Y	PY	NI	PN	N
P1	16.09	4.1	0	-	-
P2	30	16.37	-	-	-
P3	11.39	1.71	-	-	-
P4	14.58	7.18	0	-	-
P5	0	0	0	-	-
P6	**58.42**	13.2	0	-	-

RoB 1.2					
-	Y	PY	NI	PN	N
P1	0	18.29	**60.67**	0	-
P2	0	0	0	0	-
P3	0	0	0	0	-
P4	0	0	0	-	-
P5	0	0	0	-	-
P6	29.18	50.6	0	-	-

RoB 1.3					
-	Y	PY	NI	PN	N
P1	-	-	75.36	55.42	0
P2	-	-	75.36	2.83	0
P3	-	-	0	2.15	0
P4	-	-	**100**	5.16	32.67
P5	-	-	0	2.35	8.07
P6	-	-	0	0	18.28

Table 5.5: The table lists down IAA between the pair of annotators for the risk of bias signalling questions at the "response" option level $IAA_{response}$ for the risk domain 1 (bias arising from the randomization process). Highest agreement values for each signalling question are marked on bold. The lowest agreement values were always zero.

to choose "Probably no" vs "No information" when the details about the analysis used to estimate the effect of assignment to intervention are unclear.

Degree disagreement: A degree disagreement causes low $IAA_{response}$ and arises because some annotators are lenient in judging risk while others are sceptical. The lenient ones select definitive "Yes" or "No" for responding to a SQ, while the sceptical ones choose "Probably yes" or "Probably no". For example, in the corpus document 1, pair P1 selected

RoB 2.1					
Pair	Y	PY	NI	PN	N
P1	0	0	-	0	0
P2	0	0	-	0	0
P3	0	0	-	-	0
P4	**49.05**	0	-	11.76	0
P5	0	0	-	0	0
P6	0	0	-	-	0

RoB 2.2					
Pair	Y	PY	NI	PN	N
P1	0	0	0	0	0
P2	**12.98**	0	0	-	0
P3	8.16	0	0	-	-
P4	0	0	-	0	0
P5	0	0	-	-	-
P6	0	0	-	-	-

RoB 2.3					
Pair	Y	PY	NI	PN	N
P1	-	0	-	0	-
P2	-	0	-	**5.76**	-
P3	-	-	-	0	0
P4	-	0	-	0	-
P5	-	0	-	0	0
P6	-	0	-	0	0

RoB 2.4					
Pair	Y	PY	NI	PN	N
P1	-	0	-	0	-
P2	-	-	-	-	-
P3	-	-	-	-	-
P4	-	0	-	0	-
P5	-	0	-	0	-
P6	-	0	-	0	-

Table 5.6: The table lists down IAA between the pair of annotators for the risk of bias signalling questions at the "response" level $IAA_{response}$ for the risk domain 2 (biases due to deviations from intended interventions (Part I)). Highest agreement values for each signalling question are marked on bold. The lowest agreement values were always zero.

RoB 2.5					
Pair	Y	PY	NI	PN	N
P1	-	0	-	0	-
P2	0	-	-	-	-
P3	-	-	-	-	-
P4	0	0	-	0	-
P5	-	0	-	0	-
P6	0	-	-	-	-

RoB 2.6					
Pair	Y	PY	NI	PN	N
P1	33.54	0	0	0	-
P2	**63.41**	-	0	-	0
P3	34.1	0	0	-	-
P4	40	0	-	0	10.69
P5	24.32	0	0	0	0
P6	40.51	0	-	0	-

RoB 2.7					
Pair	Y	PY	NI	PN	N
P1	0	0	-	0	-
P2	0	0	-	0	0
P3	0	-	-	-	-
P4	-	-	-	0	0
P5	-	-	-	0	-
P6	-	-	-	0	0

Table 5.7: The table lists down IAA between the pair of annotators for the risk of bias signalling questions at the "response" level $IAA_{response}$ for the risk domain 2 (biases due to deviations from intended interventions (Part II)). Highest agreement values for each signalling question are marked on bold. The lowest agreement values were always zero.

RoB 3.1					
Pair	Y	PY	NI	PN	N
P1	**98.06**	3.32	-	0	0
P2	53.76	0	-	0	0
P3	19.35	0	-	0	19.83
P4	57.59	0	-	7.84	0
P5	23.15	0	-	0	0
P6	33.23	0	-	0	0

RoB 3.2					
Pair	Y	PY	NI	PN	N
P1	0	0	-	0	-
P2	-	0	-	0	0
P3	-	0	-	-	-
P4	0	0	-	0	0
P5	0	0	-	0	-
P6	-	0	-	0	0

RoB 3.3					
Pair	Y	PY	NI	PN	N
P1	-	0	0	0	-
P2	-	-	0	0	-
P3	-	-	-	-	0
P4	-	0	**94.11**	0	-
P5	-	0	-	-	0
P6	-	-	-	0	0

RoB 3.4					
Pair	Y	PY	NI	PN	N
P1	-	0	0	0	0
P2	-	-	-	0	-
P3	-	-	-	0	-
P4	0	0	-	0	0
P5	-	0	-	-	0
P6	-	-	-	-	-

Table 5.8: The table lists down IAA between the pair of annotators for the risk of bias signalling questions at the "response" level $IAA_{response}$ for the risk domain 3 (bias in the measurement of the outcome). Highest agreement values for each signalling question are marked on bold. The lowest agreement values were always zero.

RoB 4.1					
Pair	Y	PY	NI	PN	N
P1	-	-	-	0.95	12.33
P2	-	-	-	2.96	16.28
P3	-	-	-	-	6.24
P4	-	-	-	0	8.15
P5	-	-	-	0	13.07
P6	-	-	-	0	15.81

RoB 4.2					
Pair	Y	PY	NI	PN	N
P1	-	-	-	1.83	0
P2	-	-	-	0	0
P3	-	-	-	0	0
P4	-	-	-	0	0
P5	-	-	-	0	0
P6	-	-	-	0	0

RoB 4.3					
Pair	Y	PY	NI	PN	N
P1	0	0	0	0	0
P2	0	0	-	0	40.93
P3	0	0	0	65.11	0
P4	-	0	0	0	5.55
P5	0	0	0	0	76.27
P6	0	0	0	-	0

RoB 4.4					
Pair	Y	PY	NI	PN	N
P1	-	0	-	0	-
P2	-	-	-	0	0
P3	0	-	-	0	-
P4	-	0	-	0	0
P5	0	0	-	0	-
P6	0	-	-	0	-

RoB 4.5					
Pair	Y	PY	NI	PN	N
P1	-	0	-	0	-
P2	-	0	-	0	-
P3	0	0	-	0	-
P4	-	-	-	-	
P5	0	0	-	-	-
P6	0	0	-	-	-

Table 5.9: The table lists down IAA between the pair of annotators for the risk of bias signalling questions at the "response" level $IAA_{response}$ for the risk domain 4 (bias in the selection of the reported results). Highest agreement values for each signalling question are marked on bold. The lowest agreement values were always zero.

RoB 5.1					
Pair	Y	PY	NI	PN	N
P1	0	0	0	0	0
P2	0	0	0	0	0
P3	-	0	0	-	0
P4	0	0	-	-	-
P5	0	0	-	-	0
P6	-	0	-	-	0

RoB 5.2					
Pair	Y	PY	NI	PN	N
P1	-	0	0	0	0
P2	-	0	0	0	-
P3	-	0	0	0	0
P4	-	-	-	0	0
P5	-	-	-	0	0
P6	-	-	-	0	0

RoB 5.3					
Pair	Y	PY	NI	PN	N
P1	-	0	0	0	0
P2	-	0	0	0	0
P3	-	0	0	0	0
P4	-	-	-	0	0
P5	-	-	2.27	0	0
P6	-	-	-	-	0

Table 5.10: The table lists down IAA between the pair of annotators for the risk of bias signalling questions at the "response" option level $IAA_{response}$ for the risk domain 5 (bias in the selection of the reported result). Highest agreement values for each signalling question are marked on bold. The lowest agreement values were always zero.

the same sentence "Patients were randomly allocated to either intervention by a computer-generated schedule stratified by sex and attendance at a day hospital" to respond to SQ RoB 1.1. However, the more stringent annotator of the pair chose to respond with "Probably yes" and the lenient one with "Yes". In corpus document 7, annotator pair P5 selected the same information "Retention = 95.7%" and "Retention = 91.7%" to answer the signalling question 3.1. However, one annotator responded with the lenient "Probably no" and another with a definitive "No". A practical and rationally justified solution is to merge the response options "Probably yes" with "Yes" and "Probably no" with "No" to reduce the complexity of the task and increase IAA without altering the final risk judgment for this risk domain. [313] As shown in Figure 5.2 responding to any signalling question for the risk domain 2 as either "Probably yes" or "Yes" does not alter the final risk judgment for this domain (low, high, or some concerns).

Text span disagreement: A low IAA is also caused by our annotation guidelines not limiting the annotators to selecting either the phrase vs a sentence(s) vs a paragraph for answering the question leading to a text span disagreement. RoB 2 tool led to some annotators using and annotating very condensed information to come to a response. In

contrast, others used an entire paragraph to reach the same response for a SQ leading to a low token-level IAA. In corpus document 5 for pair P6, one of the annotators selected parts of a sentence "The primary outcome measure was a 0–10 NRS pain score, which reflected the average pain experienced by the patient for ten days prior to follow-up" as relevant text to answering signalling question 4.1 as "No", while another annotator selected only the phrase "a 0–10 NRS pain score". In the corpus document 7, the annotator pair P5 selected the same information "Retention = 95.7%" and "Retention = 91.7%" to answer the signalling question 3.1, but one of the annotators used additional text information ("Retention = 85.7%" and "Retention = 85.7%"Retention = 83.3%) as relevant to answering the RoB 3.1. In corpus document 4, for pair P2, one of the annotators chose an entire paragraph ("Randomisation was performed centrally by computer at the Birmingham Cancer Clinical Trials Unit, University of Birmingham. When a patient was identified as eligible for the study, and had given written, informed consent to take part, the research nurse telephoned the trials unit...") as text evidence to answer RoB 1.1 with "Yes". In contrast, the other annotator chose condensed, specific information from the same paragraph to make the same decision. This problem requires mending the annotation guidelines to precisely instruct authors to select the complete information they used to decide or the minimum necessary information to decide on a SQ. Another method is automatically extending the more condensed annotations to the broadest ones. In the guideline improvement outlined in the next section, the restriction is on marking the full sentence(s) where the relevant information is found unless otherwise instructed to mark phrases.

Disparate document section disagreement: Sometimes annotators came to a response judgment for a SQ but used different parts of the RCT text leading to disparate document section disagreement. For example, document 7, pair P5, answered RoB 2.6 as "Yes" but used different parts of RCT as evidence. One of the annotators chose this sentence "This study was guided by the HAPA, which has been widely used to address the gap between intention to change and a person's actual change in behaviour [25-27]." to reach "Yes". The other chose "intention-to-treat analysis was done with missing data substituted by the last-observation-carried-forward procedure". In corpus document 8, pair P1 chose the same response option, "Yes" for the RoB 1.1 but chose different parts of the text to answer it. One of the annotators chose the text evidence "A PHASE III INTERNATIONAL RANDOMIZED CLINICAL TRIAL" in the study's title to answer this question. In contrast, the other annotator chose the text evidence from the methods section. Such disagreements emanate from a lack of corpus annotation guidelines as well. For each signalling question, the guidelines could instruct what part of the RCT to annotate and what part to not annotate for a particular signalling question. For example, the text evidence to answer the signalling questions 1.1 and 1.2 can be found in the methods section. Applying this instruction could have retrospectively nullified this disagreement. However, the RoB 2 guidelines do not provide such instruction, and the annotator will annotate either all the places where they find text evidence to answer a question or only one of the parts of the text where they see the evidence. We noticed many SQs remained unanswered because the annotators did not understand what part of the text to annotate, even after following the RoB 2 guidelines.

There was a specific case whereby the annotators selected multiple parts of text relevant to answering the signalling question and responded to the question with different response options. For document 1, one annotator selected the text evidence, "Patients

were randomly allocated to either intervention by a computer-generated schedule stratified by sex and attendance at a day hospital" to respond to RoB 1.1 with "Yes". The same annotator selected another text evidence ("Single-blind randomized controlled trial.") from the same document to respond to RoB 1.1 with "Probably yes". For document 8, one annotator selected the text evidence, "Children were randomized at a 1 to 1 ratio and used an adaptive blocked randomization algorithm" to respond to RoB 1.1 with "Yes". The same annotator selected another text evidence ("SCATE was a Phase III, randomized, open-label, partially masked, multi-centre, international, prospective trial of hydroxyurea versus observation for children with SCA and centrally confirmed conditional TCD velocities") from the same document to respond to RoB 1.1 with "Probably yes". It has to be noticed that this issue is limited to RoB 1.1. Technically, a signalling question for a document can only have one response judgment selected, but this was not reinforced in our annotation setup. We wanted to capture as much relevant text as possible to answer a signalling question with a judgment, even if it meant selecting multiple response options for the signalling question. Though this is a beneficial signal for machine learning applications, it does not serve the purpose of good IAA.

5.3.4 Limitations and Future Work

The corpus developed in this pilot study was very small, but it traded off for i) the document length for annotation (full-text), ii) the complexity and subjective nature of the RoB annotation task, and iii) the number of the entity and document classes (22 + 5) in the RoB annotation task. However, despite the smaller scale, this work conclusively demonstrates that the RoB 2 guidelines are not suitable for serving as annotation guidelines for the RoB corpus. As detailed in the discussion section, the absence of clear annotation guidelines resulted in a low IAA across the small manually labeled corpus. However, this work does not present any annotation guidelines which are an overarching necessity for developing an annotated corpus. Clear annotation guidelines could significantly enhance annotation and have the potential to elevate the quality of a RoB annotated corpus by fostering consistent decision-making and subsequently increasing the IAA. The next section 5.4 explains how the concrete annotation instructions were developed by adapting the RoB 2 guidelines. It also details adapting these annotation instructions into visual annotation placards that could be used to annotate a large corpus.

5.4 Development of RoB Annotation Instructions and RoBuster

To develop a corpus annotated with RoB text spans, RoB 2 guidelines were adapted into comprehensive text annotation guidelines in this work [313]. These annotation guidelines were modelled in form of visual placards for ease of annotation and understanding. In addition to RoB annotation, these visual placards could also be utilized to train graduate RoB student assessors. Using the annotation guidelines in addition to the RoB 2 tool, we annotated and released a larger corpus of 41 full-text RCTs with 22 risk of bias span types which could be used to fine-tune machine learning models or LLMs. The corpus could also be used as a validation benchmark. We evaluated the performance of LLMs to automatically identify the answers to these signalling questions using prompt generation.

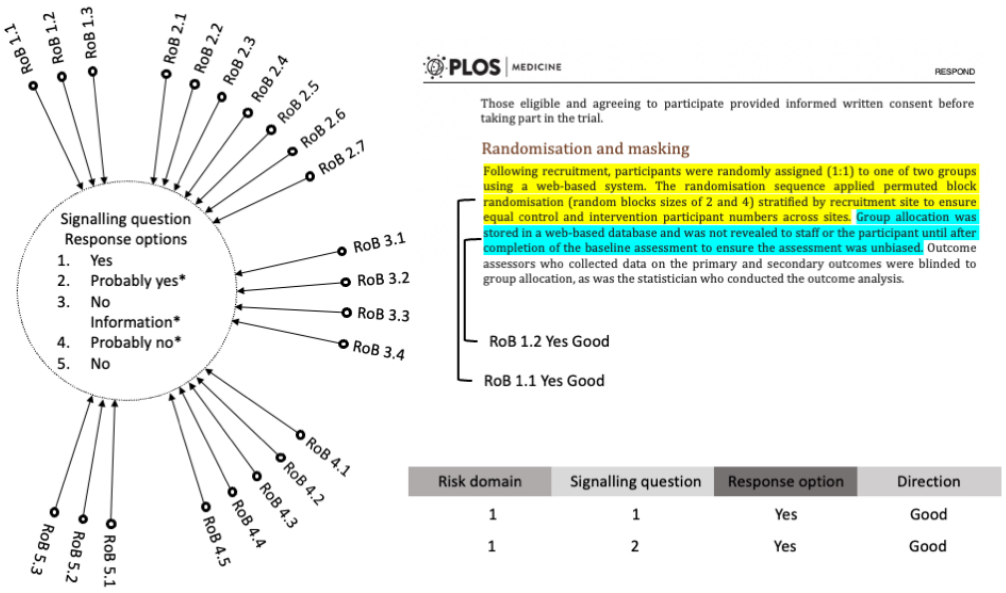


Figure 5.6: Our adapted annotation scheme. Note: No Information* = No Information label will not be manually annotated. It will be automatically considered for any SQ where the annotator did not mark any answer. Probably yes* and Probably no* = Will be collapsed with yes and no options, respectively (except for certain SQs).

5.4.1 Methods

This section provides an overview of the annotation scheme, the software tools used for annotation, and the development of visual annotation instructions. Since there were no existing annotation guidelines for the RoB span annotation task, we took the initiative to create them from the ground up by adapting the RoB 2 tool. Our team first crafted a preliminary set of visual annotation guidelines. Experts then proceeded to annotate a subset of documents with these guidelines, and any conflicts that came up during this process were used as valuable feedback to improve and refine the guidelines followed by annotating a larger subset.

5.4.1.1 Adapted Annotation Scheme

Instead of redeveloping the annotation scheme, this work adapted and enhanced the previously developed scheme in the section 5.3.1.1 as per the learning from the pilot project. This annotation scheme was directly adapted from the RoB 2 assessment procedure and hence I reiterate the how RoB 2 guidelines are structured to understand the annotation scheme. RoB 2 divides biases into five risk domains, each loosely corresponding to different parts of the trial design. Each risk domain decomposes into several SQs, each aiming to prompt the assessor to look for relevant RCT text evidence and elicit a relevant response for bias risk judgment for that SQ (refer to Table 5.1).

The response options are restricted to “Yes”, “Probably yes”, “No”, “Probably no”, or “No information” [313]. Reviewers assess these signalling questions by examining the factual evidence in the RCT. For instance, to answer the signalling question “Was the alloca-

tion sequence random?", the reviewer reads through the study to identify how participants were randomized into intervention groups. If a well-executed method of randomization is identified, the reviewer answers with "yes" (the allocation sequence is random) judging the risk of bias for this signaling question as low risk. Conversely, if a poorly executed method of randomization is found, the risk of bias is deemed high risk with response option "no".

In RoB span annotation, we mimic this assessment process by considering evidence text spans in the RCT as the main units of annotation. Each span corresponds to answering a signalling question and is annotated with the most informative label. The label incorporates information about the signalling question number and the domain it assesses (for the above example, "1.1" for the first domain and first signalling question of the domain) Additionally, the response judgement is incorporated in the label, such as "1.1 Yes Good" for a well-executed randomization (see Figure 5.6). We took the learning from our previous work *et al.* and collapsed the response options "yes" and "probably yes" into a single "yes", and "no" and "probably no" together into a single "no" to increase the inter-annotator agreement (IAA) without altering the final risk domain judgment [83]. As shown in Figure 5.2 responding to any signalling question for the risk domain 2 as either "Probably yes" or "Yes" does not alter the final risk judgment for this domain (low, high, or some concerns). Therefore, except for some special case signalling questions, these response options were collapsed as suggested. This makes it a hierarchical span annotation scheme comprising 22 entities corresponding to the 22 SQs, each with typically two response options ("Yes" or "No") and two directions ("good" and "bad"). We also remove the "No Information" response option because this was meant for the situations where actually no text evidence is found in the RCT to answer and label for a SQ. However, for selected SQs (currently only SQ 2.1), "Probably Yes", "Probably No" and "No Information" may still be acceptable. For instance, consider that an RCT uses "...random number generator and sealed envelopes for patient randomization...", but the trial provided no information on whether the envelop was "opaque" or not. In such situations, "No Information" judgment is acceptable.

5.4.1.2 Expert Team

As mentioned earlier, RoB annotation is a complex task that requires specialized expertise. It is cognitively demanding due to the need to carefully go through the entire full-text of RCTs and identify 22 different bias categories for annotation. This level of complexity would not be manageable for annotators without expertise in the field. Our annotation team consisted of two researchers specializing in RoB assessment in physiotherapy and rehabilitation domains, including an epidemiology researcher (ID: E1MA [64]) and an associate professor (ID: E2IRAA [65]) in physiotherapy. With a substantial background in both physiotherapy, advanced statistical methods and experience writing SRs, both experts possessed a deep understanding of the complexities involved in bias assessment. Two additional physiotherapy experts, two senior PhD students, were a part of developing the visual annotation guidelines and placards. Two additional researchers with expertise in natural language processing (NLP) were involved, a computational linguistics associate professor and a PhD student in computer science. Their inclusion was important because the guidelines and placards they helped create will be utilized to annotate a text corpus, serving as a benchmark for RoB text span extraction. Finally, Prof. Dr. Julian PT

[64]Expert 1 Major Annotator
[65]Expert 2 Inter-Rater Agreement Annotator

Higgins, who is the main editor of RoB 2, provided critical feedback to shape the visual annotation placards [149].

5.4.1.3 Data Collection

Different outcome categories exist in SRs: subjective, objective and mortality outcomes (a sub-category of Objective). Savović *et al.* found that trials assessing subjective outcomes are more prone to bias, therefore, had this work used only one outcome type, we would have limited label types for different risk classes [257, 293]. In context of RCTs, subjective outcomes are measurements that rely on individuals' perceptions, opinions, or feelings about their own health or well-being. These outcomes are typically self-reported by the participants in the trial and can be influenced by factors such as placebo effects, patient expectations, interpretation, and psychological factors. For example, in a study on rheumatoid arthritis, subjective outcome measures included patient-reported pain ratings [341]. Objective outcomes are measurements that are independent of individual opinions or perceptions and are based on observable and measurable data. These outcomes are typically collected by trained assessors or through laboratory tests, imaging studies, or other objective methods. For instance, in a study on peripheral artery disease, objective outcome measures included angiography and molecular imaging to evaluate the effectiveness of cell therapy [129]. Mortality outcomes refer to the occurrence of death during the course of the trial. To ensure that these various outcome types are represented in the corpus, we included 17, 17, and 7 RCTs addressing objective, subjective, and mortality primary outcomes, respectively. The epidemiology researcher (E1MA) from our team created this 41 RCTs dataset from the domain of physiotherapy and rehabilitation. PDFs of the full-text RCTs were extracted and each article was collated with its trial protocol from wherever available. Each PDF was renamed with the primary outcome that was to be examined using RoB 2 before uploading to the annotation software [66]. Corpus details are in the Appendix.

5.4.1.4 Visual Placards Development

RoB 2 tool consist of an extensive and step-by-step set of instructions to answer signaling questions and even though RoB 2 guidelines are widely used for bias assessment, there have been some research on their reliability. This reliability concern has been extensively investigated by Minozzi *et al.* [235, 236]. They formulated specific instructions on how to approach and answer the signaling questions of RoB 2. These instructions, referred to as the Instruction Document (ID), address the subjectivity present in the RoB 2 guidelines and provide clear guidance for the assessment process. Subjectivity in assessment could potentially result in different evaluators coming to disparate conclusions when analyzing the same trial. Before implementing the ID, the agreement among four expert RoB assessors was zero, but it improved after adopting the ID. Several other papers explored subjectivity and reliability of the Cochrane RoB 1.0 and 2 tools [72, 204, 235, 236]. With this in mind, precise and clear text annotation instructions using the RoB 2 tool were developed with an aim to maintain the consistency and reliability among annotators. Working closely with our team of experts, we formatted these instructions into visual instructional placards. Each placard takes the form of a flowchart and provides instructions for annotating RCT text to answer a SQ. The flowchart also provides instructions on labelling the annotated

[66] tagtog text annotation software, PAWLS annotation software

text with risk judgment. The RoB 2 tool SQs are broadly factual but leave room for subjective judgements and our visual placards aim to facilitate judgements about the risk of bias.

5.4.1.5 Annotation

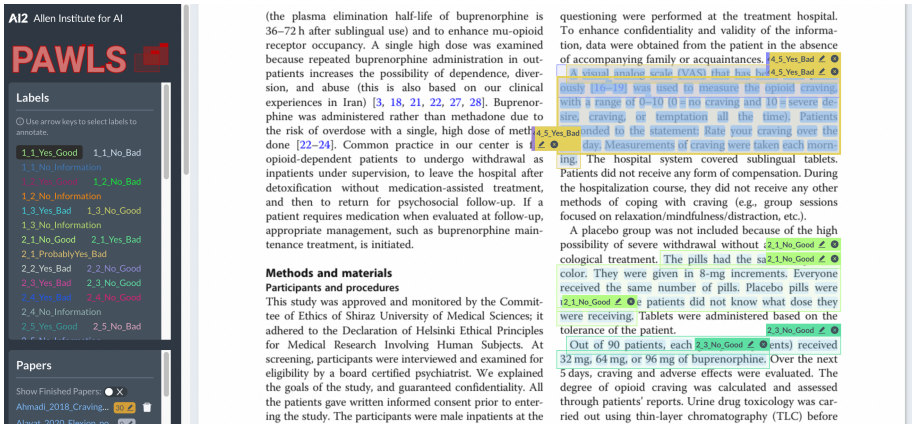


Figure 5.7: A screenshot of PAWLS interface with an example PDF and RoB annotations.

For every SQ, the annotators were guided to use the complete RoB 2 guidance document along with visual placards that were developed. They followed these instructions meticulously, going through each placard's signaling question one by one. The provided instructions directed the annotators to read specific sections of full-text RCT that needed annotation. Their task involved identifying and highlighting relevant text related to answering the signaling question. It has to be noted that the domain 2 of RoB 2 focuses on assessing the risk of bias due to deviation from the intended intervention. This domain evaluates both the effect of assignment to the intervention and the effect of adherence to the intervention. RoB 2 offers distinct sets of SQs for each aspect. The study specifically focuses on assessing Domain 2 for the effect of assignment to the intervention and consequently addresses only address the SQs corresponding to this aspect.

Tagtog [67], a commercial text annotation web application, allows for annotating PDF (Portable Document Format) documents, was used for the annotation [52]. Out of the 41 documents, 9 were doubly annotated by two experienced annotators (E1MA and E2IRAA) to calculate inter-annotator agreement (IAA) over these documents and the rest were singly annotated by E1MA. After double annotation, conflict resolution was performed to address conflicting annotations, which helped us further calibrate the visual placards. The conflict resolution was followed by annotating 51 additional RCTs.

After the annotation of 9 doubly-annotated RCTs, we switched to the PAWLS [68] annotation tool, which allows users to annotate PDFs for free [249]. We chose to annotate PDFs rather than plain text because RCT PDFs have a visual format that will be lost upon converting to text. For example the structure pertaining to sections and subsections, tables, and figures makes the annotation task quicker for the annotators and increases annotation quality. Post annotation, the feedback was taken from both the annotators, details of which could be found in the Appendix.

[67] `https://www.tagtog.net/`
[68] `https://pawls.apps.allenai.org/`

5.4.1.6 Evaluation

We report IAA at two levels checking whether the annotators agree on the text spans to answer SQs using the pairwise F1 measure. F1-measure disregards out-of-the-span tokens (unannotated tokens) during agreement calculation and is an ideal measure of annotation reliability for the token-level annotation tasks. It measures the F1 score as shown below for each pair of annotators, treating one annotator's labels as the "true" labels and the other annotator's labels as the "predicted" labels [40, 79].

$$F1 - measure = \frac{2 \times \text{True Positives}}{2 \times \text{True Positives} + \text{False Positives} + \text{False Negatives}}$$

We also check how strongly the annotators agree on the risk judgment for each SQ using prevalence and bias adjusted kappa (PABAK) κ_{pabak} and compare it with raw percent agreement. PABAK κ_{pabak} is the standard annotation reliability measure for many classification annotation tasks and is suitable to measure reliability at the risk judgment level. κ_{pabak} is an extension of Cohen's Kappa κ that takes into account prevalence and bias in the agreement. We interpret both the IAA measures as shown in the Table 5.11 [45, 68, 182, 223].

F1-measure		κ_{pabak}		Raw Agreement	
interpretation	range	interpretation	range	interpretation	range
Poor	0-0.99	No agreement	≤ 0	None	0
Slight	1 - 20.99	Slight agreement	0-0.20	Very low	1-10%
Fair	21 - 40.99	Minimal	0.21-0.39	Low	11-30%
Good	41 - 60.99	Weak	0.40-0.59	Moderate	31-50%
Substantial	61 - 80.99	Moderate	0.60-0.79	High	51-70%
Almost perfect	81 - 99.99	Strong	0.80-0.90	Very high	71-90%
Perfect	100	Almost Perfect	≥ 0.90	Perfect	>90%
		Perfect	1.0		

Table 5.11: The table details interpretation of pairwise F1-measure (Left), κ_{pabak} (Middle) and observed or raw agreement (Left)

5.4.1.7 LLM Evaluation

Our annotation guidelines and annotations were adapted for benchmarking supervised machine learning approaches and not LLMs. So even though we were annotating PDFs, we had to restrict a lot of annotations based on the assumption that PDF will be converted into text via OCR (optical character recognition) losing its structure of tables and figures, which anyway a classical ML model could not use without extensive modifications [73, 193, 194]. Recent advancements with LLMs offers a better alternative and made us rethink the evaluation. The bar for clinical applications is high and it is imperative to evaluate LLMs for the more challenging clinical tasks like RoB text span extraction [308]. The tools like ChatPDF [69] allow direct interaction between LLMs and PDFs, negating the clumsy

[69] `https://www.chatpdf.com/`

PDF to text conversion. Therefore, it is essential to evaluate LLMs instead of forcefully adapting the evaluation to a classical ML problem. LLM evaluation was formulated as a zero-shot RoB text span extraction task. This was to gauge whether an LLM encodes knowledge related to assessing trial biases. We used simple prompt constructs of the structure "Answer the $\{SQ\}$ + Action item to extract sentence supporting the answer". Consider the following example.

> *Example prompt: Question 4.3 Were outcome assessors aware of the intervention received by study participants? Provide an answer and extract the supporting sentences that you write your answer based on. Extract the sentences in JSON* [70].

The prompt serves two purposes for evaluating LLMs for correctness. LLM is prompted to 1) answering the SQ with a response option (risk judgment), and 2) extract the text evidence to support their answer. When answering a question, ChatPDF finds the most relevant paragraphs from the PDF and uses the ChatGPT API from OpenAI to generate an answer [71]. LLM is required to do the same task as human annotators and will be evaluated on the basis of correctness of the answer. If LLM answer corresponds to response option selected by the expert annotator, it is considered a correct answer for that SQ. If the text extracted by LLMs as evidence for answering the SQ fuzzy matches the text selected by the expert annotator, it is considered a correct answer. Both of these skills will be evaluated using a raw or observed agreement metrics $P_O(Extraction)$ for measuring agreement over extraction and $P_O(Response)$ for measuring agreement over response judgments and interpreted as per Table 5.11. Observed agreement is essentially the number of documents for a RoB SQ where LLM responses align with those of the human expert, divided by the total number of documents assessed [14]. In several instances, there was no information found by the expert annotator from the RCT to answer a question. For such instances, if ChatPDF correctly identifies the absence of relevant information to answer a question, it is considered a correct response. It's important to highlight that the evaluation framework is designed solely to measure the agreement of the answers between ChatPDF and the expert. LLM evaluation was manually conducted by a bias assessment expert and an NLP expert for 10 out of the 41 RCTs.

5.4.2 Results

This section outlines the visual placards, annotated dataset, reports the IAA findings, and the outcomes of the LLM evaluation.

5.4.2.1 Visual Placards

A total of 27 placards were developed to address the 22 signalling questions in RoB 2 tool. Details of the annotation guidelines and visual placards are available in the Supplementary material. Figure 5.8 presents an example placard for annotating SQ 3.1 ("Were data for this outcome available for all, or nearly all participants randomized?") which assesses the completeness of outcome data in an RCT. This question assesses whether data for the specific outcome of interest were collected and available for analysis for a high proportion of initially randomized participants. Missing data can compromise statistical power and

[70] JSON = JavaScript Object Notation

[71] ChatPDF employs GPT (Generative Pretrained Transformer) 3.5.

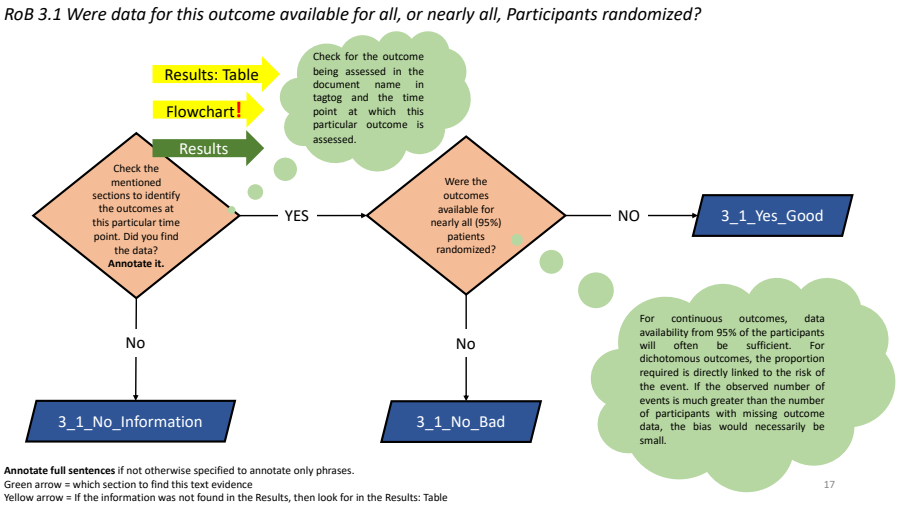


Figure 5.8: Sample annotation instruction placard for the SQ 3.1 designed and adapted using RoB 2 tool.

treatment effect estimates. The first diamond on the placard instructs annotators to check the Results section (first priority) and the flowchart and Table within the Results section (second priority) to identify outcome data at the specified time point. If outcomes data were available for at least 95% of participants, annotators mark relevant text descriptions as "3.1 Yes Good" indicating a low bias. If data were available for less than 95%, they mark it as "3.1 No Bad" indicating a high bias. Lack of information will lead to automatically assuming a "No Information" label. The yellow arrow on the first diamond suggests checking the Results section first; if not found, annotators should look in the Results section table. If still not found, annotators are instructed to check the flowchart caption, marked with a red exclamation as a last resort.

5.4.2.2 The Corpus: RoBuster

We provide key statistical information about the annotations in RoBuster in this section. The histogram in Figure 5.9 shows a visual representation of the absolute counts of annotations (tokens) for each of the RoB SQs. SQ 1.3 had disproportionately higher number of annotated tokens, while for all other SQs, the number of annotated tokens remained consistently below 2000 across the entire corpus. The only exception to this trend was for SQ 3.1 which had slightly more than 2000 annotated tokens. SQ 2.4 ("Were these deviations likely to have affected the outcome?") had only 25 annotated tokens.

Table 5.12 lists down essential information on the absolute and average annotation lengths for each RoB SQ along with the total number of documents the annotations were identified from. Annotations for the "randomization" risk domain 1 were found in an average of 35 of the 41 documents, while annotations related to answering the other risk questions were available only in a small subset of the total annotated RCTs, as also depicted in Figure 5.10 which shows the distribution of risk judgments across RoBuster. The figure highlights that, for most SQs, no information was available (indicated by yellow bars) for answering the SQs and making the risk judgment. Notably, SQs 2.2, 2.6, 3.1, 4.3, and 4.4 stood out as exceptions, with information available in more than 50% of the annotated documents. In cases where information was available, bias tended to be low, as

indicated by the prevalence of green bars with an exception for the SQs 2.2, 3.1, 4.3, and 4.4, where bias was high, as indicated by the prevalence of red bars. Check the Appendix for the references of all the studies in RoBuster.

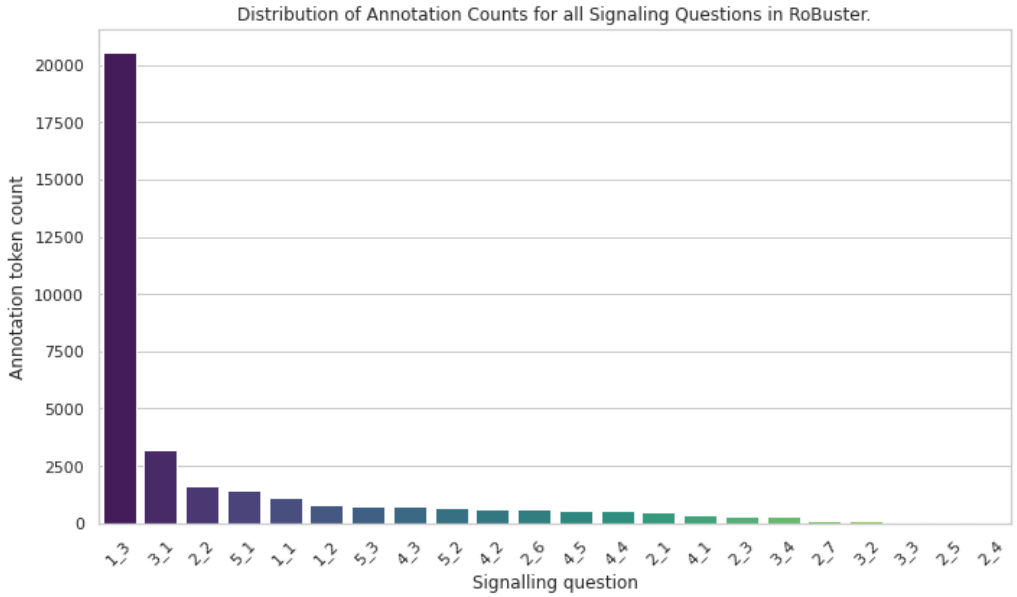


Figure 5.9: Total number of token annotations for each RoB SQ.

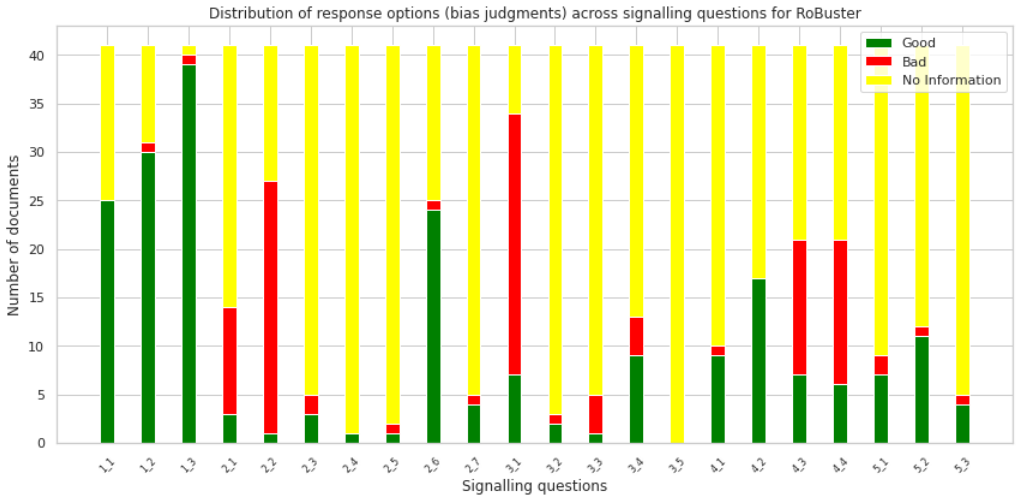


Figure 5.10: Distribution of bias judgment across RoB SQs in RoBuster.

5.4.2.3 Inter-annotator agreement

Table 5.13 illustrates the levels of F1-measure (inter-annotator agreement) between the two annotators, both before and after the development of the visual placards. The F1-measure before and after the guideline improvement were calculated on a different set of documents. The average overall F1 agreement across the corpus exhibited a modest 10.87% agreement before the introduction of visual placards. Following their implementation, this agreement

SQ	Total tokens	Average length	Total documents
Domain 1: Biases arising from the randomization process			
RoB1.1	960	32	30
RoB1.2	838	24.65	34
RoB1.3	16446	411.15	40
Domain 2: Biases due to deviations from intended interventions			
RoB2.1	455	32.5	14
RoB2.2	1502	55.63	27
RoB2.3	282	56.4	5
RoB2.4	25	25	1
RoB2.5	58	29	2
RoB2.6	544	20.92	26
RoB2.7	126	25.2	5
Domain 3: Bias due to missing outcome data			
RoB3.1	2529	74.38	34
RoB3.2	103	34.33	3
RoB3.3	75	15	5
RoB3.4	276	21.23	13
Domain 4: Bias in the measurement of the outcome			
RoB4.1	240	24	10
RoB4.2	572	33.65	17
RoB4.3	698	30.35	23
RoB4.4	585	27.86	21
RoB4.5	622	41.47	15
Domain 5: Bias in the selection of the reported result			
RoB5.1	628	69.78	9
RoB5.2	235	19.58	12
RoB5.3	628	89.71	7

Table 5.12: General statistics for the annotated corpus: This table provides an overview of the annotated corpus, including the total number of annotated tokens, the average length of token sequences, and the number of documents in which annotations were identified, out of a total of 41 annotated documents.

increased significantly by 17.14 percent points, reaching a fair 28.01%. The agreement for the "randomization" domain 1 doubled from a poor (F1 31.72 IAA) to a substantial (F1 63.30 IAA). For the "deviations from intended interventions" domain 2, the agreement increased from a slight (F1 (F1 12.76 IAA) to fair (F1 27.02 IAA). In the case of the "missing outcomes" domain 3, the agreement rose from (F1 5.89 IAA) to (F1 9.92 IAA), although it remained within the slight agreement category. For the "missing outcome measurement" domain 4 the agreement increased from (F1 4.072 IAA) to (F1 17.29 IAA). The agreement before the placard development was none and it increased to slight (F1 16.49 IAA) for the "selection of reported results" domain 5.

Improvements in agreement were observed at individual SQs level too, with considerable gains exceeding 50% points for signaling questions 1.3, 2.1, 4.1, and 4.4. The F1-measure between the two reference annotators for a total of 11 out of 22 questions remained 0 both before and after the guideline development. Notice that nine of the 11 questions for which the agreement was zero before and the after placard development are the same SQs. For two SQs, 4.4 and 5.2, the guideline improvement raised it from 0 to 56.25 and 49.49 respectively. However, for SQs 2.3 and 2.7, the agreement experienced a modest decline of 5.42 and 6.52, reaching 0 after the guideline development. For the SQ 1.2, the agreement dropped by 6.28 IAA points making it a drop in agreement for 3 SQs. Inferring from the table 5.11 and table 5.13, annotators for 11 of the 22 SQs had a poor agreement, 3 of the 22 questions had a fair agreement, 4 out of the 22 questions had a good agreement and 2 out of the 22 questions had a substantial agreement while only 2 of the 22 questions had an almost perfect agreement of beyond 81 IAA points. As expected, none of the SQ annotations had a perfect F1-measure.

We present the prevalence and bias adjusted kappa κ_{pabak} agreement, raw agreement and the percentage of κ_{pabak} agreements that stemmed from the "No Information" judgments in Table 5.14. To recall κ_{pabak} measures agreement at the SQ risk judgment level [223]. The overall κ_{pabak} agreement between the annotators stand at a weak IAA of 0.41181 IAA. The average agreement for "randomization" domain 1 was a moderate 0.629 IAA. For the "deviations due to intended interventions" (domain 2), there was a moderate agreement of 0.64 IAA and for the "due to missing outcome data" (domain 3) there was a minimal agreement of 0.388 IAA. The last two domain 4 and 5 "measurement of the outcome" and "selection of the reported result" had agreements 0.166 IAA and 0.092 IAA interpreted slight and no agreement.

The highest agreement of 1.0 is observed for SQ 2.6. It is, however, important to note that more than half of this agreement arises from the assumed "No Information" judgments for 2.6. For the remaining SQs, there were no instances of either almost perfect or strong agreement. Agreements for SQs 1.1, 2.1, 2.5, 3.1, and 4.3 fall within the moderate agreement category ranging from 0.60 to 0.80 IAA. Amongst these SQs, only for the SQ 1.1, 2.1, 3.1 and 4.3 has no substantial agreement originating from the "No Information" judgments. In contrast, 2 of the 22 signaling questions (4.2 and 5.1) exhibit agreement worse than what would be expected by chance. For SQ 5.1, one annotator did not mark any text, leading to negative IAA, while for SQ 4.2, both annotators correctly marked their answers in different parts of the text, resulting in negative agreement. The bias questions 3.3 and 4.5 show 0 κ_{pabak} agreement. The IAA was zero for SQ 4.5 because the annotators annotated mutually exclusive sets of documents, resulting in no consensus. A similar case was observed for SQ 3.3, wherein both annotators correctly marked the answer for only one document. However, since these markings were in different text parts, the agreement remained at zero.

SQ	F1-measure IAA before guideline improvement	after guidelines improvement	change
Domain 1: Biases arising from the randomization process			
RoB 1.1	24.44	55.02	+30.58
RoB 1.2	50.28	44	-6.28
RoB 1.3	20.44	90.9	+70.46
Domain 2: Biases due to deviations from intended interventions			
RoB 2.1	1.34	67.26	+65.92
RoB 2.2	7.23	38.66	+31.43
RoB 2.3	5.42	0	-5.42
RoB 2.4	-	0	0
RoB 2.5	0	0	0
RoB 2.6	68.85	83.25	+14.4
RoB 2.7	6.52	0	-6.52
Domain 3: Bias due to missing outcome data			
RoB 3.1	23.57	39.68	+16.11
RoB 3.2	0	0	0
RoB 3.3	0	0	0
RoB 3.4	0	0	0
Domain 4: Bias in the measurement of the outcome			
RoB 4.1	6.51	61.71	+55.2
RoB 4.2	0	0	0
RoB 4.3	13.85	30.21	+16.36
RoB 4.4	0	56.25	+56.25
RoB 4.5	0	0	0
Domain 5: Bias in the selection of the reported result			
RoB 5.1	0	0	0
RoB 5.2	0	49.49	+49.49
RoB 5.3	0	0	0

Table 5.13: The table displays the F1-Measure at the text span annotation level before and after the development of visual placards. The change in F1-Measure is presented in terms of absolute IAA points. Note: Dash (-) shows that one of the annotators did not annotate any text for a particular SQ.

SQ	κ_{pabak} agreement	Raw agreement	Contribution from "No Information"
Domain 1: Biases arising from the randomization process			
RoB 1.1	0.8333	88.90%	22.22%
RoB 1.2	0.5	66.70%	33.33%
RoB 1.3	0.5556	77.80%	11.11%
Domain 2: Biases due to deviations from intended interventions			
RoB 2.1	0.7037	77.80%	5.56%
RoB 2.2	0.5	66.70%	38.89%
RoB 2.3	0.5	66.70%	77.78%
RoB 2.4	0.5556	77.80%	88.89%
RoB 2.5	0.6667	77.80%	83.33%
RoB 2.6	1	100.00%	55.56%
RoB 2.7	0.5556	77.80%	77.78%
Domain 3: Bias due to missing outcome data			
RoB 3.1	0.8333	88.90%	11.11%
RoB 3.2	-	100.00%	100.00%
RoB 3.3	0	33.30%	55.56%
RoB 3.4	0.3333	55.60%	33.33%
Domain 4: Bias in the measurement of the outcome			
RoB 4.1	0.5556	77.80%	11.11%
RoB 4.2	-0.5556	22.20%	38.89%
RoB 4.3	0.6667	77.80%	27.78%
RoB 4.4	0.1667	44.40%	22.22%
RoB 4.5	0	33.30%	66.67%
Domain 5: Bias in the selection of the reported result			
RoB 5.1	-0.5556	22.20%	61.11%
RoB 5.2	0.5	66.70%	22.22%
RoB 5.3	0.3333	55.60%	72.22%

Table 5.14: Prevalence and Bias adjusted Kappa κ_{pabak} and raw agreement between annotator pairs for agreement at the risk judgment level for each SQ.

5.4.2.4 LLM Evaluation

Table 5.15, presents the observed or raw agreement between LLM and expert assessments in extracting and responding to SQ over a subset (n=10) of RoBuster. The evaluation was conducted on four RCTs reporting objective outcome, three RCTs evaluating subjective outcome and the rest three RCTs reporting mortality outcome. For the first risk domain, ChatGPT had high $P_O(Extraction)$ and $P_O(Response)$ agreements (66.6% and 55.3% IAA respectively) with experts and none of these agreements came from "No Information" responses indicating a good availability of information for bias assessment. Specifically, for individual SQs within domain 1, the $P_O(Extraction)$ was greater than $P_O(Response)$. The observed agreements for the the second domain are even higher 64.28% and 60% respectively, but 40% of these agreements emanate from "No Information" responses suggesting lower availability of information. Domain 3 exhibited a moderate average observed agreement of 47.5% for both $P_O(Response)$ and $P_O(Extraction)$, while Domain 4 demonstrated a lower observed agreement of 28% and 26% for response and extraction, respectively. For the domain 5, the observed agreement for extraction was lower than the observed agreement for response. Details about the RCTs used for LLM evaluation are in the Appendix.

5.4.3 Discussion

5.4.3.1 Visual Placards

As per [83], there were two reasons causing a low F1 IAA for annotating text span to answer a SQ. One reason was a lack of instructions whether the annotators should annotate a phrase, a sentence or sentences. While some annotators might annotate an entire paragraph as text evidence to answer a SQ, others might focus on the most informative portion of the text. To address this, our placards provide clear guidance on whether annotators should annotate a phrase, a sentence, or a combination of sentences. Another common reason for a low F1 was when annotators correctly addressed a bias SQ but annotated evidence from different parts of the full text leading to no or low agreement. To tackle this, our placards restrict annotations for a question to a specific part of the text for specific SQs, such as the Methods section, Results section, Flowchart in the Methods section, etc. Annotations in Flowcharts and Tables are restricted as last priority for all SQs except 1.3. We had this restriction since ML models face challenges in directly interpreting Tables and Figures. Reiterating the details presented in Figure 5.8, the information related to answering SQ 3.1 is found in the RCT flowchart. However, the first diamond instructs annotators to locate the information in the Results section for annotation. This decision is made to facilitate ML models, as training on text data from the Results section is deemed more effective.

The visual placards addressed various facets of RoB 2 subjectivity and also when dealing with situations lacking information for risk judgment annotation. In one trial document annotation, both annotators selected the phrase "71 allocated routine services, 67 allocated intervention service, 69 assessed at 8 weeks, 64 assessed at 8 weeks" from the PRISMA flowchart to answer signaling question 3.1 [121]. However, one annotator responded with "Yes" while the other chose "No". This question asks whether outcome data were available for all or nearly all participants randomized, but it doesn't specify the exact cutoff for participant dropouts that increase the risk. Therefore, annotators make subjective judgments based on their experience regarding what percentage of participant

SQ	Observed Agreement P_O $P_O(Extraction)$	$P_O(Response)$	"No Information"
Domain 1: Biases arising from the randomization process			
RoB 1.1	90%	70%	0%
RoB 1.2	70%	60%	0%
RoB 1.3	40%	30%	0%
Domain 2: Biases due to deviations from intended interventions			
RoB 2.1	50%	40%	0%
RoB 2.2	30%	30%	10%
RoB 2.3	60%	60%	50%
RoB 2.4	90%	90%	100%
RoB 2.5	90%	90%	100%
RoB 2.6	80%	50%	10%
RoB 2.7	50%	60%	40%
Domain 3: Bias due to missing outcome data			
RoB 3.1	30%	40%	0%
RoB 3.2	60%	40%	40%
RoB 3.3	30%	30%	30%
RoB 3.4	70%	80%	70%
Domain 4: Bias in the measurement of the outcome			
RoB 4.1	40%	30%	10%
RoB 4.2	40%	30%	20%
RoB 4.3	10%	10%	10%
RoB 4.4	10%	20%	10%
RoB 4.5	40%	40%	40%
Domain 5: Bias in the selection of the reported result			
RoB 5.1	22.22%	77.77%	0%
RoB 5.2	33.33%	44.44%	33.33%
RoB 5.3	55.55%	55.55%	44.44%

Table 5.15: LLM evaluation: Observed agreements between LLM and experts over a subset of RoBuster. Note: For the domain 5, LLM evaluation was conducted on 9 RCTs, as one of the RCTs did not have the trial registry available.

dropout is considered valid. To address this subjectivity, we introduced a threshold of 95% in the placard instructing annotation (see Figure 5.8).

5.4.3.2 The Corpus: RoBuster

The immediate points noticed in the Figure 5.9 were that SQ 1.3 had a disproportionately higher number of annotated tokens (n=16,446) while the remaining signal questions had fewer than 2600 tokens each annotated. The reason behind this is that the answer to SQ 1.3 is found in the table detailing baseline patient characteristics of the intervention groups. To ensure good F1 IAA, instructions on visual placards directed annotators to label the entire table, leading to a higher count of annotated tokens for this question. For the rest of the SQs, it is the availability of detailed description on the study design, methods and results that could have impacted the amount of tokens annotated and also the subjectivity level of bias assessment. The more information a study provided, the easier it is to evaluate a bias question. Some studies tend to not report key details making it tougher to assess certain bias questions. Feedback forms received indicated that for SQs with fewer than 100 annotated tokens, annotators consistently rated the availability of information to answer those questions as either "low" or "very low". In contrast, for the top 5 signalling questions shown in Figure 5.9, both annotators consistently rated the availability of information as "high" and "normal". It is important to interpret this qualitative feedback with caution though, as annotator ratings are influenced by the number and types of RCTs they annotated. The annotator who reviewed a greater number of documents in RoBuster experienced that it was more difficult to assess certain questions and the availability of information was lower than the other annotator who annotated fewer documents (for IAA calculation).

5.4.3.3 Inter-Annotator Agreement

Text span agreement F1 agreement was calculated before and after visual placard development showing an improvement in the agreement for choosing the RCT text spans for 10 of the 22 SQs. SQ 1.3 had the highest increase of 70.46 F1 IAA points thanks to the visual placards requesting to look for the text evidence to answer this question first priority in the table recording patient characteristics and instructs to mark the entire table along with the table caption. Earlier, the annotators would only mark a portion of the table to answer this question. The specific text selected by an annotator depended on which part of the table text they noticed first as potentially indicating bias given any of the listed patient characteristics could show imbalances between the groups. This variability in selection led to lower F1 on this SQ. Similarly, SQ 1.1 showed a 30-point increase, with placards emphasizing marking text evidence in the Methods section. The evidence to answer the SQ 1.1 could be found in both abstract and the methods section and prior to placard development, the annotators variably marked the answer to this question leading to a lower agreement. We restricted annotating this text in the methods section in the visual placards because a more detailed textual description is found in this section. In addition, prior to the placard development, the annotator marked the phrase "randomized controlled trial" to justify their decision for "1.1 Yes Good". The placards added a rule that "Yes Good" was to be marked only if the annotators found a proper method of randomization and annotated it in the text. These improvements were also the reason for an increased agreement for the SQs 2.1 and 2.2. For SQ 2.1, there was a remarkable increase of 67.26 IAA points. This can be attributed to the comprehensive instructions

provided by the visual placards, which instruct annotators to specifically identify and label text descriptions related to intervention administration and placebo administration. Earlier, the annotators were inconsistently annotating either the description of intervention administration or placebo administration, leading to lower agreement.

While the agreement drastically increased for certain SQs, for other questions it remained 0. These outcomes align with the feedback forms' findings. Almost all the RoB questions where the IAA between the annotators was zero (2.3, 2.4, 2.5, 2.7, 3.2, 3.3, 3.4, 4.2, 4.5, 5.1, 5.3) consistently received remarks from one or both annotators regarding their higher subjectivity, difficulty in assessment, and lower availability of information for evaluation and hence annotating. The poor agreement hence can be well assumed to caused by the subjectivity of the SQs, lack of information to annotate in the RCT, and the overall complex process of analysing trials.

Zero IAA could also be attributed to the theoretical nature of certain SQs like 3.4 and 4.5. For instance, SQ 3.4 assesses the "likelihood" that missing outcomes data is related to the true values of those outcomes, examining the risk of bias associated with missing outcome data. Similarly, SQ 4.5, dependent on SQ 4.4 results, aims to gauge the "likelihood" that the assessment of the outcome was influenced by knowledge of whether the intervention was received. Such "theoretical" questions require making judgments that rely on hypothetical scenarios rather than direct observations or concrete text evidence leading to higher subjectivity of assessment results. As these questions pertain to aspects of study design, conduct, or reporting that may not be explicitly addressed in the trial documentation these questions also suffer from a lack of information that could be annotated in a trial. To address this, annotators are instructed to annotate outcome and outcome measurement descriptions in the paper for these questions, ensuring a basis for judgment and the availability of annotations for training and evaluating ML models on RoBuster.

There were certain aspects where question subjectivity might have led to low agreement score. For example, for the SQ 2.4, it is asked whether deviations from the intended intervention could have affected the outcome. Assessing whether deviations could have affected outcomes requires a subjective judgment. What one person considers likely to affect outcomes, another might not. This leaves room for variability in how different assessors may interpret and score this question. In addition, annotators may have limited information available to annotate any text and to make a judgment. They may not have access to detailed data or explanations regarding the deviations, making it challenging to assess their potential impact accurately. Specifically, only 2 out of the 9 RCTs had annotations for SQ 2.4 and these annotations were made only by one of the annotators and not by both annotators lading to no agreement at the text span level. Similar subjectivity occurs for the SQ 3.3: Could missingness in the outcome depend on its true value? The SQ involves assessors making a judgment about the relationship between missing data and the true outcome, which can be challenging to determine objectively. Annotating text evidence is even more difficult for this question because authors need to annotate the reasons for missing outcomes data that are related to the true value of the outcome, for the physiotherapy outcomes like 'fatigue'. If authors describe the reasons for missing outcomes data as patients not showing up at follow-up owing to fatigue, annotating this is necessary because missingness of outcomes data in this case depends on the true value of outcome (here fatigue). However, this information is not usually available in the RCT. Placards instructed annotators to mark outcome descriptions instead, aiding judgment. Unfortunately, one annotator did not follow this instruction, resulting in 0 agreement at

the text span level. Such instances suggest a longer training and several rounds of conflict resolution and correction are required for the annotators.

Response option agreement The disagreements at response option judgments arose because of two reasons. The most apparent cause of disagreement occurred when two annotators choose a text section to address a signaling question but did not reach a consensus on the response option judgement. The other less obvious reason was when one of the annotators annotated a part of text to answer a SQ and gives it a response option judgment, but the other one did not annotate anything for this SQ leading to an automatic assumption of "No Information". Notably, 82.85% of these conflicts fall into the second category, with only about 17% involving discrepancies in response judgments. Similarly, agreements were of two types too, the agreements when both annotators chose a text span for answering a SQ and labelled it with the same response options and the agreements when none of the annotators answered a SQ leading to both their annotations being set to "No Information". The agreements were evenly distributed between these two categories. Therefore, a considerable of chunk of both agreements and disagreements came from "No Information". It is worth recalling the prevalence of the "No Information" judgment, as indicated by the yellow bar in Figure 5.10. This underscores the significance of considering whether to use RCTs from specific high-quality journals in the annotation corpus. Doing so could potentially ensure more comprehensive information reported in the paper and is available for annotation and assessment, thereby contributing to the creation of a corpus that will not have as many "No Information" judgments.

Negative κ are unlikely to occur in practice, but for two SQs (4.2 and 5.1) where the response judgment agreement was measured by κ_{pabak} was negative. The reason for lower κ_{pabak} for these questions was disagreements over seven of the nine annotated documents which was the highest number of disagreements in the subset RCTs used for kappa calculation (refer Figure 5.11). To note that these agreements came from "No Information" judgments. Though the κ_{pabak} values for these SQs are considerably smaller than -0.10. According to [223], a κ value smaller than -0.10 represents strong disagreement among the raters and the collected data are considered not meaningful.

The SQs (1.1, 2.1, 2.5, 2.6, 3.1, 4.3) with adequate agreement ($>$ 0.60 kappa) had higher number of agreements (refer Figure 5.11) not coming from "No Information" judgments. This aligned with the results of the feedback form where the difficulty and subjectivity of these questions was consistently rated as "Normal" or "very low", the availability of information was rated "normal" or "very high". For the SQs with zero and negative agreement, had a net higher number of disagreements. In a nutshell, a decrease in disagreements and an increase in agreements led to improved κ_{pabak} agreement.

Domain 5 had an average overall agreement close to zero and the reason for this could be the complexity for analysing this domain. To assess this risk domain it was imperative to attach a trial protocol with the RCT document. Consequently, the annotators needed to mark the information in both the RCT and its corresponding protocol and make a judgment after summarizing information.

5.4.3.4 Feedback on Placards

We will now address the concerns raised by Julian PT Higgins in bias assessment. To address subjectivity arising from signaling questions 2.7 and 3.1, which inquire about the potential impact of failure to analyze participants in their randomized groups and

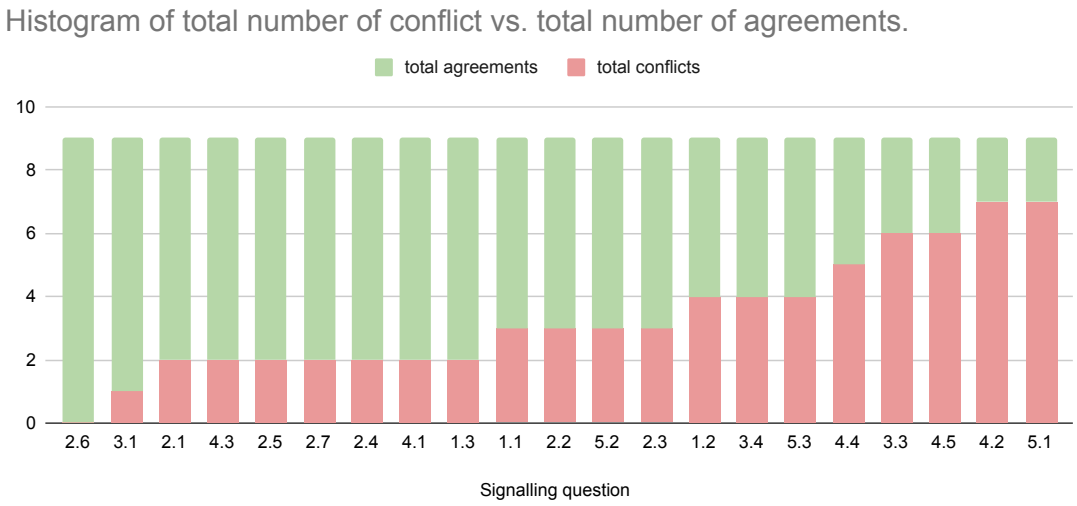


Figure 5.11: The histogram of total number of conflicts and total number of agreements in the subset of RCTs used to calculate prevalence and bias adjusted kappa.

the availability of data for all participants, respectively, the concept of a threshold was introduced in this work. Specifically, SQ 2.7 "Was there potential for a substantial impact (on the result) of the failure to analyse participants in the group to which they were randomized?". To annotate for this question, the visual placard instructs the annotators to find the data about the number of participants included and excluded from the intention-to-treat (ITT) analysis. The placard then goes to instructs that if more than 5% of participants were excluded from the ITT analysis, it was a substantial number and mark this information and label it a higher risk of bias. Similarly, for the SQ 3.1 "Were data for this outcome available for all, or nearly all, Participants randomized?", the term "all, or nearly all, participants" can be interpreted differently by different reviewers. What one reviewer considers "nearly all" might be different from another reviewer's interpretation. This ambiguity can lead to subjective judgments when applying the tool. Therefore, we introduced the 95% threshold meaning if the data were not available for at least 95% or more participants, we instructed the assessors/annotators to judge the bias risk as high. On our end, to simplify the annotation process and increase the IAA, these thresholds were chosen but enforcing a threshold deviates from the recommended RoB guidance as critiqued by Julian Higgins. In practical, thresholds are not one-solution-fits-all and are difficult to apply in different trial conditions. Thus, an alternative approach to guide annotators in assessing risk without relying on thresholds should be found. One potential solution is to provide examples illustrating instances where low numbers may still introduce bias, offering guidance on specific result figures that annotators should scrutinize.

We also received guidance to designate specific SQ placards with a "Subjective" flag to signal the presence of subjective judgments. Consequently, for several SQs, the recommendation was incorporated, adding a "subjective flag" to the placards wherever subjective judgments were necessary. Previously, a subjective judgment flag had been marked for SQ 2.4, and after the recommendation, it was also marked for SQs 3.3, 3.4, and 4.5.

5.4.3.5 LLM evaluation

Manual LLM evaluation highlighted the tricky nature of bias annotation as well as the subjective nature of LLM evaluation. LLM was tasked with the same work as human expert and was supposed to answer a SQs and extract text descriptions from RCTs that supported answer. In some cases LLM (via ChatPDF) straightforward answered a SQ with a response option - "Yes" or "No" or "No Information", but in certain cases it gave an indirect answer. For instance, in the Gleason study [124], for SQ 1.1, LLM directly and correctly answered "**Yes**, the allocation sequence was random. Extracted sentences: A random number generator (Excel, Microsoft) was used to devise a randomization series. The order of randomization was incorporated into a KEEPS database, so that treatment allocation occurred when a study identification number was assigned." In contrast, to answer SQ 5.2, LLM indirectly answered the following, "Regarding RoB 5.2, the study authors reported that to correct for multiple comparisons, they divided the p-value by the number of outcomes. For example, there were five cognitive outcomes (four domain factors and one global cognition score) and six POMS mood scores. Splitting the significance value of 0.05 equally between the five cognitive and six mood sub-domains resulted in a p-value of 0.01 for each of the two primary domains investigated (cognition and mood). However, the study authors did not provide information on whether the numerical result being assessed was selected on the basis of the results from multiple eligible outcome measurements within the outcome domain." implying it found "No Information" for answering the SQ 5.2.

The LLM evaluation results of the first risk domain SQ 1.3 saw a comparatively lower agreement between LLM and the experts because the answer to this SQ is typically found in the table text of RCT. Text in tables is read by ChatPDF, but it might have problems correlating the correct rows and columns leading to a distorted understanding of the tables. This could be the reason why ChatPDF sought to answer this question typically using RCT text rather than the table text wherever available. To elaborate, for [314], both the expert and LLM answered SQ 1.3 with "No Good", but expert used Table 2 as the evidence to answer the question while LLM used the text evidence not found in the table.

> LLM answer: No, there were no significant differences ($p < 0.05$; based on z-statistics from generalized estimating equation models) between the intervention and control groups in any of the baseline characteristics listed in Table 2 [314].

The LLM correctly extracts information about participant randomization and allocation concealment procedures pertaining to answer SQ 1.1 and 1.2 leading to a rightly good $P_O(Extraction)$ agreements. As shown in the Table, for individual SQs in domain 1, $P_O(Extraction)$ was greater than $P_O(Response)$ but in an example, LLM came to a correct response judgment without correctly extracting complete evidence. To elaborate, in [314], LLM used this part of text "'PRO-AGE Solothurn CONSORT diagram. The randomization ratio (intervention to control group) was 1:1 in the first project phase (November 16, 2000, to March 27, 2001), and 1:2 in the second project phase (March 28, 2001, to January 8, 2002), resulting in a ratio overall of 1:1.6" to answer SQ 1.1 with "Yes Good" (low risk of randomization bias) even though the extracted sentence does not contain information about the randomization method.

The results for domain two where about 40% of average P_O agreement comes from "No Information" shows a lack of availability of information in RCTs for answering the SQs especially for the SQ 2.4 and 2.5. However, in one instance for the domain 2, LLM

evaluation led to identification of incorrect label from the expert. For SQ 2.6 in [139], the expert had annotated this text ("The main analysis for each subgroup analysis was an interaction test in the regression models to determine whether the effect of treatment differed significantly across categories for that variable.") to answer the question with Yes Good, but the extracted text evidence was incorrect. LLM correctly extracted information about the intention-to-treat analysis which led to correcting the final annotation in RoBuster.

The evaluation for third domain of bias brought to light the subjectivity of bias assessment. A lenient assessor will judge a risk of bias as low in comparison to a stringent assessor judging a bias risk as high for any SQ, but it is more pronounced in subjective SQs. In [321], LLM and the expert used the same text evidence "...104 (82%) were randomized in January 2011. Five residents withdrew during Phase 1, and 99 continued participation in Phase 2..." to answer SQ 3.1 but LLM produced the response option "Yes Good" and the expert rated it as "No Bad". It was also because our placards have a stringent rule whereby if outcome data was not available for more than 95% of the participants, the assessor/annotator was instructed to judge the risk as high risk of bias without considering other factors like total number of participants randomized and the kind of study. For answering SQ 3.2 across the evaluation RCTs [243, 320], the LLM was more lenient than the experts and consistently extracted information about the sensitivity analysis carried out by the study authors to account for missing outcome data, but the authors were not explicit that there was no bias due to missing outcomes data. LLM judged signalling question 3.3 9 out of 10 times as "No information". It questions whether the missingness in outcomes depended on its true value and is quite subjective because assessor needs to contemplate on the reasons for missingness for a particular outcome and if its true value could have affected the missingness. For example, they need to assess whether the fact that data on falls is missing can be attributed to the actual occurrence of falls in the study.

The LLM exhibited a lower overall agreement with the annotators for the domain 4 despite the data for outcomes measurement was available in the RCTs. The reason for low agreement was the use of simple prompts in our work. For RoB assessment, the annotators assess bias pertaining to the pre-decided target outcome and not all the outcomes reported in the trial. However, the information about the target outcome being assessed was not available in the simple prompts used to test LLMs. Consequently, the LLMs extracted information that did not necessarily pertain to the target outcome but rather to other outcomes reported in the trial. This situation suggests a potential avenue for improvement, involving the exploration of adapted prompting styles, such as chain-of-thought (CoT) and tree-of-thought (ToT), commonly used in language generation tasks, for the RoB text extraction task [355]. While the LLM demonstrates promise in automating RoB span extraction, additional research is necessary, particularly in the domains of prompt engineering to tackle the subjectivity inherent in SQs and the scarcity of information in RCTs required to answer SQs.

5.4.4 Limitations

We have the following limitations with this work. The first limitation arises from the relative scale of our annotations, with only 41 documents undergoing the annotation process. This limitation is primarily due to financial constraints, as the availability of funds limited our ability to hire a larger number of expert annotators. Despite the limitation, the robustness of the annotations provided by the experts remains noteworthy, as they

thoroughly assessed the selected RCTs within the resources at hand. A second limitation stems from the narrow focus of annotations, concentrated exclusively on physiotherapy and rehabilitation clinical trials. This specificity arises from our reliance on domain experts for annotations, who possess the requisite expertise in these areas. Though this approach inadvertently restricts the broader applicability of the annotated corpus, it guarantees high-quality annotations within the specified domains.

The LLM evaluation was limited by the fact that we chose to work with PDFs. There are limited platforms that interact with PDFs and this restricts our choice of the models to evaluate. Another limitation is the stochastic nature of LLMs. Specifically, Google Bard is freely available tool that interacts with PDFs, but we observed that the Bard results were less deterministic than ChatPDF. Another drawback was associated with the prompts used. The prompts employed were relatively simple and lacked information about the target outcome assessed for bias. Certain SQs in the domain 4 and 5 seek information pertaining to specific target outcomes being assessed. Therefore, when LLMs are not provided this information via prompts, they tend to provide general information about outcomes reported in the RCT but did not specifically consider the target outcome of interest, as it was not instructed to do so.

5.5 Chapter Conclusions

In conclusion, the first pilot annotation project demonstrated that RoB 2 assessment guidelines cannot be directly used as RoB corpus annotation guidelines because they lead to poor inter-annotator agreements. It showed that the multi-level annotation scheme directly adapted from RoB 2 document needed improvement, as detailed in the discussion section (refer 5.3.3). This exercise gave us detailed insights into the challenging task of RoB annotation and how to develop crisp annotation guidelines to obtain consistent annotations.

In our second annotation project, RoBuster, a new, publicly-available corpus, comprising 41 full-text RCTs richly annotated with RoB span information for 22 risk of bias questions was presented. RoBuster fills the need for a corpus to evaluate RoB text span extraction using machine learning approaches. It is a comprehensive resource with detailed, fine-grained information, presenting individual RoB spans and annotator decisions on bias risk (high or low). We used RoBuster as a benchmark to evaluate LLMs for how well LLMs agree with human-led bias assessment. Developed collaboratively by bias assessors and NLP experts, RoBuster not only supports automated approaches to bias assessment but can also contribute to Living Systematic Review (LSR) systems. The work also contributes crisp RoB corpus annotation guidelines in form of visual annotation placards. Our combined work in RoB corpus annotation reaffirms the complexity of RoB assessment and the necessity of developing comprehensive instruction guidelines to increase inter annotator reliability across both tasks (assessment and annotation).

We conclude that developing the RoB annotation scheme and an annotated corpus is complex and ridden with subjective judgments but is feasible with the iterative refinement of annotation guidelines. We filled the research gap by developing visual annotation guidelines in form of placards that could be used not only for developing a RoB annotated corpus, but also for training the new RoB assessors. We also demonstrated the utility of these placards by showing an increase in the IAA between the expert annotators. Finally, we showed utility of RoBuster as a validation benchmark, by carrying out LLM evaluation on a subset of 10 annotated RCTs from RoBuster. We encountered several challenges

during the development of both RoB annotation guidelines and the actual annotation attributed to how well each RCT reports the study methodology, statistical methods and outcome information along with how subjective the assessment of each RoB question is in RoB 2 tool. In the future, we plan to refine our visual placards and extend RoBuster by adding more annotated RCT full-texts. We also plan to test the visual annotation placards to guide trainee risk of bias assessors.

Chapter 6

Thesis Result Summary

6.1 Thesis Results Summary

The demand for systematic reviews is growing, but concurrently the new evidence, primarily in the form of clinical trials, RCTs, and other primary studies, published at an unmanageable pace. The process of writing reviews is rigorous and detail-oriented. The review team starts with searching and collecting the available studies, filtering the studies based on relevance to the review question, thoroughly assessing the studies for risk of bias, performing meta-analysis, writing a manuscript and updating the review once new primary studies are available. The cost of producing a single systematic review can reach up to 300,000 US dollars. For academia and industries that generate a minimum of 100 reviews annually, the costs escalate to a staggering 30 million USD.

The thesis included eight articles, six of which have been published, one is currently under review, and one is ready for submission, spanning 2020 to 2024. In these papers, we have attempted to develop resources and explore methods for semi-automating three main stages of conducting SRs: citation screening, PICO information analysis and RoB assessment. Notice that the methods automating PICO analysis and RoB assessment support the data extraction stage by extracting relevant information. The previous automation efforts have predominantly focused on systematic reviews of medical interventions, leaving a gap in validating these methods for domains like physiotherapy and rehabilitation, particularly in a prospective scenario. In this thesis, we have examined how machine learning and NLP methods can be used to reduce this workload and how these can fit into different systematic review stages and settings, including systematic reviews in physiotherapy, rehabilitation, and pharmaceuticals.

6.1.1 Semi-Automation of Citation Screening

Chapter 3 presented two approaches to citation screening. The first approach explored static models trained on the Hilfiker dataset, a citation screening dataset for the physiotherapy domain. A static model is a fully supervised model that can only be deployed to update a systematic review, but its applicability for a *de-novo* SR tackling a completely different clinical question is limited to none. In the analysis using the static model, we highlighted class imbalance and class overlaps as primary obstacles to citation screening when treating the task as a binary classification task.

The second approach introduced an active citation screening system, evaluated specifically for *de-novo* systematic reviews in both retrospective and prospective scenarios. While

the static model was evaluated only on one dataset, the AL system was rigorously tested on 25 datasets representing domains of biomedicine, physiotherapy, and pharmaceuticals amongst others. Furthermore, the models were evaluated using two most important evaluation metrics: WSS and recall and tie breaker metrics like ROC-AUC and F1-score in case WSS and recall were consistently similar.

Most AL experimental setups saved some form of workload as measured by WSS also ensuring the AL system was only trained only using 30% of the training data. Therefore, for us recall of the "relevant" studies became the vital measure in evaluating different AL settings. To recall, the prospective / future facing scenario assumes an unlabelled citation screening dataset and was tested using hasty seed sampling triggering active classifier training as soon as one relevant citation is available. The retrospective simulation tested patient sampling technique initiating training with at least five relevant citations.

For both prospective PAL and retrospective HAL scenarios, diversity seed sampling using clustering consistently and significantly outperformed recall compared to random sampling. The best AL approaches, utilizing diversity and certainty query sampling methods, consistently outperformed random sampling. These findings reinforce the importance of selecting diverse samples for AL. The differences between certainty and diversity sampling were modest, and the choice between them may depend on deployment restrictions, considering the fact that diversity sampling took four times longer to execute compared to certainty query sampling.

We observed that be it hasty or patient sampling, the seed cost, which is the number of citations need hand labelling initially before the active training can begin, is high for low prevalence datasets. The reviewers might need to label as many as 300 citations before at least one relevant citation can be sampled. We examined if coverage was related to recall of the "relevant" citations and did not find a strong link between the both. With the focus on Coverage, we are ignoring the persistent class imbalance here. Further experiments are required to measure the impact of class imbalance for each training iteration on the recall of "relevant" citations.

6.1.2 PICO Information Analysis

In Chapter 4, we addressed a research gap by developing affordable methods for semantic PICO+ information extraction. The evaluation of these methods consistently utilized two gold-standard datasets to ensure comparability across methods: I) the EBM-PICO dataset and II) the Hilfiker dataset, both labelled with multilevel PICO annotations. EBM-PICO is representative of biomedical and clinical RCTs while Hilfiker dataset is representative of RCTs from physiotherapy. The first level of annotations include top-level coarse-grained PICO text descriptions and the second level of annotations further decompose PICO into more semantic, fine-grained categories. Firstly, we formulated the task as a multi-task learning problem allowing us to simultaneously extract fine-grained and coarse-grained information. Acknowledging the lack of a larger, more representative dataset for areas beyond biomedical and clinical fields to train deep learning models, we proposed a novel distantly supervised information extraction method. This method was validated with a proof-of-concept on "Intervention" information extraction. Finally, we introduced a weak supervision approach using generative modeling for successfully extracting PICO information, extending its utility to include "study type and design" information. The weak supervision approach showed promising performance over PICOS extraction even in the absence of any labeled data.

6.1.2.1 Multitask Learning Models

With multi-task learning, our assumption was that training simultaneously on two tasks could improve the model performance on both of them. We anticipated that the inductive bias from coarse-grained PICO (Population, Intervention, Comparison, Outcome) extraction might enhance performance in fine-grained PICO entity extraction.

The multi-task learning approach did indeed surpass the single-task learning counterpart for both "Participant" and "Intervention" extraction in the fine-grained extraction task on the Hilfiker physiotherapy corpus, showing a 2% F1 improvement. Moreover, in the EBM-PICO corpus, MTL outperformed STL for "Intervention" extraction with a 4% F1 improvement. However, effectively weighing and combining these two tasks and to design the optimal MTL architecture proved to be a non-trivial challenge.

6.1.2.2 Distant Supervision Models

The distant supervision approach for extracting "Intervention" information utilized clinicaltrials.gov as a knowledge base, generating 977,682 pseudo-annotation labels across 11 semantic types through alignment to raw text. The alignment parameter d_s ensures high-quality pseudo-annotations. Two deep BiLSTM models were trained: one solely on these pseudo-annotations and another combining them with the EBM-PICO dataset containing manually labeled "Intervention" annotations. The model based on pseudo-annotations alone exhibited a 9% increase in recall from the original baseline provided by Nye **et al.** and demonstrated a 10% improvement in F1 compared to the SciBERT model, which is a much larger model trained with 3.1 billion parameters. In contrast, the combination model, trained on both the pseudo-annotations and the hand-labeled data from EBM-PICO, outperformed the state-of-the-art by 0.93% in recall and 5.15% in F1-score, highlighting the effectiveness of this approach. Furthermore, the combined model showcased a 10% F1 improvement on the Hilfiker physio dataset compared to the model using only pseudo-labels.

6.1.2.3 Weak Supervision Approach

The distant supervision method relied on a single source for pseudo-labelling the target entity. In contrast, our weak supervision approach to PICOS extraction leveraged over 500 sources of pseudo-labelling to annotate the EBM-PICO corpus with these entities. Generative modeling was employed to derive consensus labels by combining individual labels based on the confidence of labelling sources.

The weakly-labelled EBM-PICO corpus was used to fine-tune PubMedBERT model, thus incorporating contextual information from EBM-PICO sequences. When evaluated on the EBM-PICO test set, this approach outperformed the fully-supervised model by 1.71% on the F1 score for "participant" extraction. While the performance on "intervention" extraction showed promise, it did not surpass that of the fully-supervised model.

However, the weakly supervised model under-performed in "outcome" extraction by 17.65 percentage points compared to full supervision. In the absence of any labeled data for "study type and design" information, the weakly supervised model achieved an excellent F1 score of 85.02%. Ablation experiments revealed that removing non-UMLS labelling sources improved the F1 by 3.06%, while removing rules deprecated the F1 by more than 5%, underscoring the importance of selecting representative weak labelling sources. Similar conclusions were drawn from "outcome" extraction results, where the addition of

non-UMLS ontologies degraded the F1 by 3.0%, indicating the unrepresentativeness of the selected non-UMLS vocabularies for labelling outcomes.

In addition to the representativeness of the labelling source, the class composition had a visible impact on performance. The weakly supervised model outperformed full supervision for the "participant" class due to its comparatively less heterogeneity compared to the intervention and outcomes classes. This class encompasses a more homogeneous pattern representing about patient demographics, including numerical information about age. The gender and sexual orientation could be represented by a limited dictionary, as well as ontologically well-defined disease and symptom classes. We observed a similar situation for the study type class, which is intuitively even more homogeneous than the participant class.

6.1.3 Risk of Bias Assessment

Chapter 5 addressed the research gap in the automation of RoB assessment arising from the absence of an annotated corpus. We introduced RoBuster, a corpus comprising 41 full-text RCTs annotated with 22 categories of risk of bias text spans, each classified with one of the two bias judgement classes. RoBuster to our knowledge the first of its kind dataset and our hope is that it will aid for better understanding of how reviewers perform RoB assessment and for modelling the process with automated methods.

We identified that the most challenging aspect of RoB corpus annotation was developing clear annotation guidelines given bias assessment is intricate, convoluted, and an expert-led task. The revised Cochrane risk of bias assessment tool for RCTs (RoB 2) provide a valuable foundation with comprehensive and structured bias assessment guidelines. Through pilot experiments, we discovered that Rob 2 guidelines, though being comprehensive and step-by-step, couldn't be directly employed as corpus annotation instructions. So, we collaborated with experts to adapt them for annotation purposes. These adapted instructions took the form of visual placards, which two reviewers then used to annotate 22 RoB text spans within a subset of RoBuster. Research on corpus annotation can only be as good as the annotations and the annotations can only be as good as the annotation guidelines. We quality control the annotations measuring annotator agreements over a portion of RoBuster.

6.1.3.1 Inter-Annotator Agreement

We measured annotator agreements at two levels. On level 1, the F1 agreement was measured for how well two annotators agreed on text spans indicating risk of bias. This agreement was calculated between two annotator pairs before and after the development of visual annotation guidelines. After implementing visual annotation guidelines, we observed a significant improvement in agreement for text span annotations, with an increase of over 17% F1 percentage points compared to when the guidelines were not employed.

Investigating a bit deeper, we found that the substantial increase in F1 agreement was observed in the signalling questions where the agreement was non zero before the visual placard development (except the questions 4.4 and 5.2). Nine signaling questions showed zero agreement both before and after guideline development, most likely due to limited information availability. Zero agreement often resulted from the absence of textual descriptions for annotators to evaluate.

Information related to signalling questions with non-zero agreement was mandated to be reported in clinical trials by the CONSORT statement. For example, signaling ques-

tion 1.3 demonstrated almost perfect agreement at 90.9%, as the necessary information for assessment was found in a table presenting baseline demographic and clinical characteristics for each treatment group—a requirement outlined by CONSORT guidelines [71]. These questions which are characterized by high information availability, were crucial in achieving non-zero agreement, and it was in these instances that the visual guidelines significantly enhanced agreement scores.

On level 2, k_{pabak} agreement was measured for how well two annotators chose the risk of bias level based on the selected text spans. For the first two risk of bias domains, there was moderate agreement (> 60 k_{pabak}) while for the rest the agreement was minimal to slight to low. We attribute the lower agreement k_{pabak} values to combined low information availability and subjectivity of the signalling questions. For instance, signaling question 2.3 exhibited minimal agreement of 0.5, primarily due to the subjectivity in the question asking reviewers to make a judgment based on trial context. Moreover, approximately 79% of the already minimal agreement of 0.5 was credited to "No Information" availability labels, leading us to classify this question as a low-information-availability question. The high subjectivity in these signalling questions is caused by subjectivity in the RoB 2 guidelines and a need to improve them for better reliability.

6.1.3.2 LLM Evaluation

RoBuster was utilized to evaluate GPT-3.5 where we recorded agreement between LLM and the expert on how well LLM extracted text descriptions to answer the signaling question and how well it classified the text with a risk level response option. The agreement scores showed promising but variable performance across different bias questions. LLM's extraction and response agreements varied across risk domains, with higher agreement in domains with good information availability and lower agreement in domains with subjective assessment and less available information.

In summary, agreements calculated at both annotation levels and the LLM evaluation favored more objective, high-information-availability questions and demonstrate solid utility of the visual annotation guidelines. The minimal to no agreement in low-information-availability and more subjective signaling questions signals that further research is required not only to improve the visual annotation guidelines but also to address subjectivity in RoB 2 and improve the reporting in published literature.

6.1.4 Summary

- Our investigation into the potential of active citation screening, supported by semi-supervised learning, helps bridge the gap between the current approaches and expected prospective approaches aligned with industrial automation processes.

- The generative weak and distant supervision approaches developed in this study provide a practical solution compared to fully supervised information extraction methods. Weak supervision can effectively label static datasets for new types of information where labeled data isn't easily available.

- RoBuster, our publicly available corpus annotated with risk of bias text spans, helps address the resources needed for training and evaluating NLP techniques, especially large language models.

6.2 Prospective Future Research Directions

Semi-Automatic Citation Screening:

- **Active learning for few-shot prompting LLMs:** Language models have shown improved generalization with few-shot prompting on new tasks [42, 116]. Few-shot fine-tuning requires prompting any LLM with a minimal number of labelled instances. Active learning using the diversity sampling method could be used to select these informative citations efficiently. This approach identifies informative citations representative of "relevant" and "irrelevant" classes, facilitating the few-shot learning of LLMs. Lightly.ai showed that using active learning to select instances for few-shot learning reduced 78% of labelling costs or improved the model by up to 4.6x per additional labelled batch in comparison to random selection [72] [316]. By optimizing the selection of citations for few-shot fine-tuning LLMs, active learning could enhance the generalization of LLMs in a few-shot citation screening task. It must be noted that experimentation is required to test this claim for citation screening given that a recently published study identified that few-shot prompting did not consistently improve performance, whereas 1-shot prompting did. [102].

- **Synthetic citation screening dataset generation:** The aim of this work investigating the active learning systems was to reduce the number of citations needing hand labelling for every *de-novo* SR question. AL still requires labelling a subset of the retrieved citations. Instead, using the PICO framework, LLMs could be employed to generate synthetic citation screening datasets representative of "relevant" and "irrelevant" classes. This eliminates the need for any manual annotation of citations. The synthetic dataset could be used to train active learning or machine learning models for binary classification. Abogunrin *et al.* recently carried out the first steps in this direction, testing the feasibility of using ChatGPT to synthetically generate abstracts that mimic peer-reviewed journal-looking abstracts [2].

Information Extraction Automation:

- **Ontology development**: This work tackled the lack of labelled data using weak and distant supervision approaches for PICO information extraction, but while studying the compositionality of PICO information and the requirement to extract more information beyond PICO, we propose development or compilation of ontologies for SR information extraction. Several ontologies exist in the literature, but their compilation could aid in consolidating the information that could be extracted and normalized for the automation of conducting SRs. For example, C-TrO ontology, clinical trials ontology (CTO), Ontology of Clinical Research (OCRe), Ontology of Biomedical Investigations (OBI), ontology of evidence-based medicine and meta-analysis, and RCT ontology [16, 198, 264, 290, 306].

- **LLM evaluation**: Comprehensive evaluation of LLMs for information extraction pertaining to EBM [58, 172, 319].

[72] `www.lightly.ai/post/improve-your-large-language-models-llms-with-active-learning`

RoB Assessment Automation: While this work led to the development of RoBuster, a RoB text span annotated corpus, which was used to evaluate GPT-3.5, these are still the two promising future research directions [83, 84].

- **RoB span annotated corpus in other domains**: RoBuster focuses on physiotherapy and rehabilitation domains, meaning its annotations primarily represent these areas. These annotations might not capture the nuances and patterns of bias risk descriptions from the other domains, e.g., pharmaceutical clinical trials, homoeopathy trials, and preclinical animal trials. Addressing this challenge requires a larger, more comprehensive dataset in these diverse domains. We need to identify individual RoB annotated corpora and test if these could be combined into a single compendium. For instance, Wang *et al.*'s preclinical animal trials corpus annotated with RoB classes could be combined with RoBuster to form a more representative benchmark [348]. A challenge in forming a larger compendium could be that different corpora might use different (and not necessarily compatible) Risk of Bias assessment tools.

- **LLM evaluation**: LLMs have shown to outperform on a variety of benchmark NLP tasks and showed promising results over some of the bias classes in our work (refer Section 5.4), but a thorough evaluation is required for the complex, convoluted task of RoB text span extraction. One recently published review study has proposed evaluating LLMs for extracting patterns, phrases and information relating to bias risk in clinical trials [247]. LLMs could help efficiently extract bias risk information from the trials, thereby aiding the human reviewers with faster, more accurate judgments. Only a comprehensive evaluation of LLMs across multiple RoB annotated datasets with benchmarked results could help achieve widespread adoption in real-world SR writing settings. Rose *et al.* took a step in that direction, recently publishing a proposed LLM evaluation study protocol for automatic risk of bias assessment on a larger corpus [284]. Pitre *et al.* tested the agreement between ChatGPT and risk of bias assessors over 157 clinical trials. Still, their work did not specifically dive into extracting bias descriptions from the text about individual signalling questions [265].

List of Figures

List of Tables

Bibliography

[1] M. Abaho, D. Bollegala, P. Williamson, and S. Dodd. Correcting crowdsourced annotations to improve detection of outcome types in evidence based medicine. In *CEUR Workshop Proceedings*, volume 2429, pages 1–5, 2019.

[2] S. Abogunrin, Y. Marti-Gil, M. Lane, and A. Witzmann. Can chatgpt generate synthetic data to train systematic literature review machine learning models? *Value in Health*, 26(12):S423, 2023.

[3] D. I. Adelani, M. A. Hedderich, D. Zhu, E. v. d. Berg, and D. Klakow. Distant supervision and noisy label learning for low resource named entity recognition: A study on hausa and yorùbá. *arXiv preprint arXiv:2003.08370*, 2020.

[4] J. Agnew. *Medicine in the old west: a history, 1850-1900*. McFarland, 2010.

[5] G. Aguilar, A. P. López-Monroy, F. A. González, and T. Solorio. Modeling noisiness to recognize named entities using multitask neural networks on social media. *arXiv preprint arXiv:1906.04129*, 2019.

[6] A. S. Albanna, B. M. Smith, D. Cowan, and D. Menzies. Fixed-dose combination antituberculosis therapy: a systematic review and meta-analysis. *European Respiratory Journal*, 42(3):721–732, 2013.

[7] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

[8] I. Amini, D. Martinez, D. Molla, et al. Overview of the alta 2012 shared task. In *Australasian Language Technology Association Workshop*, pages 124–129, 2012.

[9] Y. Ando and Y. Uyama. Multiregional clinical trials: Japanese perspective on drug development strategy and sample size for japanese subjects. *Journal of Biopharmaceutical Statistics*, 22(5):977–987, 2012.

[10] A. Apostolico, M. Crochemore, Z. Galil, and U. Manber. Combinatorial pattern matching third annual symposium tucson, arizona, usa, april 29–may 1, 1992 proceedings. In *Conference proceedings CPM*, page 236. Springer, 1992.

[11] S. Armijo-Olivo, R. Craig, and S. Campbell. Comparing machine and human reviewers to evaluate the risk of bias in randomized controlled trials. *Research Synthesis Methods*, 11(3):484–493, 2020.

[12] A. Arno, J. Thomas, B. Wallace, I. J. Marshall, J. E. McKenzie, and J. H. Elliott. Accuracy and efficiency of machine learning–assisted risk-of-bias assessments in "real-world" systematic reviews: A noninferiority randomized controlled trial. *Annals of Internal Medicine*, 2022.

[13] A. R. Aronson and F.-M. Lang. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.

[14] R. Artstein. Inter-annotator agreement. *Handbook of linguistic annotation*, pages 297–313, 2017.

[15] I. Augenstein, L. Derczynski, and K. Bontcheva. Generalisation in named entity recognition: A quantitative analysis. *Computer Speech & Language*, 44:61–83, 2017.

[16] A. Bandrowski, R. Brinkman, M. Brochhausen, M. H. Brush, B. Bug, M. C. Chibucos, K. Clancy, M. Courtot, D. Derom, M. Dumontier, et al. The ontology for biomedical investigations. *PloS one*, 11(4):e0154556, 2016.

[17] A. Bannach-Brown, P. Przyby?a, J. Thomas, A. S. C. Rice, S. Ananiadou, J. Liao, and M. R. Macleod. Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. *Syst Rev*, 8(1):23, Jan 2019.

[18] A. Bannach-Brown, P. Przybyła, J. Thomas, A. S. Rice, S. Ananiadou, J. Liao, and M. R. Macleod. Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. *Systematic reviews*, 8(1):1–12, 2019.

[19] M. Bauer, H. Gerlach, T. Vogelmann, F. Preissing, J. Stiefel, and D. Adam. Mortality in sepsis and septic shock in Europe, North America and Australia between 2009 and 2019—results from a systematic review and meta-analysis. *Critical Care*, 24(1):1–9, 2020.

[20] J. Baxter. A bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine learning*, 28(1):7–39, 1997.

[21] J. Beel, B. Gipp, S. Langer, and C. Breitinger. Paper recommender systems: a literature survey. *International Journal on Digital Libraries*, 17:305–338, 2016.

[22] T. Bekhuis and D. Demner-Fushman. Towards automating the initial screening phase of a systematic review. *MEDINFO 2010*, pages 146–150, 2010.

[23] T. Bekhuis and D. Demner-Fushman. Screening nonrandomized studies for medical systematic reviews: a comparative study of classifiers. *Artificial intelligence in medicine*, 55(3):197–207, 2012.

[24] T. Bekhuis, E. Tseytlin, and K. J. Mitchell. A prototype for a hybrid system to support systematic review teams: A case study of organ transplantation. In *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 940–947. IEEE, 2015.

[25] E. Beller, J. Clark, G. Tsafnat, C. Adams, H. Diehl, H. Lund, M. Ouzzani, K. Thayer, J. Thomas, T. Turner, et al. Making progress with the automation of systematic reviews: principles of the international collaboration for the automation of systematic reviews (icasr). *Systematic reviews*, 7:1–7, 2018.

[26] E. M. Beller, J. K.-H. Chen, U. L.-H. Wang, and P. P. Glasziou. Are systematic reviews up-to-date at the time of publication? *Systematic reviews*, 2(1):1–6, 2013.

[27] I. Beltagy, K. Lo, and A. Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[28] I. Beltagy, K. Lo, and A. Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[29] L. Bing, B. Dhingra, K. Mazaitis, J. H. Park, and W. W. Cohen. Bootstrapping distantly supervised ie using joint learning and small well-structured corpora. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[30] A. L. Boggiss, N. S. Consedine, J. M. Brenton-Peters, P. L. Hofman, and A. S. Serlachius. A systematic review of gratitude interventions: Effects on physical health and health behaviors. *Journal of Psychosomatic Research*, 135:110165, 2020.

[31] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146, 2017.

[32] A. Booth. Clear and present questions: formulating questions for evidence based practice. *Library hi tech*, 24(3):355–368, 2006.

[33] R. Borah, A. W. Brown, P. L. Capers, and K. 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.

[34] M. Borenstein, L. V. Hedges, J. P. Higgins, and H. R. Rothstein. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research synthesis methods*, 1(2):97–111, 2010.

[35] F. Boudin, J.-Y. Nie, J. C. Bartlett, R. Grad, P. Pluye, and M. Dawes. Combining classifiers for robust PICO element detection. *BMC medical informatics and decision making*, 10(1):1–6, 2010.

[36] F. Boudin, J.-Y. Nie, and M. Dawes. Clinical information retrieval using document and pico structure. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 822–830, 2010.

[37] F. Boudin, L. Shi, and J.-Y. Nie. Improving medical information retrieval with pico element detection. In *European Conference on Information Retrieval*, pages 50–61. Springer, 2010.

[38] C. Boudry, P. Alvarez-Muñoz, R. Arencibia-Jorge, D. Ayena, N. J. Brouwer, Z. Chaudhuri, B. Chawner, E. Epee, K. Erraïs, A. Fotouhi, et al. Worldwide inequality in access to full text scientific articles: the example of ophthalmology. *PeerJ*, 7:e7850, 2019.

[39] W. M. Bramer, D. Giustini, G. B. de Jonge, L. Holland, and T. Bekhuis. De-duplication of database search results for systematic reviews in endnote. *Journal of the Medical Library Association: JMLA*, 104(3):240, 2016.

[40] A. Brandsen, S. Verberne, K. Lambers, M. Wansleeben, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, et al. Creating a dataset for named entity recognition in the archaeology domain. In *Conference Proceedings LREC 2020*, pages 4573–4577. The European Language Resources Association, 2020.

[41] A. J. Brockmeier, M. Ju, P. Przybyła, and S. Ananiadou. Improving reference prioritisation with PICO recognition. *BMC medical informatics and decision making*, 19(1):1–14, 2019.

[42] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[43] T. Burgard and A. Bittermann. Reducing literature screening workload with machine learning. *Zeitschrift für Psychologie*, 2023.

[44] P. Butlin, R. Long, E. Elmoznino, Y. Bengio, J. Birch, A. Constant, G. Deane, S. M. Fleming, C. Frith, X. Ji, et al. Consciousness in artificial intelligence: insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*, 2023.

[45] T. Byrt, J. Bishop, and J. B. Carlin. Bias, prevalence and kappa. *Journal of clinical epidemiology*, 46(5):423–429, 1993.

[46] R. M. Califf, D. A. Zarin, J. M. Kramer, R. E. Sherman, L. H. Aberle, and A. Tasneem. Characteristics of clinical trials registered in clinicaltrials. gov, 2007-2010. *Jama*, 307(17):1838–1847, 2012.

[47] A. Capuano, A. J. Coats, C. Scavone, F. Rossi, and G. M. Rosano. Disclosure of negative trial results. a call for action. *International Journal of Cardiology*, 198:47–48, 2015.

[48] C. Cardellino, M. Teruel, L. A. Alemany, and S. Villata. A low-cost, high-coverage legal named entity recognizer, classifier and linker. In *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, pages 9–18. ICAIL Organization, 2017.

[49] R. Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[50] A. Carvallo and D. Parra. Comparing word embeddings for document screening based on active learning. In *BIRNDL@ SIGIR*, pages 100–107, 2019.

[51] A. Carvallo, D. Parra, H. Lobel, and A. Soto. Automatic document screening of medical literature using word and text embeddings in an active learning setting. *Scientometrics*, 125:3047–3084, 2020.

[52] J. M. Cejuela, P. McQuilton, L. Ponting, S. J. Marygold, R. Stefancsik, G. H. Millburn, B. Rost, F. Consortium, et al. tagtog: interactive and text-mining-assisted annotation of gene mentions in PLOS full-text articles. *Database*, 2014, 2014.

[53] S. Chabou and M. Iglewski. Pico extraction by combining the robustness of machine-learning methods with the rule-based methods. In *2015 World Congress on Information Technology and Computer Applications (WCITCA)*, pages 1–4. IEEE, 2015.

[54] S. Chabou and M. Iglewski. Combination of conditional random field with a rule based method in the extraction of pico elements. *BMC medical informatics and decision making*, 18(1):128, 2018.

[55] A. Chakrabarty, R. Dabre, C. Ding, M. Utiyama, and E. Sumita. Improving low-resource nmt through relevance based linguistic features incorporation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4263–4274, 2020.

[56] I. Chalmers. The cochrane collaboration: preparing, maintaining, and disseminating systematic reviews of the effects of health care. *Annals of the New York Academy of Sciences*, 703:156–63, 1993.

[57] I. Chalmers. Why the 1948 mrc trial of streptomycin used treatment allocation based on random numbers. *Journal of the Royal Society of Medicine*, 104(9):383–386, 2011.

[58] Q. Chen, H. Sun, H. Liu, Y. Jiang, T. Ran, X. Jin, X. Xiao, Z. Lin, H. Chen, and Z. Niu. An extensive benchmark study on biomedical text generation and mining with chatgpt. *Bioinformatics*, 39(9):btad557, 2023.

[59] Z. Chen, A. H. Cano, A. Romanou, A. Bonnet, K. Matoba, F. Salvi, M. Pagliardini, S. Fan, A. Köpf, A. Mohtashami, et al. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*, 2023.

[60] G. Y. Chung. Sentence retrieval for abstracts of randomized controlled trials. *BMC medical informatics and decision making*, 9(1):1–13, 2009.

[61] G. Y.-C. Chung. Towards identifying intervention arms in randomized controlled trials: extracting coordinating constructions. *Journal of biomedical informatics*, 42(5):790–800, 2009.

[62] A. M. Cohen. An effective general purpose approach for automated biomedical document classification. In *AMIA annual symposium proceedings*, volume 2006, page 161. American Medical Informatics Association, 2006.

[63] A. M. Cohen. 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, 2011.

[64] A. M. Cohen, K. Ambert, and M. McDonagh. Cross-topic learning for work prioritization in systematic review creation and update. *Journal of the American Medical Informatics Association*, 16(5):690–704, 2009.

[65] A. M. Cohen, K. Ambert, and M. 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.

[66] A. M. Cohen, W. R. Hersh, K. Peterson, and P.-Y. Yen. Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association*, 13(2):206–219, 2006.

[67] A. M. Cohen, N. R. Smalheiser, M. S. McDonagh, C. Yu, C. E. Adams, J. M. Davis, and P. S. Yu. Automated confidence ranked classification of randomized controlled trial articles: an aid to evidence-based medicine. *Journal of the American Medical Informatics Association*, 22(3):707–717, 2015.

[68] J. Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[69] D. J. Cook, C. D. Mulrow, and R. B. Haynes. Systematic reviews: synthesis of best evidence for clinical decisions. *Annals of internal medicine*, 126(5):376–380, 1997.

[70] T. F. Crocker, N. Lam, M. Jordão, C. Brundle, M. Prescott, A. Forster, J. Ensor, J. Gladman, and A. Clegg. Risk-of-bias assessment using cochrane's revised tool for randomized trials (rob 2) was useful but challenging and resource-intensive: observations from a systematic review. *Journal of Clinical Epidemiology*, 161:39–45, 2023.

[71] S. Cuschieri. The consort statement. *Saudi journal of anaesthesia*, 13(Suppl 1):S27, 2019.

[72] B. R. da Costa, B. Beckett, A. Diaz, N. M. Resta, B. C. Johnston, M. Egger, P. Jüni, and S. Armijo-Olivo. Effect of standardized training on the reliability of the cochrane risk of bias assessment tool: a prospective study. *Systematic reviews*, 6(1):1–8, 2017.

[73] K. Datar, M. N. Gandhi, P. Aggarwal, and M. Sohani. A review on deep learning based lip-reading. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 6(1):182, 2020.

[74] M. Dawes, P. Pluye, L. Shea, R. Grad, A. Greenberg, and J.-Y. Nie. The identification of clinically important elements within medical journal abstracts: Patient_population_problem, exposure_intervention, comparison, outcome, duration and results (pecodr). *Journal of Innovation in Health Informatics*, 15(1):9–16, 2007.

[75] B. De Bruijn, S. Carini, S. Kiritchenko, J. Martin, and I. Sim. Automated information extraction of key trial design elements from clinical trial publications. In *AMIA Annual Symposium Proceedings*, volume 2008, page 141. American Medical Informatics Association, 2008.

[76] J. De Bruin, Y. Ma, G. Ferdinands, J. Teijema, and R. Van de Schoot. SYNERGY - Open machine learning dataset on study selection in systematic reviews, 2023.

[77] P. De Matos, R. Alcántara, A. Dekker, M. Ennis, J. Hastings, K. Haug, I. Spiteri, S. Turner, and C. Steinbeck. Chemical entities of biological interest: an update. *Nucleic Acids Res*, 38(Suppl_1):D249–D254, 2010.

[78] J. J. Deeks, P. M. Bossuyt, and C. A. Gatsonis, editors. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy*. The Cochrane Collaboration, 2013. Available from `http://srdta.cochrane.org/`, accessed Aug. 2019.

[79] L. Deleger, Q. Li, T. Lingren, M. Kaiser, K. Molnar, L. Stoutenborough, M. Kouril, K. Marsolo, I. Solti, et al. Building gold standard corpora for medical natural language processing tasks. In *AMIA Annual Symposium Proceedings*, volume 2012, page 144. American Medical Informatics Association, 2012.

[80] D. Demner-Fushman and J. Lin. Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics*, 33(1):63–103, 2007.

[81] J. Devlin, M.-W. Chang, K. Lee, and K. 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.

[82] A. Dhrangadhariya, G. Aguilar, T. Solorio, R. Hilfiker, and H. Müller. End-to-end fine-grained neural entity recognition of patients, interventions, outcomes. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings 12*, pages 65–77. Springer, 2021.

[83] A. Dhrangadhariya, R. Hilfiker, K. M. Sattelmayer, K. Giacomino, R. Caliesch, S. Elsig, N. Naderi, and H. Müller. First steps towards a risk of bias corpus of randomized controlled trials. *Caring is sharing: exploiting the value in data for health and innovation: proceedings of MIE 2023*, 2023.

[84] A. Dhrangadhariya, R. Hilfiker, K. M. Sattelmayer, N. Naderi, K. Giacomino, R. Caliesch, J. Higgins, S. Marchand-Maillet, and H. Müller. Robuster: A corpus annotated with risk of bias text spans in randomized controlled trials. *JMIR Preprints*, 05 2023. Under Review.

[85] A. Dhrangadhariya, R. Hilfiker, R. Schaer, and H. Müller. Machine learning assisted citation screening for systematic reviews. In *MIE*, pages 302–306, 2020.

[86] A. Dhrangadhariya, O. Jimenez-del Toro, V. Andrearczyk, M. Atzori, and H. Müller. Exploiting biomedical literature to mine out a large multimodal dataset of rare cancer studies. In *Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications*, volume 11318, pages 77–87. SPIE, 2020.

[87] A. Dhrangadhariya, G. Manzo, and H. Müller. Pico to picos: Extracting study type and design information without labelled data. 2024.

[88] A. Dhrangadhariya, S. Millius, C. Thouly, B. Rizk, D. Fournier, H. Müller, and H. Brat. Automating quality control for structured standardized radiology reports using text analysis. In *MIE*, pages 58–62, 2020.

[89] A. Dhrangadhariya and H. Müller. Distant-cto: A zero cost, distantly supervised approach to improve low-resource entity extraction using clinical trials literature. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 345–358, 2022.

[90] A. Dhrangadhariya and H. Müller. Not so weak PICO: leveraging weak supervision for participants, interventions, and outcomes recognition for systematic review automation. *JAMIA open*, 6(1):ooac107, 2023.

[91] A. Dhrangadhariya, H. Müller, and G. Manzo. Pico to picos: Extracting study type and design information without labelled data. unpublished, 2023.

[92] A. Dhrangadhariya, S. Otálora, M. Atzori, and H. Müller. Classification of noisy free-text prostate cancer pathology reports using natural language processing. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part I*, pages 154–166. Springer International Publishing, 2021.

[93] A. Dhrangadhariya, A. Witzmann, M. Lane, H. Müller, and S. Abogunrin. Developing prospective active learning systems for citation screening and business factors in pharmaceutical living systematic reviews. *BMC Systematic Reviews*, 2024.

[94] R. Dror, G. Baumer, S. Shlomov, and R. Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, 2018.

[95] O. J. Dunn. Multiple comparisons using rank sums. *Technometrics*, 6(3):241–252, 1964.

[96] J. A. Dunnmon, A. J. Ratner, K. Saab, N. Khandwala, M. Markert, H. Sagreiya, R. Goldman, C. Lee-Messer, M. P. Lungren, D. L. Rubin, et al. Cross-modal data programming enables rapid medical machine learning. *Patterns*, 1(2):100019, 2020.

[97] A. B. Ebenezer, O. Boyinbode, and O. M. Idowu. A comprehensive analysis of handling imbalanced dataset. *International Journal*, 10(2), 2021.

[98] A. Elangovan, M. Davis, and K. Verspoor. Assigning function to protein-protein interactions: a weakly supervised biobert based approach using pubmed abstracts. *arXiv preprint arXiv:2008.08727*, 2020.

[99] A. Elangovan, Y. Li, D. E. Pires, M. J. Davis, and K. Verspoor. Large-scale protein-protein post-translational modification extraction with distant supervision and confidence calibrated biobert. *BMC bioinformatics*, 23:1–23, 2022.

[100] J. H. Elliott, T. Turner, O. Clavisi, J. Thomas, J. P. Higgins, C. Mavergames, and R. L. Gruen. Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. *PLoS medicine*, 11(2):e1001603, 2014.

[101] O. Etzioni, M. Banko, S. Soderland, and D. S. Weld. Open information extraction from the web. *Communications of the ACM*, 51(12):68–74, 2008.

[102] S. Fatemi and Y. Hu. A comparative analysis of fine-tuned llms and few-shot learning of llms for financial sentiment analysis. *arXiv preprint arXiv:2312.08725*, 2023.

[103] H. Fei, Y. Ren, and D. Ji. Recognizing nested named entity in biomedical texts: A neural network model with multi-task learning. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 376–381. IEEE, 2019.

[104] H. Fei, Y. Ren, and D. Ji. Dispatched attention with multi-task learning for nested mention recognition. *Information Sciences*, 513:241–251, 2020.

[105] G. F. Ferreira, M. G. Quiles, T. S. Nazaré, S. O. Rezende, and M. Demarzo. Automation of article selection process in systematic reviews through artificial neural network modeling and machine learning: protocol for an article selection model. *JMIR Research Protocols*, 10(6):e26448, 2021.

[106] K. Flemming. Asking answerable questions. *Evidence-based nursing*, 1(2):36–37, 1998.

[107] A. Foncubierta Rodríguez and H. Müller. Ground truth generation in medical imaging: a crowdsourcing-based iterative approach. In *Proceedings of the ACM multimedia 2012 workshop on Crowdsourcing for multimedia*, pages 9–14, 2012.

[108] L. G. Ford and B. M. Melnyk. The underappreciated and misunderstood picot question: A critical step in the ebp process. *Worldviews on evidence-based nursing*, 16(6):422–423, 2019.

[109] D. M. Fox. Systematic reviews and health policy: the influence of a project on perinatal care since 1988. *The Milbank Quarterly*, 89(3):425–449, 2011.

[110] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007.

[111] F. J. Friedlin and C. J. McDonald. A software tool for removing patient identifying information from clinical documents. *J Am Med Inform Assoc*, 15(5):601–610, 2008.

[112] J. A. Fries, E. Steinberg, S. Khattar, S. L. Fleming, J. Posada, A. Callahan, and N. H. Shah. Ontology-driven weak supervision for clinical entity classification in electronic health records. *Nature communications*, 12(1):1–11, 2021.

[113] D. Fu, M. Chen, F. Sala, S. Hooper, K. Fatahalian, and C. Ré. Fast and three-rious: Speeding up weak supervision with triplet methods. In *International Conference on Machine Learning*, pages 3280–3291. PMLR, 2020.

[114] N. Fuhr. Some common mistakes in ir evaluation, and how they can be avoided. In *ACM SIGIR Forum*, volume 51, pages 32–41. ACM New York, NY, USA, 2018.

[115] P. Gage. A new algorithm for data compression. *The C Users Journal archive*, 12:23–38, 1994.

[116] T. Gao, A. Fisch, and D. Chen. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830. Association for Computational Linguistics, 2021.

[117] V. García, R. Alejo, J. S. Sánchez, J. M. Sotoca, and R. A. Mollineda. Combined effects of class imbalance and class overlap on instance-based classification. In *Intelligent Data Engineering and Automated Learning–IDEAL 2006: 7th International Conference, Burgos, Spain, September 20-23, 2006. Proceedings 7*, pages 371–378. Springer, 2006.

[118] N. Geifman and E. Rubin. Towards an age-phenome knowledge-base. *BMC Bioinform*, 12(1):1–9, 2011.

[119] K. Giacomino, R. Caliesch, and K. M. Sattelmayer. The effectiveness of the peyton's 4-step teaching approach on skill acquisition of procedures in health professions education: A systematic review and meta-analysis with integrated meta-regression. *PeerJ*, 8:e10129, 2020.

[120] A. Giannakopoulos, C. Musat, A. Hossmann, and M. Baeriswyl. Unsupervised aspect term extraction with B-LSTM & CRF using automatically labelled datasets. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 180–188, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics.

[121] L. Gilbertson, P. Langhorne, A. Walker, A. Allen, and G. D. Murray. Domiciliary occupational therapy for patients with stroke discharged from hospital: randomised controlled trial. *Bmj*, 320(7235):603–606, 2000.

[122] M. J. Giummarra, G. Lau, and B. J. Gabbe. Evaluation of text mining to reduce screening workload for injury-focused systematic reviews. *Injury prevention*, 26(1):55–60, 2020.

[123] S. Glantz. *Primer of Biostatistics*. McGraw Hill, New York, 2012.

[124] C. E. Gleason, N. M. Dowling, W. Wharton, J. E. Manson, V. M. Miller, C. S. Atwood, E. A. Brinton, M. I. Cedars, R. A. Lobo, G. R. Merriam, et al. Effects of hormone therapy on cognition and mood in recently postmenopausal women: findings from the randomized, controlled keeps–cognitive and affective study. *PLoS medicine*, 12(6):e1001833, 2015.

[125] Y. Goldberg. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57:345–420, 2016.

[126] E. Graham-Clarke, A. Rushton, T. Noblet, and J. Marriott. Non-medical prescribing in the united kingdom national health service: A systematic policy review. *PloS one*, 14(7):e0214630, 2019.

[127] M. J. Grant and A. Booth. A typology of reviews: an analysis of 14 review types and associated methodologies. *Health information & libraries journal*, 26(2):91–108, 2009.

[128] M. W. Greaves. *Relation Extraction using Distant Supervision, SVMs, and Probabilistic First Order Logic*. PhD thesis, Carnegie Mellon University, 2014.

[129] V. Grimaldi, C. Schiano, A. Casamassimi, A. Zullo, A. Soricelli, F. P. Mancini, and C. Napoli. Imaging techniques to evaluate cell therapy in peripheral artery disease:

state of the art and clinical trials. *Clinical Physiology and Functional Imaging*, 36(3):165–178, 2016.

[130] E.-B. M. W. Group et al. Evidence-based medicine. a new approach to teaching the practice of medicine. *Jama*, 268:2420–2425, 1992.

[131] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

[132] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online, July 2020. Association for Computational Linguistics.

[133] M. M. Haby, E. Chapman, R. Clark, J. Barreto, L. Reveiz, and J. N. Lavis. What are the best methodologies for rapid reviews of the research evidence for evidence-informed decision making in health policy and practice: a rapid review. *Health research policy and systems*, 14(1):1–12, 2016.

[134] K. Hair, Z. Bahor, M. Macleod, J. Liao, and E. S. Sena. The automated systematic search deduplicator (asysd): a rapid, open-source, interoperable tool to remove duplicate citations in biomedical systematic reviews. *Biorxiv*, pages 2021–05, 2021.

[135] K. Hara and Y. Matsumoto. Extracting clinical trial design information from medline abstracts. *New Generation Computing*, 25(3):263–275, 2007.

[136] E. Hariton and J. J. Locascio. Randomised controlled trials—the gold standard for effectiveness research. *BJOG: an international journal of obstetrics and gynaecology*, 125(13):1716, 2018.

[137] L. Hartling, K. Bond, B. Vandermeer, J. Seida, D. M. Dryden, and B. H. Rowe. Applying the risk of bias tool in a systematic review of combination long-acting beta-agonists and inhaled corticosteroids for persistent asthma. *PloS one*, 6(2):e17242, 2011.

[138] H. Hassanzadeh, T. Groza, and J. Hunter. Identifying scientific artefacts in biomedical literature: The evidence based medicine use case. *Journal of biomedical informatics*, 49:159–170, 2014.

[139] L. Hassett, M. van den Berg, R. I. Lindley, M. Crotty, A. McCluskey, H. P. van der Ploeg, S. T. Smith, K. Schurr, K. Howard, M. L. Hackett, et al. Digitally enabled aged care and neurological rehabilitation to enhance outcomes with activity and mobility using technology (amount) in australia: A randomised controlled trial. *PLoS medicine*, 17(2):e1003029, 2020.

[140] Y. He, S. Sarntivijai, Y. Lin, Z. Xiang, A. Guo, S. Zhang, D. Jagannathan, L. Toldo, C. Tao, and B. Smith. Oae: the ontology of adverse events. *J Biomed Semant*, 5(1):1–13, 2014.

[141] Z. He, C. Tao, J. Bian, M. Dumontier, and W. R. Hogan. Semantics-powered healthcare engineering and data analytics, 2017.

[142] E. Heber, D. D. Ebert, D. Lehr, P. Cuijpers, M. Berking, S. Nobis, and H. Riper. The benefit of web-and computer-based interventions for stress: a systematic review and meta-analysis. *Journal of medical Internet research*, 19(2):e32, 2017.

[143] M. A. Hedderich, L. Lange, and D. Klakow. Anea: Distant supervision for low-resource named entity recognition. *arXiv preprint arXiv:2102.13129*, 2021.

[144] L. K. Henderson, J. C. Craig, N. S. Willis, D. Tovey, and A. C. Webster. How to write a cochrane systematic review. *Nephrology*, 15(6):617–624, 2010.

[145] C. A. Herrera, S. Lewin, E. Paulsen, A. Ciapponi, N. Opiyo, T. Pantoja, G. Rada, C. S. Wiysonge, G. Bastías, S. G. Marti, et al. Governance arrangements for health systems in low-income countries: an overview of systematic reviews. *Cochrane Database of Systematic Reviews*, 9(9), 2017.

[146] J. Higgins and S. Green. *Cochrane Handbook for Systematic Reviews of Interventions*. The Cochrane Collaboration, Mar 2011.

[147] J. P. Higgins, D. G. Altman, P. C. Gøtzsche, P. Jüni, D. Moher, A. D. Oxman, J. Savović, K. F. Schulz, L. Weeks, and J. A. Sterne. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ*, 343, 2011.

[148] J. P. Higgins, J. Savović, M. J. Page, R. G. Elbers, and J. A. Sterne. Assessing risk of bias in a randomized trial. *Cochrane handbook for systematic reviews of interventions*, pages 205–228, 2019.

[149] J. P. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, and V. A. Welch. *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons, 2019.

[150] R. Hilfiker, A. Meichtry, M. Eicher, L. N. Balfe, R. H. Knols, M. L. Verra, and J. Taeymans. Exercise and other non-pharmaceutical interventions for cancer-related fatigue in patients during or after cancer treatment: a systematic review incorporating an indirect-comparisons meta-analysis. *British journal of sports medicine*, 52(10):651–658, 2018.

[151] J. Hirt, J. Meichlinger, P. Schumacher, and G. Mueller. Agreement in Risk of Bias Assessment Between RobotReviewer and Human Reviewers: An Evaluation Study on Randomised Controlled Trials in Nursing-Related Cochrane Reviews. *Journal of Nursing Scholarship*, 53(2):246–254, 2021.

[152] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[153] F. Hoffmann, K. Allers, T. Rombey, J. Helbach, A. Hoffmann, T. Mathes, and D. 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.

[154] R. Hoffmann, C. Zhang, X. Ling, L. Zettlemoyer, and D. S. Weld. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 541–550, 2011.

[155] M. Honnibal and I. Montani. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1):411–420, 2017.

[156] J. Howard and S. Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.

[157] K.-C. Huang, I.-J. Chiang, F. Xiao, C.-C. Liao, C. C.-H. Liu, and J.-M. Wong. PICO element detection in medical text without metadata: Are first sentences enough? *Journal of biomedical informatics*, 46(5):940–946, 2013.

[158] K.-C. Huang, C. C.-H. Liu, S.-S. Yang, F. Xiao, J.-M. Wong, C.-C. Liao, and I.-J. Chiang. Classification of PICO elements by text features systematically extracted from pubmed abstracts. In *2011 IEEE International Conference on Granular Computing*, pages 279–283. IEEE, 2011.

[159] X. Huang, J. Lin, and D. Demner-Fushman. Evaluation of pico as a knowledge representation for clinical questions. In *AMIA annual symposium proceedings*, volume 2006, page 359. American Medical Informatics Association, 2006.

[160] Z. Huang, W. Xu, and K. Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

[161] B. L. Humphreys, D. A. Lindberg, H. M. Schoolman, and G. O. Barnett. The unified medical language system: an informatics research collaboration. *J Am Med Inform Assoc*, 5(1):1–11, 1998.

[162] Z. M. Ibrahim, M. Bader-El-Den, and M. Cocea. Improving imbalanced students' text feedback classification using re-sampling based approach. In *Advances in Computational Intelligence Systems: Contributions Presented at the 19th UK Workshop on Computational Intelligence, September 4-6, 2019, Portsmouth, UK 19*, pages 262–267. Springer, 2020.

[163] H. Imamura, I. Sato, and M. Sugiyama. Analysis of minimax error rate for crowdsourcing and its application to worker clustering model. In *International Conference on Machine Learning*, pages 2147–2156. PMLR, 2018.

[164] P. S. J. Jardim, C. J. Rose, H. M. Ames, J. F. M. Echavez, S. Van de Velde, and A. E. Muller. Automating risk of bias assessment in systematic reviews: a real-time mixed methods comparison of human researchers to a machine learning system. *BMC Medical Research Methodology*, 22(1):1–12, 2022.

[165] K. Jaseena and J. M. David. Issues, challenges, and solutions: big data mining. *CS & IT-CSCP*, 4(13):131–140, 2014.

[166] D. Jin and P. Szolovits. PICO element detection in medical text via long short-term memory neural networks. In *Proceedings of the BioNLP 2018 workshop*, pages 67–75. Association for Computational Linguistics, 2018.

[167] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.

[168] S. Jonnalagadda and D. Petitti. A new iterative method to reduce workload in systematic review process. *International journal of computational biology and drug design*, 6(1-2):5–17, 2013.

[169] A. Joshi, S. Karimi, R. Sparks, C. Paris, and C. R. MacIntyre. A comparison of word-based and context-based representations for classification problems in health informatics. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 135–141, Florence, Italy, Aug. 2019. Association for Computational Linguistics.

[170] D. Jurafsky and J. H. Martin. Speech and language processing (draft). *Chapter A: Hidden Markov Models (Draft of September 11, 2018). Retrieved March*, 19:2019, 2018.

[171] T. Kang, S. Zou, and C. Weng. Pretraining to recognize PICO elements from randomized controlled trial literature. *Studies in health technology and informatics*, 264:188, 2019.

[172] D. Kartchner, S. Ramalingam, I. Al-Hussaini, O. Kronick, and C. Mitchell. Zero-shot information extraction for clinical meta-analysis using large language models. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 396–405, 2023.

[173] S. Khangura, K. Konnyu, R. Cushman, J. Grimshaw, and D. Moher. Evidence summaries: the evolution of a rapid review approach. *Systematic reviews*, 1(1):1–9, 2012.

[174] S. Khangura, K. Konnyu, R. Cushman, J. Grimshaw, and D. Moher. Evidence summaries: the evolution of a rapid review approach. *Syst Rev*, 1:10, Feb 2012.

[175] S. N. Kim, D. Martinez, L. Cavedon, and L. Yencken. Automatic classification of sentences to support evidence based medicine. In *BMC bioinformatics*, volume 12, pages 1–10. BioMed Central, 2011.

[176] Y. S. Kim, D. Yoon, J. Byun, H. Park, A. Lee, I. H. Kim, S. Lee, H.-S. Lim, and R. W. Park. Extracting information from free-text electronic patient records to identify practice-based evidence of the performance of coronary stents. *PLoS One*, 12(8):e0182889, 2017.

[177] I. A. Knight, M. L. Wilson, D. F. Brailsford, and N. Milic-Frayling. Enslaved to the trapped data: A cognitive work analysis of medical systematic reviews. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, pages 203–212, 2019.

[178] G. Kontonatsios, A. J. Brockmeier, P. Przybyła, J. McNaught, T. Mu, J. Y. Goulermas, and S. Ananiadou. A semi-supervised approach using label propagation to support citation screening. *Journal of biomedical informatics*, 72:67–76, 2017.

[179] C. A. Kronk and J. W. Dexheimer. Development of the gender, sex, and sexual orientation ontology: Evaluation and workflow. *J Am Med Inform Assoc*, 27(7):1110–1115, 2020.

[180] Y. Kwon, M. Lemieux, J. McTavish, and N. Wathen. Identifying and removing duplicate records from systematic review searches. *Journal of the Medical Library Association: JMLA*, 103(4):184, 2015.

[181] B. Lander and E. Balka. Exploring how evidence is used in care through an organizational ethnography of two teaching hospitals. *Journal of medical Internet research*, 21(3):e10769, 2019.

[182] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[183] C. Lanera, P. Berchialla, A. Sharma, C. Minto, D. Gregori, and I. Baldi. Screening pubmed abstracts: is class imbalance always a challenge to machine learning? *Systematic reviews*, 8:1–9, 2019.

[184] L. Lansbury, B. Lim, V. Baskaran, and W. S. Lim. Co-infections in people with COVID-19: a systematic review and meta-analysis. *Journal of Infection*, 81(2):266–275, 2020.

[185] J. Lau. Systematic review automation thematic series. *Systematic reviews*, 8:1–2, 2019.

[186] R. Lauche, H. Cramer, W. Häuser, G. Dobos, J. Langhorst, et al. A systematic overview of reviews for complementary and alternative therapies in the treatment of the fibromyalgia syndrome. *Evidence-Based Complementary and Alternative Medicine*, 2015, 2015.

[187] J. Lavis, H. Davies, A. Oxman, J.-L. Denis, K. Golden-Biddle, and E. Ferlie. Towards systematic reviews that inform health care management and policy-making. *Journal of health services research & policy*, 10(1_suppl):35–48, 2005.

[188] G. E. Lee and A. Sun. A study on agreement in pico span annotations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1149–1152, 2019.

[189] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[190] I. Lerner, P. Créquit, P. Ravaud, and I. Atal. Automatic screening using word embeddings achieved high sensitivity and workload reduction for updating living network meta-analyses. *Journal of clinical epidemiology*, 108:86–94, 2019.

[191] V. I. Levenshtein et al. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union, 1966.

[192] D. Li, P. Zafeiriadis, and E. 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.

[193] M. Li, L. Zhang, M. Zhou, and D. Han. Uttsr: A novel non-structured text table recognition model powered by deep learning technology. *Applied Sciences*, 13(13):7556, 2023.

[194] P. Li, X. Jiang, and H. Shatkay. Figure and caption extraction from biomedical documents. *Bioinformatics*, 35(21):4381–4388, 2019.

[195] S. Li, S. Ge, Y. Hua, C. Zhang, H. Wen, T. Liu, and W. Wang. Coupled-view deep classifier learning from multiple noisy annotators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4667–4674, 2020.

[196] L. Liang, X. Cui, M. Feng, S. Zhou, X. Yin, F. He, K. Sun, H. Yin, R. Xie, D. Zhang, et al. The effectiveness of exercise on cervical radiculopathy: A protocol for systematic review and meta-analysis. *Medicine*, 98(35), 2019.

[197] A. Liberati, D. G. Altman, J. Tetzlaff, C. Mulrow, P. Gøtzsche, J. Ioannidis, et al. The prisma statement for reporting systematic and meta-analyses of studies that evaluate interventions: explanation and elaboration. *PLoS medicine*, 6(7):1–28, 2009.

[198] A. Y. Lin, S. Gebel, Q. L. Li, S. Madan, J. Darms, E. Bolton, B. Smith, M. Hofmann-Apitius, Y. O. He, and A. T. Kodamullil. Cto: A community-based clinical trial ontology and its applications in pubchemrdf and scaiview. In *CEUR workshop proceedings*, volume 2807. NIH Public Access, 2020.

[199] A. Y. Lin, S. Gebel, Q. L. Li, S. Madan, J. Darms, E. Bolton, B. Smith, M. Hofmann-Apitius, Y. O. He, and A. T. Kodamullil. Cto: A community-based clinical trial ontology and its applications in pubchemrdf and scaiview. In *CEUR workshop proceedings*, volume 2807. NIH Public Access, 2020.

[200] P. Lison, J. Barnes, and A. Hubin. skweak: Weak supervision made easy for nlp. *arXiv preprint arXiv:2104.09683*, 2021.

[201] J. Liu, P. Timsina, and O. El-Gayar. A comparative analysis of semi-supervised learning: the case of article selection for medical systematic reviews. *Information Systems Frontiers*, 20:195–207, 2018.

[202] S. Liu, Y. Sun, B. Li, W. Wang, F. T. Bourgeois, and A. G. Dunn. Sent2Span: Span detection for PICO extraction in the biomedical text without span annotations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1705–1715, Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics.

[203] Z. Liu, M. Yang, X. Wang, Q. Chen, B. Tang, Z. Wang, and H. Xu. Entity recognition from clinical texts via recurrent neural network. *BMC medical informatics and decision making*, 17:53–61, 2017.

[204] M. Loef, H. Walach, and S. Schmidt. Interrater reliability of rob2–an alternative measure and way of categorization. *Journal of Clinical Epidemiology*, 142:326–327, 2022.

[205] N. J. Lopez, S. Uribe, and B. Martinez. Effect of periodontal treatment on preterm birth rate: a systematic review of meta-analyses. *Periodontology 2000*, 67(1):87–130, 2015.

[206] L.-L. Ma, Y.-Y. Wang, Z.-H. Yang, D. Huang, H. Weng, and X.-T. Zeng. Methodological quality (risk of bias) assessment tools for primary and secondary medical studies: what are they and which is better? *Military Medical Research*, 7(1):1–11, 2020.

[207] X. Ma and E. Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*, 2016.

[208] K. C. Madathil, J. S. Greenstein, and R. Koikkara. An investigation of the factors that predict a healthcare consumer's use of anecdotal healthcare information available on the internet. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 58, pages 604–608. SAGE Publications Sage CA: Los Angeles, CA, 2014.

[209] E. K. Mallory, M. de Rochemonteix, A. Ratner, A. Acharya, C. Re, R. A. Bright, and R. B. Altman. Extracting chemical reactions from text using snorkel. *BMC Bioinform*, 21(1):1–15, 2020.

[210] C. Manning and H. Schutze. *Foundations of statistical natural language processing*. MIT press, 1999.

[211] G. Manzo, B. Pocklington, Y. Pannatier, C. Gay, A. Dhrangadhariya, S. Carrard, R. Hilfiker, and J.-P. Calbimonte. Towards semantic modeling of patient trajectories for rehabilitation of osteoarthritis. In *CEUR Workshop Proceedings*, 2023.

[212] M. Marrone, A. Stewart, and W. D. Dotson. Clinical utility of gene-expression profiling in women with early breast cancer: an overview of systematic reviews. *Genetics in medicine*, 17(7):519–532, 2015.

[213] I. J. Marshall, J. Kuiper, and B. C. Wallace. Automating risk of bias assessment for clinical trials. In *proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 88–95, 2014.

[214] I. J. Marshall, J. Kuiper, and B. C. Wallace. Automating risk of bias assessment for clinical trials. *IEEE journal of biomedical and health informatics*, 19(4):1406–1412, 2015.

[215] I. J. Marshall, J. Kuiper, and B. 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.

[216] I. J. Marshall, A. Noel-Storr, J. Kuiper, J. Thomas, and B. C. Wallace. Machine learning for identifying randomized controlled trials: an evaluation and practitioner's guide. *Research synthesis methods*, 9(4):602–614, 2018.

[217] A. L. C. Martimbianco, K. M. M. Sá, G. M. Santos, E. M. Santos, R. L. Pacheco, and R. Riera. Most cochrane systematic reviews and protocols did not adhere to the cochrane's risk of bias 2.0 tool. *Revista da Associação Médica Brasileira*, 69:469–472, 2023.

[218] M. K. Martinić, D. Pieper, A. Glatt, and L. Puljak. Definition of a systematic review used in overviews of systematic reviews, meta-epidemiological studies and textbooks. *BMC Medical Research Methodology*, 19(1), 2019.

[219] S. Matwin, A. Kouznetsov, D. Inkpen, O. Frunza, and P. 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, 2010.

[220] S. Matwin, A. Kouznetsov, D. Inkpen, and P. O'Blenis. 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):author reply 105, 2011.

[221] S. Matwin and V. Sazonova. Direct comparison between support vector machine and multinomial naive bayes algorithms for medical abstract classification. *Journal of the American Medical Informatics Association*, 19(5):917–917, 2012.

[222] A. T. McCray, A. Burgun, and O. Bodenreider. Aggregating umls semantic types for reducing conceptual complexity. *Studies in health technology and informatics*, 84(0 1):216, 2001.

[223] M. L. McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.

[224] S. McKeown and Z. M. Mir. Considerations for conducting systematic reviews: evaluating the performance of different methods for de-duplicating references. *Systematic reviews*, 10:1–8, 2021.

[225] K. A. McKibbon and S. Marks. Posing clinical questions: framing the question for scientific inquiry. *AACN Advanced Critical Care*, 12(4):477–481, 2001.

[226] Y. Meng, J. Shen, C. Zhang, and J. Han. Weakly-supervised neural text classification. In *proceedings of the 27th ACM International Conference on information and knowledge management*, pages 983–992, 2018.

[227] B. Merkel, H. Butzkueven, A. L. Traboulsee, E. Havrdova, and T. Kalincik. Timing of high-efficacy therapy in relapsing-remitting multiple sclerosis: a systematic review. *Autoimmunity reviews*, 16(6):658–665, 2017.

[228] A. M. Methley, S. Campbell, C. Chew-Graham, R. McNally, and S. Cheraghi-Sohi. Pico, picos and spider: a comparison study of specificity and sensitivity in three search tools for qualitative systematic reviews. *BMC health services research*, 14(1):1–10, 2014.

[229] M. Michelson and K. Reuter. The significant cost of systematic reviews and meta-analyses: a call for greater involvement of machine learning to assess the promise of clinical trials. *Contemporary clinical trials communications*, 16:100443, 2019.

[230] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[231] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[232] L. A. Millard, P. A. Flach, and J. P. Higgins. Machine learning to assist risk-of-bias assessments in systematic reviews. *International journal of epidemiology*, 45(1):266–277, 2016.

[233] T. Millard, A. Synnot, J. Elliott, S. Green, S. McDonald, and T. Turner. Feasibility and acceptability of living systematic reviews: results from a mixed-methods evaluation. *Systematic reviews*, 8:1–14, 2019.

[234] S. A. Miller and J. L. Forrest. Enhancing your practice through evidence-based decision making: Pico, learning how to ask good questions. *Journal of Evidence Based Dental Practice*, 1(2):136–141, 2001.

[235] S. Minozzi, M. Cinquini, S. Gianola, M. Gonzalez-Lorenzo, and R. Banzi. The revised cochrane risk of bias tool for randomized trials (rob 2) showed low interrater reliability and challenges in its application. *Journal of clinical epidemiology*, 126:37–44, 2020.

[236] S. Minozzi, K. Dwan, F. Borrelli, and G. Filippini. Reliability of the revised cochrane risk-of-bias tool for randomised trials (rob2) improved with the use of implementation instruction. *Journal of clinical epidemiology*, 141:99–105, 2022.

[237] M. Mintz, S. Bills, R. Snow, and D. Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, 2009.

[238] M. Miwa, J. Thomas, A. O'Mara-Eves, and S. Ananiadou. Reducing systematic review workload through certainty-based screening. *Journal of biomedical informatics*, 51:242–253, 2014.

[239] O. Mohammed, R. Benlamri, and S. Fong. Building a diseases symptoms ontology for medical diagnosis: an integrative approach. In *The First International Conference on Future Generation Communication Technologies*, pages 104–108. IEEE, 2012.

[240] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, and P. Group*. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *Annals of internal medicine*, 151(4):264–269, 2009.

[241] R. M. Monarch. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster, 2021.

[242] H. Müller, V. Andrearczyk, O. J. del Toro, A. Dhrangadhariya, R. Schaer, and M. Atzori. Studying public medical images from the open access literature and social networks for model training and knowledge extraction. In *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26*, pages 553–564. Springer, 2020.

[243] L. Myer, T. K. Phillips, A. Zerbe, K. Brittain, M. Lesosky, N.-Y. Hsiao, R. H. Remien, C. A. Mellins, J. A. McIntyre, and E. J. Abrams. Integration of postpartum healthcare services for hiv-infected women and their infants in south africa: a randomised controlled trial. *PLoS medicine*, 15(3):e1002547, 2018.

[244] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551, 2011.

[245] K. Nakamura and T. Shibata. Regulatory changes after the enforcement of the new clinical trials act in japan. *Japanese Journal of Clinical Oncology*, 50(4):399–404, 2020.

[246] Y. Nakamura, S. Hanaoka, Y. Nomura, T. Nakao, S. Miki, T. Watadani, T. Yoshikawa, N. Hayashi, and O. Abe. Automatic detection of actionable radiology reports using bidirectional encoder representations from transformers. *BMC Medical Informatics and Decision Making*, 21(1):1–19, 2021.

[247] A. J. Nashwan and J. H. Jaradat. Streamlining systematic reviews: Harnessing large language models for quality assessment and risk-of-bias evaluation. *Cureus*, 15(8), 2023.

[248] R. Navigli and F. Martelli. An overview of word and sense similarity. *Natural Language Engineering*, 25(6):693–714, 2019.

[249] M. Neumann, Z. Shen, and S. Skjonsberg. Pawls: Pdf annotation with labels and structure. *arXiv preprint arXiv:2101.10281*, 2021.

[250] A. Ng and M. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14, 2001.

[251] G. Ninot, I. Boulze-Launay, G. Bourrel, A. Gerazime, E. Guerdoux-Ninot, B. Lognos, T. Libourel, G. Mercier, A. O. Engberink, S. Rapior, et al. De la définition des interventions non médicamenteuses (inm) à leur ontologie. *Hegel*, 1(1):21–27, 2018.

[252] C. Norman. *Systematic Review Automation Methods*. PhD thesis, Université Paris-Saclay; Universiteit van Amsterdam, 2020. [Online]. Available: `https://tel.archives-ouvertes.fr/tel-03060620`.

[253] R. Nunn. Mere anecdote: evidence and stories in medicine. *Journal of evaluation in clinical practice*, 17(5):920–926, 2011.

[254] B. Nye, J. J. Li, R. Patel, Y. Yang, I. J. Marshall, A. Nenkova, and B. 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.

[255] A. O'Mara-Eves, J. Thomas, J. McNaught, M. Miwa, and S. Ananiadou. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic reviews*, 4(1):1–22, 2015.

[256] A. Páez. Gray literature: An important resource in systematic reviews. *Journal of Evidence-Based Medicine*, 10(3):233–240, 2017.

[257] M. J. Page, J. P. Higgins, G. Clayton, J. A. Sterne, A. Hróbjartsson, and J. Savović. Empirical evidence of study design biases in randomized trials: systematic review of meta-epidemiological studies. *PloS one*, 11(7):e0159267, 2016.

[258] H. Pedder, G. Sarri, E. Keeney, V. Nunes, and S. Dias. Data extraction for complex meta-analysis (decimal) guide. *Systematic reviews*, 5(1):1–6, 2016.

[259] M. Peng, X. Xing, Q. Zhang, J. Fu, and X. Huang. Distantly supervised named entity recognition using positive-unlabeled learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2409–2419, Florence, Italy, July 2019. Association for Computational Linguistics.

[260] Y. Peng, S. Yan, and Z. Lu. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*, pages 58–65, 2019.

[261] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[262] V. Piechotta, C. Iannizzi, K. L. Chai, S. J. Valk, C. Kimber, E. Dorando, I. Monsef, E. M. Wood, A. A. Lamikanra, D. J. Roberts, et al. Convalescent plasma or hyperimmune immunoglobulin for people with COVID-19: a living systematic review. *Cochrane Database of Systematic Reviews*, 5(5), 2021.

[263] D. Pinto, A. McCallum, X. Wei, and W. B. Croft. Table extraction using conditional random fields. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 235–242. ACM, 2003.

[264] D. M. Pisanelli, D. Zaccagnini, L. Capurso, and M. Koch. An ontological approach to evidence-based medicine and meta-analysis. In *The New Navigators: from Professionals to Patients*, pages 543–548. IOS Press, 2003.

[265] T. Pitre, T. Jassal, J. R. Talukdar, M. Shahab, M. Ling, and D. Zeraatkar. Chatgpt for assessing risk of bias of randomized trials using the rob 2.0 tool: A methods study. *medRxiv*, pages 2023–11, 2023.

[266] J. R. Polanin, T. D. Pigott, D. L. Espelage, and J. K. Grotpeter. Best practice guidelines for abstract screening large-evidence systematic reviews and meta-analyses. *Research Synthesis Methods*, 10(3):330–342, 2019.

[267] S. Pyysalo, F. Ginter, H. Moen, T. Salakoski, and S. Ananiadou. Distributional semantics resources for biomedical text processing. *Proceedings of LBM*, pages 39–44, 2013.

[268] X. Qi, M. Yang, W. Ren, J. Jia, J. Wang, G. Han, and D. Fan. Find duplicates among the pubmed, embase, and cochrane library databases in systematic review. *PLoS One*, 8(8):e71838, 2013.

[269] X. Qin, J. Liu, Y. Wang, Y. Liu, K. Deng, Y. Ma, K. Zou, L. Li, and X. Sun. Natural language processing was effective in assisting rapid title and abstract screening when updating systematic reviews. *Journal of clinical epidemiology*, 133:121–129, 2021.

[270] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. Improving language understanding by generative pre-training, 2018.

[271] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

[272] A. I. Ramadhan and E. B. Setiawan. Aspect-based sentiment analysis on social media using convolutional neural network (cnn) method. *Building of Informatics, Technology and Science (BITS)*, 4(4):1828–1836, 2023.

[273] S. Rani and J. Singh. Enhancing levenshtein's edit distance algorithm for evaluating document similarity. In *Computing, Analytics and Networks: First International Conference, ICAN 2017, Chandigarh, India, October 27-28, 2017, Revised Selected Papers 1*, pages 72–80. Springer, 2018.

[274] J. W. Ratcliff and D. E. Metzener. Pattern-matching-the gestalt approach. *Dr Dobbs Journal*, 13(7):46, 1988.

[275] J. Rathbone, M. Carter, T. Hoffmann, and P. Glasziou. Better duplicate detection for systematic reviewers: evaluation of systematic review assistant-deduplication module. *Systematic reviews*, 4:1–6, 2015.

[276] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 11, page 269. NIH Public Access, 2017.

[277] A. J. Ratner, C. M. De Sa, S. Wu, D. Selsam, and C. Ré. Data programming: Creating large training sets, quickly. *Advances in neural information processing systems*, 29, 2016.

[278] K. M. Richardson and H. R. Rothstein. Effects of occupational stress management intervention programs: a meta-analysis. *Journal of occupational health psychology*, 13(1):69, 2008.

[279] W. S. Richardson, M. C. Wilson, J. Nishikawa, R. S. Hayward, et al. The well-built clinical question: a key to evidence-based decisions. *Acp j club*, 123(3):A12–3, 1995.

[280] J. J. Riva, K. M. Malik, S. J. Burnie, A. R. Endicott, and J. W. Busse. What is your research question? an introduction to the picot format for clinicians. *J Can Chiropr Assoc*, 56(3):167, 2012.

[281] P. N. Robinson, S. Köhler, S. Bauer, D. Seelow, D. Horn, and S. Mundlos. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *Am J Hum Genet*, 83(5):610–615, 2008.

[282] A. Rogier, A. Coulet, and B. Rance. Using an ontological representation of chemotherapy toxicities for guiding information extraction and integration from ehrs. In *Medinfo 2021-18th World Congress on Medical and Health Informatics*, 2021.

[283] T. Rombey, K. Allers, T. Mathes, F. Hoffmann, and D. Pieper. A descriptive analysis of the characteristics and the peer review process of systematic review protocols published in an open peer review journal from 2012 to 2017. *Bmc medical research methodology*, 19(1):1–9, 2019.

[284] C. J. Rose, M. Ringsten, J. Bidonde, J. Glanville, R. C. Berg, C. Cooper, A. E. Muller, H. B. Bergsund, J. F. Meneses-Echavez, and T. Potrebny. Using a large language model (chatgpt) to assess risk of bias in randomized controlled trials of medical interventions: protocol for a pilot study of interrater agreement with human reviewers, 2023.

[285] A. Rossi, C. Friel, L. Carter, and C. E. Garber. Effects of theory-based behavioral interventions on physical activity among overweight and obese female cancer survivors: a systematic review of randomized controlled trials. *Integrative cancer therapies*, 17(2):226–236, 2018.

[286] S. Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

[287] R. Russell, M. Chung, E. Balk, S. Atkinson, E. Giovannucci, S. Ip, and J. Lau. Systematic review methods. In *Issues and Challenges in Conducting Systematic Reviews to Support Development of Nutrient Reference Values: Workshop Summary: Nutrition Research Series*, volume 2, 2009.

[288] E. Safranchik, S. Luo, and S. Bach. Weakly supervised sequence tagging from noisy rules. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5570–5578, 2020.

[289] T. Saito and M. Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432, 2015.

[290] O. Sanchez Graillet, P. Cimiano, C. Witte, and B. Ell. C-tro: An ontology for summarization and aggregation of the level of evidence in clinical trials. In *Proceedings of the Workshop Ontologies and Data in Life Sciences (ODLS 2019) in the Joint Ontology Workshops'(JOWO 2019)*, 2019.

[291] O. Sanchez-Graillet, C. Witte, F. Grimm, and P. Cimiano. An annotated corpus of clinical trial publications supporting schema-based relational information extraction. *J Biomed Semantics*, 13(1):1–18, 2022.

[292] A. Sarker, D. Mollá, and C. Paris. An approach for automatic multi-label classification of medical sentences. In *Proceedings of the 4th International Louhi Workshop on Health Document Text Mining and Information Analysis. Sydney, NSW, Australia*, 2013.

[293] J. Savović, R. M. Turner, D. Mawdsley, H. E. Jones, R. Beynon, J. P. Higgins, and J. A. Sterne. Association between risk-of-bias assessments and results of randomized trials in cochrane reviews: the robes meta-epidemiologic study. *American Journal of Epidemiology*, 187(5):1113–1122, 2018.

[294] R. M. Schmidt. Recurrent neural networks (rnns): A gentle introduction and overview. *arXiv preprint arXiv:1912.05911*, 2019.

[295] C. M. Schmucker, A. Blümle, L. K. Schell, G. Schwarzer, P. Oeller, L. Cabrera, E. von Elm, M. Briel, J. J. Meerpohl, and O. consortium. Systematic review finds that study data not published in full text articles have unclear impact on meta-analyses results in medical research. *PloS one*, 12(4):e0176210, 2017.

[296] J. Schneider, L. Hoang, Y. Kansara, A. M. Cohen, and N. R. Smalheiser. Evaluation of publication type tagging as a strategy to screen randomized controlled trial articles in preparing systematic reviews. *JAMIA open*, 5(1):ooac015, 2022.

[297] L. M. Schriml, C. Arze, S. Nadendla, Y.-W. W. Chang, M. Mazaitis, V. Felix, G. Feng, and W. A. Kibbe. Disease ontology: a backbone for disease semantic integration. *Nucleic Acids Res*, 40(D1):D940–D946, 2012.

[298] H. Schütze, C. D. Manning, and P. Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.

[299] A. L. Seidler, K. E. Hunter, S. Cheyne, J. A. Berlin, D. Ghersi, and L. M. Askie. Prospective meta-analyses and cochrane's role in embracing next-generation methodologies. *The Cochrane Database of Systematic Reviews*, 2020(10), 2020.

[300] R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units. In K. Erk and N. A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, Aug. 2016. Association for Computational Linguistics.

[301] B. Settles. Active learning literature survey, 2009.

[302] V. S. Sheng, J. Zhang, B. Gu, and X. Wu. Majority voting and pairing with multiple noisy labeling. *IEEE Transactions on Knowledge and Data Engineering*, 31(7):1355–1368, 2017.

[303] A. Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306, 2020.

[304] K. G. Shojania, M. Sampson, M. T. Ansari, J. Ji, S. Doucette, and D. Moher. How quickly do systematic reviews go out of date? a survival analysis. *Annals of internal medicine*, 147(4):224–233, 2007.

[305] B. Sibbald and M. Roland. Understanding controlled trials. Why are randomised controlled trials important? *BMJ: British Medical Journal*, 316(7126):201, 1998.

[306] I. Sim, B. Olasov, and S. Carini. An ontology of randomized controlled trials for evidence-based practice: content specification and evaluation using the competency decomposition method. *Journal of Biomedical Informatics*, 37(2):108–119, 2004.

[307] I. Sim, S. W. Tu, S. Carini, H. P. Lehmann, B. H. Pollock, M. Peleg, and K. M. Wittkowski. The ontology of clinical research (ocre): an informatics foundation for the science of clinical research. *Journal of biomedical informatics*, 52:78–91, 2014.

[308] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al. Large language models encode clinical knowledge. *Nature*, pages 1–9, 2023.

[309] A. Smirnova and P. Cudré-Mauroux. Relation extraction using distant supervision: A survey. *ACM Computing Surveys (CSUR)*, 51(5):1–35, 2018.

[310] F. Soboczenski, T. A. Trikalinos, J. Kuiper, R. G. Bias, B. C. Wallace, and I. J. Marshall. Machine learning to help researchers evaluate biases in clinical trials: a prospective, randomized user study. *BMC medical informatics and decision making*, 19(1):1–12, 2019.

[311] S. Srinivasan, T. C. Rindflesch, W. T. Hole, A. R. Aronson, and J. G. Mork. Finding umls metathesaurus concepts in medline. In *Proceedings of the AMIA Symposium*, page 727. American Medical Informatics Association, 2002.

[312] J. A. Sterne, M. Egger, and G. D. Smith. Investigating and dealing with publication and other biases in meta-analysis. *Bmj*, 323(7304):101–105, 2001.

[313] J. A. Sterne, J. Savović, M. J. Page, R. G. Elbers, N. S. Blencowe, I. Boutron, C. J. Cates, H.-Y. Cheng, M. S. Corbett, S. M. Eldridge, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ*, 366, 2019.

[314] A. E. Stuck, A. Moser, U. Morf, U. Wirz, J. Wyser, G. Gillmann, S. Born, M. Zwahlen, S. Iliffe, D. Harari, et al. Effect of health risk assessment and counselling on health behaviour and survival in older people: a pragmatic randomised trial. *PLoS medicine*, 12(10):e1001889, 2015.

[315] N. Stylianou and I. Vlahavas. Transformed: End-to-end transformers for evidence-based medicine and argument mining in medical literature. *Journal of Biomedical Informatics*, 117:103767, 2021.

[316] I. Susmelj. Improve your large language models (llms) with active learning, 2023. Accessed on 23 December 2023.

[317] S. Šuster, T. Baldwin, and K. Verspoor. Analysis of predictive performance and reliability of classifiers for quality assessment of medical evidence revealed important variation by medical area. *Journal of Clinical Epidemiology*, 159:58–69, 2023.

[318] B. Tang, X. Wang, J. Yan, and Q. Chen. Entity recognition in chinese clinical text using attention-based cnn-lstm-crf. *BMC medical informatics and decision making*, 19:89–97, 2019.

[319] R. Tang, X. Han, X. Jiang, and X. Hu. Does synthetic data generation of llms help clinical text mining? *arXiv preprint arXiv:2303.04360*, 2023.

[320] S. J. Taylor, D. Carnes, K. Homer, B. C. Kahan, N. Hounsome, S. Eldridge, A. Spencer, T. Pincus, A. Rahman, and M. Underwood. Novel three-day, community-based, nonpharmacological group intervention for chronic musculoskeletal pain (copers): a randomised clinical trial. *PLoS medicine*, 13(6):e1002040, 2016.

[321] A. N. Thorndike, S. Mills, L. Sonnenberg, D. Palakshappa, T. Gao, C. T. Pau, and S. Regan. Activity monitor intervention to promote physical activity of physicians-in-training: randomized controlled trial. *PloS one*, 9(6):e100251, 2014.

[322] T. Tian and J. Zhu. Max-margin majority voting for learning from crowds. *Advances in neural information processing systems*, 28, 2015.

[323] A. Toral and V. M. Sánchez-Cartagena. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. *arXiv preprint arXiv:1701.02901*, 2017.

[324] C. Training. Cochrane handbook for systematic reviews of interventions, version 6, 2019.

[325] M. R. Tramèr, D. J. M. Reynolds, R. A. Moore, and H. J. McQuay. Impact of covert duplicate publication on meta-analysis: a case study. *Bmj*, 315(7109):635–640, 1997.

[326] A. C. Tricco, J. Brehaut, M. H. Chen, and D. Moher. Following 411 cochrane protocols to completion: a retrospective cohort study. *PLoS One*, 3(11):e3684, 2008.

[327] A. C. Tricco, R. Cardoso, S. M. Thomas, S. Motiwala, S. Sullivan, M. R. Kealey, B. Hemmelgarn, M. Ouimet, M. P. Hillmer, L. Perrier, et al. Barriers and facilitators to uptake of systematic reviews by policy makers and health care managers: a scoping review. *Implementation Science*, 11:1–20, 2015.

[328] G. Tsafnat, P. Glasziou, M. K. Choong, A. Dunn, F. Galgani, and E. Coiera. Systematic review automation technologies. *Systematic reviews*, 3:1–15, 2014.

[329] G. Tsafnat, P. Glasziou, G. Karystianis, and E. Coiera. Automated screening of research studies for systematic reviews using study characteristics. *Systematic reviews*, 7:1–9, 2018.

[330] M. Tschopp, M. K. Sattelmayer, and R. Hilfiker. Is power training or conventional resistance training better for function in elderly persons? a meta-analysis. *Age and ageing*, 40(5):549–556, 2011.

[331] A. Tsertsvadze, Y.-F. Chen, D. Moher, P. Sutcliffe, and N. McCarthy. How to conduct systematic reviews more expeditiously? *Systematic reviews*, 4(1):1–6, 2015.

[332] A. Y. Tsou, J. R. Treadwell, E. Erinoff, and K. Schoelles. Machine learning for screening prioritization in systematic reviews: comparative performance of abstrackr and eppi-reviewer. *Systematic reviews*, 9:1–14, 2020.

[333] L. S. Uman. Systematic reviews and meta-analyses. *J Can Acad Child Adolesc Psychiatry*, 20(1):57, 2011.

[334] R. Van De Schoot, J. De Bruin, R. Schram, P. Zahedi, J. De Boer, F. Weijdema, B. Kramer, M. Huijts, M. Hoogerwerf, G. Ferdinands, et al. An open source machine learning framework for efficient and transparent systematic reviews. *Nature machine intelligence*, 3(2):125–133, 2021.

[335] S. H. van Dijk, M. G. Brusse-Keizer, C. C. Bucsán, J. van der Palen, C. J. Doggen, and A. Lenferink. Artificial intelligence in systematic reviews: promising when appropriately used. *BMJ open*, 13(7):e072254, 2023.

[336] R. van Dinter, C. Catal, and B. Tekinerdogan. A multi-channel convolutional neural network approach to automate the citation screening process. *Applied Soft Computing*, 112:107765, 2021.

[337] R. van Dinter, B. Tekinerdogan, and C. Catal. Automation of systematic literature reviews: A systematic literature review. *Information and Software Technology*, 136:106589, 2021.

[338] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[339] M. Verbeke, V. Van Asch, R. Morante, P. Frasconi, W. Daelemans, and L. De Raedt. A statistical relational learning approach to identifying evidence based medicine categories. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 579–589, 2012.

[340] C. H. Vinkers, H. J. Lamberink, J. K. Tijdink, P. Heus, L. Bouter, P. Glasziou, D. Moher, J. A. Damen, L. Hooft, and W. M. Otte. The methodological quality of 176,620 randomized controlled trials published between 1966 and 2018 reveals a positive trend but also an urgent need for improvement. *PLoS Biology*, 19(4):e3001162, 2021.

[341] J. Vollert, N. R. Cook, T. J. Kaptchuk, S. T. Sehra, D. K. Tobias, and K. T. Hall. Assessment of placebo response in objective and subjective outcome measures in rheumatoid arthritis clinical trials. *JAMA network open*, 3(9):e2013196–e2013196, 2020.

[342] E. Von Elm, G. Poglia, B. Walder, and M. R. Tramer. Different patterns of duplicate publication: an analysis of articles used in systematic reviews. *Jama*, 291(8):974–980, 2004.

[343] B. C. Wallace, J. Kuiper, A. Sharma, M. Zhu, and I. J. Marshall. Extracting PICO sentences from clinical trial reports using supervised distant supervision. *The Journal of Machine Learning Research*, 17(1):4572–4596, 2016.

[344] B. C. Wallace, A. Noel-Storr, I. J. Marshall, A. M. Cohen, N. R. Smalheiser, and J. Thomas. Identifying reports of randomized controlled trials (rcts) via a hybrid machine learning and crowdsourcing approach. *Journal of the American Medical Informatics Association*, 24(6):1165–1168, 2017.

[345] B. C. Wallace, K. Small, C. E. Brodley, and T. A. Trikalinos. Active learning for biomedical citation screening. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 173–182, 2010.

[346] B. C. Wallace, T. A. Trikalinos, J. Lau, C. Brodley, and C. H. Schmid. Semi-automated screening of biomedical citations for systematic reviews. *BMC bioinformatics*, 11(1):1–11, 2010.

[347] J. Wang and Y. Dong. Measurement of text similarity: a survey. *Information*, 11(9):421, 2020.

[348] Q. Wang, J. Liao, M. Lapata, and M. Macleod. Risk of bias assessment in preclinical literature using natural language processing. *Research synthesis methods*, 13(3):368–380, 2022.

[349] X. Wang, Y. Zhang, Q. Li, X. Ren, J. Shang, and J. Han. Distantly supervised biomedical named entity recognition with dictionary expansion. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 496–503. IEEE, 2019.

[350] Y. Wang. Indexing initiative web api: Python implementation, 2022. Accessed on 2023-11-07.

[351] Y. Wang, S. Sohn, S. Liu, F. Shen, L. Wang, E. J. Atkinson, S. Amin, and H. Liu. A clinical text classification paradigm using weak supervision and deep representation. *BMC Med Inform Decis Mak*, 19(1):1–13, 2019.

[352] Z. Wang, N. Asi, T. A. Elraiyah, C. Undavalli, P. Glasziou, V. Montori, M. H. Murad, et al. Dual computer monitors to increase efficiency of conducting systematic reviews. *Journal of clinical epidemiology*, 67(12):1353–1357, 2014.

[353] Z. Wang, T. Nayfeh, J. Tetzlaff, P. O'Blenis, and M. H. Murad. Error rates of human reviewers during abstract screening in systematic reviews. *PloS one*, 15(1):e0227742, 2020.

[354] L. Weber, K. Thobe, O. A. Migueles Lozano, J. Wolf, and U. Leser. Pedl: extracting protein–protein associations using deep language models and distant supervision. *Bioinformatics*, 36(Suppl_1):i490–i498, 2020.

[355] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[356] Z. Wen, D. Deng, R. Zhang, and R. Kotagiri. : An efficient entity extraction algorithm using two-level edit-distance. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 998–1009. IEEE, 2019.

[357] H. Wilding. Literature searching: Choosing a database, 2023. Accessed on 04.12.2023.

[358] W. E. Winkler. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage., 1990.

[359] R. Xu, Y. Garten, K. S. Supekar, A. K. Das, R. B. Altman, A. M. Garber, et al. Extracting subject demographic information from abstracts of randomized clinical trial reports. In *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems*, page 550. IOS Press, 2007.

[360] H. Yang and J. M. Garibaldi. Automatic detection of protected health information from clinic narratives. *J Biomed Inform*, 58:S30–S38, 2015.

[361] X. Yang, J. Bian, Y. Gong, W. R. Hogan, and Y. Wu. Madex: a system for detecting medications, adverse drug events, and their relations from clinical notes. *Drug safety*, 42:123–133, 2019.

[362] Y. Yang, W. Chen, Z. Li, Z. He, and M. Zhang. Distantly supervised ner with partial annotation learning and reinforcement learning. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2159–2169. Association for Computational Linguistics, 2018.

[363] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.

[364] R. Yousefi-Nooraie, S. Irani, S. Mortaz-Hedjri, and B. Shakiba. Comparison of the efficacy of three pubmed search filters in finding randomized controlled trials to answer clinical questions. *Journal of Evaluation in Clinical Practice*, 19(5):723–726, 2013.

[365] Z. Yu, N. A. Kraft, and T. Menzies. Finding better active learners for faster literature reviews. *Empirical Software Engineering*, 23:3161–3186, 2018.

[366] K. K. Yuen. The two-sample trimmed t for unequal population variances. *Biometrika*, 61(1):165–170, 1974.

[367] L. Zhang, I. Ajiferuke, and M. Sampson. Optimizing search strategies to identify randomized controlled trials in medline. *BMC Medical Research Methodology*, 6:1–10, 2006.

[368] S. Zhang, S. Jiang, Y. Yan, et al. A software defect prediction approach based on bigan anomaly detection. *Scientific Programming*, 2022, 2022.

[369] T. Zhang, Y. Yu, J. Mei, Z. Tang, X. Zhang, and S. Li. Unlocking the power of deep PICO extraction: Step-wise medical ner identification. *arXiv preprint arXiv:2005.06601*, 2020.

[370] Y. Zhang, Q. Chen, Z. Yang, H. Lin, and Z. Lu. Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific data*, 6(1):52, 2019.

[371] Y. Zhang and B. Wallace. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*, 2015.

[372] C. Zheng, Y. Cai, J. Xu, H.-f. Leung, and G. Xu. A boundary-aware neural model for nested named entity recognition. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 357–366, 2019.

[373] W. Zheng and C. Blake. Using distant supervised learning to identify protein subcellular localizations from full-text scientific articles. *Journal of biomedical informatics*, 57:134–144, 2015.

[374] H. Zhou, Z. Liu, C. Lang, Y. Xu, Y. Lin, and J. Hou. Improving the recall of biomedical named entity recognition with label re-correction and knowledge distillation. *BMC bioinformatics*, 22(1):295, 2021.